

**SENGAMALA THAYAAR EDUCATIONAL TRUST WOMEN'S COLLEGE
(AUTONOMOUS)**

SUNDARAKKOTTAI, MANNARGUDI - 614 016, TIRUVARUR (DT), TAMIL NADU.



**NATIONAL CONFERENCE ON
"EMERGING TRENDS IN ARTIFICIAL INTELLIGENCE"
(NCETAI-2024)**

19.02.2024 & 20.02.2024

PROCEEDINGS

**LET'S TALK ABOUT THE
FUTURE**

ORGANIZED BY

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE



**SENGAMALA THAYAR EDUCATIONAL TRUST
WOMEN'S COLLEGE (AUTONOMOUS)
(SILVER JUBILEE INSTITUTION)**

(Affiliated to Bharathidasan University, Tiruchirappalli)
(Accredited by NAAC & An ISO 9001:2015 Certified Institution)

**SUNDARAKKOTTAI, MANNARGUDI - 614 016.
THIRUVARUR (Dist.), TAMIL NADU, INDIA.**

Phone: 04367-255423

Website : www.stet.edu.in



**NATIONAL CONFERENCE
ON
EMERGING TRENDS IN ARTIFICIAL INTELLIGENCE
(NCETAI-2024)**

19.02.2024 & 20.02.2024

PROCEEDINGS

Organized by

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE

IN ASSOCIATION WITH IQAC

ABOUT THE INSTITUTION

Empowering the women is empowering the nation. Higher education, especially to women, empowers not only the family but also emphasized as a mission of capacity building of the nation. The significance of this aspect has been rightly recognized and there has been a paradigm shift in the scenario of women education in the urban. Nevertheless higher education to women in the rural area needs encouragement, because the main source of income is agriculture and per capita income is relatively very low. In this context, Sengamala Thayaar Educational Trust Women's College was established at Sundarakkottai, a hamlet in Thiruvarur district, in the year 1994 by Shrimathi Krishnaveni Vivekanandham, who entrusted the task of nurturing the institution to her illustrious son Dr.V.Dhivaharan now as the Correspondent with yonder vision for promoting higher education to women students in this downtrodden area. The institution offers 12 UG, 13 PG, 5 M.Phil, 4 Ph.D. and PG Diploma course. Sengamala Thayaar Educational Trust Women's College is recognized as one of the foremost educational institutions in India, affiliated to Bharathidasan University, Tiruchirappalli. It has been accredited by NAAC and it is an ISO 9001:2015 Certified Institution. The institution has attained Autonomous in the year 2019-2020. The Internal Quality Assurance Cell (IQAC) of the College is extremely active to achieve the goal of the institution.

ABOUT THE DEPARTMENT

Our Department established in the year 1994 with the course B.Sc. We offer other programmes such as M.Sc., computer science, M.Sc., Information Technology and we started research programmes M.Phil & Ph.D started from 2011. Our department going to start a new programme B.Sc., Artificial Intelligence from the next academic year. So far we have produced 3,280 Undergraduates, 1306 Postgraduates, 109 M.Phil Scholars, and 9 Ph.D., Scholars. We produced 516 University Ranks and 20 Gold Medals. At present our department students strength is 587. Department Laboratories are well equipped recent technologies. Faculty Members have been regularly presenting papers in an international and national conference and publishing papers in International and National journals with high impact factor. The department organize conference, seminars, Guest lectures, Workshops, Hands on training and Extension activities frequently we got institutional membership for ICT Academy and by using the same, the student and staff membership have been getting various opportunities for academic enrichments and conduction of Technical sessions frequently. Department has received research projects from Tamil Nadu State Council for Science and Technology (TNSCST) and received funds from various NGO's. The staff and students of our department contributed the college by developing software like Language software LearnDiff and Feedback software etc...

FELICITATION ADDRESS



I consider it to be a great honour to be a part of the inaugural session of this National Conference on “*Emerging Trends in Artificial Intelligence (NCETAI=2024)*” conducted by the PG & Research Department of Computer Science in association with IQAC, STET Women’s college, known for its yeomen services in the field of higher education for women and women empowerment. First of all, I would like to congratulate The Department of Computer Science and the Organizing Committee for organizing this National Conference, which has great potentials for advancement of research in the current and fast developing field of Artificial Intelligence, which is going to rule the world in future.

The idea of 'a machine that thinks' dates back to ancient Greece. But since the advent of computers, the evolution of artificial intelligence starts back to 1950, when Alan Turing published *Computing Machinery and Intelligence* wherein he proposes to answer the question –can machines think?|| and –systems that act like humans.|| Turing introduced a test –the Turing Test|| to determine if a computer can demonstrate the same intelligence as a human. In 1956 John McCarthy coined the term 'artificial intelligence' at the first-ever AI conference at Dartmouth College. Since then this field has undergone tremendous development.

Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind. The AI combines computer science and robust datasets, to enable problem-solving. It also encompasses sub-fields of machine learning and deep learning, which are frequently mentioned in conjunction with artificial intelligence. These disciplines are comprised of AI algorithms which seek to create expert systems which make predictions or classifications based on input data.

Over the years, artificial intelligence has gone through many cycles of hype, but even to skeptics, the release of OpenAI’s ChatGPT seems to mark a turning point. There are numerous, real-world applications of AI systems today such as Speech recognition, Customer service s such as Online virtual agents, Computer vision, Recommendation engines, Automated stock trading etc. Today, AI technology is widely used throughout industry, government, and science.

Even though the applications for this technology are growing every day and the hype around the use of AI in business takes off, there are conversations around it ethics which are also critically important.

However, one can conclude that today’s challenges faced by academia and industry are so complex that they can only be solved through the help of AI. I hope that this conference will address these by discussing the recent trends and developments in the broad topics of AI to promote exchange of ideas in various applications of AI in science and engineering that are intended to upgrade the existing knowledge in research and create deeper interest in various disciplines. Furthermore this conference will provide opportunity to young researchers to learn the current state of research & techniques, develop sound presentation skills and attitudes necessary to pursue further research in AI and its applications to the participants. Congratulations and all the Best.

Dr. K.THIYAGESAN

PRESIDENTIAL ADDRESS



Ladies and gentlemen,
Distinguished guests, scholars, researchers, and participants,
I extend a warm welcome to each one of you at the National Conference on Emerging Trends in Artificial Intelligence (NCETAI-2024). As the President of this remarkable gathering, I feel a deep sense of honour and responsibility to address you today. Our collective presence here, organized by the PG & Research Department of Computer Science in collaboration with the Internal Quality Assurance Cell (IQAC) of Sengamala Thayaar Educational Trust

Women's College, signifies not only the commitment to academic excellence but also the celebration of a significant milestone – the silver jubilee of our esteemed institution.

The theme of this conference, "Emerging Trends in Artificial Intelligence," resonates with the dynamism and evolution characterizing the field. Artificial Intelligence has become an integral part of our lives, shaping industries, economies, and societies. In the Indian context, we are witnessing a transformative journey, with AI playing a pivotal role in various sectors.

India has emerged as a hub for AI research, development, and innovation. The government's initiatives, such as the National AI Strategy, are indicative of a concerted effort to position the country as a global leader in AI. The emphasis on research and development, skill enhancement, and the creation of a conducive ecosystem has propelled India into the forefront of the AI landscape.

Recent AI cases in India illustrate the diverse applications and the potential impact of this technology. From healthcare to agriculture, AI is making significant strides. In healthcare, AI algorithms are aiding in early diagnosis and personalized treatment plans. The integration of AI in agriculture is optimizing crop yield predictions, resource allocation, and sustainable farming practices. These instances reflect not only the technological advancements but also the positive socio-economic implications of AI adoption.

However, as we embrace the promises of AI, we must also be mindful of the challenges and ethical considerations. The recent AI case involving bias in facial recognition algorithms serves as a stark reminder of the importance of responsible AI development. Ensuring fairness, transparency, and accountability in AI systems is imperative to mitigate potential harms and promote trust among users.

As we navigate through the rich tapestry of presentations and discussions over the next few days, let us collectively explore the emerging trends in AI, keeping in mind the unique perspectives and opportunities in the Indian context. Our diversity, cultural richness, and interdisciplinary collaborations can contribute to innovative solutions that address local challenges while contributing to the global discourse on artificial intelligence.

To the young minds and researchers present, I encourage you to immerse yourselves in the wealth of knowledge that will be shared during this conference. Your contributions are vital to shaping the future of AI in India and beyond. Seize this opportunity to network, collaborate, and foster partnerships that will drive innovation and positive change.

In conclusion, I express my gratitude to the organizers, sponsors, and everyone who has played a role in making NCETAI-2024 possible. Let us embark on this intellectual journey with enthusiasm, curiosity, and a shared commitment to advancing the frontiers of artificial intelligence.

Thank you, and I wish you all a fruitful and engaging conference.

Dr.M.V.SRINATH



SENGAMALA THAYAAR EDUCATIONAL TRUST WOMEN'S COLLEGE
(Autonomous)
(Affiliated to Bharathidasan University, Tiruchirappalli)
(Accredited by NAAC) (An ISO 9001:2015 Certified Institution)
SUNDARAKKOTTAI, MANNARGUDI-614 016, TAMILNADU, S.INDIA
PG and Research Department of Computer Science

PREFACE

Dr. V. GEETHA, M.Sc., M.Phil., B.Ed., Ph.D.,
Associate Professor & Head,
PG & Research Department of Computer Science



On behalf of the organizing committee, I would like to cordially welcome you to the National conference on "Emerging Trends in Artificial Intelligence (NCETAI – 2024)". It is a real honor and pleasure to welcome all, who are looking for a well-established and prestigious forum to communicate and disseminate the innovative advances of their research.


Artificial Intelligence(AI) is the transformative technology of our era and the next step in evolution of cognition. Artificial intelligence is a branch of computer science that aims to create intelligent machines. It has become an essential part of the technology industry. Research associated with Artificial Intelligence is highly technical and specialized. Artificial Intelligence enhances the speed, precision, and effectiveness of human efforts.

The current trends in AI include a wide spectrum of technologies such as Natural Language Generation, Speech Recognition, Virtual Agent, Robotic Process Automation, Deep Learning Platforms and so on. These AI trends drive sustainable growth for businesses across industries, redefining the way we work, learn, and interact with technology

Artificial intelligence is revolutionizing various sectors of health, enterprises, businesses, and the commercial sector with its smartness. Therefore, high-level events discussing solutions, trends, and the future are crucial. The NCETAI - 2024 will focus on software, applications, medical uses, solutions, trends, and more

It is a fact that AI is continuously digging its way in a vast number of diverse scientific domains. —NCETAI – 2024 aims to provide a forum for delegates from both academia and industry, who wish to share pioneer ideas, to get to know each other better and to present their research. Moreover, it aims to offer a deep glimpse into the future and to examine important AI ethical aspects. Genuine peer and expert reviews, appreciation, and recognition can take research a long way.

I hope that this Special Issue will increase the understanding of the emerging trends of Artificial Intelligences, and I would like to thank the authors for their valuable contributions, which support the scientific progress of Artificial Technology. We are trying our best to ensure that your time in the college during the conference is one of the most memorable one and you go back with rich information and as a proud AI Scientist of the NCETAI world . I thank every one of you who are contributing to the success of the conference and looking forward to see you all soon.


HEAD OF THE DEPARTMENT
PG and Research Department of
Computer Science
S.T.E.T Women's College (Autonomous)
MANNARGUDI.



ALAGAPPA UNIVERSITY

(A State University Established in 1985)

KARAIKUDI - 630 003, Tamil Nadu, India

www.alagappauniversity.ac.in

DEPARTMENT OF COMPUTER SCIENCE



2017



Accredited with A+ Grade
by NAAC (CGPA : 3.64)

2018



MHRD
Govt. of India



UGC
University Grants Commission

Graded as Category-1
& Granted Autonomy

2023



NATIONAL
INSTITUTIONAL
RANKING
FRAMEWORK

Rank : 30

2023



Asia Rank :
251-260

2024



WUR : 601 - 800

Ref. : DCS / AY 2023-24/ 81

Date: 05.02.2024

Prof. A. PADMAPRIYA

Head i/c



I am delighted to be part of this National Conference on Emerging Trends in Artificial Intelligence (NCETAI 2024) – organized by PG & Research Department of Computer Science, Sengamala Thayaar Educational Trust Women's College (Autonomous), Sundarakkottai, Mannargudi. The Conference is aimed to exchange information relating to Intelligent systems and its applications in Industry. The key areas of discussion in this conference are Computational Intelligence, Data Science & Analytics, IoT and Robotics, Telecommunications & Applications.

Through this program academicians, scientists, engineers and professionals from different universities and academic institutions, R&D organizations and industrial concerns will get the opportunity to interact, exchange ideas, knowledge, and views for enhancing collaborations. In the current scenario, it is important to network and extend cooperation to realize our potential and use the diverse talents available in our people. By interacting with the academicians and the professionals, all should mutually be able to enhance their knowledge of the state-of-the-art technology as well as the industry needs.

I congratulate all the participants of this virtual event. Get ignited and make best use of this opportunity. Also, I would like to thank wholeheartedly the Patrons, Advisors, Convenors and Organizing Committee members for their efforts and contributions to make this wonderful event happen.



BHARATA MATA COLLEGE

DEPARTMENT OF COMPUTER SCIENCE

(Affiliated to Mahatma Gandhi University Kottayam)

(Re-accredited by NAAC with 'A+' Grade, ISO 9001-2015 Certified)

Thrikkakara P O , Kochi – 682 021, Kerala, India

Tel : 0484-2425121 Fax: 0484-2427796 Email : hodcomputerscience@bharatamatacollege.in

23.01.2024



I extend my hearty congratulations to Sengamala Thayaar Educational Trust Women's College for successfully conducting the National Conference on Emerging Trends in Artificial Intelligence.

It is a great occasion and remarkable platform that brings together brilliant minds, passionate researchers and industry experts from various corners of the nation to explore into the area of cutting-edge technologies. The efforts invested by the organizers in making an event of this magnitude deserve immense appreciation.

Emerging Trends in Artificial Intelligence have transformed as the driving forces behind revolutionary advancements in numerous domains. By fostering knowledge exchange, stimulating discussions and showcasing groundbreaking research, this conference has undoubtedly played a pivotal role in shaping the future of these fields.

I would like to extend my deepest appreciation to the distinguished keynote speakers, panelists and presenters whose expertise and insights have illuminated the path toward progress. Their invaluable contributions have undoubtedly inspired the audience and ignited a spark of curiosity and innovation which is undoubtedly the exact force of progress.

Through this conference of emerging trends in Artificial Intelligence, let us reflect on the incredible strides made in these fields and channel our collective efforts towards exploring new frontiers and overcoming future challenges not only in these fields but also the various aspects in Information Technology and human life as a whole.

Once again, congratulations to the management and organisers for arranging such a remarkable event. Your dedication to promoting knowledge, fostering innovation, and shaping the future is truly commendable. Definitely it will have positive impacts in the near future in multi dimensional manners.

Best wishes for continued success in all your future endeavors. May God Bless all of us abundantly.

Sincerely,

Dr. JOHN T. ABRAHAM
PhD, MCA, MSc (ISM), MPhil, MTech (IT)
Head of the Department
Department of Computer Science
Bharata Mata College Thrikkakara
Kochi, Kerala - 682 021



A.V.C. COLLEGE (AUTONOMOUS)

Affiliated to Annamalai University - Annamalainagar
NAAC Reaccredited 'A+' Grade (CGPA = 3.46 / 4.00) in 4th Cycle
NIRF All India Ranking 2023 | College Rank Band 101 - 150
UGC Recognized "College with Potential for Excellence - Phase I & II"
MANNAMPANDAL, MAYILADUTHURAI - 609 305, TAMIL NADU

Phone : 04364 - 222264

Fax : 04364 - 222264

e-mail: avccollegeauto@gmail.com

avccsdept@gmail.com

Website : www.avccollege.net

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE

Ref. No.

Date :

8th February 2024.

Dr. K. Palanivel, M.Sc., M.Phil., Ph.D., (NET.)

Associate Professor of Computer Science

(Subject Expert in NCETAI-2024)



MESSAGE

Sengamala Thayaar Educational Trust Women's College has certainly come a long way and has provided educational platform to innumerable women students who are enthusiastic to accomplish their dreams and ambition especially from the village background. National or International conferences are very significant in the development and growth of an Institution.

As the Subject Expert of "National Conference on Emerging Trends in Artificial Intelligence (NCETAI-2024)" organized in the Department of Computer Science of STET Women's College on 19 & 20th February 2024, I congratulate and convey my wishes to the Management, the Organizers of this conference and the Faculty members of Department who dedicatedly involved in this conference.

I am delighted to note that the accepted papers are being published in an ISBN-Edited Volume. I wholeheartedly appreciate all the sincere efforts of the entire team of the conference.

I feel very proud that this conference definitely would develop and promote the research at a higher level in the field of Computer Science, especially in Artificial Intelligence.

I hope this conference will hold a series of intellectually interactive sessions and intensive deliberations by scholars and technical experts participating in it. This Conference will be the eye opener for the researchers, students and faculty to show the various avenues in the field of Computer Science.

My best wishes to this noble endeavour of Department of Computer Science of STET Women's College. I wish the conference a grand success.


(K. Palanivel)

**NATIONAL CONFERENCE
ON
EMERGING TRENDS IN ARTIFICIAL INTELLIGENCE
(NCETAI-2024)**

**Organized by
PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE
IN ASSOCIATION WITH IQAC
ORGANIZING COMMITTEE**

CHIEF PATRON

**Dr.V.Dhivaharan, M.Sc., D.E.M., Ph.D., CORRESPONDENT,
STET Group of Institutions**

PATRON

**Dr. N. Uma Maheswari , M.Sc., Ph.D., B.Ed.,
PRINCIPAL,**

S.T.E.T Women's College (Autonomous)

CONVENOR

**Dr.V.Geetha, M.Sc., M.Phil., Ph.D.,
Associate Professor & Head,**

PG and Research Department of Computer Science

CO CONVENOR

**Dr.M.V.SRINATH, M.C.A., Ph.D.,
Research Advisor**

PG and Research Department of Computer Science

EXECUTIVE COMMITTEE

Dr. R. Anitha	M.Sc., M.Phil., Ph.D
Ms.R. Manjupargavi	M.Sc.,M.Phil.,B.Ed.,SET
Ms.N. Subhalakshmi	M.Sc., M.Phil
Ms.K. Savima	M.Sc.,M.Phil
Ms.C.Jasmine	M.Sc(IT)., M.Phil
Ms.V. Mangaiyarkarasi	M.C.A.,M.Phil
Ms.G. Geetha	M.Sc., M.Phil.,M.Tech
Ms.S. Keerthika	M.Sc.,M.Phil
Ms. R.Akilandeswari,	M.Sc.,M.Phil
Ms.V.Akilandeswari	M.Sc.,M.Phil
Ms. A.Srilekha	M.C.A.,M.Phil
Ms. S.L.Savitha	M.C.A.,M.Phil
Ms. V.Suganya	M.Sc., M.Phil
Ms.P.Nagajothi	M.Sc., M.Phil
Ms.A.Indhuja	M.C.A.,M.Phil

SENGAMALA THAYAR EDUCATIONAL TRUST WOMEN'S COLLEGE (AUTONOMOUS)



SILVER JUBILEE INSTITUTION

Affiliated to Bharathidasan University, Tiruchirappalli
(Accredited by NAAC) (An ISO 9001:2015 Certified Institution)

Sundarakkottai, Mannargudi- 614016, Thiruvarur(Dt), Tamil Nadu, South India.



A TWO DAY

NATIONAL CONFERENCE ON EMERGING TRENDS IN ARTIFICIAL INTELLIGENCE (NCETAI - 2024)

ORGANIZED BY

PG AND RESEARCH DEPARTMENT OF COMPUTER SCIENCE IN ASSOCIATION WITH IQAC

VENUE : CONFERENCE HALL, VIVEKANANDHAM BLOCK

DAY I - 19.02.2024

INAUGURAL SESSION (10.00 a.m - 11.00 a.m)

TAMIZH THAI VAZHTHU

LIGHTING THE LAMP : DIGNITARIES

WELCOME ADDRESS : **Dr. V. GEETHA**, ASSOCIATE PROFESSOR & HEAD, DEPARTMENT OF CS

HONOURING & RELEASING THE ABSTRACT

PRESIDENTIAL ADDRESS : **Dr. M.V. SRINATH**, RESEARCH ADVISOR, DEPARTMENT OF CS

INAUGURAL ADDRESS : **Dr. N. UMA MAHESWARI**, PRINCIPAL,
STET WOMEN'S COLLEGE (AUTONOMOUS)

FELICITATION ADDRESS : **Dr. K. THYAGESAN**, ACADEMIC ADVISOR,
STET WOMEN'S COLLEGE (AUTONOMOUS)

CHIEF GUEST INTRODUCTION : **MS.R. MANJUPARGAVI**, ASSISTANT PROFESSOR, DEPARTMENT OF CS

KEY NOTE ADDRESS : **Dr. MANIKANTAN POONKUNDRAN**,
DIRECTOR & CLIENT ENGAGEMENT LEADER, CAPGEMINI LTD, CHENNAI

TEA BREAK (11.00 a.m to 11.20 a.m)

CHIEF GUEST INTRODUCTION : **MS.V. MANGAIYARKARASI**, ASSISTANT PROFESSOR, DEPARTMENT OF CS

INVITED LECTURE I : **Dr. JOHN T. ABRAHAM**, PROFESSOR, BHARATA MATA COLLEGE, BHARATHURAI
(11.20 a.m to 12.00 p.m)

TECHNICAL SESSION - I - PAPER PRESENTATION (12.00 p.m to 1.00 p.m)

LUNCH (1.00 p.m to 2.00 p.m)

TECHNICAL SESSION - II - PAPER PRESENTATION (2.00 p.m to 3.00 p.m)

DAY II - 20.02.2024 (10.00 a.m)

WELCOME ADDRESS : Ms. K.SAVINA , ASSISTANT PROFESSOR, DEPARTMENT OF CS

CHIEF GUEST INTRODUCTION : MS.R. AKILANDESWARI, ASSISTANT PROFESSOR, DEPARTMENT OF CS

**INVITED LECTURE II : Mr.M. SIVA SANKAR , ASSOCIATE INFORMATION SECURITY CONSULTANT,
(10.15 a.m to 11.15 a.m) QSEAP INFO TECH PVT LIMITED, BANGALORE**

TEA BREAK (11.15 a.m to 11.30 a.m)

TECHNICAL SESSION - III - PAPER PRESENTATION - (11.30 a.m to 1.00 p.m)

LUNCH (1.00 p.m to 2.00 p.m)



VALEDICTORY SESSION (2.00 p.m to 3.00 p.m)

**WELCOME ADDRESS : MS.C. JASMINE, ASSISTANT PROFESSOR,
DEPARTMENT OF CS**

**VALEDICTORY ADDRESS : Dr. A. PADMAPRIYA, PROFESSOR,
DEPARTMENT OF COMPUTER SCIENCE,
ALAGAPPA UNIVERSITY, KARAUKUDI**

**REPORT OF THE CONFERENCE: Ms. N.SUBHALAKSHMI,
ASSISTANT PROFESSOR, DEPARTMENT OF CS**

CERTIFICATE DISTRIBUTION

**VOTE OF THANKS : Dr. R. ANITHA, ASSOCIATE PROFESSOR,
DEPARTMENT OF CS**

NATIONAL ANTHEM

CONTENTS

S.No	Title	Author	Page No
1.	PROVISIONING COMPUTATIONAL RESOURCES FOR ONLINE BASED E-LEARNING PLATFORMS USING DEEP LEARNING TECHNIQUES	Dr. V.GEETHA	1
2.	DEVELOPING AN EFFICIENT ONE TIME PASSWORD GENERATOR WITH ENCRYPTION AND DECRYPTION FOR SECURE FILE ACCESS IN CLOUD ENVIRONMENTS UTILIZING MODIFIED ARTIFICIAL NEURAL NETWORK	R. MANJUPARGAVI Dr. M.V.SRINATH	11
3.	A SECURE ELECTRONIC HEALTH FRAMEWORK TO PROTECT HEALTH RECORDS USING NATURAL LANGUAGE PROCESSING WITH MULTI LEVEL DATA ENCRYPTION IN CLOUD	N. SUBHALAKSHMI Dr. M.V.SRINATH	24
4.	DEVELOPING NEXT GENERATION CAMPUS USING IOT DEVICES WITH ENHANCED SECURITY	K.SAVIMA Dr. M.V.SRINATH	36
5.	PREDICTION OF DIGITAL CURRENCY ETHEREUM PRICES USING DATA ANALYTICS	R.AKILANDESWARI	44
6.	HYBRID GENERAL ADVERSARIAL NETWORK(HGAN) FOR CLOUD WORKLOAD PREDICTION AND TREND CLASSIFICATION	V.MANGAIYARKARASI	51
7.	APPLICATION USAGE OF DRONES IN AGRICULTURE	R.PRAGADEESHWARI	64
8.	WIRELESS TECHNOLOGY TRENDS FOR 5G	S. ARISTA	69
9.	RESEARCH ON CLOUD STORAGE AND ITS TECHNOLOGIES	Ms. C.TAMILMANI	76
10.	GREEN IOT FOR ECO-FRIENDLY SMART CITIES	SANTHI V	82
11.	A REVIEW OF IMPROVING PRIVACY PROTECTION USING ANOMNIZED	VINAYAKAN K,	89

S.No	Title	Author	Page No
	ROUTING PROTOCOL IN MOBILE AD-HOC NETWORKS	SHEIK ABDHUL KHADIR	
12.	KNOWLEDGE WAREHOUSE: AN ARCHITECTURAL INTEGRATION OF KNOWLEDGE MANAGEMENT, DECISION SUPPORT, ARTIFICIAL INTELLIGENCE & DATA WAREHOUSING	Dr.C.LALITHA	98
13.	AN OVERVIEW OF DATA ANALYSIS AND ALGORITHMIC TOOLS IN EDUCATIONAL DATA MINING	Dr. P. AMUTHA	102
14.	ARTIFICIAL GENERAL INTELLIGENCE: AUTONOMOUS SYSTEMS REACHING HUMAN-LEVEL INTELLIGENCE	Dr.N.SATHYA JEY PRABU	108
15.	BLOCKCHAIN AND ITS CHALLENGES IN REAL ESTATE	Dr. K.UMAMAHESWARI	113
16.	DEEP LEARNING FOR OCEAN OIL SPILL DETECTION	P.SHANTHI, HARI HARA SUDAN A HARISH MRITHYUNJAYAN	117
17.	THE ROLE OF ARTIFICIAL NEURAL NETWORK IN THE PROOCES OF VARIOUS CANCER PREDICTION- A COMPARATIVE STUDY	Dr. SRINAGANYA G SHOBANA Y	120
18.	A SURVEY IN CLOUD COMPUTING USING ADAPTIVEALGORITHMS	GAYATHRI.M Dr. SRINAGANYA.G	126
19.	GENERATIVE AI IN VIRTUAL REALITY	Dr.S.SARANYA, SIVAPRABHA C, SUBHIKSHA	133
20.	SUSPECT PREDICTION CRIME ANALYSIS INVESTIGATION TRACKER THE CRIME	Dr. C. PREMILA ROSY K. NARMADHA	136
21.	A STUDY OF ENHANCING BIOMETRIC AUTHENTICATION TECHNIQUES AND THEIR	Dr.S.JAYAPRAKASH J.P.KEERTHANA	140

S.No	Title	Author	Page No
	APPLICATIONS IN FINGERPRINT SENSORS		
22.	DETAILED INVESTIGATION OF IN-VITRO FERTILIZATION: A SIMULATION-BASED STUDY UTILIZING HYPER-SPECTRAL PARAMETERS AND ADVANCED DATA MODELS	P.MALATHI M.GOMATHI	148
23.	MACHINE LEARNING TECHNIQUES IN EMERGING CLOUD COMPUTING INTEGRATED PARADIGMS	MS.J.VINITHA	157
24.	REAL TIME FACE RECOGNITION AND DETECTION USING DEEP LEARNING ALGORITHM	D. GAYATHRY Dr. R. LATHA	164
25.	THE ROLE OF ARTIFICIAL NEURAL NETWORK IN THE PROOCES OF VARIOUS CANCER PREDICTION- A COMPARATIVE STUDY	Dr. SRINAGANYA G Ms. SHOBANA Y	175
26.	SOCIAL IMPACT OF AI : AN OVERVIEW	Dr. S. ARULSELVARANI M.SARANYA,	181
27.	THE FUTURE OF SURVEY ANALYSIS: AI AUTOMATION IN MARKETING	R.THARANESWARI	186
28.	VISION-BASED ROBUST LANE DEDUCTION AND TRACKING IN CHALLENGING CONDITIONS	B. KAVIYA D. SHANTHINI ²	192
29.	A HEARISTIC APPROACH TO BUILD TRUST MODEL IN PEER TO PEER	V. AKILA	200
30.	A NOVEL APPROACH TO PREDICT BLOOD GROUP IDENTIFICATION SYSTEM	R. JANANI	207
31.	A TRUST FEEDBACK MODEL FOR CONTEXT AND CONTENT PATTERNS GENERATION IN WEB MINING	M. VAISHNAVI DEVI	213

S.No	Title	Author	Page No
32.	ADAPTIVE QUERY GENERATOR USING FACETS IN DATA MINING	A. MANIMOZHI	218
33.	ADAPTIVE WILDCARD RULE CACHE MANAGEMENT FOR SOFTWARE DEFINE NETWORKS	S. DHANARAKSHANA	227
34.	AN IMPROVED OCCLUSION THERAPHY FOR AMBLYOPIC PATTERN TECHNIQUE	R. YOGA	234
35.	CONGESTION CONTROL MECHANISM FOR BACKGROUND DATA TRANSFERS WITH LATENCY REDUCERY	P. SATHYAPRIYA	241
36.	CONSTRAINT-BASED TEMPORAL TASK SCHEDULER FOR PROFIT MAXIMIZATION	D. SWETHA	247
37.	DETECTION OF NODE FAILURE TECHNIQUES USING CHURN RESILIENT PROTOCOL	K. ABARNA	253
38.	ENABLING IDENTIFIED-BASED AUDITING	P.PUNITHA	260
39.	IMPROVED PRIVACY PRESERVING CONTENT BASED RETRIEVAL IN CLOUD RESPOSITORIES	S. ROJA	268
40.	ONLINE CRIME REPORTING AND MANAGEMENT SYSTEM	K. SARATHA	274
41.	RANK BASED FRAUD DETECTION FOR MOBILE APPLICATION	B. SIVAPRIYA	280
42.	REGULATING DOCUMENT STREAMS ON TOP-K USING MONITERING SYSTEM	S. SOWMIYA	287
43.	SENTIMENT ANALYSIS OF FOOD REVIEW USING NEURAL NETWORKS	M. KEERTHANA	293

S.No	Title	Author	Page No
44.	SMART CARD BASED PATIENT HEALTH MONITORING SYSTEM	G. VINTHIYA	301
45.	SMART GRID ELECTRICITY PRICE FORECASTING IN BIGDATA ANALYTICS	K. ABI	307
46.	A COMPREHENSIVE FAULT PREDICTION BASED ON K-MEANS CLUSTERING ALGORITHM	M.NALINI	313
47.	A NEW INTELLIGENCE-BASED APPROACH FOR COMPUTER-AIDED DIAGNOSIS OF DENGUE FEVER	C. MONISHA	320
48.	A NOVEL APPROACH TO PREDICT BLOOD GROUP IDENTIFICATION SYSTEM	R. JANANI	326
49.	ADVANCED DETECTION AND MONITORING SYSTEM FOR SPAM ZOMBIES IN COMPUTER NETWORKS	N. APOORVAPRIYA	333
50.	AN APPROACH TOWARDS COEFFICIENTS BASED LOSSY COMPRESSION AND ITERATIVE RECONSTRUCTION IN IMAGE PROCESSING	B. SATHIYASRI	341
51.	AN ATTRIBUTE BASED CLUSTER MODEL FOR FEATURES SUBSET SELECTION ALGORITHM IN WEB MINING	N. NANDHINI	347
52.	APPROACH OF SECURED FILE DETECTION USING DATA MINING	M. BHUVANESHWARI	355
53.	AUTOMATED CHATBOT ASSISTANT SYSTEM USING ARTIFICIAL INTELLIGENCE	E. AARTHIKA	361
54.	BLOCK MISBEHAVING MODEL ANONYMIZING NETWORKS	P. SWETHA	368

S.No	Title	Author	Page No
55.	BUDGET BASED SECURE ROUTING PROTOCOL DESIGN FOR WIRELESS SENSOR NETWORKS	G. VISHALINI	374
56.	CLASSIFICATION OF SKIN CANCER DETECTION USING MACHINE LEARNING	S. ADHISRI	381
57.	CROWD MANAGEMENT USING AUTOMATED PATH FINDING ROBOT	A. SARIKA	387
58.	CUSTOMER SENTIMENT ANALYSIS BASED ON LATENT DIRICHLET ALLOCATION (LDA) TECHNIQUE	M. SIVASAKTHI	393
59.	DETECTION AND CLASSIFICATION OF FRUIT DISEASES USING IMAGE PROCESSING TECHNIQUES	R. SOORYA	400
60.	DETECTION OF MENTAL DISORDERS IN SOCIAL MEDIA NETWORK	R. ELAKKIYA	409
61.	EMBED IMAGE ENCRYPTION USING 126- BIT BASE RANDOM BIT SHIFT TECHNIQUE	S. KANNIKA	417
62.	FINANCIAL ANALYSIS AND LOAN PREDICTOR USING DATA MINING TECHNIQUES	M. KOWSIKA	425
63.	HEART DISEASE PREDICTION USING CONVOLUTIONAL NEURAL NETWORK	C. AARTHIKA	432
64.	IMAGE FORGERY DETECTION USING CONVENTIONAL NEURAL NETWORK IN NEURAL NETWORKS	M. SNEHA	439
65.	IMAGE RE-RANKING BASED GEOSPATIAL DATA NETWORKS	T. TAMILNILAVU	446

S.No	Title	Author	Page No
66.	IMPLEMENTATION OF COST EFFECTIVE MULTI CLOUD STORAGE IN PUBLIC CLOUDS	M. SAMEERA BEGOM	454
67.	IMPROVED MECHANISM FOR FAKE CURRENCY DETECTION	S. HARISREE	461
68.	IMPROVED SOCIAL MEDIA OPTIMIZATION (SMO) AND SPAM DATA FILTERING IN SOCIAL NETWORKS	S. VALLEESWARI	468
69.	INTEGRATED QUERY REORGANIZATION PROCESS FOR EFFICIENT BOOLEAN INFORMATION RETRIEVAL	T. SHAMILA	474
70.	MULTISTAGE TRUST MECHANISM IN PRIVACY PRESERVING DATA MINING	A. NIVETHA	481
71.	NUMERICAL OPTIMIZATION ON CENTROID BASED ACTIONABLE 3D SUBFACE CLUSTERING	R. THILAGAVATHI	487
72.	OPTIMIZATION OF RESPONSE TIME OF M-LEARNING COMPUTING ENVIRONMENT USING FIRE FREE APPROACH	V. DIVYA	496
73.	PARALLEL JOB SCHEDULER WITH REPLICATION STRATEGY IN GRID COMPUTING	G. MANISHA	504
74.	PATTERN BASED ASSOCIATION RULE MINING	G. SANTHAPRIYA	509
75.	PRODUCT ASPECT RANKING AND ITS APPLICATIONS USING SENTIMENT CLASSIFIER FRAMEWORK	S. DHARANISRI	517
76.	PUBLIC RELIABILITY AND AUTHENTICATION FOR CLOUD	V. ABARNA SREE	523

S.No	Title	Author	Page No
	RESOURCE AGAINST AUDITORS BEHAVIOR USING BLOCK CHAIN		
77.	QUALITY OF SERVICE AWARE WEB SERVICE RECOMMENDATION IN CLOUD NETWORKS	A. NIVETHA	531
78.	RELAY NETWORK BASED QUERY PROTOCOL RECOMMENDATION IN UNSTRUCTURED PEER TO PEER NETWORKS	M. NANDHINI	537
79.	SERVICE LEVEL OBJECTIVE MODEL FOR CLOUD STORAGE SERVICE ACROSS DISTRIBUTED CLOUD ENVIROMENT	R. PARKAVI	547
80.	SOFTWARE VULNERABILITY CLASSIFICATION USING SVM AND KNN TECHNIQUES	R. GOMATHI	554
81.	TWEET DATA SEGREGATION AND SEGMENTATION USING BATCH MODE PROCESS	M. CHITRA	560
82.	INTERNET OF THINGS	V ABINAYA	567
83.	ADVANCEMENTS IN NATURAL LANGUAGE PROCESSING	B.SUBASRI	574
84.	A SURVEY ON ROBOTIC PROCESS AUTOMATION	S.AAKASH	581
85.	A SURVEY OF INTERNET OF THINGS	S.AARTHI	585
86.	A SURVEY ON IOT FLOOD MONITORING & ALERTING SYSTEM	M.ARTHI	589
87.	A SURVEY ON DECISION MANAGEMENT GOVERNMENT	N.ARUNA	593
88.	A SURVEY ON ARTIFICIAL INTELLIGENCE	S.AYYAPPAN	597

S.No	Title	Author	Page No
89.	A SURVEY OF NATURAL LANGUAGE PROCESSING TECHNOLOGY	S.M.BALAJI	601
90.	A SURVEY ON DATA ANALYTICS	T.DHARANI	605
91.	A SURVEY ON SPEECH RECOGNITION TECHNOLOGY	M.GOPINATH	609
92.	A SURVEY ON DATA ANALYTICS USING BIG DATA TOOLS	J S JAYASURYA	613
93.	A SURVEY ON ARTIFICIAL SUPER INTELLIGENCE (ASI)	G.JAYASRI	617
94.	A SURVEY OF POWERFUL VIRTUAL AGENTS	B.KALAIYARASI	621
95.	A SURVEY ON NATURAL LANGUAGE GENERATION (NLG)	S.KAVIYA	624
96.	A SURVEY ON CYBER SECURITY	KAVYA ARJUNAN	627
97.	A SURVEY ON ARTIFICIAL INTELLIGENCE	R.KEERTHANA	631
98.	A SURVEY OF ATTENDANCE MANAGEMENT SYSTEM USING FACE RECOGNITION	S. KIRANISHA	635
99.	EDGE COMPUTING	K.KIRUTHIGA	638
100.	A SURVEY ON NATURAL LANGUAGE PROCESSING	G.MADHUMITHA	641
101.	A SURVEY ON EXPERT SYSTEMS FOR CONSTRUCTION PROJECT	K.MALATHI	644
102.	A SURVEY ON MACHINE LEARNING PROJECT	NANCY.D	648
103.	A SURVEY ON BIOMETRICS	R. NISHANTHINI	651
104.	A SURVEY OF PEER-TO- PEER	S.PRIYANKA	655
105.	A SURVEY OF DEEP LEARNING PROCESSING	M.RAJALAXMI	658
106.	A SURVEY ON NAVIGATING THE WORLD OF ARTIFICIAL INTELLIGENCE	V.RAJESH	661
107.	A SURVEY ON AI- ARTIFICIAL INTELLIGENCE	T.R.VARSHIKA	664

PROVISIONING COMPUTATIONAL RESOURCES FOR ONLINE BASED E- LEARNING PLATFORMS USING DEEP LEARNING TECHNIQUES

Dr.V.GEETHA^[1]

^[1]Associate Professor & Head, PG and Research Department of Computer Science, S.T.E.T Women's College (Autonomous), Mannargudi.
Email: kkmannai@gmail.com

Abstract— Research focus on E-learning, or numeral erudition receipts convert gradually prevalent in current eons as a method to distribute erudition plus preparation using numeral skills besides the internet. Provision organization provision aimed at e- learning takes enthused on the online. For the purpose, that effectual provisioning of the possessions aimed at such platforms. This project offerings an archetypal for forecasting the practise of computational possessions for the e-learning podiums Legislative Classification, Deterioration Procedure there are castoff of this provisioning computational possessions on behalf of online-based e-learning podiums consuming deep learning techniques. Cutting-edge this both pupil performed an arrangement of the activities same a theory besides real-world while linked toward server and application. In this offered perfect achieves high accuracy. The imminent new impression on this mission is open, interoperable e-learning gateway construction (OEPortal), consuming e-learning schemes incorporation expertise to the organisation running test consequences display that to the pupils can browse resources over their mobile terminal devices. Aimed at this unfathomable erudition methods stay castoff this application achieves to the predictable design goals, besides this one stays conducive to improving that excellence the coaching E-learning scheme incorporation expertise designed at races, workflows, demonstrations besides the real-world lab. This E- learning podium helps near conjecture to the slow apprentices in particular domain besides of recommend to give more importance proceeding the respective domain.

Keywords— Electric Appliances, Channel Controlling Bluetooth, Random Forest, Supervised Technique.

INTRODUCTION

An edification scheme established on dignified coaching then thru of the aid on automated capitals are an acknowledged as E-learning. Whereas coaching be able near erected hip before of the show on the coaching space, practise to the mainframes besides the Intranet customs the foremost element on E- learning. E-Learning, or else microelectronic wisdom, there nigh be transmission on culture then drill concluded numeral properties. Even though eLearning stays an established of pompous knowledge, that stays a lengthy as concluded microelectronic plans are same as mainframes, medicines besides smooth cellular receivers there is attached near on intranet. The primary modification amongst eLearning and wired tutoring are the aggregate on communication. Pupils besides trainers cooperate extra through accessible culture, whereas eLearning is a further self-paced. In an unpretentious dialectal, e-learning is a demarcated voguish abode of -erudition that stays aided automatically. Stereotypically, e- learning stands showed of the Intra-net, someplace apprentices stay capable to entrance their wisdom things online next to some dwelling besides while. Through e-learning, apprentices be capable to procure by our peculiar stride, beginning somewhere then on some period. Finished conveyance approaches same a knockouts plussocietal broadcasting, e- learning likewise varieties to the wisdom progression extra immersive besides cooperative. Furthermore, e-learning empowers moderately speedier passage rotations. E-learning be capable to yield several unlike arrangements, counting online passages, cybernetic laboratories, webinars, alphanumeric replications, edifying knockouts, then itinerant wisdom submissions. That one permits beginners near admittance enlightening possessions besides act together thru tutors plus nobles since everyplace to the ecosphere, on every while. Habitually, an apprentice's stay simplified near entree verified sermons gamble there lost a lecture. Numerous academies bargain online passages that partake lithe then calm measures. Cramming popular schoolrooms necessitates succeeding an appropriate tedious. Apprentices devise near stay prompt whereas joining programs.

Objective of the study : By leveraging workplace technologies, e-learning is bridging the gap between learning and work. Workers can integrate learning into work more effectively because they use the same tools and technology for learning as they use for work.

LITERATURE REVIEW

Extracting the focal pieces of e-Learning gameness assessment on behalf of Iraqi universities. Now the post- corona epoch, Iraqi academes executed intermingled wisdom in 2021 and 2021. However, given the uncertainty regarding the pandemic's potential recurrence or the country's acquaintance near another crunch of a similarnature affecting the enlightening process, it is required to weigh Iraqi universities' genuine probable to approve thee-learning structure as a substitute educational system? The persistence of that sense stays to mental the basic aspects plus the extents and features to weigh the gameness of e-learning flat arranged behalf of Iraqi academies. To fix this, the Fuzzy Delphi Method (FDM) was practical to quotation the vibrant facets for gaging e-learning inclination, and lastly, 2 proportions and 12 aspects take stood extracted from Iraqi specialists in e- learning and didactic scheme. The fallouts of that tabloid exhibited that "Infrastructure" has essential upshot and extra imperative than other dimensions, and "Technological" factor over a bulk of 0.751 takes the utmost outcome proceeding e-learning eagerness than the breather factors. Today, the e- learning scheme stays vital near the instructive system [1].

The Adeptness of E-Learning Deal Superiority hip Swaying E-Learning Pupil Gratification plus Devotion next to Telkom University. Merged wisdom process concluded the DMS arrangement by Tel Campus is quite fresh. This varieties an immense revolution hip the culture route, exclusively thru the Co-19 plague which involves available wisdom. Online culture afore e-learning rears aces too ploys by Tel Campus. Is revision habits an assessable mode and fate thru SsM PMS as per the inquiry practise. Specimen skills castoff are purposive specimen thru an illustration of 334 active Tel Campus pupils spread thru very majors starting since the Class of 2015 - 2020. The The verdicts display that the calibre if e- learning offerings affects its performance for the e-learning platform, overall competence of instructors plus content for e- learning, its worth of folks among automated culture aids,plus others. The calibre for virtual tutoring offerings takes an 86.8% upshot on the happiness of the Tel Campus e- learning consumers. In Tellkom Campus pupils, e-learning operator happiness profits a strong effect goingarranged e-learning operator loyalty by 90%. The custom over very the e-learning arrangement determines the fulfilment besides allegiance of TeLlkom Campus apprentices for the cause that the tactic of philosophy is approved vacant done DMS thus that students experience easy admission toward supplies also absorb independently. The expansion of Sign plus Announcement Expertise (EAE) takes pretentious all aspects, single of which stays the biosphere of tutelage. Solitary are the wisdom mock- ups is developing the Evidence also Announcement Equipment (EAE) based wisdom exemplary that has many terms same as per electronic, web- based, online, plus detachment culture. Telkom Academy stays solitary of the Gen Technology-based Universities that adapts en route for the e-learning erudition prototypical consuming the WAS (wisdom administration structure) platform..

A Comparative Analysis between Data-Driven vs Data- Centric Buildings With E-Learning Solutions. To diminish the stride of transmission are the original Covid-19 virus, institutions have shifted toward e-learning to substitute lectures plus duties hip the laboratory minus presence fully ready and technologically equipped. By tactic of an outcome, institutions must identify and pay the utmost suitable facts manner castoff for their universities. For this resolution, this revision major finds 119 e-learning elucidation habit bags, gathers their 963 consumer journals later the e-learning commerce, plus formerly tags them rendering near their facts buildings for fitting link thru the settled intangible agenda. The verdict spectacles that data-driven plus data-centric stood the solitary buildings castoff via e-learning keys, it auxiliary acclaims data-centric same as per a top apt for e-learning. For take the co-19 rampant, e-learning takes lately gained popularity gained popularity.

However, there takes stood around later 1998 [1]. Many canvassers take defined e- learning, then the American society of drill plus education's definition are a supreme regularly used. It outlines E-learning stays distinct that the practise of pupils knowing thru the consumption of interactive media e.g., the interanet, corporate networks, computers, satellite broadcasts, audiotapes, videotapes, interactive television, plus solid disk [2]. A more recent definition is by [3], who defines e-learning for specimen a category of wisdom vogueish which technology mediates

the wisdom course, tutelage are delivered entirely done the internet, plus apprentices plus tutors stays not obliged near be available thru the alike while too location. Per the dawn exist the fresh Covid-19 contagion, institutions stood required near close all direct wisdom facilities and transition near e-learning vogue abode of the solitary realistic alternative for continuing the nation manner while limiting the virus's spread [4]. The critical change fallouts modish a noteworthy surge hip ordinal wisdom files.

Faculty members including disaster e-learning: Findings after a Polish trial study examining views on e-learning, e-learning events, or plans for utilising e-learning techniques thru the inclusive epidemic. The study's goal hunted to depict the realities from emergency e-learning in Polish thru the viewpoints of deans at universities. The reading stole abode with a web-based analysis between 94 workers of Poland's largest public pedagogic institution. The poll utilised measures to weigh views on e-learning, expectations of e-learning habit formerly the epidemic had ended, as fit as familiarity wit e-learning erstwhile to fluid to disaster e-learning. The revision stole out during April and November of 2020, & its quantitative findings is below: 1) A hefty percentage of edifying professors sight the prospects provided thru e-learning favourably. 2) A popular of the partakers believe that integrated edification is their most appropriate method. 3) Nearly a third of the whole respondents state that e-learning agrees entirely are the established wisdom intentions near stay achieved; 4) Evaluation are e-learning stays not a homogeneous phenomenon, so teachers differentiate between elements linked a assessing are worth the e-learning; 5) The merely feature that differentiates the helpfulness are the valuation a e-learning stays seniority (weak correlation); 6) Thru the survey, over two-thirds of folks who participated having no previous involvements pending e-learning outbreak are the pandemic; 7) Evaluation Mid the sorts for emergency e-learning includes an influence upon the utilisation thereof e-learning-based solutions later a pandemic. To address these issues, institutions stood required near pick between two files buildings (FB) that could see the demands of e-learning: data-driven design (DDD) and data-centric building (DCB), raising a query in which a top suited by e-learning. The co-19 illness triggered exponential vagaries in a lives of individuals by fit as social groups.

Related Work

E-Learning provides scalability which helps in providing training. All students can receive the same type of syllabus, study materials, and train through E-Learning. Through E-Learning, you can save time, and money and reduced transportation costs. So, E-Learning is cost-effective compared to traditional learning. This article is structured as follows: following a brief discussion on e-learning a literature review results of some previous studies on e-learning from different parts of the world are presented. This is followed by exploratory research into perceptions and intentions of students regarding e-learning. Besides, the results of focus groups results of survey among students are discussed. In the end implications, limitations and opportunities for future research are explained. This study uses mixed methods to obtain answers to the research questions. First, results of two focus groups are reported, followed by results of survey on 104 respondents. There are two broader research objectives pursued in order to answer these research questions. First, based on empirical study involving students in higher education identify students' knowledge and perception of e-learning, along with their attitudes and experience with it. Second research objective is to assess readiness of students to engage in e-learning and determine their willingness to pay for it.

For are M =Database shaped by smearing piercing found to M ;

If ending opinion touched pointed by the pathway, formerly S' = make greenery lump and tag with suitable class;

Else S' = DTBUMILD(M);

S = add S' to arc;

METHODOLOGY PRE-YEAR METHODOLOGY

1.A. Approach

The Verdict Shrub taxation stays lone have the procedures that might stay castoff for unproven erudition. Deterioration and cataloguing questions potency stay cracked spending the assessment, dissimilar others. It stays imaginable near exist theory a guess archetypal by consuming had Verdict Shrub (training data). When envisaging

a best's class marker consuming verdict bushes, it flinch through the source stay the tree. This single stays ended by linking the crib feature's charge thru the evidence's feature. The ensuing lump is grasped by subsequent the outlet accompanying thru the charge.

B. Algorithm: THE DECISION OF TREE ALGORITHM Input:

Exercise Dataset = J Output:

Conclusion Tree =T Steps:

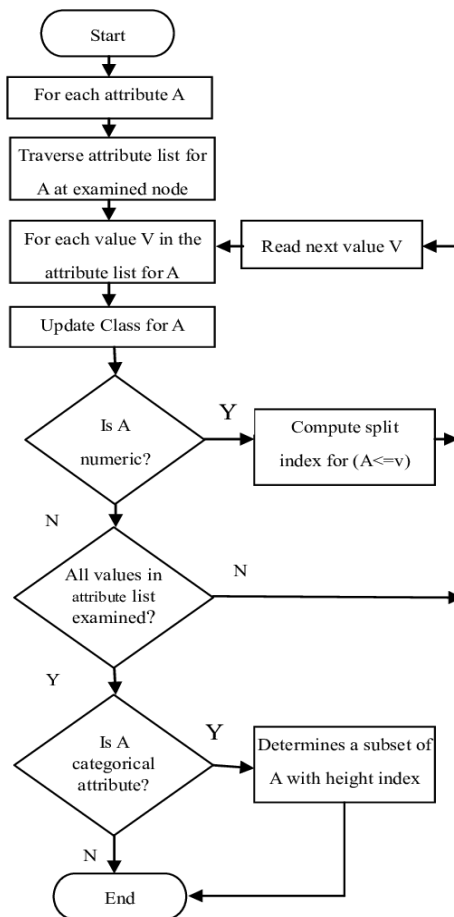
DTBUMLD(*D)

J=φ;

S=Make root manner and sticker with piercing feature:

S=Enhance stay to cause bulge for every divided establish and tag;

C. Architectural View



4.1 The Decision Of Tree Algorithm

2.A. Approach

Naive Bayes is exploited here an extensive variability stay an institute difficulties. Created on Bayes' proposition, this one is an ordering attitude. Owing near its effortless and usefulness, the Naive Bayes Classifier simplifies arranged the making of mechanism wisdom reproductions that exist speedy and precise here the wildlife. Equally an upshot, the taxation ensures not trust arranged an only progression toward remain organise structures. These practises of slice exist a longer intimate. . Though, in practical subjects, now stay regularly several X influences near stay deliberate. Bayes Tenet valour be long toward pardon is branded as Naive Bayes next a physiognomies are self-governing. Here the situation, the X's are supposed near be autonomous of unique extra, hereafter the stretch "Naive". The site exists a durable method, no substance what it's baptised.

B. ALGORITHM: NAIVE ON BAYES ALGORITHM

INPUT

TEACHING DATASET T.

$F=(L_1,L_2,L_3,\dots,L_M)$

//WORTH EXIST AFORECASTER MUTABLE IN DIFFICULT DATASET.

OUTPUT

A LECTURE OF DIFFICULT DATASET.

STEP

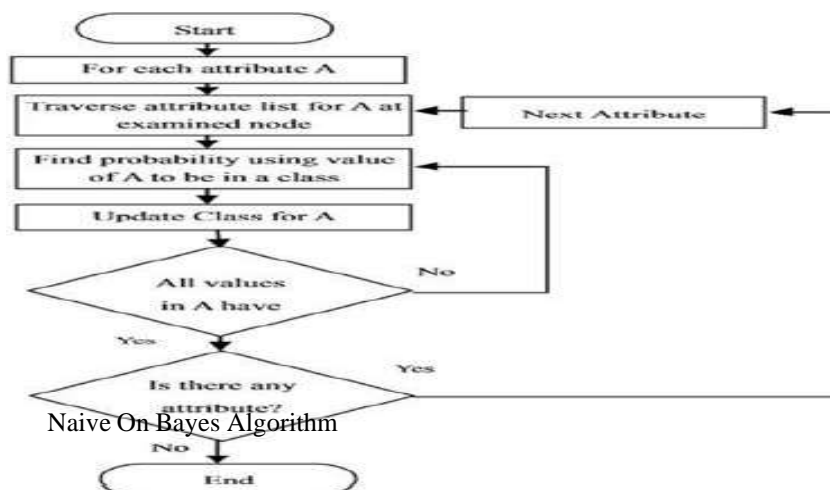
1. RECITE THE EXERCISE DATASET R;
2. ANALYSE THE UNKIND AND NORMAL ECCENTRICITY OF A PREDICTOR VARIABLES HERE A PIECE CLASS;
3. REPEAT

ESTIMATE THE LIKELIHOOD HERE L_1 CONSUMING THE GAUSS DENSITY EQUATION HERE A PIECE CLASS;

PENDING THE LIKELIHOOD OF ENTIRELY PREDICTOR VARIABLES (L_1,L_2,L_3,\dots,L_M) TAKES STOOD CALCULATED.

4. GOVERN THE PROSPECT FOR EACH CLASS;
5. SECURE THE UTMOST LIKELIHOOD;

C. Architectural View



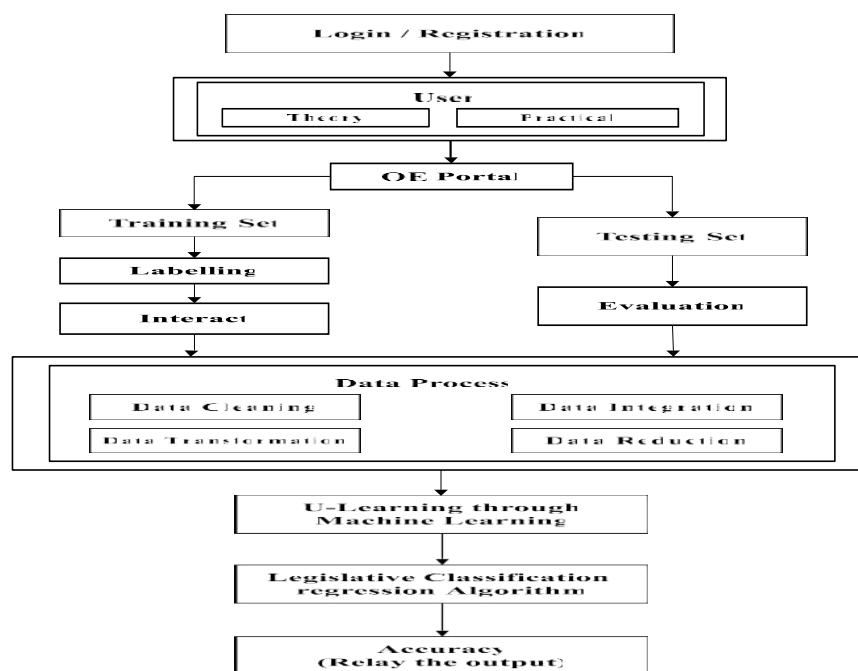
3. A. Approach

Voguish the rejoinder exist the overhead hitches, it are bestowing an innovative inkling arranged the exposed, interoperable U- Learning gateway construction (OE Portal), consuming U-Learning scheme addition expertise castoff their the amenities, workflows plus formerly appearances. A Human-Computer Interaction (HCI) based apparatus wisdom attitude under Exposed plus interoperable U-Learning porch manner (OE Portal). The residue stay the investigation stays ordered equally per shadows Human-Computer Collaborationconstructed appliance learning are discussed in presents of solution. The stay information which partake stood uploaded near the u-learning co-ordination can sincerely stay castoff while plus again. The practise of repetitive wi-fi devices live records discern how near remain clean to do per the practice of reinvention. It makes are the progression of tracking the prevailing issues tough. Hip tallying the motive of a guise by is near bargain are accurate technique hip mechanism erudition facsimiles in classifying wi-fi gadgets stay information consistent with topic categories plus subtopics correctly and quickly. So, are inquiries in u-mastering discernhow near stay identified as per returned precisely as wished mechanically. research is toward discover the exact technique to procedure the version castoff their apparatus studying insidethe process was classifying questions or pupil monitoring or devices accurately.in edict near utilized hip abundant gainingknowledge of could stay located easily and quickly.

B. Algorithm: THE CLASSIFICATION OF QUESTIONS MACHINE LEARNING FOR THROUGH YOU LEARNING APPROACH

- Step1: Live Facts N equally Input.
- Step2: Assemble records starting Tools insider or Arrest $S=0$.
- Step3: Infinite numeral of wireless devices oftenly occurs.
- Step4: me Equivalent just before zero, I Less then N and Aincrement.
- Step5: First Pullman the statistics then labeling a figures.
- Step6: Testing the records $S=S+I$.
- Step7: WHEREAS P Countless than N.
- Step8: End Radio Entrance Resistor than Established gesture metier Suggestion Fewer than one.
- Step9: ELSE Deal Fixed Identifier than Conventional IndicatorMetier Suggestion identical near one.
- Step10: Categorize the Preprocessor using Appliance Wisdom Attitude like Records Attack, Records Assimilation, Records Makeover plus Files Reduction.
- Step11: Convert Extensible markup verbal to Interpersonal Catalogue Organization System.
- Step12: Any where besides several while records tell how near halt tracking hip global Learning.
- Step13: If Dawn assessment Minus than five, Interruption Uncovering Scheme Found.
- Step14: ELSE Inception Significance countless than five,Imposition Recognition Scheme Found.
- Step15: Correctness of graphical yield spectacles the Result.

C. Architectural View



PROPOSED METHODOLOGY

A. Approach

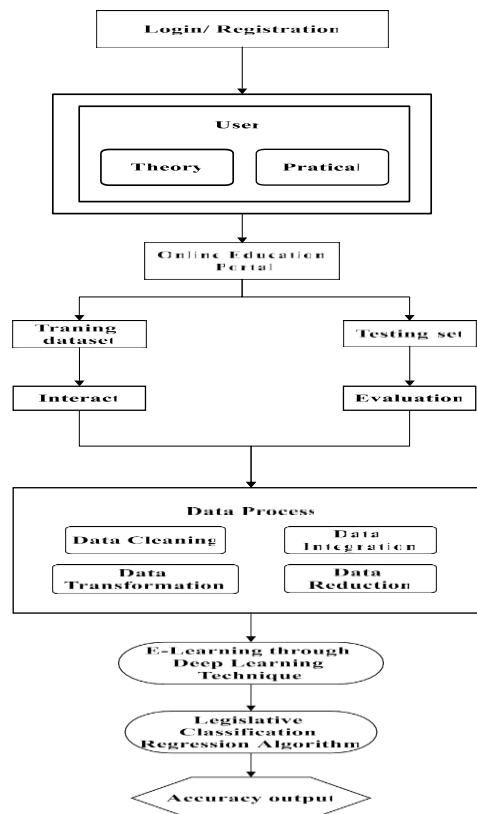
Deterioration Procedures stay castoff thru nonstop facts. Classification Isolated facts stays utilised with computers. Theroute thru the chief fit in extrapolation stays one that exists peak accurately guessing the outcome. When classifying data,we look their on choice of border that caught separate the fixed of facts mad nearly separate categories. Continuous quantities are produced via categorisation. Constant values are produced using extrapolation. This technique aids hip the division of gen among several categories supplied an agreed had variables. It Extrapolation stays one sort of controlled wisdom practice in DI which helps in bridging a likelihood relationship among labelling plus the evidence pieces; the peak usually said regression ways hip ML employ one utility called the mapping function to translate variables to output that stays continuous are linear, polynomial, logistic, stepwise, etc. Particular algorithms have the word -regression in their name, such as direct relapse plus logistic relapse, which caught make thingsconfusing because direct lapse stays a lapse procedure whereas logistic relapse stays one sorting procedure.

B. Algorithm: REGRESSION OF CLASSIFICATION ALGORITHM

```

For i=0 toward p - 2 do
adjust  $\gamma_i$  and  $\beta$  so that survival speed on vigorous-i files is  $q_i = 2 - g/h$ 
for i=1 near k ensure
set 5 near stay a maximum vigorous of queries choosing  $Dic[i]$  for  $l=1$  to  $\max \gamma_i$  do
if  $j \leq \gamma_i$  then  $M[i,j] = E_{pk}(1)$  else
 $M[i,j] = E_{pk}(0)$  FileFilter(run thru a cloud)
foreach file  $F_j$  hip was haze do for  $k=1$  near sail through  $\gamma_i$  do for  $i=1$  near L ensure
 $C_{j,k} = \{ [Dic[i] \in F_j \mid M[i,k]; e_{j,k} = C|F_j|/(j,k) \}$  Map  $(C_{j,k}, e_{j,k})$  once near a barrier of size  $\beta$ 
    
```

C. Architectural View



EXPERIMENTAL RESULT

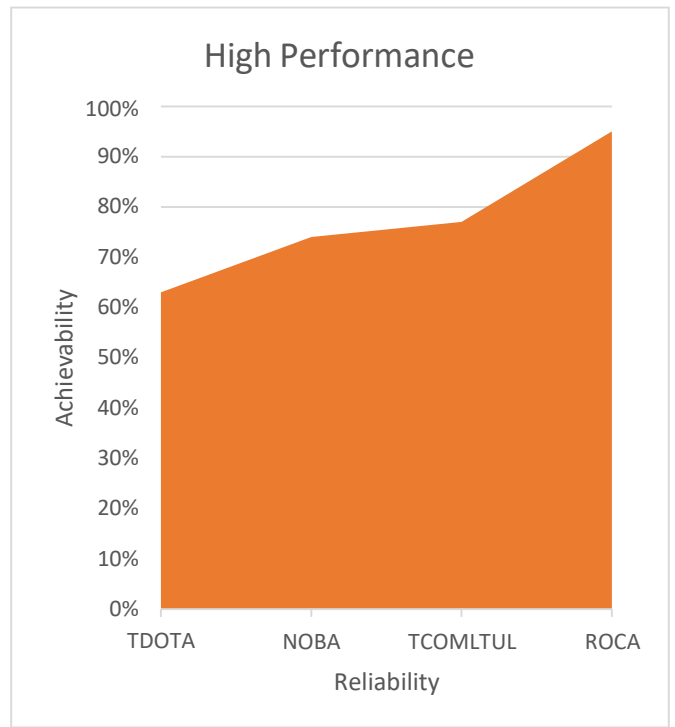
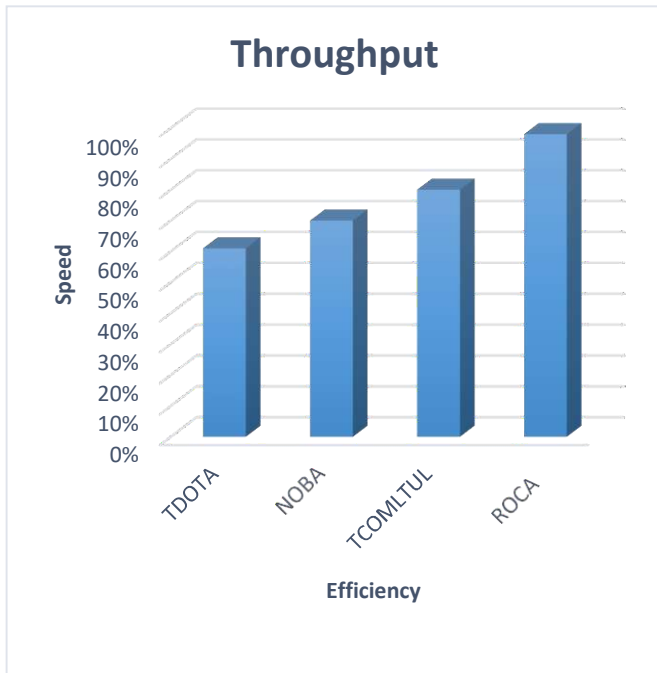


Fig. 5.1 Throughput & Fig 5.2 Cost

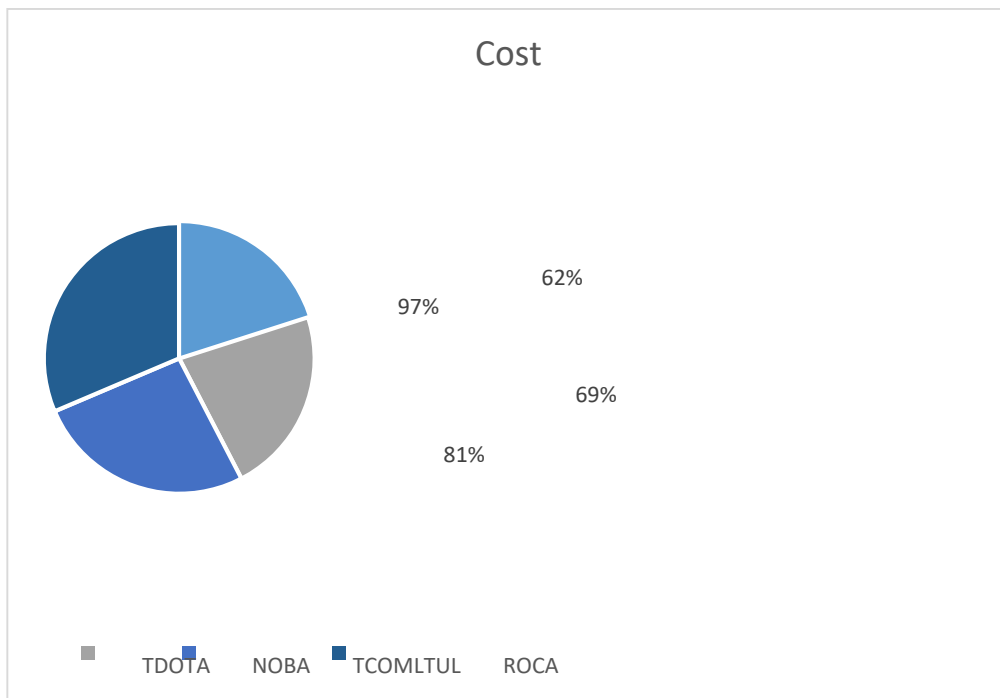


Fig. 5.3 High Performance

CONCLUSION

Based on the above study, it can be concluded that

- Hip assumption, one mission "Provisioning Computational Properties castoff their Cloud-Based E-Learning Podiums By resources on Profound Wisdom Practices" discourses has want for scalable hip tallying effectual reserve administration for accessible wisdom podiums.
- In mission proposes the habit of yawning wisdom skills near elevate resource allocation, diminish outlays, plus expand the inclusive routine nearly e- learning podiums.
- By leveraging the clout of yawning wisdom, in venture wishes near afford one further seamless and high-quality user experience their virtual learners, ultimately enabling other folk's adjacent ingress feature tutelage wired.
- These venture takes the likely near revolutionizethese manner e-learning panels' provision computational resources, making virtual tutelage more accessible, affordable, and effective.
- Overall, the habit of yawning wisdom practices castoff these endowment computational funds there cloud-based e-learning boards has significant allegations castoff in place of preek on future of wired tutelage, plus the venture are of footstep hip adetailed track towards achieving an added capable and scalable e-learning eco system.

FUTURE WORK

- ✓ Based on the studies carried out, the following areas have been identified for future research
- ✓ Now summary, tutelage devises a bright future plus resolve see be ration of fascinating new innovations.
- ✓ The online instruction milieu are graceful near variation as the effect on AI, VR/AR, small-scale learning, gamification, which peer wisdom, plus collaborations more fetching, cooperative, plus personalized than endlessly earlier.
- ✓ E-learning consumes stood annoying on behalf of centuries currently near foil a tactic we absorb toward brand this one extra current plus measurable. With edification past an interanet, both their instructor plus the pupil may choose our specific speed for instruction, where worth near are likewise the additional freedom of choosing a timetable there the whole thing designed them everyone's agenda.
- ✓ There an upshot, consuming a wired instructive podium tolerates in lieu fair a thriving balance of drudgery plus studies, so there's no must toward give anything up.

REFERENCES

- [1] Extracting the main aspects of e-Learning readiness assessment for Iraqi universities: Yasser Kareem Al-Rikabi, Gholam Ali Montazer_2023
- [2] The Effectiveness of E-Learning Service Quality in Influencing E-Learning Student Satisfaction and Loyalty at Telkom University: Marheni Eka Saputri, Fitriani NurUtami, Devilia Sari_2022
- [3] A Comparison of Data-Driven and Data-Centric Architectures using E-Learning Solutions: Ismaila Idris Sinan, Vivian Nwoacha, Jules Degila, Saidat Adebukola Onashoga_2022
- [4] University teachers and crisis e-learning : Results of a Polish pilot study on: attitudes towards e-learning, experiences with e-learning and anticipation of using e- learning solutions after the pandemic: Łukasz Tomczyk, Katarzyna Potyrała, Nataliia Demeshkant, Karolina Czerwiec_2021
- [5] A Scalable Heterogeneous Big Data Framework for e- Learning Systems: David Otoo- Arthur, Terence L. van Zyl_2020
- [6] Methodology for the Development of an Ontology based E-Learning Platform: Monika Sharma, Supriya

Lamba Sahdev, Gurinder Singh_2020

- [7] Effective E-learning utilizing Internet of Things: Mohammad Hadi Zahedi, Zeinab Dehghan_2019
- [8] Visually Enhanced E-learning Environments Using DeepCross-Medium Matching: Mozhdeh Dokhani, Babak Majidi, Ali Movaghar_2019
- [9] Relationship modeling for OSN-based E-Learning Deployment: Quadri Noorulhasan Naveed, Naim Ahmad, Shamimul Qamar, Nawsher Khan_2019
- [10] Deep E-School-Nurse for Personalized Health-Centered E-Learning Administration: Tannaz Karimi, Babak Majidi, ohammad Taghi Manzuri_2019

DEVELOPING AN EFFICIENT ONE TIME PASSWORD GENERATOR WITH ENCRYPTION AND DECRYPTION FOR SECURE FILE ACCESS IN CLOUD ENVIRONMENTS UTILIZING MODIFIED ARTIFICIAL NEURAL NETWORKS

R.MANJUPARGAVI^[1] Dr.M.V.SRINATH^[2]

^[1]Research Scholar & Assistant Professor , ^[2]Research Supervisor ,S.T.E.T Women's College (Autonomous), (Affiliated to Bharathidasan University, Tiruchirappalli)Sundarakkottai, Mannargudi - 614016, Thiruvarur Dt., Tamil Nadu, India

^[1]Email : mprithishni5@gmail.com Mobile No. 9789072371 ^[2]Email : sri_induja@rediffmail.com Mobile No. 9710944476

Abstract- Cloud computing is a advanced area of Information and Communication Technology (ICT) that enables the storage, access, and execution of data, programs, and other information-related services over the Internet. However, the objective of this research paper is to develop a security method using fingerprint recognition to identify a person before accessing cloud services. Fingerprint recognition or fingerprint authentication refers to the automated process of verifying the match of two human fingerprints. Fingerprints are unique, reliable, and relatively easy to obtain. This study explores fingerprint recognition for personal identification. Gabor filtering, ridge region segmentation, normalization, morphological thinning, and local orientation estimation of fingerprint ridges are all used in approaches to improve fingerprint identification images. This OTP (private key) generation or creation mechanism can be used to protect documents from unauthorized users. Second, it provides data encryption to protect your data files from cloud providers. To encrypt files, the encryption library uses a symmetric approach based on the same key for encrypting and decrypting files. The Cryptography package introduced in this work provides integrated functionality for key generation, ciphertext encryption, and ciphertext decryption using both encryption and decryption methods. The Fernet module ensures that data encrypted with the Fernet module cannot be edited or read in the absence of the key. This makes it easy to share public data with other authenticated cloud users. The processing time for secure authentication with 4-digit, 6-digit, 8- digit, and 10-digit OTP creation is reduced to 0.0016, 0.0028, 0.0035, and 0.0042 seconds respectively, and encryption and decryption such as reverse cipher, Caesar cipher, and ROT13 Algorithm elapsed time is 0.00025, 0.00065, 0.00207 seconds.

Keywords: Fingerprint recognition, Cloud, Security, Authentication, Minutiae, One Time Password (OTP)

INTRODUCTION

Cloud is a rapidly growing field that "allows pervasive, convenient, on-demand network access" by utilizing cloud resources. This technology makes use of cloud resources to provide "omnipresent, easy, on-demand network connectivity." In other words, storage, network servers, services and applications are all shared resources. In cloud computing, the definition of "cloud" is provided as [1], a distributed and parallel system made up of a collection of virtualized and networked computers that are energetically provided and shown as one or more integrated computing resources established on service-level agreements between users and service providers. Service providers are dispersed around the globe under this condition, and the number of these service providers is rapidly expanding in order to offer users with sufficient cloud storage space resources. Users in cloud can obtain services depending on their needs without understanding how the services are supplied or where they are located. In most situations, while discussing internet authentication, people still refer to passwords. One of the most serious flaws in current authentication methods is that each user has too many password account pairs, which makes user to forget or use the same username and password across different sites. The use of biometrics might be a feasible answer to this challenge.

Biometrics refers to the technology that allows for the electronic authentication and identification of a person centered on their biological and behavioral features [2].

A biometric engine is commonly used in biometric devices. A biometric engine is a piece of software that interacts with biometric system hardware. Its goal is to keep track of biometric data during the extraction, capture, and matching stages [3]. Biometric technology research has enhanced the identity, identification, and verification needs have greatly improved in terms of dependability and security. [4]. Due to its novelty, history's longevity, peculiarity, popular acceptance, and all of these reasons, as well as the little danger of personal invasion that fingerprint matching entails, should be considered. Fingerprints are being widely and effectively used to help in evidence of identification. [5]. A fingerprint is a pattern impression on the surface of a finger [6]. One of the most extensively utilized biometric technologies is fingerprint-based identification as well as authentication since each person's fingerprint is unique and does not deteriorate with age. Apart from being comparatively cheap to implement, the uniqueness, dependability, and relative simplicity of acquisition are the key motives due to which fingerprint identification has developed as the most regularly utilized biometric authentication technology, more than half of all current recognition systems are based on this technology [7]. The fingerprint identification technology has attracted a large number of researchers in recent years due to its numerous benefits. One of the most significant advantages is that its use for personal identification is recognized by the legal community. This method of identification is simple to use, accurate, inexpensive, and very simple to identify. The graphical pattern of fingerprint ridge features is shown in figure.1.

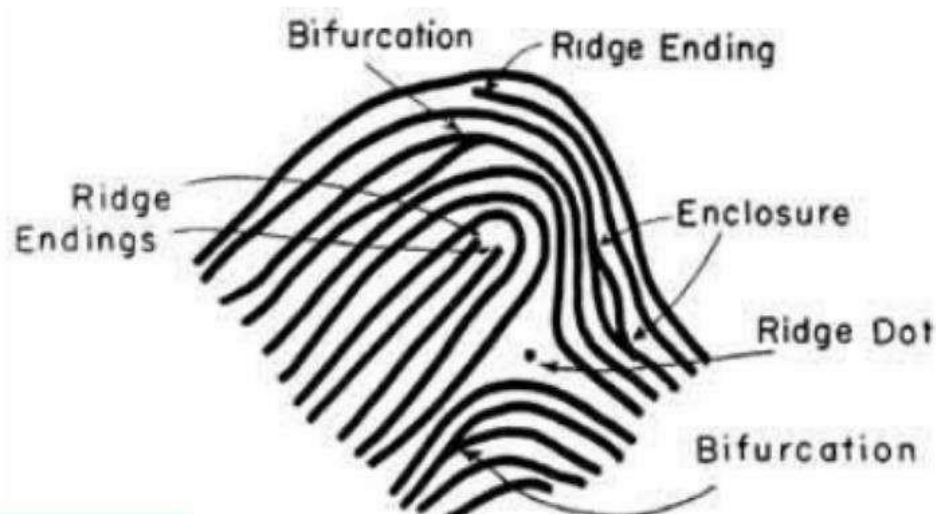


Figure.1 various ridge features on fingerprint image recognition

Furthermore, this recognition method is well-known for its accuracy in authentication, and the chances of two people having the same fingerprint are low. As a result of the issues, we have Type-I and Type-II mistakes in recognition. A Type-I error or false rejection is expected when a recognition system rejects an authorized user. On the other hand, may be accepted by the system, resulting in unlawful access and a false acceptance or Type-II error. The mistakes drive the development of a system which successfully mitigates these two sorts of flaws. The uniqueness of a fingerprint may be determined by examining minute points, ridges, and furrow patterns. As no two fingerprints are same, it's critical to recognize details and patterns in order to recognize them. Fingerprint-based biometric authentication is used in cloud computing in this suggested hypothesis. Through fingerprint authentication, the user can gain access to the specific file that has been requested. When a requested user asks to access a data in a computer's secured folder, particular user uses cloud computing to be authenticated by an authorized user.

Cryptography protects users by allowing them to encrypt data and verify the identity of other users. Data encryption is well recognized for its ability to keep information safe from prying eyes. It employs an encryption key to transform data from one format, known as plaintext, to another known as cipher text. At the moment, compression and encryption are carried out independently. Before the current era, cryptography was essentially identical to encryption, or the conversion of data from a readable to what appears to be a nonsense state. Data encryption comprises a random string of bits designed specifically to scramble and unscramble data. Data encryption employs procedures to ensure that each key is unique and random. Cryptography employs both asymmetric and symmetric keys. The most popular are symmetric keys, which employ the same key for both ciphertext's decryption as well as encryption. This is referred to as a secret key. The two most prevalent forms of secret-key ciphers are stream ciphers as well as block ciphers. It encrypts a block of data at the same time with a private key and algorithm, whereas a stream cipher does it one bite at a time. The manual decryption procedure and a technique of decryption utilizing the appropriate codes or keys. Encryption is the process of transforming plain text data into ciphertext. Decryption is the process of converting ciphertext to plaintext. This paper explains how to create a three-level security mechanism that uses a mix of biometrics and a one-time passcode to verify users and devices signing in. As a result, the focus of this research is on three levels of security techniques that provide the CSP with high authentication for accessing data in the cloud environment.

The paper is organized as follows, Section 2 summarizes the associated efforts in cloud platform based on OTP and symmetric key generation, and Section 3 describes the methodology for OTP method based on random function, identification of fingerprint recognition and MANN and key generation process based on a symmetric key algorithm, Section 3 discusses the results and discussion and Section 4 end with the conclusion.

LITERATURE REVIEW

In cryptography, pseudo-random bit streams are generated [8]. Nevertheless, it accounts for only a small portion of the OTP used in the bit stream. As a result, using an algorithm to extract parts is required by the OTP technique. The purpose of the study is to demonstrate the security and performance of the different security ideas, notably Advanced Encryption Standard (AES), Blowfish, and Twofish [9]. The safety aspects and security concerns of the blowfish, Twofish, and AES algorithms are discussed in this work. Then, using the three methods, encrypt a range of data types such as text, picture, and audio files to determine encryption speed. A Software Guard Extensions (SGX)- Unified Access Control (UAM), a unique unified access management strategy grounded on Intel software guard extensions, using OTPs (SGX) is presented [10]. The findings show that for a single login request, SGX-UAM takes roughly the same amount of time as OpenID and OAuth2.0 and that it performs consistently while processing several login requests. A technique for generating OTP using Pseudo-Random Number Generators (PRNGs), which are inherent to many operating systems and computer languages [11].

The Image-Based Password System (IBPS) produces an OTP based on the image selected by the user. In this work, random numbers have been generated using image properties and used as an OTP, which serves as a strong authentication factor [12]. A variety of cryptographic systems, including AES, is one of the most powerful. Confidentiality, authenticity, integrity, and nonrepudiation are all features of today's information security system. On the World Wide Web, communication security is a critical concern. It is all about maintaining confidentiality, integrity, and authentication while accessing or altering sensitive internal documents [13]. The picture steganography is achieved using a decoding Rivest-Shamir-Adleman (RSA) message and Hash-LSB encoding or file decryption and encryption, and chaos utilized an image decryption encryption approach [14]. A CC security strategy based on multilayer cryptography is presented. The model takes a hybrid approach to key cryptography, combining symmetric and asymmetric approaches. In this case, the Data Encryption Standard (DES) and RSA are utilized to deliver multi-layer encryption and decryption

on both the sender and receiver sides, therefore boosting cloud storage security.

In comparison to the present system, this approach raises data security to the highest level possible and takes less time to upload and download text files. A paradigm for reducing cyber-attacks on cloud data secrecy is presented. This method is used to increase the strength of an encryption key by using chaotic random noise, as the strength is decided by the unpredictability of the input noise rather than the length of the key.

K. Cao et al. [15] recommend combining two transform and minutiae models to increase problem identification accuracy. This study presents a fingerprint identification methodology based on transform-minutiae fusion. The suggested design performed better than current approaches on the FVC 2002 dataset, having an accuracy measuring 100 percent compared to 96.67 and 98.55 percent for the existing methods. According to Krish et al. [16], protecting user agents against security assaults is substantially more complex. To combat security and privacy threats for cloud computing services, the author proposes SSO-based biometric fingerprint authentication architecture. Fingerprint-based authentication, on the other hand, is still susceptible since users may leave their fingerprints on any surface, which attackers can exploit for authentication. According to Tertychnyi et al. [17], data distribution and storage amid numerous users is extremely accessible and cost-effective, but it causes data idleness and safety difficulties. Security employing bio key generation is used in our suggested work to allow users to access data safely. The major goal in a cloud server is to eliminate idleness besides using XOR operation with Gabor filters for biometric authentication. The two most essential goals achieved in this study were biometric cryptographic security and restricting data de-duplication in cloud storage. The suggested methodology offers more secure and quick encryption methods. W. Wang et al. [18] investigate the use of fingerprints in ATM security, as well as the typical validating ways of inputting the client's fingerprints, which are subsequently transferred and checked twice by the administrator. The protection function is substantially enhanced for the strength and steadiness of client's identity. The entire system is built on a fingerprint approach, which ensures that the procedure is reliable, secure and easy to use. In digital money or electronic transactions, this is going to be the most beneficial technology.

Kouamo et al. [19] used a large number of trials to demonstrate how off-the-shelf CNNs can do fingerprint classification. Future research might focus on hyper-parameter optimization utilizing a grid-search or metaheuristics on various parameters (e.g., filter size, strides, and activation functions) [20]. As a result, we opted to maintain the CNN hyper-parameters and settings as they were in the original implementations; in fact, we concentrated on running a large number of tests including all possible combinations of three CNNs, two datasets, and three different classification tasks. Jagtap et al. [21] published a ground-breaking deep learning-based study to improve assessment measure performance and minimize fingerprint rejection rates. Gabor features are paired with CNN feature strength to build a unique feature map that has a substantial impact on the intelligence system's growth. The feature map is subjected to the dimensionality reduction approach to increase accuracy.

RESEARCH METHODOLOGY

The cloud side of our proposed approach is secure, and the user data is secure as well. This research work presented architecture based on which can guarantee security to both the cloud environment and the user. However, this research aims at building a security technique that combines the use of fingerprint identification with OTP to identify users and applications before using the cloud services. This algorithm will mandate the user/device forward both the finger-print recognition based Modified Artificial Neural Network (MANN) validates the user at the time of log-on to access the cloud services and the one-time passcode passed to the phone as SMS before starting to use the cloud services (invoking any cloud-based API). This is required in addition to the key, the user gets received while signing up with software deployed as a service. The proposed system encrypts both the key and the fingerprint using the OTP code provided and forwards it to the Cloud

Service Provider (CSP) and logs in with a password from CSP is allocated using a CSP key for the user in terms of a cloud-based authenticator for validation.

Moreover, this system mandates the reset of the CSP key on regular periodic bases for accessing the data or software in the cloud environment.

OTP GENERATION

The proposed OTP algorithm based on the randomization mechanism is described in this section. The Fig.2 depicts a broad overview of the proposed approaches. This system utilizes any secure random function that exists on your operating system or programming languages, such as CryptGen Random function on Windows OS, /dev/random on Linux OS or Java Random(), and Python Random() classes. OTP is a password that is only valid for one login session or transaction on a computer or digital device. OTPs are now utilized in nearly every service, such as Internet Banking and online commerce. They usually consist of a mix of four or six numeric numbers or a six-digit alphanumeric code.

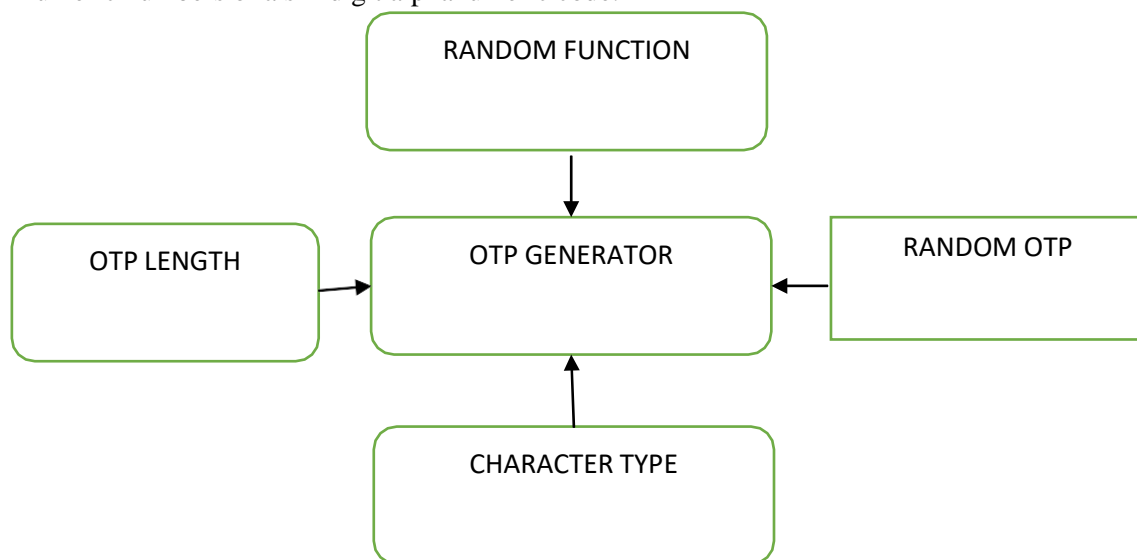


Fig.2. Proposed OTP method based on a random function

The requirements for generated OTP are as follows:

- **Length of OTP:** Required length of OTP must be between 6 and 10. This is a standard requirement for the length of OTP [2, 3].
- **Character type:** Generated OTP may consist numeric, upper alphanumeric and lower-upper alphanumeric characters: [0, ... ,9], [0, ... ,9, A, ... , Z] and [0, ... ,9, a, ... , z, A, ... , Z]. Special characters (for example, @,?,!) are not suitable for OTP because time is limited to input or the process may be unpleasant for users.
- **Used Function:** a) random.random(): Familiar function in Python and other machine learning applications. This method is mainly used to collect the random numbers ranging from 0 to 1. b) math.floor(): The mathematical functions used to convert the floor of the return floating-point number convert into an integer number.

Using the aforesaid code, select a random index from a string array containing all potential possibilities for a certain digit of the OTP.

Algorithm to generate OTP for 4-digit

Step 1: Import the library of random() and math()

Step 2: Generate the OTP function

Step 3: Declare a digit variable

OPT_range = "0123456789"; OTP = ""

Step 4: The password's length can be altered

Step 5: Changing value based on corresponding range for i in range(4) :

Step 6: `OTP += OPT_range[math.floor(random.random()*10)]`

`return OTP`

Step 7: Driver the code execution based on main and other functions

Step 8: Start the system clock using the stop watch/ counter functions

Algorithm to generate OTP for 6-digit

Step 1: Import the library of random() and math()

Step 2: Generate the OTP function

Step 3: Declare a string variable

Step 4: `String = "0123456789abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ"`

`OTP_string = ""`

Step 5: Corresponding password's length can be altered

Step 6: Based on range the corresponding values can be change

Step 7: Driver the code execution based on main and other functions

Step 8: Start the system clock using the stopwatch/counter functions

`t1_clockstart = process_time()`

`print("OTP of length 6:", generateOTP())`

`t1_clockstop = process_time()`

`print("Elapsed time during the whole program in seconds:",t1_clockstop-t1_clockstart)`

Algorithm to generate String and Digit OTP

Step 1: Importing the packages for random suction to generate the sequence

Step 2: OTP generation using random string sequence

Step 3: string library function

Step 4: Takes random choices from the initial string OPT

Step 5: OTP considers only ascii_uppercase, ascii_lowercase, and digits

Step 6: `generate_OTPpassword`

Step 7: Drive the code execution based on main and other functions

`if __name__ == "__main__" :`

Step 8: Start the system clock using the stopwatch/counter functions

`password = rand_pass(8)`

`print("OTP for alpha numeric of length 10:",password)`

`t1_clockstop = process_time()`

`print("Elapsed time during the whole program in seconds:",t1_clockstop-t1_clockstart)`

USING PYTHON ENCRYPTION AND DECRYPTION OF FILES

Only on entering this OTP do users get access to the requested file. Then the encrypted key is obtained from the server. Encryption is the process of decoding a communication so that it can only be read by the intended receivers. To encrypt a file, then utilize the cryptography library employs a symmetric technique to encrypt the file. The same key is used to encrypt and decode the file in symmetric algorithms like Reverse cipher, Caesar Cipher, and ROT13 Algorithm. The cryptography package's fernet module contains built-in functions for generating the key, using `encrypt ()` and `decrypt ()` methods. This function

convert encrypts plain text into cipher text and decodes cipher text into plain text. All actions will be performed using the nba.csv file.

Installation: Using the command below, you may install the cryptography library:

Generate Key to encrypt the file: In the cryptography library, encrypting the file using a cryptography algorithm is known as fernet.

Once the file is encrypted then open the file containing the key. Create and save the Fernet object in the fernet variable. Examine the original document. Put the file inside an object and encrypt it. Then, in the same file nba.csv, write the encrypted data. Now write the following code for the file encryption:

```
# Import OTP-related packages and required module
Step 1: From cryptography.fernet import Fernet
The function used to generate the key and store:
Step 2: Fernet key generation python function
Key = Fernet.generate_key()
Step 3: Corresponding string key is stored in a file
The above code delivers the following output
J64ZHFpCWFIS9zT7y5zXuQN1Gb09y7cucne_EhuWyDM=
```

Algorithm for encryption of file using the key generation

- Step 1:** Open the file that contains the output key
- Step 2:** Initialize the Fernet object and its corresponding output and store it in the related variable using Fernet(key)
- Step 3:** Read the corresponding file
 - #opening the source file which is used to encrypt the corresponding text with open(„OutputFile.csv“, „rb“) as file:
- Step 4:** The process encrypts the given file and converts it into an object to save the file with open(„OutputFile.csv“, „rb“) as file:
- Step 5:** The given encrypt function writes the encrypted data to store the same file
 - # encrypted file open using the write mode
 - # writing method used to write the encrypted data

The Fig.3 shows the nba.csv file before and after running the aforementioned software. The Fig.4 shows the OutputFile.csv file after running the above program.

	A	B	C	D	E	F	G	H	I
1	Name	Team	Number	Position	Age	Height	Weight	College	Salary
2	Avery Brac	Boston Ce	0	PG	25	06-Feb	180	Texas	7730337
3	Jae Crowd	Boston Ce	99	SF	25	06-Jun	235	Marquette	6796117
4	John Holla	Boston Ce	30	SG	27	06-May	205	Boston University	
5	R.J. Hunte	Boston Ce	28	SG	22	06-May	185	Georgia St	1148640
6	Jonas Jere	Boston Ce	8	PF	29	06-Oct	231		5000000
7	Amir John	Boston Ce	90	PF	29	06-Sep	240		12000000
8	Jordan Mi	Boston Ce	55	PF	21	06-Aug	235	LSU	1170960
9	Kelly Olyn	Boston Ce	41	C	25	7-0	238	Gonzaga	2165160
10	Terry Rozi	Boston Ce	12	PG	22	06-Feb	190	Louisville	1824360
11	Marcus Sn	Boston Ce	36	PG	22	06-Apr	220	Oklahoma	3431040
12	Jared Sulli	Boston Ce	7	C	24	06-Sep	260	Ohio State	2569260
13	Isaiah Tho	Boston Ce	4	PG	27	05-Sep	185	Washingtc	6912869
14	Evan Turn	Boston Ce	11	SG	27	06-Jul	220	Ohio State	3425510
15	James Yol	Boston Ce	13	SG	20	06-Jun	215	Kentucky	1749840

Fig.3. File for uploading encryption

	A	B	C	D	E	F	G	H
1	gAAAAABf9YzFUxFGmExwMVR9RDlkFmvgJiZjMREB1FroJbr14X9Y6f5pPw1HFxu-HaLnf8ha							
2	GslNm0RDHxgkswX3AltXrJsJ-VVwcuDJOPGBIAh2cbc3NWR_ZM5ZD8DHjC7YF_24vj4cn3um							

Fig.4. After the encryption process

Algorithm for decryption of file using the key generation

Step 1: Initialize the required objects and variables related to fernet function.

```
fernet = Fernet(key)
```

Step 2: Read the encrypted file

```
#Use Open mode to opening the corresponding encrypted file
with open(„OutputFile.csv“, „rb“) as enc_file:
    Encrypted = enc_file.read()
```

Step 3: Decrypt the OutputFile and save it an object

```
# decrypting OutputFile
decryptedFile = fernet.decrypt(encrypted)
```

Step 4: Use write mode and decrypted input data into OutputFile.csv.

```
# Use write mode and writing decrypted information into OutputFile.csv
with open(„OutputFile.csv“, „wb“) as dec_file:
    dec_file.write(decryptedFile)
```

The Fig.5 shows the OutputFile.csv file before and after running the aforesaid application to retrieve the original data. Fig.6 shows the nba.csv file after decrypting it to acquire the original data.

	A	B	C	D	E	F	G	H
1	gAAAAABf9YzFUxFGmExwMVR9RDlkFmvgJiZjMREB1FroJbr14X9Y6f5pPw1HFxu-HaLnf8ha							
2	GslNm0RDHxgkswX3AltXrJsJ-VVwcuDJOPGBIAh2cbc3NWR_ZM5ZD8DHjC7YF_24vj4cn3um							

Fig.5. Before the decryption process

	A	B	C	D	E	F	G	H	I
1	Name	Team	Number	Position	Age	Height	Weight	College	Salary
2	Avery Brac	Boston Ce	0	PG		25	06-Feb	180 Texas	7730337
3	Jae Crowd	Boston Ce	99	SF		25	06-Jun	235 Marquette	6796117
4	John Holla	Boston Ce	30	SG		27	06-May	205 Boston University	
5	R.J. Hunte	Boston Ce	28	SG		22	06-May	185 Georgia St	1148640
6	Jonas Jere	Boston Ce	8	PF		29	06-Oct	231	5000000
7	Amir John	Boston Ce	90	PF		29	06-Sep	240	12000000
8	Jordan Mi	Boston Ce	55	PF		21	06-Aug	235 LSU	1170960
9	Kelly Olyn	Boston Ce	41	C		25	7-0	238 Gonzaga	2165160
10	Terry Rozi	Boston Ce	12	PG		22	06-Feb	190 Louisville	1824360
11	Marcus Sn	Boston Ce	36	PG		22	06-Apr	220 Oklahoma	3431040
12	Jared Sulli	Boston Ce	7	C		24	06-Sep	260 Ohio State	2569260
13	Isaiah Tho	Boston Ce	4	PG		27	05-Sep	185 Washingtc	6912869
14	Evan Turn	Boston Ce	11	SG		27	06-Jul	220 Ohio State	3425510
15	James Yol	Boston Ce	13	SG		20	06-Jun	215 Kentuckv	1749840

Fig.6. After the decryption to get the original file

The proposed system provides additional security to the data on the cloud. Here proposed research work provides security for accessing the files stored on the cloud by providing OTP to the authenticated person only if the data owner accepts to share his data. The proposed system shows that the user who wants to access the file has to initially request the data owner for file access then if the request is accepted then only OTP is sent on authenticated user number. The user then has to enter the same OTP then only the access is granted or else access is rejected. The program on the cloud server then encrypts the submitted files using the symmetric algorithm and the encrypted files can be accessed by the users only if permission is granted by the data owner.

Then Data owner uses the symmetric key to decrypt the ciphertext of data files. Hence in this way the research work by providing two-way security to the data stored on the cloud by integrating the above two mechanisms.

3.3 The steps of the fingerprint enhancement algorithm

3.3.1. Image Normalization and segmentation

The mean and variance of an input fingerprint image are predetermined before it is normalized. To define the ROI, blockwise coherence is used to construct a mask that separates the ridges from backdrop. Although there are several strong approaches for segmentation, I selected to use grey level variance calculation. The image is divided into (W W)-sized sub-blocks, and the variance is calculated for each block using equation 1.

$$V = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} (I(i,j) - M)^2 \dots \dots \dots (1)$$

The image's global variance is then used to compare the root of each block's variance to a threshold T; the related block is regarded to represent the image's backdrop if the value generated is less than the threshold and will be omitted from further processing. The block will be regarded as an important component of the image. $T = 0.2 * \text{std}(\text{image})$ is the chosen threshold value for this repository, and $W = 16$ is the chosen block size. This stage allows you to reduce the size of the image's useful region and, as a consequence, improve the biometric data extraction phase.

3.3.2. Orientation field

To determine the field's direction, Sobel filters are utilized. The gradients in horizontal and vertical directions are obtained using the 3 by 3 operators G_x and G_y . Thus, the following equations 2 and 3 estimate the indigenous path in region of V, in the path of lines ($V_x I_j$), and columns ($V_y I_j$).

$$V_x(i, j) = \frac{1}{2} (u_{i-1} - 2u_i + u_{i+1} + v_{j-1} - 2v_j + v_{j+1}) (G_x u, v) \dots \dots \dots (2)$$

$$V_y(i, j) = \frac{1}{2} (u_{i-1} - 2u_i + u_{i+1} - v_{j-1} + 2v_j - v_{j+1}) (G_x u, v) \dots \dots \dots (3)$$

In equation 4, the evaluation of the local environment in the neighborhood V is $\theta(i, j)$

$$\theta(i, j) = \frac{1}{2} \tan^{-1} \left(\frac{V_x(i, j)}{V_y(i, j)} \right) \dots \dots \dots (4)$$

$W1 = 16$ offers a good orientation estimate while reducing processing time since the usual width of the ridge or valley in a fingerprint image is five to eight pixels.

3.3.3. Frequency map

In order to create the Gabor filter, we must first estimate the frequency map locally as well as the directional map. The picture frequency map is generated by computing the local frequency of each pixel's streaks. The following are the steps in the frequency estimate stage:

1. Cut the image into WW-sized chunks.
2. The pixels inside each block that run in opposite direction of local ridge orientation. The ridges in the fingerprint correspond to the local minimum points, forming an almost sinusoidal- shaped wave.
3. Computing the amount of pixels amid successive minimum points in the estimated waveform may be used to compute the ridge spacing. A block with less than two maxima has its period adjusted to zero, and the block is classified as noise. The maxima are the intersections of the streaks, whereas the minima are the intersections of the valleys. Extrema are a set of maximum and minimum values that occur in a certain order.
4. Divide the distance between the first and last peaks by the wavelength (T); if the wavelength exceeds the permitted limits, the frequency image has been turned off. Use the ratio $(1 / T)$ to get the frequency, where T is the time interval between two successive extrema.

3.3.4. Gabor filter

Filtering is the technique of altering the pixel values of an image to improve its appearance. In practice, this entails generating a new image from the original's pixel values in order to identify the collection of frequencies

that make up the Fourier domain area.

Equation 5 specifies the Gabor filter, which has even symmetry and is oriented at 0 degrees.

$$hb(x,y;\theta, f) = e^{-12(x^2+y^2)} \cdot \cos(2\pi f x) \dots\dots\dots (5)$$

The filter may have a number of preferred orientations depending on the distinct blocks of the image. In this case, the ultimate filter is a group of fundamental filters organized in each direction. The final picture is a spatial convolution of the original (normalized) image with one of the two-directional and frequency maps base filters in terms of direction and local frequency.

3.3.5. Minutiae extraction

Finding ridge bifurcations and ends is simple using the crossing number method. The Crossing Number approach will be used to analyze 3x3 pixel blocks. Equation 6 is used to compute the CN value.

$$CN(P) = 12 \sum_{i=1}^8 |P_i - P_{i-1}| \dots\dots\dots (6)$$

There are additional methods for extracting details from gray-scale images that do not need binarization or thinning. These considerations have influenced my decision:

- During the linearization process, data may be lost.
- Thinning and Binarization take a long time.

3.3.6. Singularity Detection

Let G stand for the field that corresponds to a fingerprint orientation, and [i,j] denote element's location. The following is how it is calculated.

- A closed route is defined as an ordered succession of particular D items with [i,j] as an internal point.
- The total of the orientation differences between C's neighboring components in algebraic form given in equation 7 yields PG, C(i,j).

$$PG, C(i,j) = \sum_{k=0}^7 \text{angle}(d_k, d_{k+1} \pmod{8}) \dots\dots\dots (7)$$

The Poincaré index adopts just one of the discrete values for closed curves: 0°, 180°, or 360°, as is widely known and easily shown. When it comes to fingerprint singularity, there are a few things to keep in mind:

- 0° does not belong to a certain area.
- 360° belongs to a unique entire area
- 180° is a unique area of loop type.
- 180° is a unique area of delta type

This method entails pre-processing (binarization, thinning, filtering, and so on) in order to refine the image and extract the most useful characteristics. We will offer methods that deal solely with the appearance of standardized categorization as part of this effort. On the basis of details, a comparison is made (Image mapping) because it is based on particular characteristics in each person, it is the most extensively utilized strategy. Minutiae have been retrieved from two prints and characterized as a collection of points in a 2D plan based on model coordinates (type of minutiae, angle and coordinates). The objective of the assessment is to determine a correct arrangement of two fingerprint minutiae (F1 and F2) that results in the most comparable pairs of minutiae possible. The suggested MANN is then used to validate two fingerprint minutiae and determine whether they match or not.

The proposed neural network architecture is a locally associated neural network, that would categories input image sub blocks using probabilistic calculations based on Henry's classification. MANN is used to find the fingerprint matches effectively. The MANN combined to the back propagation algorithm and also ANN. The MANN is a multilayered NN which include the feed-forward approaches. It is used to grouping and mapping the compound regions for the targeted portions. The main objective for the proposed system is utilized to determine the functional link between the sources of input and the intended output's destination.

The net function and active function is expressed in equation.8 and equation.9

$$net_{xj} = \sum_i Y_{ij} Z_{ai} \dots\dots\dots (8)$$

Y_{ij} represents threshold values of each pixel

$$active_{j} = net_{xj} + u_{hj} \dots\dots\dots (9)$$

Zai represents input of source for iterations

The sigmoid function is expressed in equation.10

$$Zai = 1/(1 + e^{-k_1 * active_j}) \dots \dots \dots (10)$$

uhj represents hidden threshold, Zai represents output layer is expressed in equation.11

$$netak = Y_{jk}Z_{aj} \dots \dots \dots (11)$$

Include the activation function is expressed in equation.12 and equation.13

$$active_{Ok} = netak + U_{ok} \dots \dots \dots (12)$$

$$Zak = 1/(1 + e^{-k_2 * active_{Ok}}) \dots \dots \dots (13)$$

The output of Oak with target output ak is expressed in equation.14

$$Yak = tak - Zak^2 \dots \dots \dots (14)$$

Weighted vector Wjk using the equation.15 and equation.16

$$\Delta Y_{jk} = 2k_2 yak Z_{aj} Zak(1 - Zk) \dots \dots \dots (15)$$

$$\Delta W_{ij} = k_1 Z_{ai} Z_{aj} (1 - Zai) Yak Z_{jk} \dots \dots \dots (16)$$

The proposed equation is expressed in equation.17

$$\Delta uhj = k_1 (1 - Z_{aj}) yak Y_{jk} \dots \dots \dots (17)$$

Using the function $Y_{jk} + Y_{jk}$, the threshold and weights is determined. This threshold must decide if two patterns match or not, and the back propagation method and ANN algorithm must be used to train the network.

RESULTS AND DISCUSSIONS

Following the implementation of the proposed methodology, authors have concluded that cloud security can be improved by using a model of secure authentication with OTP based on 4-digit, 6-digit, 8-digit, and 10-digit OTP lengths, along with data encryption and decryption using Reverse cipher, Caesar cipher, and ROT13. The data sent/received by the user is extremely important and must be treated with caution. Using encryption and decryption technologies, we can minimize the processing time for OTP creation and key generation.

Table.1. OTP creation using digits and time

OTP EXAMPLES	OTP Digits	Time (s)
1234	4	0.0016
FeDh34	6	0.0032
FRDEf12gt	8	0.0046
NMHKU6567h	10	0.0048

The Table.1 have shown the time taken for OTP generation for each digit number. The string with an alphanumeric of length 10 takes the time of 0.0042 seconds which is more compared to the string with an alphanumeric of length 8. The time taken for OTP generation of 4-digit number has 0.0016 seconds which is lesser than the string of 6-digit number. The overall time taken for OTP generation of 4-digit numbers consumes lesser than the OTP generation for 6, 8, and 10-digit numbers.

Table.2. Encryption and decryption using processing times in seconds

Algorithm Name	Time(s)
Reverse Cipher	0.00025
Ceaser Cipher	0.00065
ROT13	0.00307

The time it takes to decrypt cipher text into plain text and encrypt plain text into cipher text using the decrypt () and encrypt () procedures are shown in Table 2. The Reverse Cipher algorithm takes the time to encrypt

and decrypt the file 0.00025 seconds which is lesser compared to the time taken for encrypting and decrypting the file using Ceaser cipher as 0.00065 seconds and ROT13 as 0.00207 seconds.

A database (named BDAL) containing 500 fingerprint images, each measuring 248 X 338 pixels in 8-bit gray scale, from 100 distinct persons, each having five images of fingerprints (The fingerprints of the same person are obtained and considered for the said work). The International NIST database 4 is a fingerprint recognition database. It's made up of 2000 pairs of 512 X 512 fingerprint images, each having an 8-bit grey scale. Figure 7 depicts the input fingerprint image. Figure.8 represents binary image of fingerprint pattern. In order to execute fingerprint-matching, strategies for fingerprint identification image enhancement where the ridges are highly sharp and clear are presented in figure 9.



Figure.7 Input fingerprint image Figure.8 Binary image Figure.9 Enhanced fingerprint Image

CONCLUSION

The proposed technique combines the use of biometrics, a one-time passcode, and the usual key to validate the user/device at the time of log-on to access the cloud services. Thus, this research is focused on three-level security mechanisms that assist the CSP with high authentication for accessing the data present in the cloud environment. This system may access the owner task environment by entering the correct owner password and completing the owner task. In this paper two of the most secure encryption, decryption, and OTP algorithms. In comparison to the prior technique, the two security approaches make our framework more secure. The OTP and symmetric key generation are offered as a novel approach for cloud service authentication; this solution is more secure and simple to use. Because the need for cloud computing is growing in today's world, the security of the cloud and its users is a major problem. This paper outlines the security concerns that cloud computing faces in general, as well as the mitigating strategies that have been offered to address the issues. This process successfully developed the above-mentioned system and has concluded that by utilizing this recommended method we may more effectively attain improved security in cloud computing. The major goal here is to reduce idleness in a cloud server and to recognize fingerprints utilizing Gabor filters. The research also includes a MANN-based fingerprint recognition method for authentication. The findings of the various strategies investigated demonstrate that employing blocks of processed fingerprint images allows for low FAR identification in the range of 0.6% to 4.8%. The suggested MANN model achieves 95.3% and 94.3% accuracy on the NIST and BDAL databases, respectively, demonstrating a considerably superior performance than previous techniques. This research identifies cloud storage as a means of accessing data or files from a Personal Computer (PC) via a mobile device from any location. However, Fingerprint-based authentication still has a vulnerability, because users' fingerprints may be left on any surface, which attackers can exploit for authentication. In future, building a security technique which encrypts both the key and the finger-print using the OTP code provided and forwards to the CSP and login with password from CSP is allocated using CSP key for the user in term of cloud based authenticator for validation. In the future present various comparisons using our technique and results to demonstrate the efficacy of our proposed framework.

REFERENCES

[1] Ankita Patil, Kiran Zambare, Preeti Yadav, Pankaj Wasulkar and Nisha Kimmatkar, "Integration of Encryption of File and One Time Password for Secure File Access on Cloud", *International Journal of Advances in Computer Science*

and *Cloud Computing*, Vol. 3, No. 1, pp. 1-13, 2015.

- [2] V. Mohammadi, A.M. Rahmani, A.M. Darwesh and A. Sahafi, "Trust-Based Recommendation Systems in the Internet of Things: A Systematic Literature Review", *Human-Centric Computing and Information Sciences*, Vol. 9, pp. 1-21, 2019.
- [3] Y.S. Jeong and J.H. Park, "Security, Privacy, and Efficiency of Sustainable Computing for Future Smart Cities", *Journal of Information Processing Systems*, Vol. 16, pp. 1-5, 2020.
- [4] J.Y. Park and E.N. Huh, "A Cost-Optimization Scheme Using Security Vulnerability Measurement for Efficient Security Enhancement", *Journal of Information Processing Systems*, Vol. 16, pp. 61-82, 2020.
- [5] J. Kang, "Mobile Payment in Fintech Environment: Trends, Security Challenges, and Services", *Human-Centric Computing and Information Sciences*, Vol. 8, pp. 1-16, 2018.
- [6] H.W. Kim and Y.S. Jeong, "Secure Authentication-Management Human-Centric Scheme for Trusting Personal Resource Information on Mobile Cloud Computing with Blockchain", *Human-Centric Computing and Information Sciences*, Vol. 8, pp. 1-11, 2018.
- [7] D.R. Stinson and M. Paterson, "*Cryptography: Theory and Practice*", CRC Press, 2018.
- [8] Hyunki Kim, Juhong Han, Chanil Park and Okyeon Yi, "Analysis of Vulnerabilities That Can Occur When Generating One-Time Password", *Applied Sciences*, Vol. 10, pp. 1-12, 2020.
- [9] M. Robinson Joel, V. Ebenezer, M. Navaneethakrishnan and N. Karthik, "Encrypting and Decrypting Different Files Over Different Algorithm on Cloud Platform", *International Journal of Emerging Trends in Engineering Research*, Vol. 8, No. 4, pp. 1-5, 2020.
- [10] Liangshun Wu, H. J. Cai, and Han Li, "SGX-UAM: A Secure Unified Access Management Scheme with One Time Passwords via Intel SGX", *IEEE Access*, Vol. 9, pp. 38029-38042, 2021.
- [11] J. Galbally, R. Haraksim, and L. Beslay, "A study of age and ageing in fingerprint biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1351–1365, 2019.
- [12] T. Sabhanayagam, V. Prasanna Venkatesan, and K. Senthamaraikannan, "A comprehensive survey on various biometric systems," *International Journal of Applied Engineering Research ISSN*, vol. 13, no. 5, pp. 2276–2297, 2018, <http://www.ripublication.com>.
- [13] S. Gu, J. Feng, J. Lu, J. Zhou, and S. Member, "Efficient rectification of distorted fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 156–169, 2018.
- [14] S. R. Borra, G. J. Reddy, and E. S. Reddy, "A broad survey on fingerprint recognition systems," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET*, pp. 1428–1434, Chennai, India, March 2016.
- [15] K. Cao and A. K. Jain, "Learning fingerprint reconstruction: from minutiae to image," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 104–117, 2015.
- [16] R. P. Krish, J. Fierrez, D. Ramos, F. Alonso-Fernandez, and J. Bigun, "Improving automated latent fingerprint identification using extended minutia types," *Information Fusion*, vol. 50, pp. 9–19, 2019.
- [17] P. Tertychnyi, C. Ozcinar, and G. Anbarjafari, "Low-quality fingerprint classification using deep neural network," *IET Biometrics*, vol. 7, no. 2, pp. 550–556, 2018.
- [18] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "Security and accuracy of fingerprint-based biometrics: a review," *Symmetry*, vol. 11, no. 2, p. 141, 2019.
- [19] Kouamo, S. and Tangha, C, "Handwritten Character Recognition with Artificial Neural Network. Distributed Computing and Artificial Intelligence", *Advances in Intelligent and Soft Computing*, 2012, vol- 151, 535-543, http://dx.doi.org/10.1007/978-3-642-28765-7_64
- [20] Sakshica, Gupta, K., Vidyapith, B., Campus, J. and Jaipur, "Handwritten Digit Recognition Using Various Neural Network Approaches" *International Journal of Advanced Research in Computer and Communication Engineering*, 2015, vol- 4.
- [21] Jagtap, V.N. and Mishra, S.K, "Fast Efficient Artificial Neural Network for Handwritten Digit Recognition" , *International Journal of Computer Science and Information Technologies*, 2014, vol-5, pp-2302-2306.

A SECURE ELECTRONIC HEALTH FRAMEWORK TO PROTECT HEALTH RECORDS USING NATURAL LANGUAGE PROCESSING WITH MULTI LEVEL DATA ENCRYPTION IN CLOUD

N. SUBHALAKSHMI^[1] Dr.M.V.SRINATH^[2]

^[1]Research Scholar, ^[2]Research Supervisor ,S.T.E.T Women's College (Autonomous), (Affiliated to Bharathidasan University,Tiruchirappalli)Sundarakkottai, Mannargudi - 614016, Thiruvarur Dt., Tamil Nadu, India

^[1]Email :subha.stet@gmail.com Mobile No. 9095598954 ^[2]Email :sri_induja@rediffmail.com Mobile No. 9710944476

Abstract-- Big data is a set of a massive quantity of large datasets with data volume. With the growing number of data, the demand for big data storage will increase. By setting the records inside the cloud, that data is to be available to anybody from anywhere. Cloud computing is an evolving, carrier-centric framework for performing distributed and parallel computing on large datasets. As the benefits of cloud computing increase in terms of cost, storage space, and scalability, all data providers and institutions are also focusing on offloading data from local servers to remote cloud servers. Medical records are essential and most important because the government retains additional data on the medical history of the data and medical professionals can provide the most appropriate and effective remedies or support for their concerns. It is also useful for diagnosing viable illnesses, identifying family hereditary and possible illnesses, allergic reactions, past and present dosing, and vaccination statistics. The proposed work aims to develop a three-tier framework to protect the privacy of records stored in big data environment and analyses the document about the protected text and breaks the protected content into separate documents. This research work categorizes, distributes, and stores health-related content using a combination of Natural Language Processing (NLP) and text mining algorithms. After associating the distributed content with the original parent document, it encrypts the attribution information of the patient's history and saves in the clouds for future.

Keywords: *cloud computing, big data, encryption, Natural Language Processing (NLP), patients' history*

I. INTRODUCTION

Cloud computing technology provides convenient on-demand system access to shared pools of configurable computing resources (networks, servers, storage, applications, services, etc.) for rapid provisioning and sharing. Cloud computing [1] platforms are divided into Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) based on the services they provide. Information privacy and safety are one of the biggest reluctances of cloud computing, as it is an open environment with very restricted user control. Cloud computing has developed as one of the most significant changes in the recent Information and Communications Technology (ICT) era [2]. Cloud Computing security is a new extent of computer safety that states to a set of strategies, controls, and cryptographic primitives to protect connected information, system applications, and substructure. The National Institute of Standards and Technology (NIST) [3] explains the definition of CC technology as follows: "A concept that allows appropriate on-demand system access to a common pool of configurable computing resources (supplied by a third-party provider known as a cloud service provider), with minimal administrative labour or services, interactive provisioning, and shared providers". Safety is one of the major obstacles to cloud growth and the use of big data in cloud environments.

Big data security in the cloud is the most exciting area that causes numerous safety issues such as network security, information security, application layer security, and data protection. Security and Privacy issues are exacerbated by the speed, dimensions, and variation of big data, including huge cloud organizations, different data bases and formats, the flowing nature of data collection, and high-volume inter-cloud relocation [4]. Computing and big data are interconnected. Big data has attracted the attention of academic circles, government and medical industry. In addition, it transforms health care, knowledge, manufacturing, economics, business, and ultimately societies [5]. Big Data is used for large collection of information that is massive in length and developing exponentially with time. The data is being generated from several sources which include social media, utilization of Search engines, Sensors [6], medical systems, Banking transactions, financial packages etc., and that statistics may be established, unstructured or semi-established. Big records are so large and complicated that no typical record-keeping or processing system can keep or process them effectively. Big data is a set of technology and knowledge that necessitate new methods of incorporation to expose great concealed value from diverse, composite, large and very large datasets. Data develops big data when discrete data becomes useless and only huge collections of data or analyses consequent from them have more value. Many Big data analysis technologies can be used to derive insights that permit better decision-making in key growth areas such as healthcare, monetary efficiency, energy and natural disaster prediction. Safety and confidentiality problems are exacerbated by the amount, accuracy, variety, and speed of big data, such as large infrastructure, different data sources and setups, the nature of streaming data collection, and large volumes of inter-cloud migration [7]. Recently, large amounts of data have become a hot topic with sizeable influence, transforming industries around the world. Companies and governments see massive records analytics as a cutting-edge and valuable way to analyze complex, historical data and find patterns that help them make meaningful decisions. Big data plays an important role in managing and operating fate records in various business sectors, in addition to healthcare, manufacturing, retail, traffic control, banking, meteorology, education and transportation. Several advantages of big data packages have been discovered after extensive investigation [8]. However, latest literature surveys performed in the subject matter of huge information protection specify that malicious attackers pointing massive facts have been on the upward push. However, the main issues and solutions surrounding security risks and privacy have not yet been fully investigated in the vast archive area. These challenges motivate new innovations and research activities to uncover open issues that pave the way for future research and practice. The paper outlines a security and confidentiality of large-scale sensitive medical information of the patients in a cloud computing environment. Identify new advances in cloud provider arrangement, source controller, and cloud service administration layers. The latest encryption and decryption security technologies and outlines additional privacy protection approaches for processing sensitive data for processing big data in cloud computing using Natural Language Processing (NLP) [9] are also evaluated. The objectives of the proposed system are listed as follows:

1. To develop an approach to secure and protect a document that will be stored in a Big Data and Cloud environment.
2. To analyse and eliminate the unwanted words and punctuations from the medical history of the patients through Natural Language Processing (NLP) and Decision Tree (DT) algorithms.
3. To secure the records with encryption and decryption techniques and finally saves them in cloud.

II. SECURITY AND PRIVACY OF MEDICAL RECORDS

Privacy, security, and confidentiality are common problems that need to be addressed in health record systems. Security and confidentiality are inextricably linked, but they are essentially dissimilar. Confidentiality denotes to the precision of somebody to regulate when, how, and at what level access to individual's private data is transmitted or shared by others. Security refers to the restricted access to an individual's private data [10] and is granted to authorized persons. Unauthorized transmission or distribution of complex medical data can result in data breaches. Confidentiality can be conceded in unavoidable systemic credentialing that occur throughout the health infrastructure [11], as well as centralized technologies and parties who watch the activity of healthcare workers and patients saved in the cloud. Still, in some cases, there may be good reason for governments, managers, pharmacological concerns, researchers, and test center [12] to access and obtain data on patient health records, and by doing so, healthcare providers can accidentally abuse access to health records purposefully.

The three simple information generation safety requirements are confidentiality [13], integrity and availability. Confidentiality can be described as limiting data to persons that aren't authorized to access data [14] throughout both storages, transmitting or when they are being dealt with. Confidentiality can be achieved thru technological approach such as information encryption or thru controlling having access to the systems. Confidentiality is likewise completed through working on ethical inclinations consisting of expert silence. However, it turned into found out by using the fact that encryption is generally used for health information that are dispatched across uncovered networks [15], it's miles much less carried out to records that is saved in cell gadgets and other garage media. The want for confidentiality is a response to privateness concerns that are also very critical inside the fitness care zone because of the very sensitive facts concerning sufferers and clients that they carry. Confidentiality ensures that the data remains covered from unauthorized deletion or modification and undesired change by legal users [16]. On the alternative hand, availability ensures that a gadget may be accessed and is absolutely running at any second that a licensed individual is in want of the usage of them. Availability method a number of factors from scalability to resilience and to recoverability of information in case the information is misplaced for any cause. Physicians are frequently concerned that an unauthorized individual could gain access to patient records contained in an electronic medical data device and misuse the information, resulting in a felony complication as a result of a breach in the confidentiality of the patients' information. Physicians are very eager on the security and confidentiality concerns greater than the sufferers themselves. The majority of doctors who use electronic scientific information select paper data more than electronic clinical records because they agree with that paper facts are much greater stable and private. This is an indication that the issue of privacy and security on medical records are taken very seriously. If the sufferers aren't assured privacy [17], they might determine to withhold the records to prevent inappropriate use.

III. PROPOSED METHODOLOGY

The basic flow of the proposed system is shown in figure 1. The data collected is the history of the patients. The medical history is divided into sub parts and this process is analyzed using decision tree algorithm. Then, the medical data is saved in cloud after NLP. For the security and authentication, the

medical history will be encrypted and decrypted. The decision tree that protects privacy is produced from the dataset. The dataset is distributed across multiple members without mutual disclosure. They usually involve the use of cryptography, secret sharing schemes, or other cryptographic algorithms. In this research, a decision tree and three different encryption algorithms (caesar cipher, reverse cipher and ROT13) are used to help multiple hospitals collaborate on the cloud to build a Secure Electronic Health Framework (SEHF) classification model without revealing patient histories. This research focuses on protecting the privacy of datasets when building decision trees. This system encrypts the mapping information after mapping the distributed content to the original parent document. The proposed research work aims at developing a three-tier framework [18] to protect the privacy of documents that is to be stored in the Big Data environment [19]. It parses the document for text that requires protection and separates the content that should be protected into separate documents. It classifies the data that requires protection using a combination of Natural Language Processing (NLP) and text mining methods. It stores the content in distributed fashion.

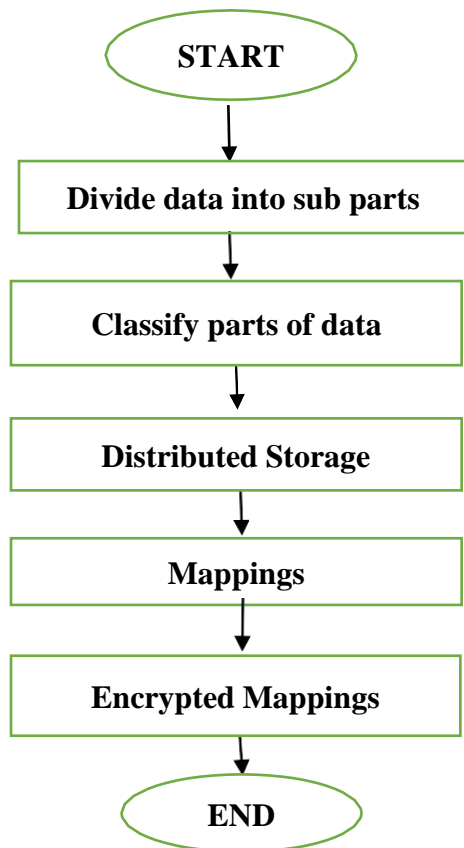


Figure 1 Basic Flow of Proposed System

3.1 Decision Tree

Machine learning decision trees [20] are part of the classification system and also use classification rules (from root to leaf node) to provide results to regression problems. Its construction is like a flow chart where each core node denotes a functional test (for example, if the arbitrary sum is superior than a number)

and each leaf node is cast-off to denote a class label (results are calculated). After all the decisions have been made, the branch denotes the aggregation of the function that leads to the session specification. Machine learning decision trees have a wide range of fields in the current world. Being a predictive model, decision tree analysis is done through an algorithmic approach that conditionally divides the dataset into subsets. The name itself indicates that it is a tree-like model in the method of if-then-else statements. Table1 shows the details of the medical history of the patients. The attribute node split is used to generate the decision tree. Information gain and the Gini coefficient are two often employed divisional criteria. In order to place the split subgroups in the same category, they choose the split attribute by making each split subgroup as "pure" as feasible. Some of the example decision tree determined in this research are shown in figure 2a and 2 b.

Table 1 Medical history of the patients

	age	sex	BP	cholesterol
0	70	1	130	322
1	67	0	115	564
2	57	1	124	261
3	64	1	128	263
4	74	0	120	269
...
265	52	1	172	199

The key persistence of the algorithm is to build a decision tree from a dataset of instances and their classes. The algorithm follows a divide-and-conquer model and attempts to discover the best attributes to divide the dataset at each step. To do this, two values are calculated: entropy and information gain. Information gain trials the amount of information a single function provides about a class. The way it functions as a master key when creating decision trees. The feature with the main information gain is split first. Decision trees always maximize information acquisition. The entropy of an instance varies when nodes are used to partition it into minor subsections. Entropy is the degree of the indecision or impureness [21] of an arbitrary inconstant. Entropy determines how the decision tree divides the data into subsections. The information gain is measured with the equation (1).

$$Information\ gain = entropy\ (parent) - [weighted\ average * entropy\ (children)] \dots\dots\dots (1)$$

The Gini coefficient is a metric that determines how often randomly selected items are misidentified. This clearly shows that attributes with a lower Gini coefficient take precedence. The equation (2) is the Gini index representation. Where $p(X)$ is the probability of root X.

$$\text{Gini Index: } 1 - \sum i = n * p(X)^2 \text{----- (2)}$$

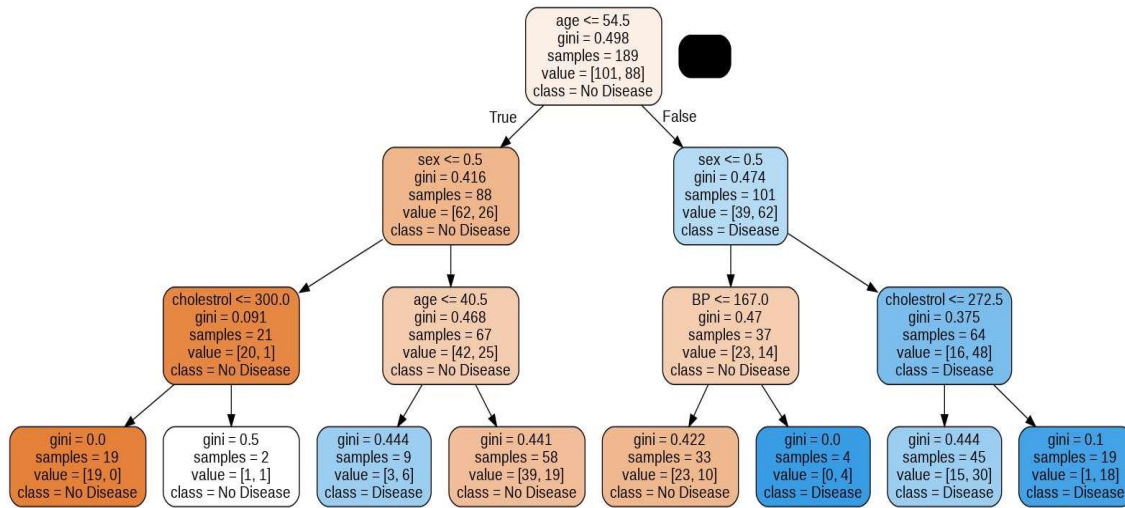


Figure 2a. Decision Tree for Medical History

Decision tree Splitting create a break up, first, we want to calculate the Gini score. The records are break up the usage of a listing of rows having an index of a characteristic and a split price of that characteristic. After the proper and left dataset is observed, we are able to get the break-up value with the aid of the Gini rating from the primary part. Now, the break up cost will be the decider where the attribute will be living. The next part is evaluating all the splits. The first-class feasible value is calculated by means of evaluating the cost of the break up. The high-quality cut up is used as a node of the DT.

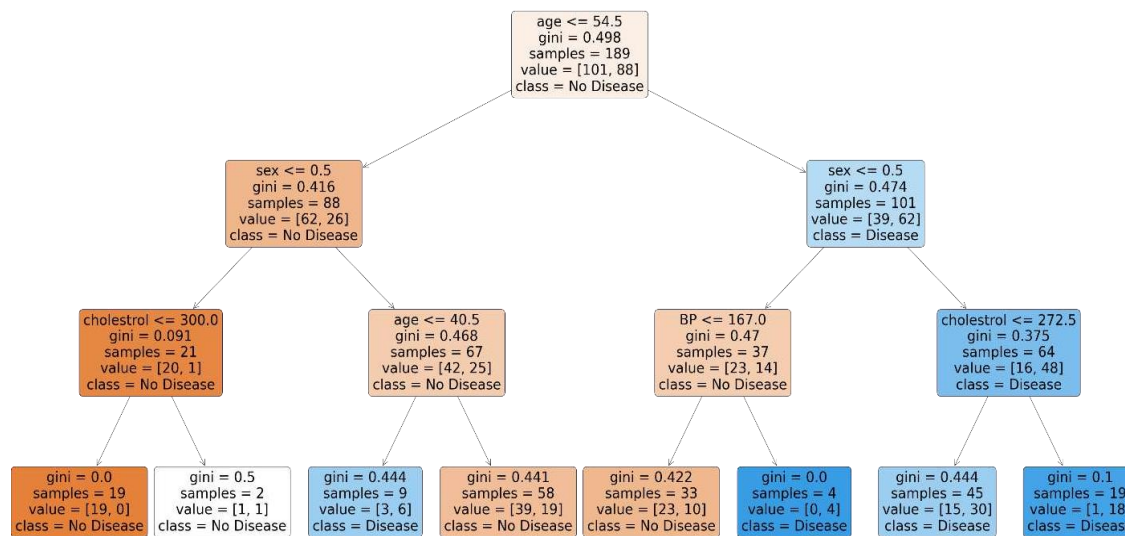


Figure 2b Sample Decision Tree for Medical History

3.2 Natural Language Processing (NLP)

Named Entity Recognition (NER) is a subtask of NLP that identifies and categorizes named entities encountered in unstructured text. Rule based matching is carried out and then looping through all transcriptions to extract the text matching the pattern are determined. Tokenization [22] of words in the patient's history, sentence segmentation of words in the patient's history summarization, stop words identification eliminating punctuations, repeated words and tokenization of words are done. Looping through all the transcriptions and extracting the text matching the pattern are determined. Then, check whether the token is an alphabet character, a digit, lower or upper case, currency, email, number, URL, etc., Finally, trained pipelines are used to make predictions about tokens based on their context. Here, trained pipelines are a type of pipeline which uses statistical models.

3.3 Encryption Algorithm

In order to safeguard the data in the medical history, encoding (i.e., secret key) algorithms are used in this research work. The three algorithms used are caesar cipher, reverse cipher and ROT13. The steps involved in the encryption are shown in algorithm 1.

Algorithm 1: Encryption Algorithms

Input: sample transcription (plain text) **Output:** Encrypted text

1. Start

2. Ceaser Cipher method Def encrypt (textual content,s):

```

end cipheroutput = "" for i in variety(length(input_text)):
    input_char = input_text [i]
    if (char.Isupper()):
        end cipheroutput += chr((ord(input_char) + s-sixty_five) % 26 + 65)
    else end cipheroutput += chr((ord(input_char) + s - ninety_seven) % 26 + ninety_seven)

```

3.In Reverse Cipher

```

output_translate = ""
    i = len (input_message) -1
    while i>= 0:
        output_translate = output_translate + input_message [i]
        i = i-1
    print ("The given ciphertext is:", output_translate)

```

4. ROT13 def rot13(input_text) #Rotate by 13

```

Return text. translate (rot13trans)

```

```
def main()
    output_txt = "ROT13-Algorithmus"
    print rot13 (output_txt)
```

- 5. Print encrypted text with secret key
- 6. End

Steps involved in Encryption algorithms:

Step 1: Begin with sample predicted medical transcription

Step 2: First method, caesar cipher will replace each plaintext letter with a fixed number of letters below the alphabet. Transverse the obvious textual content, encrypt the uppercase characters in plain textual content and then encrypt lowercase characters in undeniable textual content

Step 3: Second method, reverse cipher uses a pattern that reverses a plaintext string and converts it into a ciphertext. The encoding and decoding processes are the same. To decrypt the ciphertext, reverse the ciphertext to get the plaintext.

Step 4: Third method, each character is shifted by 13 digits to encrypt or decrypt the text in ROT13.

Step 5: Final outcome with secret key is generated.

Step 6: End

IV. RESULTS

The final outcome proposes a framework that protects the privacy of the document in the BigData environment by protecting parts of the document that requires protection. Exploratory analysis specifies the matplotlib functions for charting the data. We chose columns with 1 to 50 unique values for presentation and plotted them as a graph in figure 3. X axis and Y axis is measured in terms of units.

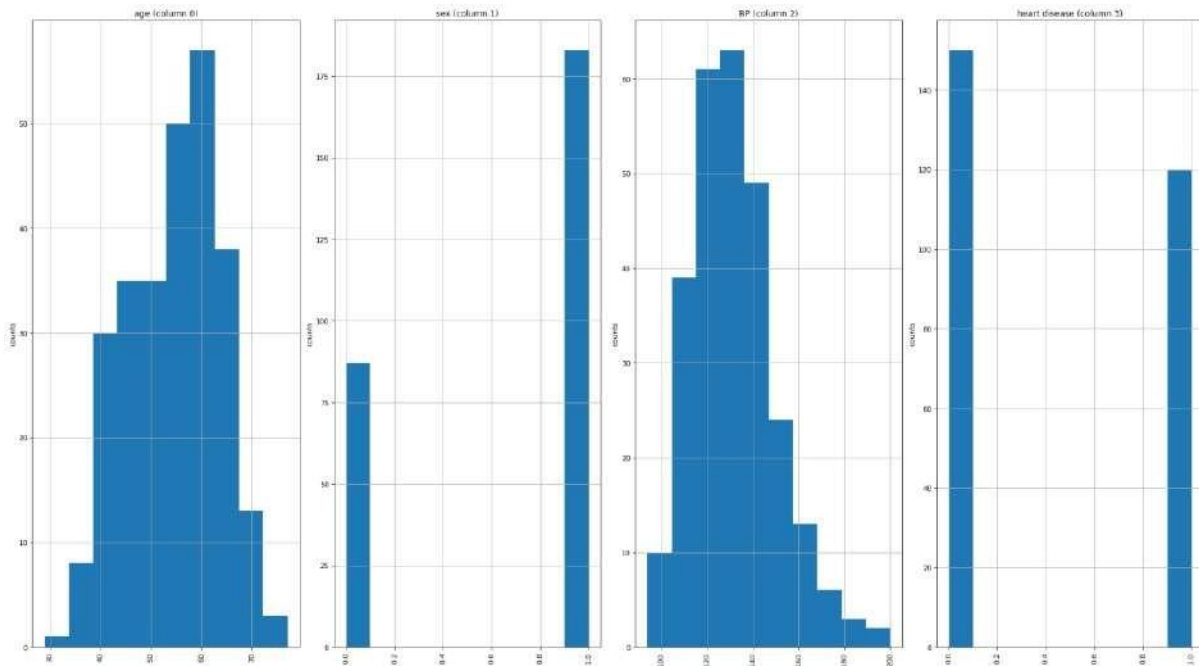


Figure 3 Distribution graph of column data between 1 to 50

Correlation matrix for different patients' history such as heart disease, anaemia, Blood Pressure (BP) are shown in figure 4. X and Y coordinated are measured in terms of seconds. Null values are removed from this matrix, leaving only numerical values. In kernel density charts, it reduces the column for matrix.

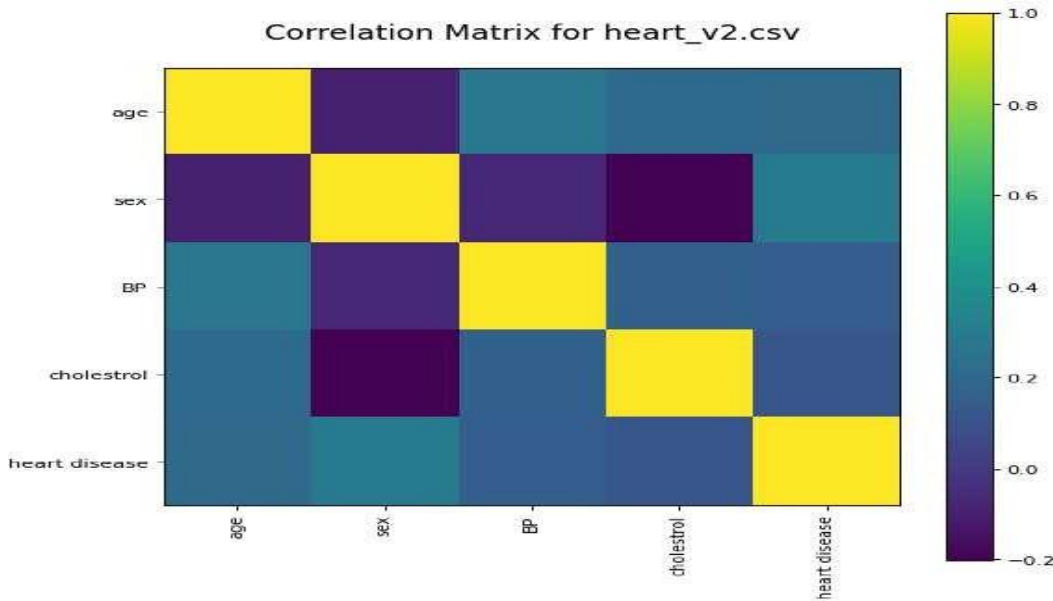


Figure 4 Correlation matrix using decision tree algorithm

Encryption:

Encryption is carried out using Caesar Cipher, reverse cipher and ROT13 algorithm. A block cipher is a type of symmetric cipher that is processed (usually) by a fixed allocation (usually) through a block of information (often 64-bit or 128-bit). "Lightweight" block ciphers differ from block ciphers in that they use algorithms that require less computing power. Table 2 shows a comparative analysis of three different algorithms over time.

Table 2. Comparative output of Cipher executions with respect to time

Name of Algorithm	Elapsed time	Output
Caesar Cipher With key=13	0.000362920994	.rewopgnitupmocsseleriugertahstmshtirogliehtsesutitahtoskcol ftnereffidsirehpickcolb "thgiewthgiL" .)stib 821 ro 46 olbnoitamrofnsissecorp)yllausu(gnippamtnatsnochguorhth piccirtemmys a fodnik a sirehpickcolB
Reverse cipher	0.00020399599408	oerjbcatavgchczbpaffryarevhdreagnugafzugvebtynarugafrrhagv faxpbyoarugazbesagarersvqafvaerucvpaxpbyoacgutvrjgutvYca ysraebauwaargsbiafxpbyoabvgnzebsavafrrfrpbecajlyynhfhiata angfabpauthbeugaupvujamerucvpapvegrzzlfanasbaqavxanafvae pbyO Elapsed time: 1.839298437 1.839094441
ROT13	0.00200601400	Oybpxpvcurevf n xvaqbs n flzzrgevppv

	86	puguebhtupbafgnagznccvat (hfhnyyl) cebprffrfvasbezngvbao a 64 be 128 tugjrvtug"oybpxpvcurevfqvsreragsebzguroybpxfb gung gurnytbevzuzf gung erdhveryrffpbzchgvatcbjre
--	----	---

V. CONCLUSION

With increased use of Big Data for storing various information, data safety and confidentiality is becoming a major challenge on a day-to-day basis. This research aims to develop a framework that combines the power of natural language processing and text mining with data encryption methods to protect sensitive data in documents. This approach will speed up processing since only the mapping information is encrypted and not the content of the document itself. Patient data and medical information from health histories are relatively straightforward to distribute. Information can be obtained and updated while the patient is being treated. However, such systems are heavily affected by safety and confidentiality concerns. According to statistics, patients can face substantial issues when sophisticated data is shared with third parties. Validated from the research and based on the security area, it is clear that various guidelines and values are interrelated. Medical records are saved in clouds for the sake of privacy and security. Though, such classifications need to be consistent to determine potential conflicts and contradictions between values. Encryption algorithms have been proposed by different time lapse helps in finding the appropriate security of the system. It is extremely endorsed to use a well-organized encryption system and easy to use by both medical specialists and doctors. Priority accesses the controller model for the patient recording using Natural Language Processing (NLP) through decision tree splitting.

In future, the data analyzed for the research is very limited and can be compared with more patient’s history. There are many encryption and decryption techniques are available can be compared for the better security of the system.

REFERENCES

[1] J. Shen, T. Zhou, D. He, Y. Zhang, X. Sun and Y. Xiang, "Block Design-Based Key Agreement for Group Data Sharing in Cloud Computing", in IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 6, pp. 996-1010, 1 Nov.-Dec. 2019, doi: 10.1109/TDSC.2017.2725953.

[2] N. M. Ibrahim and A. Zainal, "A Model for Adaptive and Distributed Intrusion Detection for Cloud Computing", Seventh ICT International Student Project Conference (ICT-ISPC), 2018, pp. 1-6, doi: 10.1109/ICT-ISPC.2018.8523905.

[3] F. Fowley, C. Pahl, P. Jamshidi, D. Fang and X. Liu, "A Classification and Comparison Framework for Cloud Service Brokerage Architectures", in IEEE Transactions on Cloud Computing, vol. 6, no. 2, pp. 358-371, 1 April-June 2018, doi: 10.1109/TCC.2016.2537333.

[4] Z. Chunlei, J. Yin and X. Qianli, "The Workload Assessment of National Grid Big Data Projects Based on Content Recommendations and Text Classification", IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2020, pp. 482-490, doi: 10.1109/ICCCBDA49378.2020.9095612.

- [5] MuhnedHussam, Ghassan H. Abdul-majeed, Haider K. Hoomod, "New Lightweight Hybrid Encryption Algorithm for Cloud Computing (LMGHA-128bit) by using new 5-D hyperchaos system", Turkish Journal of Computer and Mathematics Education Vol.12 No.10, 2021, 2531-2540.
- [6] Mohammed Nazeh Abdul Wahid, Abdulrahman Ali, BabakEsparham and Mohamed Marwan, "A Comparison of Cryptographic Algorithms: DES, 3DES, AES, RSA and Blowfish for Guessing Attacks Prevention", Journal of Computer Science Applications and Information Technology, ISSN Online: 2474-9257, 2018.
- [7] C.G. Thorat, V.S. Inamdar, "Implementation of new hybrid lightweight cryptosystem", Applied Computing and Informatic, 2018, 2210-8327 doi: <https://doi.org/10.1016/j.aci.2018.05.001>.
- [8] Zaid M. Jawad Kubba1 and Haider K. Hoomod, "Modified PRESENT Encryption algorithm based on new 5D Chaotic system", IOP Conference Series: Materials Science and Engineering 928 (2020) 032023 IOP Publishing doi:10.1088/1757-899X/928/3/032023.
- [9] F. Pallas, D. Stauffer and J. Kuhlenkamp, "Evaluating the Accuracy of Cloud NLP Services Using Ground-Truth Experiments", IEEE International Conference on Big Data (Big Data), 2020, pp. 341- 350, doi: 10.1109/BigData50022.2020.9378188.
- [10] AliGholami and Erwin Laure, "Big Data Security and Privacy Issues in the Cloud", International Journal of Network Security & Its Applications (IJNSA) Vol.8, No.1, January 2016.
- [11] Soleimany, H., "Self-similarity cryptanalysis of the block cipher", Institute of Engineering and Technology Information Security research article, 2015, Vol. 9, Issue 3, pp.179-184.
- [12] R. Kumar and M. P. S. Bhatia, "A Systematic Review of the Security in Cloud Computing: Data Integrity, Confidentiality and Availability", IEEE International Conference on Computing, Power and Communication Technologies (GUCON), 2020, pp. 334-337, doi: 10.1109/GUCON48875.2020.9231255.
- [13] N. A. Patel, "A Survey on Security Techniques used for Confidentiality in Cloud Computing", International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018, pp. 1-6, doi: 10.1109/ICCSDET.2018.8821135.
- [14] S. A. Oli and L. Arockiam, "Confidentiality Technique to Encrypt and Obfuscate Non-Numerical and Numerical Data to Enhance Security in Public Cloud Storage", World Congress on Computing and Communication Technologies (WCCCT), 2017, pp. 176-180, doi: 10.1109/WCCCT.2016.51.
- [15] A. Mondal, S. Paul, R. T. Goswami and S. Nath, "Cloud computing security issues & challenges: A Review", International Conference on Computer Communication and Informatics (ICCCI), 2020, pp.1-5, doi: 10.1109/ICCCI48352.2020.9104155.
- [16] M. Elsayed and M. Zulkernine, "Towards Security Monitoring for Cloud Analytic Applications", IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International

- Conference on Intelligent Data and Security (IDS), 2018, pp. 69-78, doi: 10.1109/BDS/HPSC/IDS18.2018.00028.
- [17] Eslam w. affify, Abeer T. Khalil, Wageda I. El sobky, RedaAboAlez, "Performance Analysis of Advanced Encryption Standard (AES) S-boxes", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9, Issue-1, May 2020, DOI:10.35940/ijrte.F9712.059120
- [18] SyihamMohdLokman, ChuahChaiWen, NurulHidayahBinti Ab. Rahman, IsredzaRahmiBinti A. Hamid, "A Study of Caesar Cipher and Transposition Cipher InJawi Messages", Journal of Computational and Theoretical Nanoscience, March 2018 DOI: 10.1166/asl.2018.11130
- [19] Priti V. Bhagat, Kaustubh S. Satpute, Vikas R. Palekar, "Reverse Encryption Algorithm: A Technique for Encryption & Decryption", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 2 Issue 1 January 2013, ISSN: 2278-621X
- [20] X. Wu, X. Xu, F. Dai, J. Gao, G. Ji and L. Qi, "An Ensemble of Random Decision Trees with Personalized Privacy Preservation in Edge-Cloud Computing", 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), 2020, pp. 779-786, doi: 10.1109/iThings-GreenCom- CPSCom-SmartData-Cybermatics50389.2020.00134.
- [21] Yasir Nawaz and Lei Wang, "Block Cipher in the Ideal Cipher Model: A Dedicated Permutation Modeled as a Black-Box Public Random Permutation", December 2019, Journal of Symmetry 2019, 11, 1485; doi:10.3390/sym11121485
- [22] V. K. Soman and V. Natarajan, "An enhanced hybrid data security algorithm for cloud", 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), 2017, pp. 416-419, doi: 10.1109/NETACT.2017.8076807.

DEVELOPING NEXT GENERATION CAMPUS USING IOT DEVICES WITH ENHANCED SECURITY

SAVIMA K ^[1] Dr.M.V.SRINATH ^[2]

PG and Research Department of Computer Science,Sengamala Thayaar Educational Trust Women's College (Autonomous),(Affiliated to Bharathidasan University, Tiruchirappalli) Sundarakkottai, Mannargudi – 614001.Email: savimastet@gmail.com¹, sri_induja@rediffmail.com²

ABSTRACT

With the advent of IoT, and smart technologies evolving rapidly, many campuses are recognizing their significance in optimizing student and faculty success. The number of IoT connected devices is expected to skyrocket to over 75 billion by 2025. Today, connected devices, cameras, sensors, and machines – all embedded with smart technology – are increasingly used on campuses throughout the nation. So much so that we are starting to call some of them smart campuses. Similar to smart cities, a smart campus utilizes smart technologies to create new experiences and services. These smart technologies, which are connected to the Internet and AI-driven – can improve various aspects of the student and faculty experience on campus such as:

- Enhancing the financial aid process
- Improving student services
- Reducing wait times
- Mitigating compliance mistakes
- Reducing human errors
- Automating workflows
- Conserving energy and resources

KEYWORDS

Artificial Intelligence, IoT, Sensor, Cryptography, Smart Campus

1. INTRODUCTION

The Internet of Things (IoT) has brought many positive changes in our lives. In addition to areas such as industry, Smart cities, Transport, Health, Agriculture and various other areas, IoT will also have a large organization at universities or colleges. In this age of digital transformation, IoT technology can be used to maintain a smart and secure environment in a university or college. The purpose of this project is to focus on how the IoT plays an important role in building a smart Education campus and studying the challenges in its implementation. In the proposed project sensor gets the temperature of the students automatically through wireless connection via Bluetooth, the hardware consists of a microcontroller and module, software cloud services are monitored.

Coronavirus found mainly in Wuhan, China, has spread rapidly to many countries, including India, the world's second most populous country with a population of over 134 billion [20], [21], [22]. India could have a problem in avoiding the spread of coronavirus. Facial masks and sanitizers are the most active ways to reduce infections. To reducing the transmission of disease, this has exposed positive outcomes. Most of the time, it

is broadcast indirectly in many places. The development time can be very high, ranging from 10 to 14 days in severe cases, and the virus can spread straight (from person to person) through respirational dewdrops [2]. Governments have used a variety of preventive and preventive measures to reduce the spread of disease, including segregation of people, forced indoor face mask, confinement, restrictions on civilian movement within national and foreign borders, segregation, segregation and the withdrawal of major events and public meetings [10]. All kinds of games are all affected by the COVID-19 epidemic [4]. People with the flu should not be allowed to enter common areas since they are at more hazard of contamination and the spread of the virus and so wearisome a mask is important. At the doors of any city, workrooms, supermarkets, and hospital entries, infection tests are also required. This proposed method automatically displays a person's body temperature on entry is improved. Improved vision is used in this arrangement to several factors including fever identification.

The COVID 19 epidemic is causing a lifelong widespread. The most effective protection is a sensitive face cover in common areas. COVID 19 has required regulators around the world to use lockdown locks to prevent the spread of the virus. Based on research reports, wearing a face mask in open areas minimizes the threat of further spread. In this paper, an intelligent educational center powered by IoT uses a model to assess fever identification. This can be used by educational institutions. As a result a commercial and reliable way to use AI and nerves to create a strong atmosphere. In addition, each student's body temperature is monitored using an unaffected temperature sensor. This proposed system could identify COVID 19 users by assisting Internet of Things (IoT) technology.

Colleges and universities can benefit in many ways, including:

- **Easy access to student information as well as temperature.** An effective document management system significantly reduces the time spent by staff searching for information. Student records are organized, and easily available. Staff have more time to focus on helping students which improves the student experience.
- **Increased security.** The records management system allows colleges to protect important student records and other sensitive information. Participants can easily assign appropriate users and block unauthorized users.
- **Reduced costs.** With Enrolment, colleges need to find new ways to reduce their operating costs. The document management system removes huge costs for paper, ink, and storage. And since documents are easily accessible, there are no inefficiencies and additional costs for obtaining or reproducing a lost document.
- **Environment.** Many students are committed to natural causes and urge their institutions to have an eco-friendly nature. With a text management system, no paper is spilled. And with cloud-based document management and / or business process management (BPM) solutions, students significantly reduce their energy use.

2. Related Work

The essential of body temperature testing in clinical analysis and treatment cannot be exaggerated [23], [24], [25]. There are other disadvantages, which include low measurement exactness and long measurement time. Indigenous performance measures make it hard to find a patient's body temperature robotically and exactly. To solve the problem, they introduced a centralized monitoring scheme that is used to find body fever.

Immediate statistics is important in the area of public health [11]. This paper explains how to track a person's heart rate per second and find the normal body temperature away. Reduce the power consumption of the device by starting the device with the remote-control instruction from the delivery PC [26], [27], [28].

The function of the Radiocommunication Neighborhood Network in a growing use that includes sensory vessels, medicinal uses, accommodation monitoring, and seismic investigation was examined [5, 17]. Wireless Sensor Network recently focused on local tracking and market utilization. The effectiveness of the WSN models created by the PIC is established in this project. To establish sensory events, common temperature sensor nodes are used in networks.

By performing tasks such as present chasing and post-event study, video statistics enhance video surveillance resources [6]. People will save time and money, and the performance of the monitoring system will improve. The paper describes the concepts behind these processes, as well as the most commonly used human acquisition and facial recognition methods. This new approach to problem solving has created a very simple solution that can be implemented in real time. The success of the algorithm in checking the video sequence provides important information to improve the efficiency of hidden face detection.

The Haar Cascade algorithm for online features of low-cost features was used using the Raspberry Pi method [13]. It is an advanced access control system.

A. *New incident studies show two home safety precautions and resource planning as evidence of performance during a deadly epidemic. Compatible applications for most mobile platforms are built using the Application Sheet Framework [20].*

1) *3. Hardware Requirements*

3.1. Raspberry Pi

The Raspberry Pi is a inexpensive price computer that links to a computer or television monitor as well as works with a standard input devices as shown in Fig. 1 [8]. It is a small useful device focused on coaching individuals about writing languages such as Scratch and Python. It will do all computer activities, such as browsing the Internet and playing games. It is used on a number of digital peripherals together with chirping birdhouses, musical instruments, and detectors, along with weather stations and IR cameras as they are able to communicate with the external location.



Figure 1. Raspberry Pi

3.2 R pi Cam (Raspberry Pi Camera)

The 8-megapixel Raspberry sensor Pi camera is used for this task. This contains 1080p30, 720p60, and 640x480p90 video support and adjustment support for 3270×2444 pixels. Figure 2 shows the Raspberry Pi camera module. Stable lens and Sony IMX219 image sensor designed for R Pi as a board extension. The Pi module is connected to the RPi by one of the small board ports in the upper part, and uses a special CSI GUI, designed specifically for camera communication.



Figure 2. Raspberry Pi Camera

3.3 IR sensor

IR sensors are used to calculate and display the quantity of persons entering and leaving a room. The operating voltage of the IR sensor is 5VDC. Fig. 3 shows an InfraRed sensor that incorporates a in-built light sensor with an ascending hole, and an modifiable sensor.



Figure 3. IR sensor

3.4 Temperature Sensor

Temperature sensor operates as an unobtrusive IR temperature reader that reads temperature without contact. A sound reduction thermometer and a powerful DSP unit are used that help achieve greater accuracy.



Fig 4. Temperature Sensor

CHALLENGES INVOLVED IN IOT ARCHITECTURE

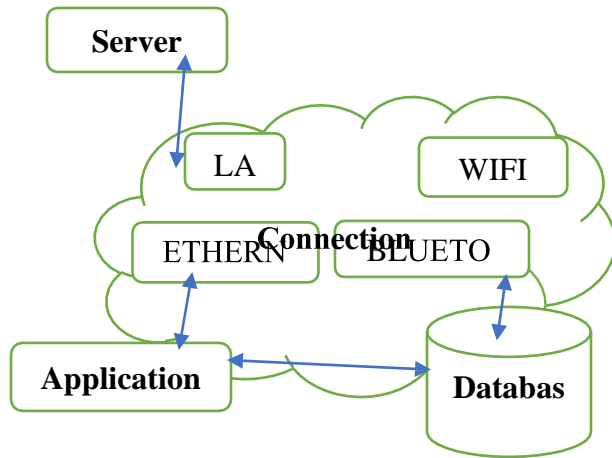
- Use of IoT – based applications are continuously growing vertically as well as horizontally. Due to evolving technology, there will be a need to update the devices & equipment. This will increase the cost, so Universities need to come with new ideas for finance including research in low-cost technology.
- Privacy & Security – IoT environment stores data on Internet based network. So, privacy and security issues become very crucial. The private information about the financial background of the family, medical records, and student’s progress should not be disclosed in any case. There is a need of high level of Encryption techniques to avoid the data from hacking.
- Energy efficiency - Lastly, additional vital challenge of IoT- enabled smart campuses is the power source and energy efficiency. Smart grid and solar based system can be used for efficient management and control on energy usage.

BRIEF EXPLANATION OF THE CAMPUS ARCHITECTURE

- **Server Module** - The first module contains a server, a status display board - The server can be considered the heart of the campus management system. The server captures, collects, stores and processes data generated in sensitive mode. The generated data will be stored using a specific data management system. The Dash Board can be used as a notice board where all the required information is displayed.
- **Connection module** - LAN, Wi-Fi, Bluetooth, mobile network. The IoT system can use an individual or combination of network technology to transfer data to a server. This data is processed and the microcontroller sends a signal that operates through these networks.

- **control frequency module** - has a variety of sensors, transmitters and actuators. Hundreds of continuous sensory devices will monitor the environment e.g. Temperature sensors, Motion sensors, RFID, IR sensor, Cameras.

- **Application module** - Application layer is the final layer that will receive and process information for various applications specific to specific applications.



CONCLUSION

Today's students need experience in a seamless college enrolment program and in need of sensor temperature automatically in this pandemic situation. In an increasingly competitive and dynamic environment, colleges and universities need to find new ways to improve the college enrolment system. Eliminating manual and paper-based documentation processes helps colleges save money, boost productivity, increase compliance, and provide superior student experiences. Organizations can eliminate these costs with automation technologies such as an electronic document management system and business process automation solution. These solutions allow colleges to store student records in one secure and centralized location.

REFERENCES

[1] A. Rosebrock, "COVID-19: Face Mask Detector with OpenCV, Keras/TensorFlow, and Deep Learning", May 4, 2020, <https://www.pyimagesearch.com/2020/05/04/covid-19-face-mask-detector-with-opencv-keras-tensorflow-and-deep-learning/>.

[2] A. Hidayat, Subono, V.A. Wardhany, A.S. Nugroho, S. Hakim, M. Jhoswanda, "Designing IoT-Based Independent Pulse Oximetry Kit as an Early Detection Tool for Covid-19 Symptoms ", 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE).

[3] Carlo Alberto Boano Matteo Lasagni Kay Romer Tanja Lange. "Accurate Temperature Measurements for Medical Research using Body Sensor Networks".

[4] Cristina S.....-C....., B-P BUTUNOI and Calin C.....," IoT Based Intelligent Building Applications in the Context of COVID-19 Pandemic", 2020, 10.1109/ISETC50328.2020.9301124

- [5] F.H. Yahaya, Y.M. Yusoff, H.Z. Abidin, R.A. Rahman, Development of a PIC-based wireless Sensor node utilizing XBEE Technology, IEEE International Conference on Information Management and Engineering, 2010.
- [6] Gayatri Deora, Ramakrishna Godhula and Dr. Vishwas Udpikar "Study of Masked Face Detection Approach in Video Analytics, IEEE Conference on Advances in Signal Processing, 2016.
- [7] <https://www.electronicwings.com/arduino/servo-motor-interfacing-with-arduino-uno>.
- [8] <https://www.robocarstore.com/products/raspberry-pi-4-model-b-board-with-1gblpddr4-sdram>.
- [9] K Baskaran, Baskaran P., N. Kumaratharan, Rajaram V., "IoT Based COVID Preventive System for Work Environment", the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) IEEE Xplore Part Number: CFP200SV-ART; ISBN: 978-1-7281-5464-0.
- [10] Lim, M.G., & Chuah, J.H. (2018). Durian Types Recognition Using Deep Learning Techniques. 2018 9th IEEE Control and System Graduate Research Colloquium(ICSGRC).doi:10.1109/icsgrc.2018.865753.
- [11] L. Tang and H. Kiourmars, Wireless Network for Health Monitoring Heart Rate and Temperature Sensor", 10.1109/ICSensT.2011.6137000.
- [12] T. Meenpal, A. Balakrishnan, A. Verma, Facial Mask Detection using Semantic Segmentation, 2019 4th International Conference on Computing, Communications and Security (ICCCS), 2019, doi:10.1109/cccc.2019.8888092.
- [13] Ana.nya Pa.ndey Mrudula, Kruthi.ka Dinesh, P. Reethika, Smart Door Unlocking System, Int. Res. J. Eng. Technol. (IRJET) Volume: 07 (2020) Issue: 05 May. [14] N. Ozkaya, S. Sagioglu, Intelligent face Mask Prediction System, IEEE International Joint Conference on Neural Networks, 2008.
- [15] Petrovic, N. & Radenković, M. & Nejkovic, V.. (2020). Data-Driven Mobile Applications Based on AppSheet as Support in COVID-19 Crisis. https://www.researchgate.net/publication/343678867_DataDriven_Mobile_Applications_Based_on_AppSheet_as_Support_in_COVID-19_Crisis.
- [16] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth," SSDMNv2: A real-time DNN-based face mask detection system using single shot multibox detector and MobileNetV2", Published online 2020 Dec 31.
- [17] Rit.wik Biswas, Av.ijit Roy, Real-Time Temperature Graph using MATLAB and Arduino, Int. J. Eng. Res. Technol. (IJERT) 09 (05) (2020) (May 2020) <https://www.ijert.org/real-time-temperature-graph-using-matlab-and-arduino> .
- [18] Sam.my V. Militante, Na.nette V. Dionisio, Deep Learning Implementation of Facemask and Physical Distancing Detection with Alarm Systems, in: Vocational Education and Electrical Engineering (ICVEE) 2020 Third International Conference on, 2020, pp. 1–5. <https://ieeexplore.ieee.org/document/9232610>.
- [19] S.S.. Vedaei, A. Fotovvat, M.R. Mohebbian, G.M.E. Rahman, K.A. Wahid, P. Babyn, H.R. Marateb, M. Mansourian, And Ramin Sami, "An IoT-Based System for Automated Health Monitoring and Surveillance in Post-Pandemic Life", date of publication October 12, 2020.
- [20] P. Subramani, G.B. Rajendran, J. Sengupta, R. Pérez de Prado, P.B. Divakarachari, A block bi-diagonalization-based pre-coding for indoor multiple-input-multiple-output-visible light communication system, Energies 13 (13) (2020) 3466.

- [21] T. Nguyen, B.H. Liu, N. Nguyen, B. Dumba, J.T. Chou, Smart grid vulnerability and defense analysis under cascading failure attacks, *IEEE Trans. Power Delivery* (2021).
- [22] V. Vinitha., V. Velantina., COVID-19 facemask detection with deep learning and computer vision, *Int. Res. J. Eng. Technol. (IRJET)* Volume: 07 (2020) Issue: 08 Aug.
- [23] S. Rajendrakumar, V.K. Parvati, Automation of irrigation system through embedded computing technology, in: *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 289–293.
- [24] S. Vadivel, S. Konda, K.R. Balmuri, A. Stateczny, B.D. Parameshachari, Dynamic Route Discovery Using Modified Grasshopper Optimization Algorithm in Wireless Ad-Hoc Visible Light Communication Network, *Electronics* 10 (10) (2021) 1176.
- [25] Yu. Chen, H. Zhang and NA. Wang, "Body Temperature and Alarm System Used in Hospital-Based on 1-wire and Wireless Communication Technology", 2008 International Workshop on Education and Training & 2008 International Workshop on Geoscience and Remote Sensing, pp. 401-404, 2008.
- [26] M.A. Naeem, T.N. Nguyen, R. Ali, K. Cengiz, Y. Meng, T. Khurshaid, Hybrid Cache Management in IoT-based Named Data Networking, *IEEE Internet of Things J.* (2021).
- [27] S.I. Chu, C.L. Wu, T.N. Nguyen, B.H. Liu, Polynomial Computation Using Unipolar Stochastic Logic and Correlation Technique, *IEEE Trans. Comput.* (2021).
- [28] Y. Sun, J. Liu, K. Yu, M. Alazab, K. Lin, "PMRSS: Privacy-preserving Medical Record Searching Scheme for Intelligent Diagnosis in IoT Healthcare", *IEEE Trans. Ind. Inf.*, doi: 10.1109/TII.2021.
- [29] IoT smart campus review and implementation of IoT applications into education process of university | *IEEE Conference Publication* | *IEEE Xplore*

ACTION OF SECURE MESSAGES USING AES ALGORITHM AND BLOCKCHAIN TECHNOLOGY

R.AKILANDESWARI^[1]

^[1]Assistant Professor, PG and Research Department of Computer Science, Sengamala Thayaar Educational Trust Women's College (Autonomous), Mannargudi, Tamilnadu, India, akil30kmr@gmail.com.

Abstract: In recent years, there has been a rapid increase in the use of mobile devices and messaging applications for communication, leading to a growing concern about the security of text messages exchanged through these platforms. This study proposes a novel method that uses the AES algorithm and Blockchain to secure text messages in messaging applications on mobile devices. The AES algorithm is selected due to its faster encryption and decryption processes, which are superior to asymmetric cryptography algorithms. On the other hand, Blockchain is chosen for its inherent security properties that only allow data addition and cannot be altered. This study aims to achieve both speed and security to prevent cybercrime in text messages. The Avalanche Effect calculation and Processing Time measurement are used as the analysis methods to evaluate the proposed approach. The results show that the computation time of the message delivery process using Blockchain and AES algorithm has an average total process time of 33.59 milliseconds. Additionally, the testing results of the Avalanche Effect value show that the AES algorithm has a value of 50% for character lengths up to 16 characters and below 50% for character lengths greater than 16 characters. Based on these testing results, the proposed combination of the AES algorithm and Blockchain is an effective method for securing text messages in messaging applications. This method can offer a secure and efficient way of exchanging text messages on mobile devices and can adopt as a standard approach for messaging applications.

Keywords: AES Algorithm, Avalanche Effect, Blockchain, Chatting, Cryptography

INTRODUCTION

The development of information technology continues to increase, resulting in increased human needs as well. Communication is essential to facilitate daily human performance so that community productivity does not decrease and community performance is maintained. Chat Messenger is message communication that mobile smartphone users widely use, but the messages sent are not necessarily safe from cybercrime or hacker. Therefore, cryptography is critical to prevent hackers from eavesdropping on the messages sent. Cryptography is a knowledge to maintain the security of text messages (plaintext) by encrypting messages into a form that is difficult to read (ciphertext). When decrypting, a ciphertext is converted into plaintext. This encryption and decryption protect messages from unauthorized parties viewing the messages contents

Cryptography has several methods, including the Advanced Encryption Standard (AES) algorithm. AES is a block cipher algorithm that uses a key when encrypting and decrypting. AES algorithm has various block sizes, such as 128-bit, 192-bit, and 256-bit. The differences between the three versions of AES affect the number of keys and their rounds

Blockchain is a reliable technology that allows participating nodes to exchange data reliably without a central management system. Blockchain uses a point to point method called "flooding" as a data transfer mechanism, where the data is represented as a transaction and a block. In order to avoid needless spread in such a transmission system, all nodes in the block chain search 'memory pool.' There is no limit to the rate of output of transactions in the current blockchain. Many creative uses of blockchain technology are being investigated, including crypto currencies, smart contracts, health care, communication systems, IoT, financial systems, censorship resistance, electronic voting and distributed provenance. These applications benefit from an append-only approach, in which "transactions" acknowledged in Blockchain cannot be updated, thanks to Blockchain's transparent as well as fully distributed peer-to-peer design. The Blockchain's transparency allows for the storage of publicly verifiable and incontrovertible records

LITERATURE REVIEW

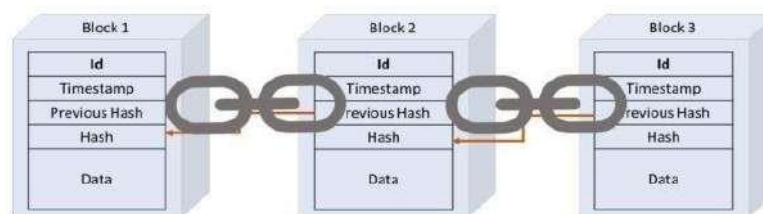
1. Blockchain

A blockchain is a list of records called blocks and is interconnected and secured by cryptographic methods Blockchain is a decentralized ledger, not a central authority. In other words, the term blockchain itself is a blockchain. A block with a data structure contains data and some other attributes. Blocks can be linked with other blocks to form a blockchain. Block basic components:

1. Hash: Unique identifier of the current block with a unique value.
2. Timestamp Value: This takes the hash of each block of items to be timed and publicly publishes the hash.
3. Previous Hash: Hash of the previous block.
4. Data: Contains data based on the type of Blockchain.

A hash can be treated like a fingerprint to identify a block. Figure 1 shows a representation of a blockchain. The first block is often called the genesis block in the blockchain world and does not reference the previous block (Das, 2020).

Figure 1. Representation of Blockchain (Das, 2020)



One of the mechanisms of Blockchain is the proof-of-work mechanism. With this mechanism, the node that can add data to the Blockchain is the node that solves the math puzzle first. When a node successfully solves a puzzle, the Blockchain protocol automatically issues a new puzzle for the existing nodes to solve. This mechanism is known as mining

This mechanism is useful for implementing peer-to-peer distributed server timestamps and involves scanning for values that have been hashed. For example, like SHA-256, the hash starts with a zero bit number. The average effort required is exponential to the number of zero bits required and can be verified by executing a single hash.

In this timestamp network, the proof-of-work mechanism is implemented by an incremental sequence of the nonce in the block until it gets a value that can give the hash of the block the required number of zero bits. Once the computational power is increased to meet the proof-of-work requirements, the block cannot be changed without restarting the process. When the blocks are connected afterwards, attempts to change the block require repeating the creation of the blocks after that as well

2. AES (Advanced Encryption Standard)

Advanced Encryption Standard (AES) algorithm is a block cipher algorithm with a symmetrical nature that uses a symmetric key during the encryption and decryption process. AES algorithm has various key lengths, namely 128 bits, 192 bits, and 256 bits. The difference between the three variations is the length of the key which affects the number of rounds. The following is a comparison of the processes of each variation of AES:

Table 1. AES Algorithm Data Sequence

	length	length	r of Rounds
128			
192			
256			

AES 128-bit encryption process can be done by selecting the block size and key so that the number of processes passed can be determined. There are 4 round transformations carried out in the encryption and decryption process:

1. SubBytes: replaces the contents of the bytes using the substitution table.
2. ShiftRows: performs the process of shifting blocks per line in the state array.
3. MixColumn: performs the randomization process of data in each state array.
4. AddRoundKey: combine state array and round key with XOR.

In the AES decryption process:

1. InvShiftRows: performs a right shift of the bits in each row block
2. InvSubBytes: maps each element to the state with an Inverse S-Box table.
3. InvMixColumn: multiply each column in the state by the AES matrix.
4. AddRoundKey: combine state array and round key with XOR.

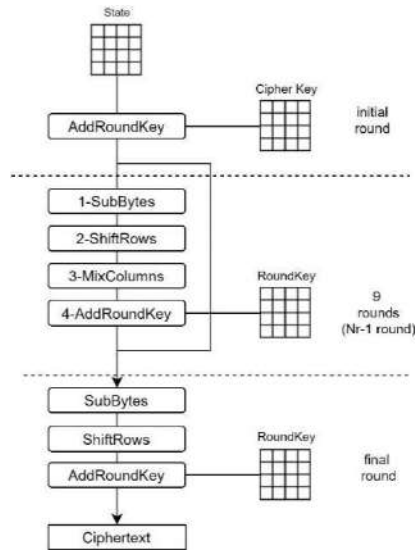


Figure 2. AES Encryption Process

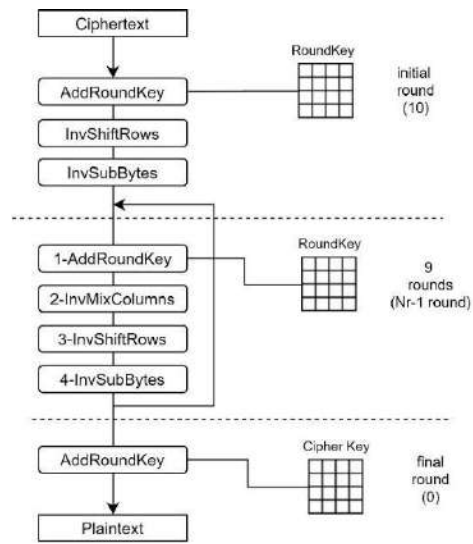


Figure 3. AES Decryption Process

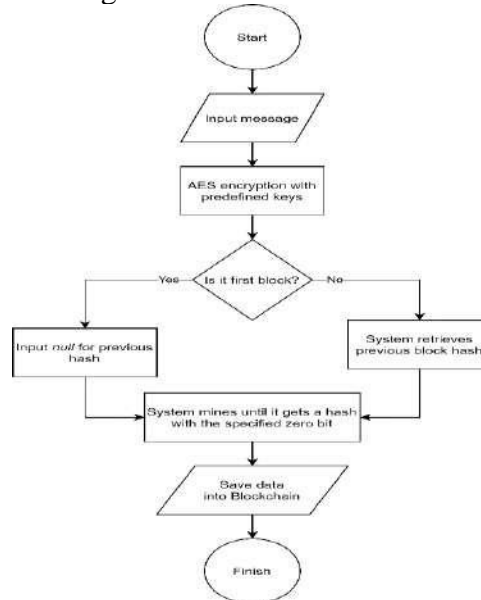
METHOD

Dataset

The data used in this study was obtained from a text data source site called <https://lipsum.com/>. The dataset used in this study was 1,500 characters long. The character combines Latin words and phrases with no meaning or relevance to the research question. The dataset is common in fake texts used in many applications.

Data Processing

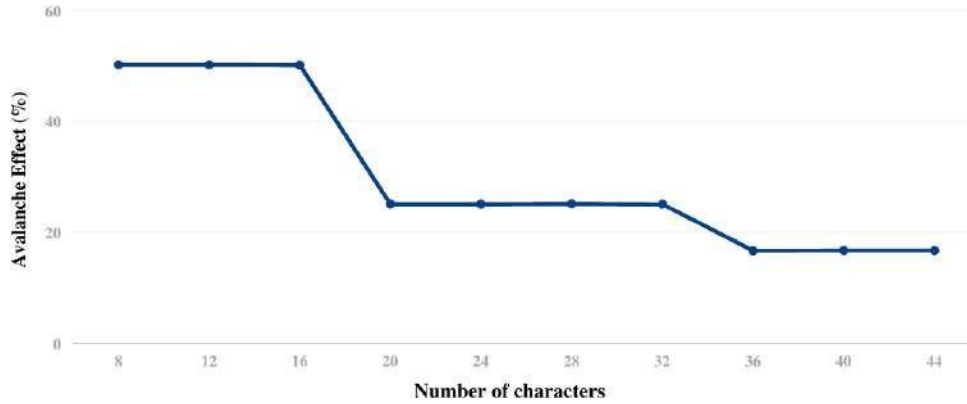
The process of implementing the AES algorithm and Blockchain in this study is depicted



AES Method and Blockchain Implementation Process

Avalanche Effect

This test is carried out by calculating the Avalanche Effect obtained from the text that has been encrypted and has a difference of one bit. This test is carried out on a dataset of 1,500 characters and then divided into blocks with a length of 8 to 44 characters in each block, after which 1-bit changes will be made in each block

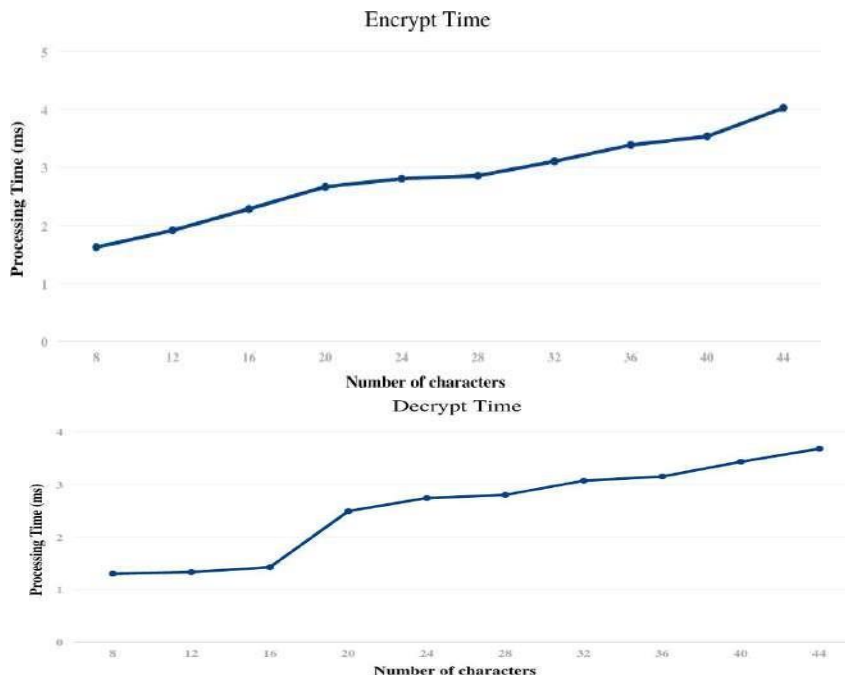


to produce a test of 12,609 times. The data results from the Avalanche Effect test can be seen in Figure 5.

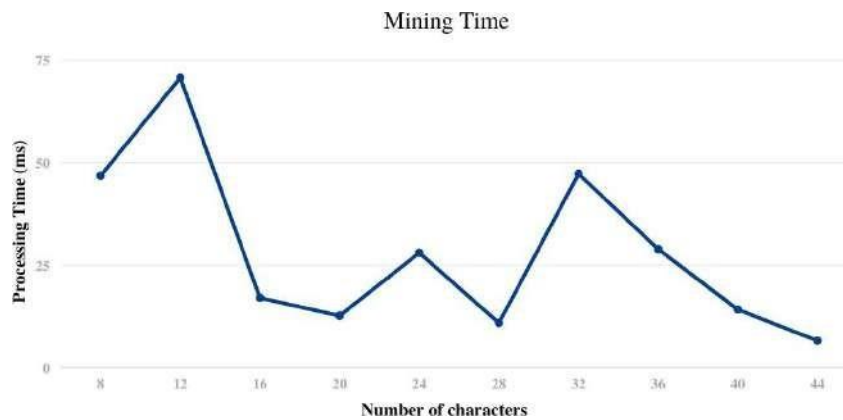
The avalanche effect above shows that the AES algorithm method produces percentage values around 50% for 8, 12, and 16 characters. However, character lengths above 16 characters will result in a lower Avalanche Effect value because the AES-128 algorithm can only encrypt up to 128 or 16 characters, so the longer the character of a text, the result the percentage value of the Avalanche Effect will decrease.

Processing Time

The Processing Time calculation calculates the computational time required to secure text messages using Blockchain and the AES algorithm. The graph is summarized in the line graph below.



It can be concluded from Figure 6 and Figure 7 that the cryptography algorithm for securing messages using Blockchain and the AES algorithm produces different times according to the length of the character. The longer the character of a text, the longer the Processing Time required. In the message encryption and



decryption process, the processing time increases as the length of a text message increases, with an average encryption and decryption time of 2.83 ms and 2.54 ms.

CONCLUSION

Based on the results of this study, the avalanche effect test applied to measure the level of security in the implementation of the AES algorithm obtained a calculation with the avalanche effect percentage value of 50% for messages with sizes of 8, 12, and 16 characters. Based on the processing time result with the AES algorithm and Blockchain method, the message security process can maintain the security and confidentiality of the message with the time required for the encryption and decryption process, which increases every time the character length of the text message is increased. However, in contrast to the encryption and decryption time, the mining time on the Blockchain does not increase based on the size of the character length and produces an indeterminate time with the lowest time of 6.52 ms, the highest time of 70.72 ms, and the average of 28.25 ms.

REFERENCES

- [1] Alfajar, F., & Akbar, M. (2021). Implementasi Keamanan Chat Realtime Menggunakan Aes- Cbc Dan Base64.
- [2] *Journal of Information System and Artificial Intelligence*, 1(2), 2-4.
- [3] Ali, A. H., & Sagheer, A. M. (2017). Design of Secure Chatting Application with End to End Encryption for Android Platform. *Iraqi Journal for Computers and Informatics (IJCI)*, 43(1), 23-25. doi:http://dx.doi.org/10.25195/2017/4315
- [4] Argani, A., & Taraka, W. (2020). Pemanfaatan Teknologi Blockchain Untuk Mengoptimalkan Keamanan Sertifikat Pada Perguruan Tinggi. *ADI Bisnis Digital Interdisiplin Jurnal*, 1(1), 12-15.
- [5] Astuti, N. I., Arfani, I., & Aribowo, E. (2019). Analysis of the security level of modified CBC algorithm cryptography using avalanche effect. *IOP Conference Series: Materials Science and Engineering*, 5-6.
- [6] Basri. (2016). Kriptografi Simetris dan Asimetris dalam Perspektif Keamanan Data dan Kompleksitas Komputasi. *Scientific Journal of Computer Science, Faculty of Computer Science, Al Asyariah Mandar University*, 2(2), 18-20.
- [7] Das, S. K. (2020). Secure Messaging Platform Using Blockchain Technology. *International Journal of Research in Engineering and Science (IJRES)*, 8(12), 27.
- [8] Harahap, A. K., Oktari, N. S., Kartini, A., Agung, A. A., & K, R. B. (2020). Perbandingan ROI

- Metode Konsensus Proof of Work, Proof of Stake, dan Proof of Service (Masternode). *Jurnal Teknologi Informasi dan Manajemen*, 2(2), 2-5.
- [9] Ilham, L. I., & Widyassari, A. P. (2021). Pengembangan Aplikasi Pesan Instan Terenkripsi Menggunakan Algoritma Kriptografi AES (Advanced Encryption Standard). *Jurnal Teknik Elektro Smart*, 1(1), 1-4.
- [10] Kurnia Hu, S. D., Palit, H. N., & Handojo, A. (2019). Implementasi Blockchain: Studi Kasus e-Voting. *Jurnal Infra*, 7(1), 184-185.
- [11] Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. *Bitcoin*, 1-8.
- [12] Prameshwari, A., & Sastra, N. P. (2018). Implementasi Algoritma Advanced Encryption Standard (AES) 128 Untuk Enkripsi dan Dekripsi File Dokumen. *EKSPLORA INFORMATIKA*, 8(1), 52-54.
- [13] Putri, A. E., Kartikadewi, A., & Rosyid, L. A. (2020). Implementasi Kriptografi Dengan Algoritma Advanced Encryption Standard (AES) 128 Bit Dan Steganografi Menggunakan Metode End Of File (EOF) Berbasis Java Desktop Pada Dinas Pendidikan Kabupaten Tangerang. *Applied Information Systems and Management*, 3(2), 70-77.
- [14] Randi, A., Lazuardy, K., Chandra, S., & Dharma, A. (2020). Implementasi Algoritma Advanced Encryption Standard pada Aplikasi Chatting berbasis Android. *JIKOMSI Jurnal Ilmu Komputer dan Sistem Informasi*, 3(1), 2-4.
- [15] Takale, A. P., Vaidya, C. V., & Kolekar, S. S. (2018). Decentralized Chat Application using Blockchain Technology. *International Journal for Research in Engineering Application & Management (IJREAM)*, 92-93.
- [16] Verma, R., & Sharma, A. K. (2020). Cryptography: Avalanche effect of AES and RSA. *International Journal of Scientific and Research Publications*, 10(4), 119-122.
- [17] Yusfrizal. (2019). Rancang Bangun Aplikasi Kriptografi pada Teks Menggunakan Metode Reverse Cipher dan Rsa Berbasis Android. *Jurnal Teknik Informatika Kaputama*, 3(2), 30-33.

HYBRID GENERAL ADVERSARIAL NETWORK(HGAN) FOR CLOUD WORKLOAD PREDICTION AND TREND CLASSIFICATION

V.MANGAIYARKARASI^[1]

^[1]Assistant Professor , Sengamala Thayaar Educational Trust women's College (Autonomous),
Mannargudi, Tamilnadu, India. mangaiyarkarasi2906@gmail.com Mobile: 9751630620

ABSTRACT

Hybrid generative adversarial network (HGAN) for which we can enforce data density estimation via an autoregressive model and support both adversarial and likelihood framework in a joint training manner which diversify the estimated density in order to cover different modes. I propose to use an adversarial network to \textit{transfer knowledge} from an autoregressive model (teacher) to the generator (student) of a GAN model. Efficient resource management approaches have become a fundamental challenge for distributed systems, especially dynamic environment systems such as cloud computing data centers. These approaches aim at load-balancing or minimizing power consumption. Due to the highly dynamic nature of cloud workloads, traditional time series and machine learning models fail to achieve accurate predictions. In this paper, we propose novel hybrid VTGAN models. Our proposed models not only aim at predicting future workloads but also predicting the workload trend (i.e., the upward or downward direction of the workload). Our results show that VTGAN models outperform traditional deep learning and hybrid models, such as LSTM/GRU and CNN-LSTM/GRU, concerning cloud workload prediction and trend classification.

Keywords: density, autoregressive model, dynamic environment, cloud workload

INTRODUCTION

The QoS (Quality of Service) of various sorts of apps is duty of the Internet Service Provider (ISP). In SDN (Software Defined Networking), centralised network controller allows for the instillation of intelligence in the management of network resources based on QoS criteria. By recognising various traffic flow types and categorising them according to various application/classes, a fine-grained QoS Traffic Engineering may be achieved. Previous approaches, such as port-based classification and DPI are found to be ineffective and computationally intensive. As a result, machine learning (ML)-based traffic classifiers have gotten a lot of attention from research community, as seen by a rise in the number of papers published [1]. Artificial intelligence (AI) has advanced rapidly in recent years, particularly with breakthroughs of ML and DL. It is well known in the study of networks that applying machine learning techniques to real-world problems is both possible and promising. To bridge the network with ML, some pioneering work has been done. ML is a technique that allows a computer to analyse data and derive knowledge. It is more than just learning or extracting information; it also entails putting that information to use and developing it through time and with experience. Purpose of machine learning is to find as well as exploit hidden patterns in "training" data. Unknown data is analysed using the patterns learned so that it can be grouped together or mapped to recognised groups. Traditional programming model, in which programmes are developed to automate processes, is shifted as a result of this. The programme that fits data is created using machine learning. ML has recently regained popularity. [2].

SDN is currently viewed as a promising networking paradigm that has potential to drastically enhance network resource usage, simplify network management, lower operating costs and foster

innovation as well as development [3]. The separation of control plane and data plane, which is a key component of SDN, necessitates rethinking and revamping traffic engineering (TE) solutions in order to fully leverage SDN's promise qualities, such as openness, programmability, and global visibility[4]. Because SDN handles network traffic based on "flows," the traffic categorization (TC) engine's accuracy and efficiency are critical in SDN[5]. On the one hand, in SDN, delivering appropriate QoS for various traffic flows is an important aspect of TE. As a result, TC engine must determine QoS class for specific traffic flows in order to select the best routing paths. Traditional traffic classification systems, on the other hand, which seek to classify specific application of every network flow, are ineffective in identifying the QoS class of traffic flows since several applications may belong to same QoS class, which has comparable QoS criteria. Furthermore, with so many new applications appearing every day, keeping a real-time update of list of all applications available on Internet is both time demanding and unfeasible[6].

This paper's contribution is as follows:

- To enhance the scalability and security based on machine learning techniques integrated with routing and intrusion based system
- The scalability of the computer network has been enhanced using hybrid General Adversarial network with cognitive routing protocol in which the HGAN will classify the network traffic with higher data transmission by integrating cognitive based routing protocol.
- The network breach was detected using an authenticated cryptographic intrusion detection system, which improved the system's security by applying cryptographic-based encryption and decryption during data transmission.
- The experimental results shows network scalability and network security based parameters in terms of accuracy, precision, end-to- end delay, scalability, network throughput.

RELATED WORKS

Despite the fact that researchers have contributed and laid the groundwork for developing an intrusion detection system using various techniques, much of the previous work in network intrusion detection has focused on developing a predictive model [7] for detecting normal or abnormal network traffic. The authors in [8] offered a SOM-based solution. They relied on flow information gathered from forwarding devices. Parameters utilized to train SOM are viewed from perspective of switch, not the controller. They comprise average number of average bytes per flow, packets per flow, percentage of pair flows, average length per flow, single flow growth and single port growth. This approach has a number of drawbacks. Because intelligence is built on the data plane, it violates SDN principles. Furthermore, the technique's efficiency must be improved. Another technique was proposed with the goal of detecting DDoS assaults in the SDN control plane by computing an entropy value[9]. The authors of this system employed an experimentally set threshold. The method has significant limits and may not be suitable for all situations. As a result, the given method suffers from a lack of stability as well as adaptability to changes in network status. In, a method for detecting DoS attacks is provided that operates on top of NOX controller[10]. This module uses authorization data to take extra precautions when determining a packet's origin. has a low detection accuracy that needs to be enhanced The source MAC address, source IP address, source port number and source switch ID are

all recorded. Normal packets are allowed, however malicious traffic is rejected by this module. The second module is the classification operation. This module uses a feature vector to determine the class of network traffic using an artificial neural network approach called SOM. The proposed method is simple to use and successful in detecting attack. However, it suffers from a lengthy training period that must be cut. It also [11]. The authors implement a two-stage detection approach for DoS attacks in SDN in their paper [12]. The packet rate is calculated in the first stage using the controller's information on a flow. If the packet rate exceeds a predetermined threshold, the ML-based model recognises the attack. This method is successful in detecting the attack. It does, however, have significant limits. Because the attack is carried out in two stages, it takes a considerable time to notice it. Its detection accuracy also needs to be improved. In their paper [13], the authors describe a method for detecting anomalous SDN streams in SDN architecture. They used their methodology to detect anomalies in DDoS attacks. As the basic algorithm for flow categorization and attack detection, they used DPTCM-KNN approach. Technique shown is effective, however detection accuracy as well as classification performance gained are improved. We learned from the literature that a variety of intrusion detection systems have been built utilising various machine learning algorithms. For example, several research papers use single learning techniques to construct intrusion detection models, such as self-organizing maps [14], neural networks [15], genetic algorithms [16], decision trees [17], and pattern matching algorithms [18]. On the other hand, certain intrusion detection systems, such as the hybrid approach or ensemble techniques [19], are created by mixing several machine learning algorithms and ensemble classifiers [20]. All of the aforementioned strategies are built as a predictive model in order to detect or categorise whether incoming network data is intrusion or legitimate access. There is, however, no attempt to create a scalable and adaptable intrusion detection solution.

Proposed machine learning techniques with routing protocol integrated with authentication system:

This section discuss about the proposed model for enhancing the network scalability and security based on machine learning techniques and authentication based security systems. The QoS of network has been improved by design of HGAN-CRP in traffic classification and data transmission. The machine learning based technique will predict the traffic of the network based on TPR and FPR from classification results of HGAN based CRP. When the network traffic is high, then based on prediction of intrusion in network, the ACIDS has been established to detect the intruder and enhance the network security. The overall architecture is given in figure-1.

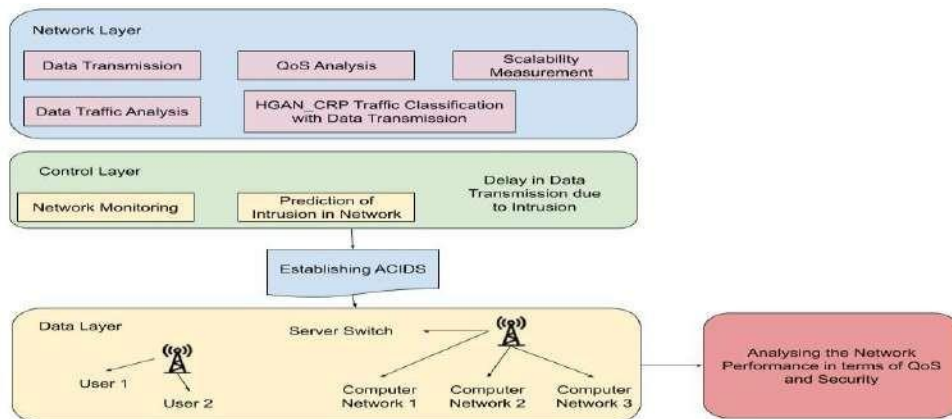


Figure-1 overall proposed system design

Hybrid General Adversarial network with cognitive routing protocol:

G captures the dataset's probability distribution in its entirety. D has no way of distinguishing between the produced and genuine samples. Due to the incorporation of adversarial principles, GAN can transform unlabeled datasets into labelled data (actual samples). As a result, GAN samples created by G can be considered fraudulent samples. Assume that x is a data sample. G's input is represented by p(z), which is a type of noise data with a Gaussian distribution. The mapping of noise data to produced samples is represented by G(z). The likelihood that x is a real sample rather than created samples is represented by D(x). As a result, GAN are thought of as a issue of maximisation and minimization of G and D. Objective function is given by eqn (1):

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \tag{1}$$

In this formula, p_{data}(x) and p_z(z) are probability distribution and prior distribution of real sample, respectively. After produced sample passes D, D(G(z)) is chance of being judged as a real sample. G's goal is to make produced sample look like a real sample. D(G(z)) declines as well. Goal of D is to evaluate x as a representative sample. D(x) will be close to 1 and D(G(z)) will close to 0 since created samples will be judged as no genuine samples. A model is fixed during the specific training procedure. Then, using alternate iteration training, another parameter is updated to maximise opponent's error and establish Nash equilibrium. Adaptive training allows both G and D to reach the ideal value. The total loss function is broken into two sections: supervised learning

$H_{supervised}$ and unsupervised learning loss function $H_{unsupervised}$ to make full use of unlabeled data to aid in supervised learning training categorization. Goal of $H_{supervised}$ is to predict that D will be accurately classified based on multiclass label sample data's true probability distribution p_{data}(x, y) given by eqn (2).

$$H = -E_{x,y \sim p_{data}(x,y)} [\log_{p_{model}}(y|x)] - E_{x \sim G} [\log_{p_{model}}(y = K + 1|x)] = H_{supervised} + H_{unsupervised}$$

$$H_{unsupervised} = -E_{x,y \sim p_{data}(x,y)} \log_{p_{model}}(y|x, y < K + 1) \tag{2}$$

For unlabeled samples, $H_{unsupervised}$ is unsupervised learning loss function. $P_{model}(y = K + 1|x_j)$ indicates sample and x_j is producing sample probability. When $D(x) = 1 - P_{model}(y = K + 1|x_j)$ is satisfied, then $H_{unsupervised}$ is given as eqn (3):

$$H_{unsupervised} = - \left[E_{x \sim p_{data}(x)} \log D(x) + E_{x \sim G} \log_{p_{model}}(1 - D(G(x))) \right] \tag{3}$$

The classification flow chart of HGAN is given in figure-2.

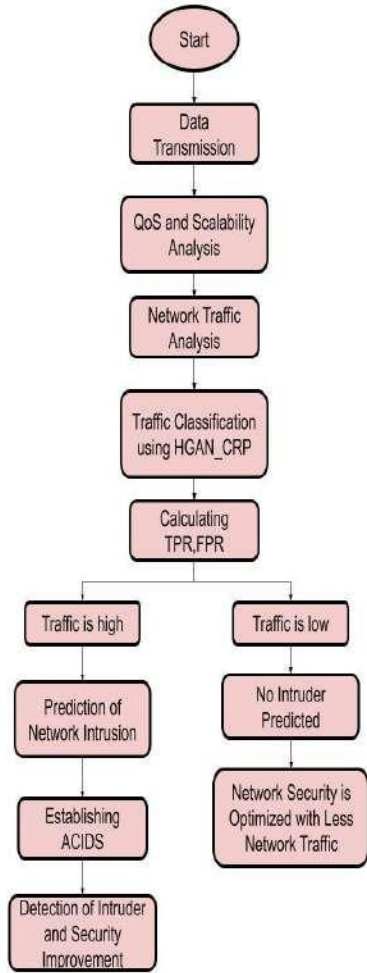


Figure-2 classification flow chart of HGAN

Hybrid generative-adversarial networks (fig. 3) emerged as a result of the development of generative modelling, and they are made up of two NN:

Generator:This NN is required for development of new features X to be learned. Features must be formed in such a way that discriminator cannot identify them as false features. Discriminator is a criterion for determining generator's quality.

Discriminator:Its major role is to distinguish genuine X indications from false X signs. When the discriminator model receives a genuine sign, it outputs a number labelled "real" (in the range of 0 to 1), and when it receives a generated sign, it outputs a number labelled "fake." The discriminator's level of confidence that input feature belongs to a real or artificial class is represented by these numerical numbers. Discriminator class output "real" should incline to 1 when input is a real sign, forexample.

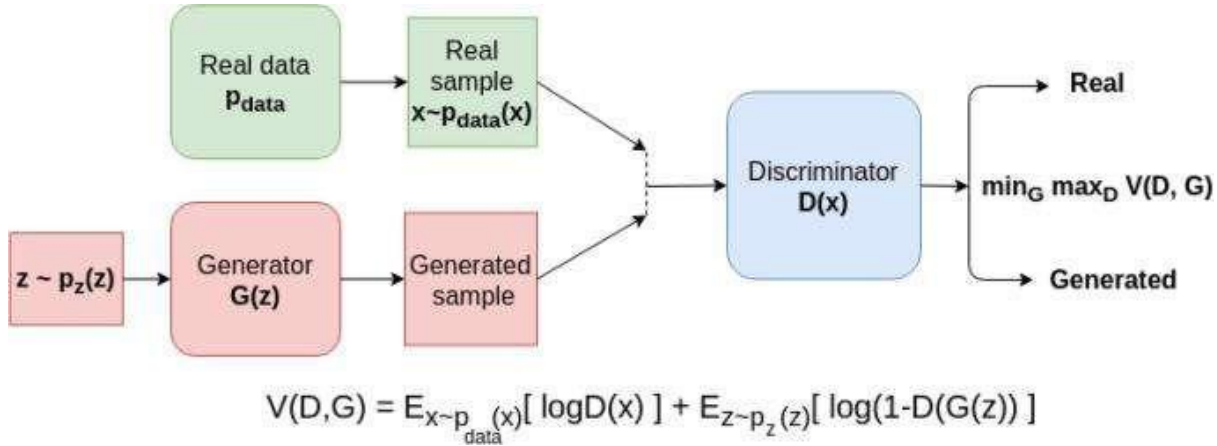


Figure-3 Hybrid Generative adversarial networks (HGAN) traffic classification model

The adversarial samples are introduced to HGAN training to improve classification robustness by taking advantage of HGAN's deep learning advantages. In the meantime, HGAN's optimization efficiency can be increased. G and D have more experience, and the model is more generalizable. The G and D will achieve their objectives after HGAN has been better trained. The loss function of G is designed as follows:

$$L_{pb}(G) = E_{z \sim p_{z,c_{fake}} \sim P_c} [\log(1 - D(G(z, c_{fake})))]$$

$$L_G = \omega L_{pb}(G) + \theta L_{fl}(G) + \chi L_{sh}(G)$$

$$L_{fl}(G) = E_{z \sim p_{z,c_{fake}} \sim P_c} [L_D(c_{fake} | G(z, c_{fake}))] \tag{4}$$

The generator G's function is as follows: given random noise z and fake category data cfake, produced sample xfake is output. Loss function is a function that calculates how much money. Weighted sum of three loss functions is LG of G. They contain xfake's discriminant loss Lpb (G), which D uses to determine whether a sample is real or manufactured. Lfl (G) is the classification loss of xfake, which is classed and forecasted. The input datum for xfake is the confrontation loss Lsh (G). In Eqs. (6–8), ω, θ and χ are weights of three loss functions.

As a result, weight are set so that G increase created samples aggressiveness. D also has a better ability to distinguish the samples. With X*real as input data, X* real is output feature sample data from feature selector. Weighted sum of four loss functions is defined as loss function L(D). There are four sections to loss function: the discriminant loss L * pb (D) that X * real is judged as real or false sample by D. xfake is classified by D to discriminate classification loss L fake fl (D). The four loss functions are expressed by Eqs

$$L_{pb}^*(D) = - E_{x_{real} \sim P_{real}, f_y} [\log D(f_y(x_{real}))]$$

$$L_{fl}^*(D) = - E_{x_{real} \sim P_{real}, f_y} [L_D(c_{real}^* | f_y(x_{real}))]$$

$$L_{pb}^{fake}(D) = - E_{z \sim p_{z,c_{fake}} \sim P_{yy}} [\log(1 - D(G(z, c_{fake})))]$$

$$L(D) = L_{pb}^*(D) + L_{fl}^*(D) + L_{pb}^{fake}(D) + L_{fl}^{fake}(D) \tag{5}$$

D is trained for classification of different samples under guidance of $L_{pb}^*(D)$, $L_{fl}^*(D)$, $L_{pb}^{fake}(D)$ and $L_{fl}^{fake}(D)$. The batch training technique was used in the training procedure and loss function was optimised using Adam optimizer. G and D should keep a balance of confrontation in the training proportion setting. In the training ratio setting, they should be kept in a combative balance.

Because distance between CUs and PUs can affect link's availability, CRP considers link quality, link's available duration and load for each node. The length of time a CU is accessible for relay before it enters the PU's interference area is indicated by the link availability duration. The expected transport of a packet from present node to destination is referred to as link quality (ETX). Lower ETX, closer node is to target. Given mobile equipment's limited energy and capability, limiting the load is critical to avoid quickly depleting the battery's power and drastically reducing performance.

For SRP, each node in network has a priority in terms of connection availability, ETX, and the node's load. A successful packet relay for a forwarder means that the packet is received by a higher priority neighbour than sender. Forwarder checks packet's successful relay by listening in on the neighbor's reception. The forwarder finds packet by PSN. Scenario for cognitive routing protocol in the network is shown in below figure-4.

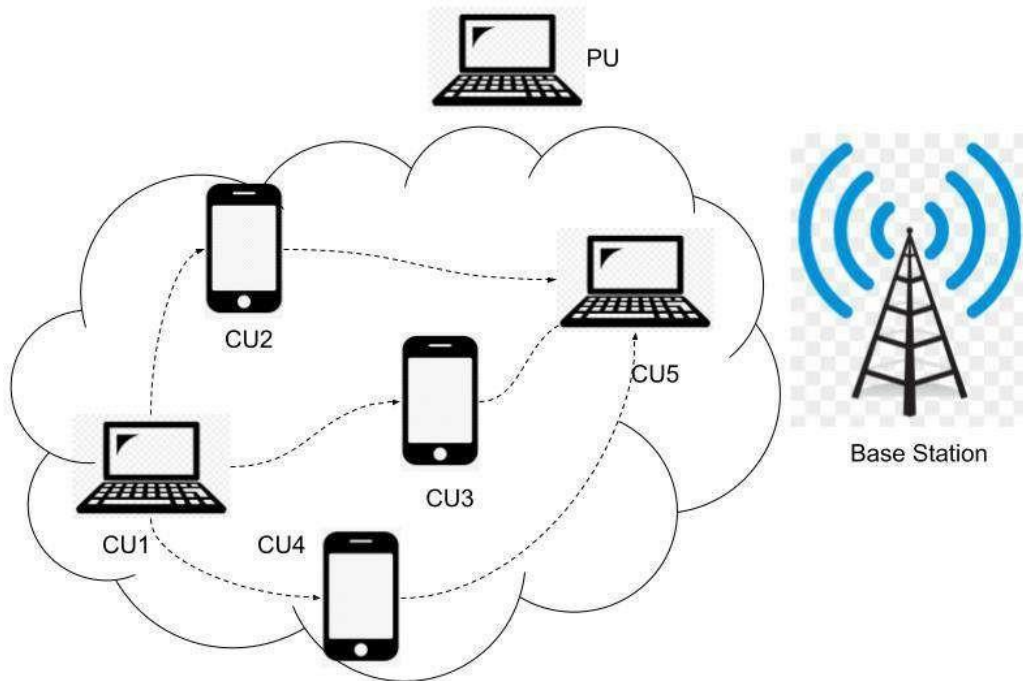


Figure-4 scenario for cognitive routing protocol

In this work, the essential approach for predicting link availability is to calculate link availability time T for two nodes. Further by considering changes in velocities, another critical parameter L (T) may be obtained by evaluating chance that link will be available until end of T. Initial distance between node A and node B at time t_0 is d_0 and detect distance between two nodes at time $t_1 = t_0 + T_1$ and $t_2 = t_0 + T_2$. For forecasting distance at time t_i , suppose that at t_1 distance is d_1 , distance is d_2 at time t_2 and distance is d_i at time t_i . For $T_j = t_i - t_0$, then by eqn (6)

$$d_i^2 = (m + v_a T_i)^2 + (n + v_b T_i)^2 - 2 \cos \phi (m + v_a T_i)(n + v_b T_i) \tag{6}$$

Since v_a, v_b, m, n and ϕ are constant against T_1 during T, we can get that $\partial^2 d_i^2 / \partial^2 T_i = 0$. So, d_i can be simply expressed as

$$d_i^2 = AT_i^2 + BT_i + C \tag{7}$$

Where A, B and C are constant. By setting $t_0 = 0$ A, B and C are given by eqn (8)

$$\begin{cases} A = \frac{(d_1^2 t_2 - d_2^2 t_1) - d_0^2 (t_2 - t_1)}{t_1 t_2 (t_1 - t_2)} \\ B = \frac{(d_1^2 t_2^2 - d_2^2 t_1^2) - d_0^2 (t_2^2 - t_1^2)}{t_1 t_2 (t_1 - t_2)} \\ C = d_0^2 \end{cases} \tag{8}$$

Therefore d_i at any $t_i < T$ can be calculated combining equation (2) and (3). Define D as greatest communication distance between two nodes. If no change in velocities occurs, calculate link available duration using equation (2) (accounted from 2t) as

$$T_d = \frac{\sqrt{B^2 + 4AD^2 - 4AC} - B}{2A} - t_2 \tag{9}$$

Authenticated cryptographic intrusion detection system (ACIDS):

For each node, we assumed that a public key, private key, and secret key were generated and disseminated in advance. The suggested model's ACIDS consists of two processes:

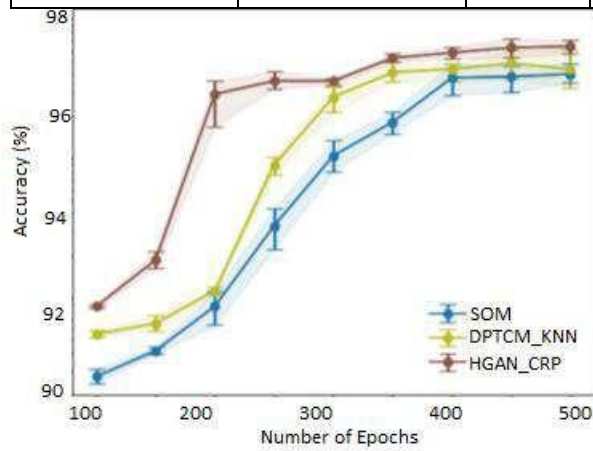
- At the Destination Node, the covert acknowledgment packet is formed.
- Secrete Acknowledgement packet Verification Process at Source Node

The destination node must transmit an acknowledgement packet after receiving the data packet from the source node. The generation of Secrete Acknowledgement at the Destination Node is the first step in our suggested hybrid cryptography technique. The concealed Acknowledgement packet is created

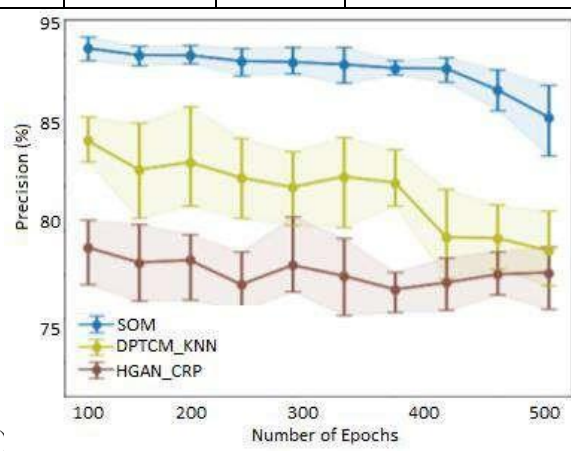
by encrypting the acknowledgment packet using the secret key and then digitally signing it with the private key.

Table-1 Comparison of parametric analysis for various dataset

Datasets	Techniques	Accuracy	Precision	End-End Delay	Scalability	Network Throughput
NSL_KDD	OM	6.5	0	3	4.5	3
	TCM-KNN	7	5	2	5.6	6.4
	AN-CRP	7.8	5	1	7.5	7
UNSW-NB15	OM	3	4	4	5	5
	TCM-KNN	4	7	3	6	6.5
	AN-CRP	6	5	2	7	7.8

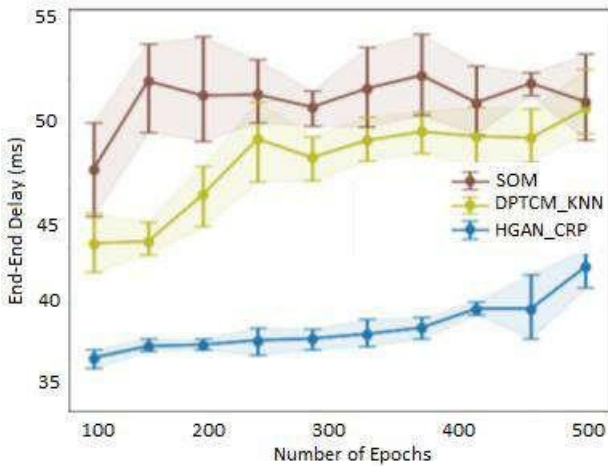


(a)



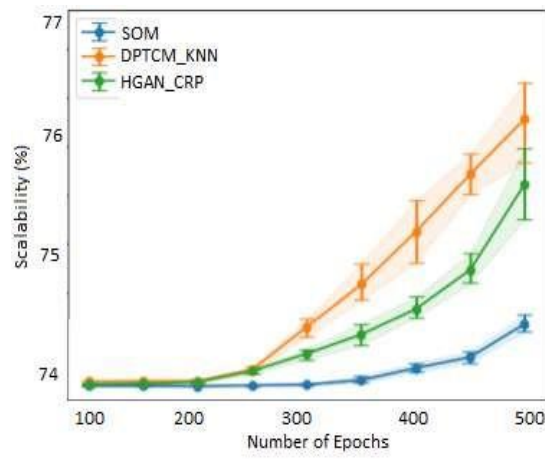
(b)

Accuracy



(c)

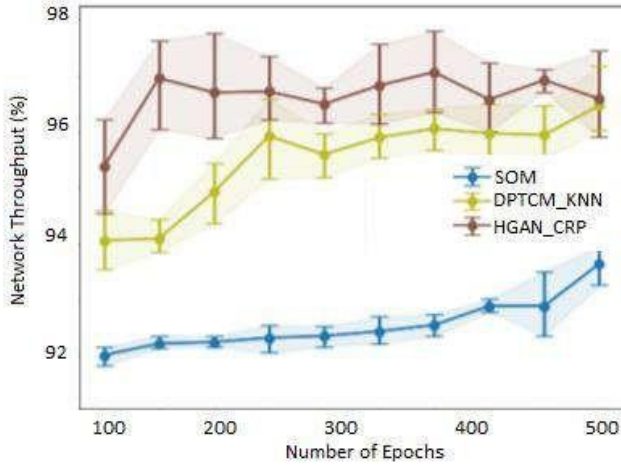
recision



(d)

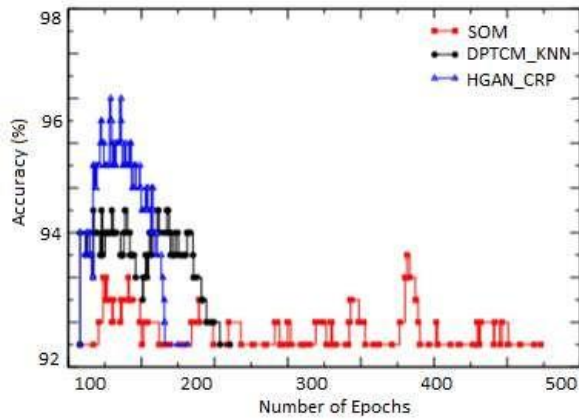
End-end Delay

calability

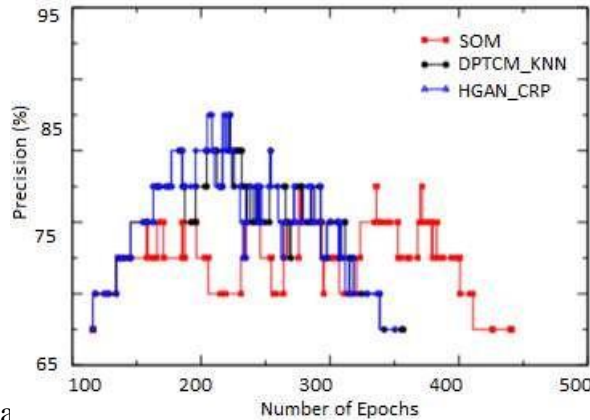


e) Network Throughput

Figure-5 Parametric comparison for NSL_KDD dataset in terms of (a) accuracy, (b) precision, (c) end-to-end delay, (d) scalability, (e) network throughput.

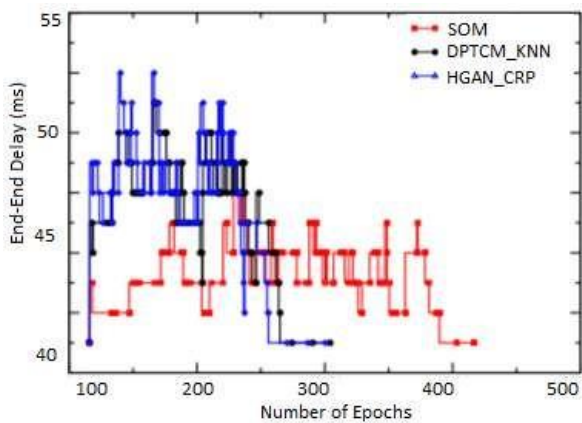


(a)



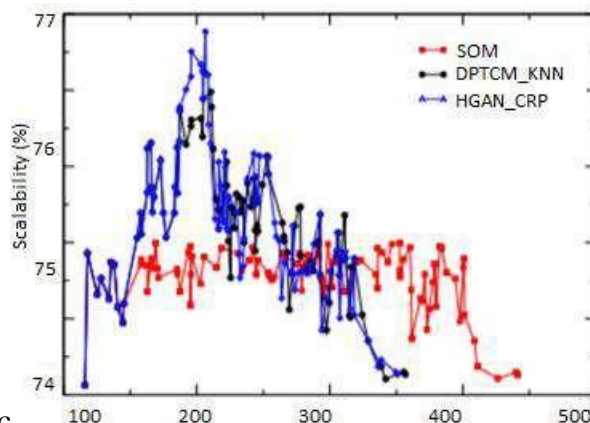
(b)

Accuracy



End-end Delay

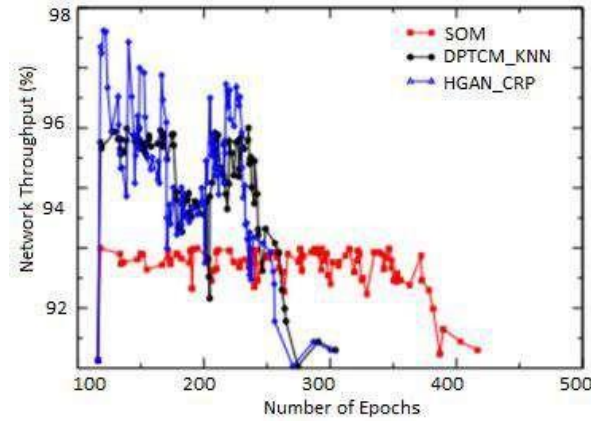
recision



(c)

calability

(d)



e) Network Throughput

Figure-6 Parametric comparison for UNSW-NB15 dataset in terms of (a) accuracy, (b) precision, (c) end-to- end delay, (d) scalability, (e) network throughput

The above figure-5 and 6 shows comparative analysis for parameters in terms of (a) accuracy, (b) precision, (c) end-to- end delay, (d) scalability, (e) network throughput for NSL-KDD and UNSW-NB15 dataset. From above comparative analysis, proposed technique in enhancing the network scalability and security based on machine learning techniques and authentication with security systems has improved the network scalability by increasing throughput and end-end delay of the network. By improving the network throughput, QoS has been enhanced with minimal end to end delay. For both dataset analysed in simulation part proposed technique in improving scalability and security. The increased throughput and minimum end to end delay shows efficient detection of intrusion has been carried out by proposed routing protocol.

CONCLUSION

This paper proposed novel technique in in enhancing the network scalability and security based on ML methods and authentication. Aim of this proposed design is enhance the scalability and security based on machine learning techniques integrated with routing and intrusion based system. Thescalability of the computer network has been enhanced using hybrid General Adversarial network with cognitive routing protocol in which the HGAN will classify the network traffic with higher data transmission by integrating cognitive based routing protocol. Intrusion of the network has been detected using authenticated cryptographic intrusion detection system in which the security of the system has been enhanced by this cryptographic based encryption and decryption process while data transmission. The experimental results shows network scalability and network security based parameters in terms of accuracy, precision, end-to- end delay, scalability, network throughput.

REFERENCES:

- [1] Aladaileh, M. A., Mohammed, A., Iznan, H., Yung-weychong, H., and Yousef, K. S. (2020). Detection Techniques of Distributed Denial of Service Attacks on Software-Defined Networking Controller—A Review. DOI: 10.1109/ACCESS.2020.3013998.
- [2] Kumar, S. D. and Mahbubur, M. R. (2019). Effects of Machine Learning Approach in Flow-Based Anomaly Detection on Software-Defined Networking. MDPI: *Symmetry*, 12, 7.1-21.
- [3] Amaral, P., Dinis, J., Pinto, P., Bernardo, L., Tavares, J., & Mamede, H. S. (2016, November). Machine learning in software defined networks: Data collection and traffic classification. In *2016 IEEE 24th International conference on network protocols (ICNP)* (pp. 1-5). IEEE.
- [4] Yungaicela-Naula, N. M., Perez-Diaz, J. A., Zareei, M., & Vargas-Rosales, C. (2021). Towards the security automation in Software Defined Networks. *Computer Communications*.
- [5] Ahmad, A. A. Solution Model for Intrusion Detection in Software Defined Networking (SDN) using Machine Learning.
- [6] Amin, R., Rojas, E., Aqdu, A., Ramzan, S., Casillas-Perez, D., & Arco, J. M. (2021). A survey on machine learning techniques for routing optimization in SDN. *IEEE Access*.
- [7] Gadallah, W. G., Omar, N. M., & Ibrahim, H. M. (2021). Machine Learning-based Distributed Denial of Service Attacks Detection Technique using New Features in Software-defined Networks. *International Journal of Computer Network and Information Security (IJCNIS)*, 13(3), 15-27.
- [8] Ma, W., Zhang, Y., Guo, J., & Li, K. (2021). Abnormal Traffic Detection Based on Generative Adversarial Network and Feature Optimization Selection. *International Journal of Computational Intelligence Systems*, 14(1), 1170-1188.
- [9] Asyaev, G. D. (2021, November). Prospects for Generative-Adversarial Networks in Network Traffic Classification Tasks. In *Journal of Physics: Conference Series* (Vol. 2096, No. 1, p. 012174). IOP Publishing.
- [10] Zhang, Y., Guan, J., Xu, C., & Zhang, H. (2012, December). The stable routing protocol for the cognitive network. In *2012 IEEE Globecom Workshops* (pp. 1090-1095). IEEE.
- [11] Babatunde, A. O., Adewole, K. S., Abdulraheem, M., & Oniyide, S. A. (2014). A Network-based Key Exchange Cryptosystem Using Elgamal Algorithm.
- [12] AL-Twajre, B. A., & Jeberson, W. (2014). Hybrid Cryptography Enhanced Adaptive Acknowledgment (HCEAACK) Intrusion Detection System.
- [13] Audah, M. F., Chin, T. S., Zulfadzli, Y., Lee, C. K., & Rizaluddin, K. (2019, August). Towards Efficient and Scalable Machine Learning-Based QoS Traffic Classification in Software-Defined Network. In *International Conference on Mobile Web and Intelligent Information Systems* (pp. 217-229). Springer, Cham.
- [14] Chiche, A., & Meshesha, M. (2021). Towards a scalable and adaptive learning approach for network intrusion detection. *Journal of Computer Networks and Communications*, 2021.
- [15] Sarumi, O. A., Adetunmbi, A. O., & Adetoye, F. A. (2020). Discovering computer networks intrusion using data analytics and machine intelligence. *Scientific African*, 9, e00500.
- [16] Mohammed, A. R., Mohammed, S. A., & Shirmohammadi, S. (2019, July). Machine learning and deep learning based traffic classification and prediction in software defined networking. In *2019 IEEE International Symposium on Measurements & Networking (M&N)* (pp. 1-6). IEEE.

- [17] Wang, P., Lin, S. C., & Luo, M. (2016, June). A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs. In *2016 IEEE International conference on services computing (SCC)* (pp. 760-765). IEEE.
- [18] O.A. Sarumi, C.K. Leung, Scalable data science and machine learning algorithm for gene prediction, In: *The 7th International Conference on Big Data Applications* (2019) 118–126.
- [19] Anithaashri, T. P., &Azaruddin, S. (2020). Enhancing Intrusion Detection System In Computer Networks Using AI Techniques. *European Journal of Molecular & Clinical Medicine*, 7(8), 2055-2058.
- [20] Dawoud, A., Shahristani, S., &Raun, C. (2018). Deep learning and software-defined networks: Towards secure IoT architecture. *Internet of Things*, 3, 82-89.

APPLICATION USAGE OF DRONES IN AGRICULTURE

R.PRAGADEESHWARI^[1]

^[1]Assistant Professor, Department of Computer Applications, Bon Secours College for Women, Thanjavur.

Abstract

Drone technology has gotten most of the recognition in the industry because of its diversity and considered the future for the agrarian community. The population is increasing tremendously and with this increase the demand of food. The traditional methods which were used by the farmers were not sufficient enough to fulfil these requirements. Thus, new automated methods (Drone technology) were introduced. Drones don't merely enhance overall performance but also encourage farmers to solve other assorted barriers and receive plenty of benefits through precision agriculture^[1]. Drones technologies saves the excess use of water, pesticides, and herbicides, maintains the fertility of the soil, also helps in the efficient use of man power and elevate the productivity and improve the quality. The objective of this paper is to review the usage of Drones in agriculture applications. This paper summarizes the current state of drone technology for agricultural uses, including crop health monitoring and farm operations like weed management, evaporation, spraying etc. The research article concludes by recommending that more farmers invest in drone technology to better their agricultural outputs.

Keywords: Drone, crop health monitoring, evaporation, spraying

1. INTRODUCTION

Drones are easier to use and can be used by the farmers in order to get the accurate and real-time data. By mapping, localization and the high resolution images captured by the drone will help for the efficient crop management. Satellite images are used for applications such as to identify the shrub lands and the grasslands for monitoring with 79% and 66% accuracy respectively.

However to satisfy all these needs drones must be used. By using the drones we can measure the distance and adjust from the terrain, measure the water level of the crops and many other applications. Thus a drone when equipped to adequate tools will make the agriculture as a precision agriculture. As a result it will benefit the farmers in order to look after the crops.

However, their hardware implementation purely depended on critical aspects like weight, range of flight, payload, configuration and their costs.

Drones have long been thought of as expensive toys. Drones can fly autonomously with dedicated software which allows making a flight plan and deploying the system with GPS and feed in various parameters such as speed, altitude, ROI (Region of Interest), geo-fence and fail-safe modes. The Japanese manufacturer Yamaha created the first agricultural drone in 2000. It was known as the R-50 and was made for field analysis and crop mapping. The GPS and camera on the R-50 enabled it to gather information on the size, health, and development of crops.



2. DRONES AS CROP HEALTH MONITORS

Synergy of technology and farming practices has given birth to the concept of precision agriculture^[3,4]. At the forefront of this revolution are drones equipped with a diverse array of sensors, allowing farmers to gain unprecedented insights into the health and condition of their crops.

Capturing a Bird's-Eye View

Drones, also known as Unmanned Aerial Vehicles (UAVs), have become the eyes in the sky for farmers.

Unveiling Crop Secrets

One of the most remarkable aspects of drone technology is its ability to harness spectral analysis. By capturing images beyond what the human eye can perceive, drones reveal a hidden world of plant health indicators.

Data-Driven Decisions

The integration of drone-derived data into agricultural practices has empowered farmers to make highly informed decisions.

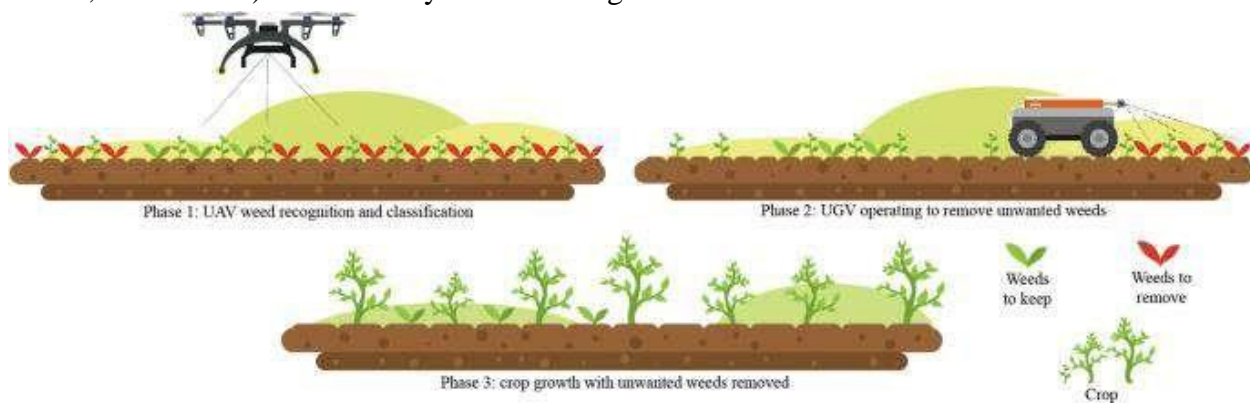
3. DISEASE MONITORING

One of the challenges farmers face is effectively monitoring expansive fields for pest infestations. Drones rise to this challenge by covering large areas quickly, giving farmers a comprehensive overview of their crops' health^[5]. Traditional farming related to disease detection relied on naked-eye observation, which is time-consuming and expensive and requires a lot of expertise. Currently, there are many methods to detect diseases in agricultural crops. Two main categories can be distinguished: direct and indirect detection methods. Direct detection methods consist of polymerase chain reaction, fluorescence in-situ hybridization, enzyme-linked immunosorbent assay, immunofluorescence and flow cytometry. Indirect detection methods consist of thermography, fluorescence imaging, hyperspectral techniques and gas chromatography. The focus of this study lies within the indirect detection methods, especially thermography and hyperspectral techniques that are supported by drones. RGB^[5, 6] and multispectral images have been preferred methods for acquiring information about the studied areas, but hyper spectral and thermal images^[7] have also been tested.

4. WEED CONTROL

Weeds are not desirable plants, which grow in agricultural crops and can cause several problems. They are competing for available resources such as water or even space, causing losses to crop yields and in their growth. Yield losses due to weed in India: Rice (10-100%), Wheat (10-60%), Maize (30-40%), Sugarcane (25-50%), Vegetables (30-40%), Jute (30-70%), Potato (20-30%) etc. ^[8]. The use of herbicides is the dominant choice for weed control. In conventional farming, Farmers uprooted weeds after post emergence and the most common practice of weed management is to spray the same amounts of herbicides over the entire field, even within the weed-free areas.

However, the overuse of herbicides can result in the evolution of herbicide-resistant weeds and it can affect the growth and yield of the crops. Agro-drone application for weedicide spray useful for pre-emergence & post emergence weed control. Spraying is possible in any field condition (muddy, weeds, insects etc.) also in sunny and drizzling condition.



5. Evaporation

Evaporation is an important process by which water is transferred from the land to the atmosphere by evaporation from the soil and by transpiration from living plants. Estimates of potential evaporation are used by professionals in the fields of hydrology, agriculture, and water management. Estimating evaporation has been one of the most important researches in agriculture recently because of water scarcity, growing population, and climate change. Many kinds of unmanned aerial vehicles are used on different research purposes for evaporation estimation. Typically, there are three different UAV platforms, aircraft, fixed-wings, and quad copter.

Aircraft is usually expensive, but it can fly longer and carry heavy sensors^[4]. Compared with aircraft, fixed-wings and quadcopter are less expensive. Fixed-wings can usually fly about 2 hours, which is suitable for a large scale of field. Quadcopter can fly about 30 minutes, which is used for short flight mission in a small scale of field. Being used as a remote sensing platform, UAVs also arouse new research problems, such as drone image processing, and flight path planning. A fixed-wing UAV to collect thermal data to estimate ET with two source energy balance models^[9]. Compared with other satellites based remote sensing methods, UAV platform and light weight sensors can provide better quality, higher spatial and temporal resolution images.^[10]

6. Spraying

Drones offer a multitude of benefits in various applications, including farm spraying, mosquito spraying, algae control in ponds, and locust control. They are approximately 40 times faster than manual labor, significantly improving ^[11]operational efficiency. Drones can reach inaccessible areas with ease,

enabling comprehensive coverage of vast or difficult-to-reach locations. They provide uniform spraying, ensuring consistent application and better control over the target area. Additionally, drone-based spraying eliminates harmful side-effects on manual labor, as operators are not exposed to potentially hazardous chemicals^[12]. With their ability to operate efficiently, cover larger areas in less time, and minimize human risk, drones have become a valuable tool in agricultural.



7. CONCLUSION

Drones have great potential to transform Indian agriculture. With the advancement of technology in the future, the production of drones is expected to become economical. Agricultural drone have the potential to improve the crops and helps in providing an insight about the disease management technique through imaging and sensors. It will also provide help in the monitoring of irrigation and water supply by predicting the availability of water. Agricultural drone can help the farmers to transform the agriculture industry.

8. REFERENCES

- [1] Aditya S Natu, Kulkarni SC. Adoption and Utilization of Drones for Advanced Precision Farming: A Review. published in International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169. 2016; 4(5):563-565
- [2] N. Robert Colwell Determining the prevalence of certain cereal crop diseases by means of aerial photography Hilgardia, 26 (5) (1956), pp. 223-286, [10.3733/hilg.v26n05p223](https://doi.org/10.3733/hilg.v26n05p223)
- [3] Cisternas, I. Velásquez, A. Caro, A. Rodríguez Systematic literature review of implementations of precision agriculture Comput Electron Agric, 176 (2020), [10.1016/j.compag.2020.105626](https://doi.org/10.1016/j.compag.2020.105626)
- [4] Van Blyenburgh P. UAVs: an overview. Air & Space Europe. 1999; 1(5-6):43-47
- [5] Altas Z, Ozguven MM, Yanar Y. Determination of Sugar Beet Leaf Spot Disease Level (Cercospora Beticola Sacc.) with Image Processig Technique By Using Drone. Curr. Investig. Agric. Curr. Res. 2018; 5:621-631.
- [6] Dang LM, Hassan SI, Suhyeon I, Sangaiah AK, Mehmood I, Rho S et al. UAV Based Wilt Detection System Via Convolutional Neural Networks. Sustain. Comput. Inform. Syst., 2018.
- [7] Calderón R, Navas-Cortés J, Lucena C, Zarco-Tejada P. High-resolution Hyperspectral and Thermal Imagery Acquired From UAV Platforms for Early Detection of Verticillium Wilt Using Fluorescence, Temperature and Narrow-band Indices. In Proceedings of the Workshop on UAV-basaed Remote Sensing Methods for Monitoring Vegetation, Cologne, Germany, 2013, 7-14p.
- [8] Gharde Yogita, Singh PK. -Yield and Economic losses due to weeds in India, ICAR-DWR, Jabalpur.
- [9] Hoffmann H, Nieto H, Jensen R, Guzinski R, Zarco-Tejada P, Friberg T. Estimating

- evapotranspiration with International Journal of Chemical Studies <http://www.chemijournal.com>
thermal UAV data and two source energy balance models. Hydrology & Earth System Sciences
Discussions, 2015
- [10] Niu, Haoyu et al. Estimating evapotranspiration with UAVs in agriculture: A review 2019
ASABE Annual International Meeting. American Society of Agricultural and Biological
Engineers, 2019.
- [11] Sarghini F, De Vivo A. Analysis of preliminary design requirements of a heavy lift multi rotor
drone for agricultural use Chemical Engineering Transactions. 2017; 58:625-630.
- [12] Kedari S, Lohagaonkar P, Nimbokar M, Palve G, Yevale P. Quadcopter-A Smarter Way of
Pesticide Spraying. Imperial Journal of Interdisciplinary Research, 2016, 2(6).

WIRELESS TECHNOLOGY TRENDS FOR 5G

S. ARISTA^[1]

^[1]Assistant Professor, Department of Computer Applications, Bon Secours College for Women,
Villar Bypass, Thanjavur.

Abstract:

5G—connect anytime, anywhere, anyhow|| promising everywhere network access at high speed to the end users, has been a topic of great interest mainly for the wireless telecom industry. The new upcoming technology of the fifth generation wireless mobile network is advertised as lightning speed internet, everywhere, for everything, for everyone in the nearest future. There are a lot of efforts and research carrying on many aspects, e.g. millimetre wave (mmW) radio transmission, massive multiple input and multiple output (Massive-MIMO) new antenna technology, the promising technique of SDN architecture, Internet of Thing (IoT) and many more. In this brief survey, we highlight some of the most recent developments towards the 5G mobile network. 5G seems to be the solution for the growing user necessities of wireless broadband access and the boundaries of the existing wireless communication system. The wireless industry is busy with the standardization of the 4th generation (4G) cellular networks

Keywords—5G; millimetre wave (mmW); Internet of Thing (IoT); SDN; massive multiple input and multiple output (Massive- MIMO)

INTRODUCTION

There is a protracted journey of wireless communication from past four decades from 1st generations (1G) to Fifth generation (5G). Fifth generation wireless communication technology, terribly high information measure that nobody will expertise before. It's expected from 5g that it's most powerful than alternative wireless technology with new advanced options. In Current days' mobile technologies square measure victimization Third and Fourth generation (3G and 4G) mobile networks. However, the future mobile network that is 5G during which there'll be a mobile multimedia system internet networks during which there's utterly wireless communication with no limitations, that makes the globe good, wireless. Over Fourth generation and Fifth generation ought to build profit to the globe and might add additional services.

DIFFERENT GENERATION OF MOBILE TECHNOLOGY

A. FIRST GENERATION (1G)

1G emerged in The Eighties. 1G contains Analog System that it had been typically referred to as the mobile phone. The mobile technologies MTS, IMTS, AMTS, PTT were employed by 1G. It uses analog radio radiation of frequencies 150MHz and used technique FDMA (Frequency Division Multiple Access) for voice decision modulation. It's poor and low capability voice links with no security.

Disadvantages: Less data rate, using analog cellular technology, FDMA multiplexing and using circuit switching.

B. SECOND GENERATION (2G)

2G was first introduced in the late 1980s. It has a speed of 64 kbps and uses the digital signal for voice transmission. It offers SMS (Short Message Service) facility and uses the bandwidth of 30KHz to 200KHz. It uses Packet Switching and Circuit Switching technique and offers data speed up to 144 kbps.

Disadvantages: Unable to handle complex data, required a strong signal, weak digital signal, short network coverage.

C. THIRD GENERATION (3G)

Wide Brand Wireless Network is being used by Third Generation (3G) because of which clarity of the signal is increased. Packet Switching is the technology through which the data are sent, through Circuit Switching Technique voice calls are interpreted. It works at a range of 2100MHz and has a bandwidth of 15MHz to 20MHz which is used for high-speed internet service, video calling, etc.

Disadvantages: Required higher bandwidth, expensive services.

D. FOURTH GENERATION (4G)

Fourth Generation (4G) offers a 100Mbps downloading speed and more. It provides the same feature as 3G with further [2] services or options like live T.V programs with High Definition (HD) quality and sends information a lot of quicker than the previous generation. Future Evolution (LTE) is taken into account as 4G technology. 4G has been being developed to avail Quality of Service (QoS) and like wireless broadband access, transmission electronic communication Service (MMS), Digital Video Broadcasting (DVB), lowest services like voice and information, and alternative services that utilize information measure.

Disadvantages: Implementation difficult, required complicated hardware, battery drainage fast.

E. FIFTH GENERATION (5G)

5G is a Fifth Generation broadband technology. 5G is a new network system that has very high data rate and reliable, and low- latency than the previous generation. 5G build on the foundation created by 4G, the technologies to be used in 5G, are still being defined. The 5G networks use encoding type called as OFDM. The air interface designed for much lower latency and greater flexibility. 5G networks can use low frequencies or high as -millimetre wave and that frequency can transmit large amounts of data, but few blocks at an instant of time. 5G networks are more likely to be networks of small cells like as size of a home router than to be large towers, it is to expand network capacity. The aim is to have high speed available and high capacity at low latency than 4G. The latency rate of 4G is around 50 milliseconds, but 5G reduces to about one millisecond. This is particularly important for driverless cars and industrial applications. The aim of 5G is to get high speed to 20Gbps, which is 40 times faster than 4G network. And its speed has been being tested up to 7.5Gbps and uninterrupted 1.2Gbps while traveling 100km/h [5]. 5G network is set to provide up to a million of connections per square kilometer.

COMPARISON OF ALL DIFFERENT GENERATION OF MOBILE TECHNOLOGY

Table. I

Generation/ Features	1G	2G	3G	4G	5G
Initiated	1980	1990	1998	2008	After 2019
Data Bandwidth	2 kbps	64 kbps	2 mbps	150 mbps-1gbps	More than 1gbps
Technology	Analog cellular	Digital cellular	CDME, EDGE	LTE	WWW
Service	Voice	Digital voice, SMS, MMS	Internet, audio and video streaming	Dynamic information access, Wearable devices, HD mobile TV	IoT, Device to Device
Switching Technique	Circuit	Circuit, packet	Packet	All packet	All packet
Core Network	PSTN	PSTN	Packet network	Internet	Internet
Hands off	Horizontal	Horizontal	Horizontal	Horizontal and Vertical	Horizontal and Vertical
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	CDMA
Drawbacks	Poor capacity, bad voice connection, not secure	Need strong digital signal to help cellular phones	Required to accommodate higher network capacity	Being deployed	Yet to be implemented

TECHNOLOGIES USED IN 5G

A. FLAT IP BASED NETWORK

The basic concept of the 5G mobile technology is designed by taking into account the users requirement. The user has been given the top most priority instead of the operator-centric concept as in 3G or the service-centric for 4G. The 5G has been made up of OSI model consisting of different layers from a physical layer to the application layer [1].

The network layer at 5G networks can be divided into several sublayers to provide all IP connectivity anywhere and anytime [1]. The use of the Internet Protocol (IP) in the network layer is inevitable, given the IP system is the best and most used system to support and expand the network layer nowadays [1].

All-IP Network system has started very well from the development of LTE. Flat IP architecture is a one of the key concept, used to make acceptable to all kinds of technology. Evolve Packet Core is also a part of Flat IP architecture.

- IP implementation does not require any additional proprietary middleware or gateways.

B. COGNITIVE RADIO

A. Cognitive radio is a very smart communication which takes into the accounts the surroundings and uses the methodology to understand by developing to learn from the environment and alter its systems to make correspondence in the real time to different operating parameters with two primary objectives in the mind that is highly reliable communication, efficient utilization of radio spectrum. By that terminal Cognitive terminal is a very smart terminal with the intelligence to choose the appropriate network from all the existing networks. The choice is making on some information such as resources, demand and time. The 5G technology proposes a universal terminal which includes all of the radio predecessors features into a single device.

5G NETWORK LAYERS

In 5G network layer, terminal contain vertical handoff as they have access to all different wireless networks. The vertical handoffs will be debarred by the 5G technology as an invite and developing lot of new technologies. The 5G technology will have the facilities error control schemes along with modulation techniques.

Fig.: OSI Layers and 5G Network Layers

Application Layer	Application (Services)
Presentation Layer	
Session layer	Open Transport Protocol (OTP)
Transport Layer	
Network layer	Upper network layer
	Lower network Layer
Data link Layer(MAC)	Open Wireless Architecture (OWA)
Physical Layer	

A. PHYSICAL/MAC LAYERS

The physical layer, OSI layer-1 has a major responsibility for coordination. It co-ordinates the functions required to carry a bit stream over the physical layer. The transmission and interface media have some electrical and mechanical specifications which are determined by the physical medium. The devices and interfaces on the physical layer have to perform some functions and procedure for the transmission to occur which are also governed by the physical layer. The physical layer takes into consideration that the physical characteristics of devices, interfaces, and medium, voltage fluctuations, bits representation, physical data rate, synchronization of bits, configuration of the line, transmission mode, and physical topology.

B. NETWORK LAYER

In the network layer, the limitations of number connection which were carried out by Internet Protocol version 4 (IPv4) had been solved by IPv6 but required bigger packet header. The Network layer, OSI layer-3 is responsible for the packet delivery from the source to the destination through multi- channel networks. The Network layer is applicable only when their source to destination delivery arises or require.

The 5G devices will maintain the virtual multi-wireless networks environment. For that purpose, network layer should be separated into two sub-layers in 5g devices i.e.: Lower network layer for an interface of each, and Upper network layer for the device terminal. This is due to the very first design of the Internet, where all the routing is based on the IP addresses which is different in each IP network worldwide. The middleware between the Lower and Upper network layers should maintain address translation from Lower network layer IP addresses (IPv4 or IPv6) to different Upper network

layer IP address (IPv6), and vice versa.

C. OPEN TRANSPORT PROTOCOL (OTP) LAYER

The transport layer, layer-4 in the OSI model is the layer through which various wireless networks vary from each other. Transport layer performs process-to-process delivery of the entire message. The transport layer converts the packets that it received from the layer-3 network layer into segments and assures that the segments are ready for transmission. It brings out the relationship between the two packets which are delivered and ensure that the entire messages arrives as it was overseeing. The Transport layer also responsible for error control, flow control, control connection, segmentation and reassembly and service-point addressing.

For the 5G device terminals, it is very relevant to have transport layer this is possible to download and install. Such devices have the possibility to download version which is targeted to a specific wireless technology installed at the base stations [11]. This is called Open Transport Protocol (OTP).

D. APPLICATION LAYER

The application layer, layer-7 in the OSI model is responsible for providing services to the user and user interface. The application layer enables the user to access the networks, whether human or software too. The application layer provides and supports services such as file transfer and remote file access, e-mail, shared database management access, and other types of information services. The application layer also consists of a virtual network terminal which allows the user to create use any Host ID. Nowadays, in the devices, the users manually select the wireless interfaces for specific Internet service without having the possibility to use Quality of Service history to select the best wireless connection.

The Quality of Service (QoS) parameters, like jitter, delay, bandwidth, reliability, will be stored in a database in the 5G devices with objective to be used by brilliant algorithms running in the device terminal as the system processes, at the end will provide the best wireless connection upon required Quality of Service.

5G NETWORK ARCHITECTURE

The Fifth generation network system is all IP based network model for wireless networks and mobile networks ability. The All-IP network is capable of fulfilling all the rising demand in the market cellular communications. The Fifth Generation technology is a common platform for all radio access technologies. The All-IP network uses packet switching technique and its continuous evolution provides improved performance and cost. The Fifth Generation Architecture consist of a number of an independent, autonomous radio access technologies (RAT) and the user terminal. In Fifth Generation network architecture all-IP based mobile applications and services such as mobile banking, mobile commerce and etc, are offered through Cloud Computing Resources (CCR). Cloud computing is a model for beneficial on-demand network access configurable computing resources such as storage, servers, applications, services and networks. Cloud computing allows the users to use applications without any installation and access their personal data at any mobiles or computers with internet access.

CCR links the Reconfigurable Multi-Technology Core (RMTC) with remote reconfiguration data from RRD attached to Reconfiguration Data models (RDM). The main challenge for an RMTC is to deal with increasing different radio access technologies. The core is a convergence of the

nanotechnology, cloud computing, and radio, and based on All IP Platform. Core changes its communication functions depending on the status of the network and/or user demands. RMTC is connected to different radio access technologies ranging from 2G/GERAN to 3G/UTRAN and 4G/EUTRAN in addition to 802.11x WLAN and 802.16x WMAN. Other standards are also enabled such as IS/95, EV-DO, CDMA2000...etc. Interoperability process criteria and mechanisms enable both terminal and RMTC to select from heterogeneous access systems. [2,4]

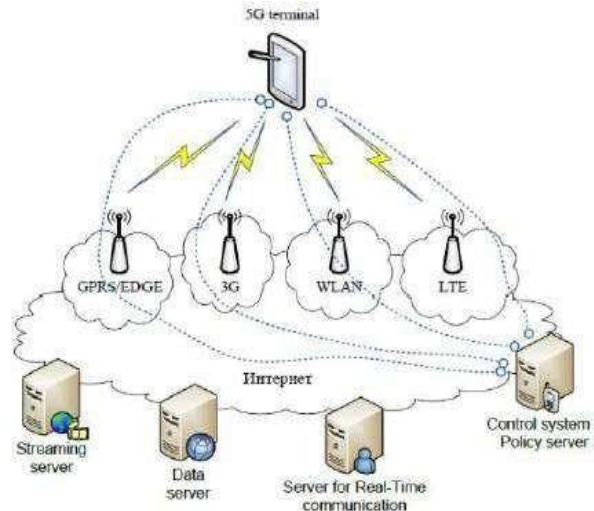


Fig.: 5G Network Architecture [4]

CONCLUSION

In this paper, we have surveyed 5G mobile technology. The 5G mobile technology is designed as an open platform on different layers. 5G technology is about to begin and going to give tough completion to mobiles and computers. In the world of mobile communication technology from 1G,2G,3G, and 4G to 5G. The 5G devices will have access to different wireless technologies at the very same time. 5G offers high speed and resolution for passionate mobile phone consumer.

FUTURE SCOPE

The future enhancement of 5G technology will be incredible as it combines with the Artificial Intelligent. The 5G will use the technology Internet of Things (IoT). The IoT will be used to connect all the devices to the internet through the sensors. The use of IoT technology will lead to building the smart cities and country.5G network technology will bring a new era in the mobile wireless communication technology. The 5G devices will have access to various wireless technologies at the same time and the terminal should be able to combine various flows from various technologies. 5G technology offers high resolution for the for the crazy phone user. We can monitor any place of the world and observe the space, and watch HD channels in tablets and mobile from anywhere. The 5G will use the technology Internet of Things (IoT). The IoT will be used to connect all the devices to the internet through the sensors. The use of IoT technology will lead to building the smart cities and country. The future enhancement of 5G technology will be incredible as it combines with the Artificial Intelligent.

REFERENCES

- [1] -A Review of 5G Technology|| by Suvarna Patil, Vipin Patil and Pallavi Bhat published in

- International journal of Engineering and Innovative Technology Issue 1, January 2012
- [2] –Architecture of 5G Technology in Mobile Communication|| by Sheetal J published in 18th IRF International Conference, 2019, Pune, India,
 - [3] –Key Concept and Network Architecture for 5G Mobile Technology|| by Sapana Singh and Pratap Singh published in International Journal of Scientific Research Engineering & Technology (IJSRET) Volume 1, Issue 5, pp 165-170, August 2018
 - [4] –5G Mobile: A Review of Innovation Technology|| by Santhana Mani.J, Karthik.S and Sasikala.K published in International Research Journal of Engineering and Technology (IRJET) Volume 2, Issue 6, September 2017
 - [5] 5G WIRELESS TECHNOLOGIES-Still 4G action not over, but time to start talking 5G|| by Ms. Neha Dumbre, Ms. Monali Patwa and Ms. Kajal Patwa published in International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 2,
a. Dec 2018
 - [6] –5G Technology of MobileCommunication: A Survey|| by Asvin Gohil, Hardik Modi and Shobhit K Patel published in 2019 International Conference on Intelligent Systems and Signal Processing
 - [7] –Study of Recent Development in 5G Wireless Technology|| by Dheeraj Gandla published in International Journal of Electronics and Communication Engineering & Technology Issue 5, September–October2020.

RESEARCH ON CLOUD STORAGE AND ITS TECHNOLOGIES

C.TAMILMANI^[1]

^[1]Assistant Professor in Computer Applications, Bon Secours College for Women, Thanjavur
Email:tamilsadhana@gmail.com

ABSTRACT:

Cloud computing is a revolutionary mechanism that changing way to enterprise hardware and software design and procurements. Because of cloud simplicity everyone is moving data and application software to cloud data centers there still exist significant issues that need to be considered before shifting into cloud. Security stands as major obstacle in cloud computing. This paper gives an overview of the security issues on data storage along with its possible solutions many companies are not ready to implement cloud computing technology due to lack of proper security control policy and weakness in protection which lead to many challenge in cloud computing. An independent mechanism is required to make sure that data is correctly hosted in to the cloud storage server. In this paper, we will discuss the different techniques that are used for secure data storage on cloud. This paper also provides a process to avoid Collusion attacks of server modification by unauthorized users. Keywords: Introduction, Cloud Computing

KEYWORDS: Introduction, Cloud Computing And Cloud Storage, Cloud Storage Security, issues, solution, conclusion.

1.INTRODUCTION

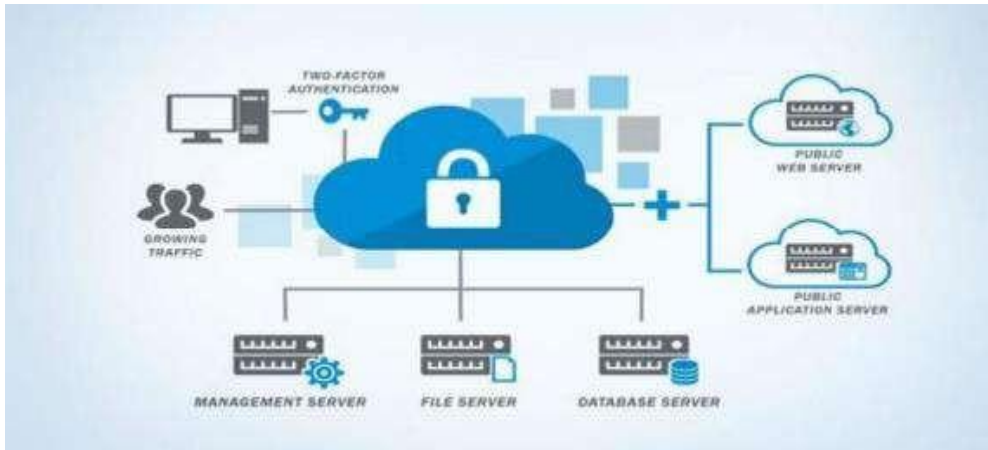
Cloud computing is the combination of many pre-existing technologies that have matured at different rates and in different contexts. The goal of cloud computing is to allow users to take benefit from all these technologies.

Many organizations are moving into cloud because it allows the users to store their data on clouds and can access at anytime from anywhere. From small to large enterprises poignant towards cloud computing to increase their business and tie-ups with other enterprises Security and privacy stands as major obstacle on cloud computing i.e. preserving confidentiality, integrity and availability of data.

Cloud computing has given a new dimension to the complete outsourcing arena(SaaS, PaaS and IaaS) and they provide ever cheaper powerful processor with these computing architecture.

The major thing that a computer does is to store in the available space and retrieve information whenever requested by the authenticated user As simple solution encrypt the data before uploading it onto the cloud.

This approach ensures that the data are not visible to external users and cloud administrators but has the limitation that plain text based searching algorithm are not applicable.



2.CLOUD COMPUTING AND CLOUD STORAGE

Cloud computing definition Cloud computing arises from the combination of the traditional computer technology and network technology, such as grid computing, distributed computing, parallel computing, utility computing, virtualization. One of the core concept of cloud computing is reducing the processing burden on user’s terminals through continuously enhancing the clouds’ handling capacity. Eventually user’s terminals are simplified into a simple input and output devices. Users can use the powerful computing and processing function on clouds and they can order their service from the cloud according to their own needs.

Users can use the powerful computing and processing function on clouds and they can order their service from the cloud according to their own needs.

Cloud storage is essentially one of the simplest applications of cloud technology available. Google Drive, Microsoft OneDrive, Dropbox; all of these services are cloud storage services. They do one thing and one thing only: they allow you to store data on the cloud.

Meanwhile, a company like Google, Microsoft, Amazon, Dropbox, or one of the many other cloud service providers has much, much more money and resources at their disposal. They can provide petabytes of storage for near-trivial amounts of money, with redundancy and backups in case of hardware failure. All you need to do is pay to access it.

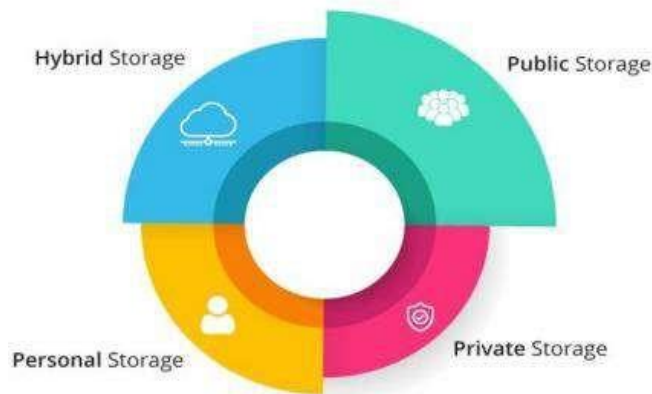


Fig2.cloud computing and cloud storage

Cloud storage definition and its architecture Cloud storage is a system that provides functions such as data storage and business access. It assembles a large number of different types of storage devices through the application software which are based on the functions of the cluster applications, grid techniques, distributed file systems, etc.

Cloud storage can be simply understood as the storage in cloud computing, and also can be considered to be a cloud computing system equipped with large capacity storage. Cloud storage system architecture mainly includes storage layer, basic management layer, application interface layer and access layer .

- **Personal Cloud Storage** It is also known as mobile cloud storage. In this type storage, individual's data is stored in the cloud, and he/she may access the data from anywhere
- **Public Cloud Storage** In Public cloud storage the enterprise and storage service provider are separate and there aren't any cloud resources stored in the enterprise's data centre. The cloud storage provider fully manages the enterprise's public cloud storage
- **Private Cloud Storage** In Private Cloud Storage the enterprise and cloud storage provider are integrated in the enterprise's data centre. In private cloud storage, the storage provider has infrastructure in the enterprise's data centre that is typically managed by the storage provider.
- **Hybrid Cloud Storage** It is a combination of public and private cloud storage where some critical data resides in the enterprise's private cloud while other data is stored and accessible from a public cloud storage provider



3. CLOUD STORAGE TECHNOLOGY OF ENTERPRISES

3.1 GFS : System Architecture

A GFS cluster consists of a single master, multiple chunkservers and multiple clients. Each of these is typically a commodity Linux.

- **GFS Master:** Master manages all file system metadata and the files directory structure. GFS uses single master policy which means in the same time only one master providing services so that it can avoid extra costs for coordinating between multiple masters synchronously. A client interacts with the master only for metadata, and interacts with the chunkservers directly for all other data.

- **Chunkserver:** GFS files are divided into fixed-size chunks stored on each chunkserver and the default block size is 64M. Each chunk is identified by an immutable and

globally unique 64 bit chunk handle assigned by the master as soon as the chunk is created. Each block is replicated on three chunkservers. Users can set different replication levels for each regions of the file namespace. As shown in Figure, there are four chunkservers and five chunks as C0-C4. Each chunk is saved on three chunkservers.

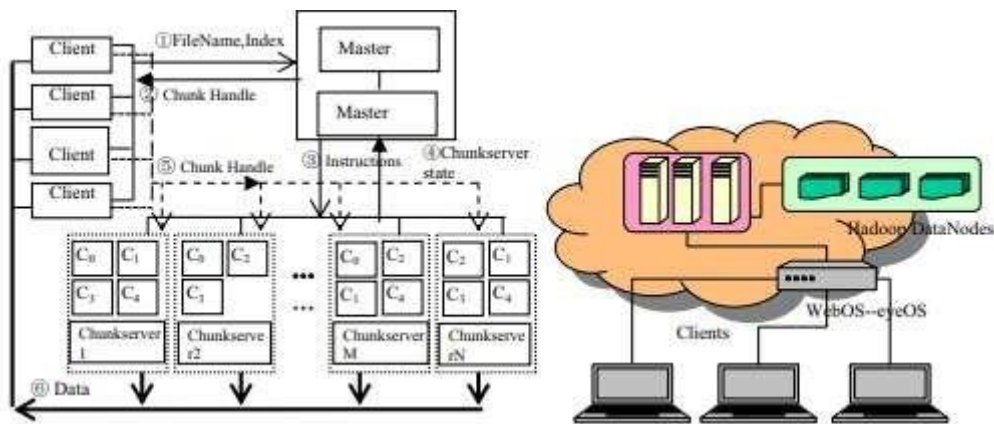
- **Client:** GFS client code linked into each application implements the file system API and communicates with the master and chunkservers to read or write the master for metadata operations, but all data-bearing communication goes directly to the chunkservers[1].

3.2 Workflow thin solid lines represent the control information between clients and master or between master and chunkservers, thick solid lines represent the data communication between chunkservers and client, dashed lines indicate the control information between clients and chunkservers. Kun Liu and Long-jiang Dong / Procedia Engineering 29 (2012) 133 – 137 Author name / Procedia Engineering 00 (2011) 000–000 3135 Firstly, clients compute chunk index from files structure and chunk size, then send file name and chunkindex to master .

Secondly, master sends chunk handle and chunk locations to clients. Thirdly, clients send chunk handle and byte range to the nearest chunkserver.

Finally chunkserver sends data to client. Once clients get chunk locations from master, clients do not interact with master any more. Master does not permanently save the mapping from chunkserver to chunk.

Instead, it asks each chunkserver about its chunks at master startup or whenever a chunkserver joins the cluster. The master periodically communicates with each chunkserver in HeartBeat message to give it instructions and collect its state.



3.3 HDFS

The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture

Hadoop is hosted by the Apache Software Foundation, which provides support for a community of open source software projects. Although Hadoop is best known for MapReduce and its distributed file system (HDFS), the other subprojects provide complementary services, or build on the core to add higher-level abstractions. The detailed contents refer to document.

The full name of HDFS is Hadoop Distributed File System. HDFS is run on large clusters of commodity hardware and is like GFS of Google. The architecture of HDFS is master/slave and a HDFS cluster has one namenode and multiple datanodes. Namenode is the central server, equivalent to master in GFS. It is responsible for the namespace operation of file systems.

Datanode is similar to chunkserver of GFS which is responsible for managing storage on datanodes, creating block, deleting block, copying block and etc. The files in HDFS are divided into one or multiple blocks which are stored in datanodes. Namenode and datanodes can be run on the low-cost Linux computer. HDFS is developed by java language.

3.4 Network storage technology

Network storage is a special private data storage server, which can provide cross-platform file sharing function. Network storage usually occupies its own node on a LAN without the need for application server intervention to allow users to access data on the network.

In this configuration, network storage centrally manages and processes all data on the network, unloads the load from the application or enterprise server, effectively reduces the total cost, and protects the user's investment.

P2P storage technology Peer-to-peer technology (P2P), also known as peer interconnected network technology, is a new network technology that relies on the computing power and bandwidth of the participants in the network, rather than relying on a small number of servers.

A pure point-to-point network has no concept of a client or server, it only has equal peer nodes, and ACTS as a client and server for other nodes on the network. P2P networks can be used for many purposes, such as various file sharing software, real-time mediabusiness, etc.

3.5 System implementation

Architecture, The storage system includes clients, web operating system eyeOS, cloud server (NameNode), cloud storage center (DataNode).

- **Clients:** Each client is only pre-installed with web browser and users log in this cloud storage through web browser. Clients are the interface between users and cloud storage system.

- **Web Operating System:** Web operating system receives users' access requests, verifies the users' validity, and interacts directly with the clients. It is based on eyeOS which offers a large number of applications to users. Users can download their required applications and achieve a personalized system. EysOS is also the file access interface for users and files can be saved in the cloud storage clusters by this interface.

- **Cloud server (Cloud NameNode):** Cloud storage cluster based on Hadoop includes cloud server (NameNode) and cloud storage center (DataNode). Cloud server is the namenode in Hadoop which manages file system namespace, computes the mapping from files to data nodes, allocates data nodes to save file blocks, and controls external clients' access.

- **Cloud Storage center (Cloud Data Node):** Cloud storage center is data node in Hadoop. It is in charge of saving files, realizing file distributed storage, ensuring loadbalancing, files fault-tolerant and etc.

Operation Process Users' operations based on eyeOS are writing files and reading files. When reading a file, we download the file to the local computer, then handle or display the file using the application software in web operating system. When the files are modified and saved, web operating system uploads them to cloud storage system from local computer.

Reading files process:

1. Users log in the web OS from client through clients' browser and double click a file icon on the web OS. Then eyeOS requests the file from the Hadoop name node.

2. Namenode finds the related information of files, and computes the file's location. Data nodes which saved the blocks of the file send the blocks to the clients.

3. Clients download the file blocks from the data nodes and merge these blocks into a file. 4. Applications associated with the file in the web operating system auto start and display the file.

Writing files process:

1. Users log in web OS from client's web browser modify and save files using the selected application. EyeOS requests uploading files to Hadoop name node.

2. Namenode allocates storage space to data nodes according to the file size and the data nodes' storage condition after it received the uploading request.

3. Clients upload file. Name node divides it into one or multiple blocks and saved in the allocated datanodes.

CONCLUSION:

Cloud computing is the inevitable product with the development of the internet, and it also brings more rich applications to the internet. Cloud data storage technology is the core area in cloud computing and solves the data storage mode of cloud environment. In this paper, we introduce the related concepts of cloud computing and cloud storage. Then we pose a cloud storage architecture based on eyeOS web operating system in our computers. Experiments verified the system is well.

REFERENCES:

- [1] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google file system[C]. Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2003:29-43.
- [2] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters[C]. Proceedings of the 6th Symposium on Operating System Design and Implementation. New York: ACM Press. 2004:137-150.
- [3] Fay Chang, Jeffrey Dean, et al. Bigtable: A Distributed Storage System for Structured Data[J]. ACM Transactions on Computer Systems. 2008,26(2):1-26.
- [4] Tom White. Hadoop: The Definitive Guide[M]. United States of America: O'Reilly Media, Inc. 2009.
- [5] Dhruba Borthakur. The Hadoop Distributed File System: Architecture and Design [EB/OL]. (2008-09-02)[2010-08-25]. http://hadoop.apache.org/common/docs/r0.16.0/hdfs_design.html.
- [6] Hbase Development Team. HBase: Bigtable-like structured storage for Hadoop HDFS[EB/OL]. (2010-08-10) [2010-08-25]. <http://wiki.apache.org/hadoop/Hbase>.

GREEN IOT FOR ECO-FRIENDLY SMART CITIES

SANTHI V^[1]

^[1]Assistant Professor, Department of Computer Applications, Bon Secours College for Women, Thanjavur
santhinivasbon@gmail.com

ABSTRACT

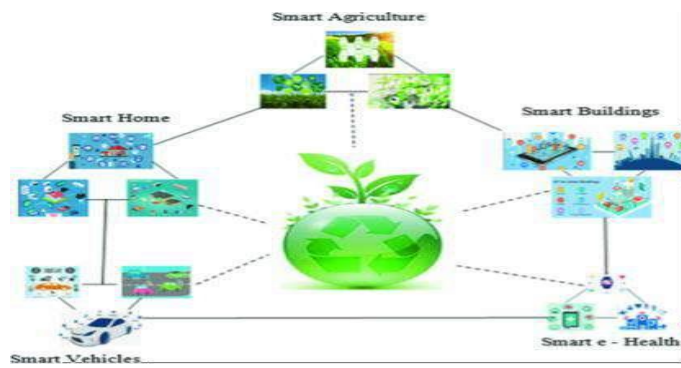
The way we live and work has changed as a result of the advancement of Internet of Things (IoT) technology and its incorporation into smart cities, enriching our civilization. Nevertheless, there are a number of drawbacks to IoT technologies, including higher energy usage, hazardous pollutants, and the creation of e-waste in smart cities. Applications for smart cities must be eco- friendly, which calls for a shift to green IoT. Green IoT creates a more sustainable ecosystem that is more suited for smart cities.. Adopting energy-efficient standards to lessen the environmental impact of IOT applications is known as Green IOT. IoT devices need to be energy-efficient in order to reduce the impact of CO2 emissions. Its primary objective is to maximize the technology's carbon footprint while utilizing IOT sustainably to minimize its detrimental effects on the environment. An interesting future where green networks deeply integrate our physical surroundings is hinted at by green IoT. It is a given that IoT green networks with sustainable designs would lower operational expenses, energy consumption, and pollution levels in the environment. IoT will surely improve and simplify our lives, and it will have a big impact on how we handle certain problems in our daily lives. Green IoT aims to make IoT devices more energy- efficient so that a more sustainable environment can be created.

Keywords -Green IoT, Sustainability, Carbon footprint, Energy efficiency, Smart Citites

INTRODUCTION:

Since 2020, as digital technology has been widely used, the use of IoT devices has grown exponentially. By 2030, 100 billion of these networked devices will have been produced. It has resulted in a concerning increase in energy usage. A significant investment is anticipated in the Internet of Things (IoT) space as a result of the government's goal to construct smart cities. It is anticipated that IoT will be crucial to the management of numerous finite resources[6].

Green IoT is a hardware and/or software energy-efficient procedure. This refers to linked devices working together in an energy-efficient manner to minimize CO2 emissions, power consumption, and the consequences of greenhouse gas pollution by utilizing network-based designs with minimal energy consumption and high bandwidth utilization, communication protocols, and green computing units. Energy-efficient and sustainable design is a fundamental component of the Internet of things.



Enabling technologies, leveraging technologies, and design technologies are the three concepts that make up green IoT[1]. Design technologies include device energy efficiency, communications protocols, network designs, and linkages. An IoT product has to go through a closed process that includes green design, green production, green usage, and greendisposal/recycling in order to be considered green.

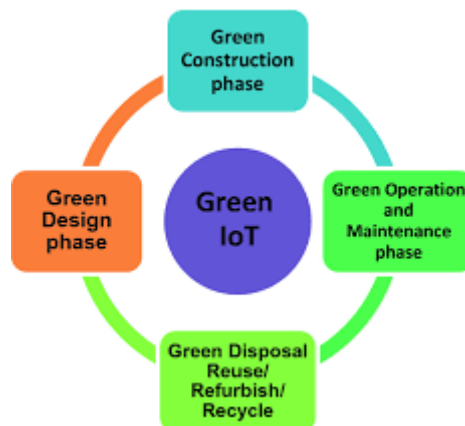
In smart city arrangement, IoT is going to look after the effective utilization of the resources. The application of IoT in real time requires a high level of connectivity between a large number of IoT devices. Every activity has an associated cost with it[10]. Experts predict that in the upcoming years, the cost of these devices in terms of their environmental impact will skyrocket. Green and renewable technologies are becoming more and more important in the technology sector due to the high emission and environmental effect scenario. This gives rise to the idea of the "Green Internet of Things" and its uses. [2]. IoT needs to be made cleaner and greener by increasing its energy efficiency in order to lessen its negative effects on the environment.



LITERATURE REVIEW

The goal of the present project is to create smart cities by utilizing IoT devices to enhance environmental health, safety, and sustainability. The present study aims to investigate every tactic for enhancing sustainable urban environments. The Internet of Things' (IoT) energy efficiency is the primary focus of "green IoT" [8]. Green IoT mitigates greenhouse gas emissions [9]. This document discusses many scientific solutions for GreenIoT that scholars have offered, together with all the specifics and summaries of these tactics. A work plan for the Green IoT's structural execution is presented in [5].

1) The Green IoT lifecycle



The Green IoT life cycle spans the entire IoT product lifecycle – from design to production to deployment, and to recycling – to have minimum negative impact on the environment by facilitating reduction in energy consumption and carbon emissions.

Potential of Green IoT

Green IoT encompasses a variety of technologies, including green cloud computing networks, green sensing networks, and green RFID tags.

2) *Green IoT techniques based on software*

Although they need to be made more energy-efficient for IoT to be practicable, data centers can play a crucial role in an energy-efficient Internet of Things network. A repository for data management, distribution, and storage is the Green Data Center (GDC). Things, systems, users, etc. produce this data. an architecture that uses a client-server model with an Orchestration Agent (OA) to evaluate servers' contexts for resource efficiency and data center management. The servers that were carefully chosen then deliver the processed data back to the client devices. However, this design necessitates the installation of OA on each client device, as well as the need for backup servers to maintain dependability, which may result in significant energy consumption. As a result, it must have a context-aware sensing platform that employs Selective Sensing to maximize energy efficiency. A solution is needed to decrease communication latency, which may be done by duplicating data closer to the users on cloud services.

3) *Green IoT techniques based on hardware*

In an IoT network, the design of integrated circuits (IC) is critical for energy conservation. Green IoT enhances IoT network architecture by integrating sensors and processing power on a single chip to minimize total infrastructure energy consumption, carbon footprint, traffic, and e-waste. Processor-based with an energy-efficient processor that can be split into 2 cores^[10]. One core is for low-computing tasks, the other for high-computing tasks. Use the scheduling framework to assign tasks to reduce energy consumption.

4) *Awareness based*

Campaigns to raise awareness are an important part of lowering energy use. Smart Metering Technology can be used to provide homeowners with real-time feedback on their energy consumption from various sources of their buildings, offices, and homes. Then, based on that real-time data, green IoT can advise customers on how to regulate and decrease their energy use.

5) *Policy-based*

Policies based on real-time data from IoT devices can contribute to large-scale energy savings. There are several phases to developing rules for energy efficiency, including automation, user feedback, data management, and monitoring.

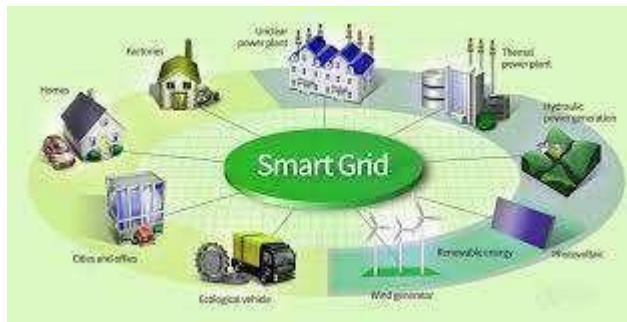
6) *Recycling based*

The use of recyclable materials in the manufacture of devices in an IoT network can contribute to the network's environmental friendliness. Mobile phones are produced from some precious natural resources, including non-biodegradable materials, plastic and copper, and can contribute to the greenhouse effect if not properly disposed of after use.

GREENER SOLUTIONS

1. SMART GRID

The smart grid is a communication network that collects data from many electrical system sensors to help control energy sources and consumer consumption. It refers to the grid’s capacity to dynamically adjust and re-adjust dynamically in order to supply electricity at the best possible quality and price^[3]. This helps to coordinate the amount of energy consumed to ensure the safety and security of the power network. The Smart grid uses many renewable energy sources, including thermal power plants, solar power, wind power plant, and nuclear power plant.



B. SMART CITIES

IoT may be defined as the efficient use of energy to enable a sustainable smart world. Equipping large cities with eco-friendly IoT devices is a viable way to transform them into sustainable living environments^[15]. Benefits like effective energy utilization, lower CO2



emissions, improved security, and practical solutions to aid people will result from this. The elements that form a smart city are as follows:

- **Waste management:** the Green IoT-based waste management system collects data from many sensors and then transfers them to a common data center, where the data will be analyzed with the aim of optimizing the collected waste.
- **Intelligent transportation system:** building smart roads with green IoT devices will give suggestions on the most suitable route when there are obstacles ahead, weather reasons or traffic accidents, traffic jams.
- **Safe city:** this is a combination of public announcement systems, fire control management, and digital video monitoring.
- **Smart building:** efforts focus on lowering energy usage by redefining building functions including lighting, heating, and air conditioning.
- **Smart street lights:** this involves controlling the transformers to adjust the lighting level appropriately at night or during the day with changing weather conditions.
- **Smart parking** uses a real-time vehicle tracking system to make suggestions for the nearest available parking in the city.

FUTURE OF GREEN IOT

In the future, with the combination of sensors, IoT devices along with the connection to 5G networks and the strong support of AI, this will provide users with green solutions. This means that IoT devices will also be context-aware and be able to perform certain functions, suggesting green forms of communication between people and things and between themselves. The improvement in power in the design phase will make Green IoT popular soon. In fact, IoT devices should sleep when not in use and use routing algorithms during data exchange.

Currently, the areas that Green IoT focuses on include Green RFID Technology, Green Wireless Sensor Network Technology, Green Cloud Computing Technology, Green Machine to Machine Technology, Green Data Center Technology, Green Communication, and Networking, Green Internet Technologies. In the future, the potential of Green IoT will grow if the following improvements are made:

CONCLUSION

Future Green IoT will bring about major improvements that will eventually result in a green environment. The 21st century has seen tremendous advancements in a variety of technology, which have enhanced living conditions in smart cities. IoT technology has shown increased promise recently for improving our quality of life in smart cities. However, the creation of new technologies requires a lot of energy and is associated with unintended emissions of pollutants and e-waste. This study looked at ways to make cities safer, smarter, greener, and more sustainable without sacrificing the quality of our lives. We emphasized the green IoT in particular for effective resource use, building a sustainable, cutting energy use, cutting pollution, and cutting e-waste. Every day, we witness sensors, devices, drones, and other things interacting with one another to carry out tasks in an intelligent manner for the environment that is sustainable and green. Green IoT has a direct impact on the IoT sector in

addition to assisting other businesses in protecting the environment and growing sustainably. Green IoT will be a sustainable design and technology in the future.

REFERENCES

- [1] Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. *ComputNetw* 54(15):2787–2805
- [2] Minerva R, Biru A, Rotondi D (2015) Towards a definition of the Internet of Things (IoT). *IEEE Internet Initiative* 1(1):1–86
- [3] Perera C, Zaslavsky A, Christen P, Georgakopoulos D (2014) Context aware computing for the internet of things: a survey. *IEEE Commun Surv Tutor* 16(1):414–454
- [4] Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (iot): a vision, architectural elements, and future directions. *Future Gen Comput Syst* 29(7):1645–1660
- [5] Tellez M, El-Tawab S, Heydari HM (2016) Improving the security of wireless sensor networks in an iot environmental monitoring system. In: *Systems and information engineering design symposium (SIEDS) IEEE*. IEEE, Conference Proceedings, pp 72–77
- [6] Shah J, Mishra B (2016) Iot enabled environmental monitoring system for smart cities. In: *Internet of things and applications (IOTA), International conference on*. IEEE, Conference Proceedings, pp 383–388
- [7] Chen X, Ma M, Liu A (2018) Dynamic power management and adaptive packet size selection for iot in e-healthcare. *Comput Electric Eng* 65:357–375
- [8] Kong L, Khan MK, Wu F, Chen G, Zeng P (2017) Millimeterwave wireless communications for iot-cloud supported autonomous vehicles: overview, design, and challenges. *IEEE Commun Mag* 55(1):62–68
- [9] POPA D, POPA DD, CODESCU M-M (2017) Reliability for a green internet of things. *Buletinul AGIR nr* 45–50
- [10] Prasad SS, Kumar C (2013) A green and reliable internet of things. *Commun Netw* 5(01):44
- [11] Pavithra D, Balakrishnan R (2015) Iot based monitoring and control system for home automation. In: *Communication technologies (GCCT) global conference on*. IEEE, Conference Proceedings, pp 169–173
- [12] Kodali RK, Jain V, Bose S, Boppana L (2016) Iot based smart security and home automation system. In: *Computing, communication and automation (ICCCA) international conference on*. IEEE, Conference Proceedings, pp 1286–1289
- [13] Gu M, Li X, Cao Y (2014) Optical storage arrays: A perspective for future big data storage. *Light Scie Appl* 3(5):e177
- [14] Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of big data on cloud computing: Review and open research issues. *Inf Syst* 47:98–115
- [15] Syed F, Gupta SK, Hamood Alsamhi S, Rashid M, Liu X (2020) A survey on recent optimal techniques for securing unmanned aerial vehicles applications. *Trans Emerg Telecommun Technol* e4133

REVIEW OF IMPROVING PRIVACY PROTECTION USING ANOMNIZED ROUTING PROTOCOL IN MOBILE AD-HOC NETWORKS

Dr. VINAYAKAN K^[1] SHEIK ABDHUL KHADIR^[2]

^[1]Assistant Professor, ^[2]Associate Professor & Head, PG & Research Department of Computer Science, Khadir Mohideen College, Adirampattinam. (Affiliated to Bharathidasan University)

Abstract: Mobile Ad Hoc Networks (MANETs) use anonymized routing protocols for security reasons. The protocol hides the original identities of alien nodes, so observers cannot compromise the network. Anonymous communication technology in MANET can be broadly classified into three types: reactive technology, proactive technology, and anonymous routing technology. Other backend routing techniques include hop-by-hop encryption and redundant traffic routing. Routing methods are either costly or fail to provide complete anonymity protection for sources, destinations, data, and routes. Mobile ad-hoc networks (MANETs) use various anonymous routing protocols to protect source, destination, and data anonymity. Therefore, an anonymous location-based efficient routing protocol (ALERP) is proposed to ensure a high level of anonymity protection. The idea behind ALERP is to dynamically partition the network field into zones and randomly select nodes within the zones as intermediate relay nodes to create untraceable anonymous routes.

Keywords: MANET, ad-Hoc, GPSR, ALERP.

I. INTRODUCTION

The rapid development of mobile ad-hoc networks (MANETs) has given rise to many wireless applications that can be used in many areas. It has a self-organizing and independent infrastructure and is used for communication, information exchange, etc. With its self-organizing and independent infrastructure, MANET is an ideal alternative for applications such as information exchange and communication. Due to the decentralized and open nature of the MANET, it is generally undesirable to limit the inclusion of nodes within the network. Ad-hoc mobile network nodes are vulnerable to malicious entities that falsify and analyze data and traffic by eavesdropping or attacking routing protocols. Anonymity should not be a fundamental requirement for civil claims. However, it becomes important in military applications, such as soldier communications. Consider a mobile ad-hoc network environment deployed in a battlefield military arena. Enemy nodes that can intercept transmitted packets can attack the commander node, and through traffic analysis, it can also involve relay nodes to block data transmission. Therefore, to provide secure communication, anonymous routing protocols play an important role in her MANET, hiding node identities and also preventing traffic analysis attacks from external observers.

MANET includes anonymity regarding data identity and location, source, destination, and itinerary. It is very difficult for sources and destinations to get the real IDs and exact locations of other nodes. Even on the road, there is anonymity, adversaries on or off the path cannot trace a packet flow to its source or destination, and nodes have no information about the actual identities and locations of intermediate nodes along the path, don't have to isolate the relationship between sender and receiver (i.e., an unobservable relationship [1]), but they can establish an anonymous path between two endpoints, allowing nodes along the way to discover the location of the endpoints. It is important to avoid devices that may be equipped. In MANET, the following existing anonymous routing protocols can be classified into two

categories. Redundant traffic [8] and hopper encryption. Public key encryption and high traffic lead to high costs. Many approaches have constraints by a focus on anonymity at the cost of valuable assets. In addition, many LAMT approaches cannot provide all of the above anonymity protections. With existing protocols, ALARM cannot protect the anonymity of source and destination locations [10]. SDDR protects source and destination location anonymity, but not route anonymity. ZAP [11] only protects the anonymity of the destination. Many anonymous routing algorithms [4] are based on geographic routing protocols (e.g., greedy perimeter stateless routing GPSR, which forwards packets to the node closest to the destination). Analyze the traffic by looking at the source and destination. The MANET's complex routing and tight channel resource constraints impose severe limits on system capacity. Furthermore, the recent growth of multimedia applications (such as video transmission) has increased the demands on routing efficiency.

However, existing anonymized routing protocols generate much higher costs and exacerbate the MANET resource limitation problem. Poor quality of service in the transmission of voice and video data can cause catastrophic delays in military operations. Anonymous Location Based and Efficient Routing Protocol (ALERP) was previously used for mobile ad-hoc networks to employ costly anonymous routing on the battlefield to protect source, destination, data, and route anonymity at low cost. proposed a new protocol, as the idea of ALERP is to dynamically divide network fields into groups. We call this a "zone" and randomly select nodes within the zone as intermediate relay nodes to create nameless, untraceable routes. Specifically, at each routing stage, the sender or data sender splits the network field to separate it and the destination into two different zones. It then randomly selects a node from another area as the next relay node and uses GPSR [4] to send data to the relay node. In the final step, the data is sent to k nodes present in the target region, and the nodes anonymously deliver k-nodes to the target. It also supports ALERP (Anonymous Location Based and Efficient Routing Protocol). Hide data initiators under a set of initiators to increase anonymity protection for source nodes. ALERP also provides protection against cross and point attacks [13]. The contributions to this work are:

1. Routing anonymously. ALERP provides identity, travel anonymity, source anonymity, and destination anonymity.
2. Low cost. Instead of relying on sequential over-the-top encryption and redundant traffic, ALERP uses random encryption, routing a message copy to ensure anonymity is protected.
3. Resilience to point strikes and intersection strikes. ALERP has a strategy for effective intersection avoidance attacks.

II. LITERATURE SURVEY

L. Zhao and H. Shen, "ALERT: An Anonymous Location- Based Efficient Routing Protocol in MANETs," IEEE. Mobile Informatics Transactions, Vol. 12 No. 6 June 2013[1]

Older anonymized routing protocols based on hop-by-hop encryption or redundant traffic. cause high costs. Also, certain protocols are incapable of fully preserving source, destination, and route anonymity. ALERP is characterized by source and low cost, route, and destination protection. It uses dynamic hierarchical range partitioning and random relay node selection to make it difficult for intruders to detect incoming nodes and ends. Alert packets contain the source and destination regions, but not their locations, to protect the anonymity of the source and destination. It also improves source and destination anonymity protection by hiding the

data originator/recipient from various data originator recipients. There is a "notify and run" mechanism to anonymize the source. Also, local broadcasting is used to ensure target anonymity. ALERP also provides an effective solution for mitigating attacks at intersections. Z.Zhi and Y.K. Choong, "Anonymizing Ad Hoc Geographic Routing to Preserve Location Data", Proc. 3rd International Workshop on Mobile Distributed Computing (ICDCSW), 2005. [2]

In this document, Zhi, Z. states: On the other hand, a related claim about the existence of unresolved burial grounds concerns site confidentiality that has not been adequately addressed. This article tries to keep the website confidential based on the idea of keeping the website away from the user's information and identity. We propose an algorithm for anonymous geographic routing that retains efficiency while protecting identity and location.

V. Patak, D. Yao and L. If node, "Securing Location-Aware Services over VANETs Using Geo-Secure Path Routing," Proc. IEEE International Conference. Vehicle Electronics and Safety (ICVES), 2008 [3]

V. Patak, D. Yao and L. If node provide secure location services on Ad Hoc Vehicle Networks (VANET) using the Secure Geographical Routing Protocol (GSPR). GSPR is an unrestricted geographic infrastructure routing protocol that resists interference from faulty or malicious nodes. Geographic locations of anonymized nodes are authenticated, ensuring location authentication and location confidentiality at the same time. This protocol also authenticates the routing paths used by individual messages. This document presents the structure of the Secure Geographic Routing Protocol (GSPR).

III. ROUTING PROTOCOLS IN MANETS

Various routing protocols have been suggested or used for MANET. A fundamental research challenge in MANET is routing, which has to deal with restrictions including high power consumption, high bandwidth consumption, high error rates, and unpredictable node displacement. Current Manet protocols can be broadly divided into the following categories:

A. Proactive (Table-Driven):

Proactive routing protocols are identical to current Internet routing protocols, including RIP (Routing Information Protocol), DV (distance-vector), OSPF (Open Shortest Detour Route First), and link-state. They try to keep the latest routing information for the entire network consistent. Everyone and every node has to maintain one or more tables to store its routing information and must respond to changes in network topology besides broadcasting the above-mentioned latest information. A few existing pro-active ad hoc routing protocols are: DSDV (Destination Sequenced Distance-Vector), WRP (Wireless Routing Protocol), CGSR (Cluster head Gateway Switch Routing), GSR (Global State Routing), FSR (Fisheye State Routing), HSR (Hierarchical State Routing), ZHLS (Zone-Based Hierarchical Link State), and STAR (Source Tree Adaptive Routing)

B. Reactive (Source-initiated, demand-driven):

These protocols attempt to eliminate traditional routing tables so that they no longer need to be updated to track changes in network topology. Whenever a source needs to travel to a destination, it must go through a path-by-path discovery process and perform some kind of path maintenance procedure to maintain it until the path is no longer needed or is inaccessible. Finally, demolition is done using the root deletion method. Some existing active routing protocols are [14]. DSR (Dynamic Source Routing), ABR (Associativity Based

Routing), TORA (Temporally Ordered Routing Algorithm), SSR (Signal Stability Routing), PAR (Power-Aware Routing), LAR (Location Aided Routing), CBR (Cluster-Based Routing), and AODV (Ad Hoc On-Demand Distance Vector Routing). In proactive routing protocols, routes are always available, consuming signalling traffic and power. In contrast, reactive protocols cause longer delays during the route discovery process. Both types of routing protocols have been enhanced to be more secure, scalable, and support a higher quality of service.

C. Hybrid Protocols:

Hybrid routing protocols [14, 15] aggregate large numbers of nodes into zones within a topology. The network is then divided into several numbers. A proactive approach is used within each zone to manage routing information. A reactive routing approach is used to forward packets between different zones. However, in the hybrid approach, routes to destinations located in the same zone are established without delay, while destinations located in different zones require route maintenance and route-finding procedures. The Zone Routing Protocol (ZRP) and Zone-based Hierarchical Link State (ZHLS) routing protocols solve the scalability problem. These routing protocols can also provide a better trade-off between communication overhead and latency, but zone dynamics and zone size have an impact on this correlation. Therefore, the hybrid routing protocol approach is an excellent candidate for routing large networks. At the network layer, routing protocols are used to find routes for packet transmission. The main benefits of routing protocols can be explored using qualitative and quantitative metrics that can be used to measure their suitability and performance. MANET has many desirable qualitative properties. These include loop-free, distributed operation, on-demand operation, security, sleep-time operation, proactive operation, and support for unidirectional connections. Some quantitative metrics that are useful for evaluating routing protocol performance are out-of-order delivery percentage, throughput, route acquisition time, end-to-end delay, and efficiency. Key parameters that need to be changed are network connectivity, link capacity, network size, topology change rate, traffic patterns, unidirectional link percentage, mobility percentage, and sleep node frequency.

IV. ANONYMOUS ROUTING PROTOCOLS DESCRIPTION

Anonymous routing protocols are very important in MANET to ensure secure communication by hiding node identities and preventing traffic analysis attacks from external observers. Anonymity in MANET includes data source identity and location anonymity (which also includes root anonymity). Precise location of source and destination Regarding route anonymity, route or not, an attacker cannot trace the flow of a packet to its source or destination, nor can either node along the route. It has no information about the true identity and location of intermediate nodes. To isolate the connection between the source and destination, it is important to form an anonymous path between the two endpoints so that nodes in the middle do not know where the endpoints are. Mainly on the MANET, where location devices are located.

V. EXISTING ANONYMOUS ROUTING PROTOCOLS ALARM

Anonymous Location-Aided Routing in Suspicious MANETs is one of the anonymous routing protocols in MANETs. ALARM detects MANET problems. It also provides secure, anonymous routing on your network. To do this, it uses a link-state routing protocol. LSRs get the current positions of nodes, send and create topology snapshots, and transfer data. For security, ALARM uses advanced cryptography and provides non-traceability, node authentication, and data integrity and privacy features. It also provides security against active and passive attacks. However, the problem with ALARM is that it cannot protect the anonymity of the location of the source and destination nodes.

ASR

Another anonymous routing protocol is the Anonymous Secure Routing (ASR) protocol. This protocol offers some additional properties of anonymity. B. Strong location privacy and identity anonymity. At the same time, it also ensures the security of discovered roots against various passive and active attacks. However, the ASR protocol has a problem with route anonymity.

AO2P

AO2P is one of the most important anonymous routing protocols. This is an ad-hoc, on-demand location-based private routing algorithm. This protocol is proposed primarily for anonymity-related communications. Node positions are used for route detection instead of node IDs.

VI. ALERP

The Anonymous Location-Based and Efficient Routing Protocol (ALERP) [1] provides high anonymity protection for data, sources, destinations, and routes at a low cost. It dynamically divides network arrays into zones and randomly selects nodes for each zone. B. An intermediate relay node that creates untraceable anonymous routes and destinations in two different zones. It then randomly selects a node in another zone as the next relay node and uses the GPSR algorithm to send the data to the relay node. In the final stage, the data is sent to k nodes in the target zone, giving target k anonymity. Additionally, ALERP [1] has the ability to hide data initiators among a large number of initiators to enhance source anonymity. ALERP is also immune to timing attacks. Consider a MANET deployed over a large field that uses geographic routing for node communication to reduce communication delays. A message's sender's location is revealed simply by revealing the direction of transmission. What is needed, therefore, is an anonymous communication protocol that can provide resilience and strictly guarantee sender anonymity when communicating with the other side of the field. A malicious observer also intercepts packets on the node's no. Knowing the direction of data transmission also allows for node tracking, or traceback to the sender. Therefore, the route should be undetectable and untraceable. A malicious observer could attempt to discover the destination node through traffic analysis by launching an intersection attack. So the target node also needs anonymity protection.

Capabilities:

Eavesdropping allows adversary nodes to analyze routing protocols and obtain information about nearby communication packets and the locations of other nodes in the network. You can also track data transfers on-the-fly as your node communicates with other nodes and

record historical communication for your nodes. Attacks can be launched against specific vulnerable nodes to control behavior, such as denial of service (DOS) attacks.

Ability: The Attacker does not perform powerful active attacks like Black Hole. They can only infiltrate all nodes. Computing resources are limited. So, both the symmetric key and the public/private key cannot be brutally decrypted in a reasonable amount of time. Encrypted data is therefore reasonably secure as long as the key is not known to the attacker.

VI. RELATED WORKS

	Category		Name	Identity Anonymity	Location Anonymity	Route Anonymity
Reactive	HOP-by-HOP Encryption	Topology	MASK[15]	Source	n/a	Yes
			ANODR[16]	Source, Destination	n/a	Yes
			Discount-ANODR[17]	Source, Destination	n/a	Yes
		Geographic	Zhou et al.[19] Pathak et al.[4] AO2P[20]	Source, Destination Source,	Source, Destination	No
	Redundant Traffic	Topology	ADA[8]	Destination	n/a	Yes
		Geographic	ASR[11] ZAP[13]	Source, Destination	Source, Destination	No
Proactive	Redundant Traffic	Topology	ALARAM [1]	Source, Destination	Source	No
Middleware	Redundant Traffic	Geographic	MAPCP[1]	Source, Destination	n/a	Yes

TABLE 1

Summary of Existing Anonymous Routing Protocols

Anonymous routing schemes for MANETs have been studied in recent years. On-demand or reactive routing methods [8], [16], [17], [15], [3], [4], [11], [10], [13], and proactive routing methods [5] can be categorized based on the way topological data can be used. Also, there is anonymous middleware working between the network layer and the application layer [9]. Since topology routing does not need node location information, location anonymity protection is not necessary. Table 1 shows the classification of the methods along with their anonymity protection. The table gives a finer classification of different anonymity approaches, including hop-by-hop encryption [8], [16], [17], [15], [3], [4], [11], [10], and redundant traffic routing [8], [11], [13], to clearly demonstrate the highlighted anonymity protection through various reactive routing methods.

In hop-by-hop encryption routing, a packet is encrypted in the transmission of two nodes en route, preventing adversaries from tampering with or analyzing the packet contents to interrupt the communication or identify the two communicating nodes. Hop-by-hop encryption routing is divided further into onion routing and hop-by-hop authentication. In onion routing, packets are encrypted at the source node and decrypted layer by layer (i.e., hop by hop) along the routing path. It is used in Aad [8], ANODR [16], and Discount-ANODR [17] topological routing. ad [8] integrates onion routing information, multicast messages and protocol coding concepts for ensuring destination and route anonymity. The onion used in ANODR [16] is called trapdoor boomerang onion (TBO), which uses a trapdoor function instead of public key-based encryption. ANODR needs onion construction in both route discovery and return routing, generating high costs. To deal with this problem, the authors further proposed Discount-ANODR, which constructs onions only on the return routes. Hop-by-hop authentication is used to prevent adversaries from participating in the routing to ensure route anonymity [15], [3], [4], [11], [10], [7], [18]. MASK [15] topological routing uses neighbourhood authentication in routing path discovery to ensure that the discovered routes consist of legitimate nodes and are anonymous to attackers. The works in [3], [4], [11], and [10] are based on geographic routing. In GSPR [3], nodes encrypt their location updates and send them to the location server. GSPR, on the flip side, fails to guarantee route anonymity because packets always take the shortest paths possible utilizing geographic routing, hence, the route can be identified by adversaries in a long communication session. In [4], a mechanism called a geographic hash is used for authentication between two hops en route, but the anonymity is compromised because the location of each node is known to nodes in the vicinity. In the AO2P [20] geographic routing algorithm, pseudonyms are used to protect nodes' real identities, and a node chooses the neighbourhood that can reduce the greatest distance from the destination. Since AO2P does not provide anonymity protection to destinations, the authors further improve it by avoiding the use of destinations in deciding the classification of nodes. The improved AO2P selects a position on the line connecting the source and destination that is further from the source node than the destination and replaces the real destination with this position for distance calculation. ASR [11] conducts authentication between the source and the destination before data transmission. The source and each forwarder embed their public keys in the messages and locally broadcast the messages. The destination responds to the source in the same way. In each step, the response is encrypted using the previous node's public key so that only the previous forwarder can decrypt the message and further forward it. However, such public key dissemination in routing makes it possible for attackers to trace source/destination nodes. Ariadne [7] uses TESLA [18] to conduct broadcasting-style authentication between two neighboring hops en route. Although it uses symmetric key cryptography for authentication, a high amount of traffic is inevitably incurred in broadcasting. SEAD [18] uses low-cost one-way hash functions rather than asymmetric cryptographic operations in conducting authentication for lower costs. However, all of these hop-by-hop encryption methods generate high costs due to the use of hop-by-hop public-key cryptography or complex symmetric key cryptography.

VII. CONCLUSION AND FUTURE WORK

Existing anonymous routing protocols are expensive because they rely on either hop-by-hop encryption or redundant traffic. Also, some protocols do not provide complete destination, source, and route anonymity protection. ALERP stands out by protecting source, low cost, destination, and route anonymity. ALERP uses dynamic hierarchical zone partitioning and random relay node selection to make it difficult for intruders to see her two endpoints and nodes in transit. Each ALERP packet contains a source and destination zone instead of a location to provide source and destination anonymity protection. ALERP also has the ability to protect source and destination anonymity by hiding data triggers/receivers under a set of data triggers and receivers.

REFERENCES

- [1] L. Zhao and H. Shen, -ALERT: An Anonymous Location-Based Efficient Routing Protocol in MANETs,|| IEEE Transactions on Mobile Computing, June 2013.
- [2] Pfitzmann, M. Hansen, T. Dresden, and U.Kiel, -Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Managementa Consolidated Proposal for Terminology, Version 0.31,|| technical report, 2005.
- [3] Sk.Md.M. Rahman, M. Mambo, A. Inomata, and E. Okamoto, -An Anonymous On-Demand Position-Based Routing in Mobile Ad Hoc Networks,|| Proc. Intl Symp. Applications on Internet (SAINT), 2006.
- [4] Z. Zhi and Y.K. Choong, -Anonymizing Geographic Ad Hoc Routing for Preserving Location Privacy,|| Proc. Third Intl Workshop Mobile Distributed Computing (ICDCSW), 2005.
- [5] Y.-C. Hu, A. Perrig, and D.B. Johnson, -Ariadne: A Secure On- Demand Routing Protocol for Ad Hoc Networks,|| Wireless Networks, vol. 11, pp. 21-38, 2005.
- [6] I Aad, C. Castelluccia, and J. Hubaux, -Packet Coding for Strong Anonymity in Ad Hoc Networks,|| Proc. Secure comm and Workshops, 2006.
- [7] K.E. Defrawy and G. Tsudik, -ALARM: Anonymous Location- Aided Routing in Suspicious MANETs,|| Proc. IEEE Intl Conf. Network Protocols (ICNP), 2007.
- [8] X. Wu, -AO2P: Ad Hoc On-Demand Position-Based Private Routing Protocol,|| IEEE Trans. Mobile Computing, vol. 4, no. 4, pp. 335-348, July/Aug. 2005.
- [9] T. Camp, J. Boleng, and V. Davies, -A Survey of Mobility Models for Ad Hoc Network Research,|| Wireless Communications and Mobile Computing, vol. 2, pp. 483-502, 2002.
- [10] K. El-Khatib, L. Korba, R. Song, and G. Yee, -Anonymous Secure Routing in Mobile Ad-Hoc Networks,|| Proc. Intl Conf. Parallel Processing Workshops (ICPPW), 2003.
- [11] S. Ratnasamy, B. Karp, S. Shenker, D. Estrin, R. Govindan, L. Yin, and F. Yu, -Data-Centric Storage in Sensornets with GHT, a Geographic Hash Table,|| Mobile Network Applications, vol. 8, no. 4, pp. 427-442, 2003.
- [12] X. Hong, M. Gerla, G. Pei, and C.C. Chiang, -A Group Mobility Model for Ad Hoc Wireless Networks,|| Proc. Second ACM Intl Workshop Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM), 1999.

- [13] J. Raymond, -Traffic Analysis: Protocols, Attacks, Design Issues, and Open Problems,|| Proc. Int'l Workshop Designing Privacy Enhancing Technologies: Design Issues in Anonymity and Unobservability (WDIAU), pp. 10-29, 2001.
- [14] Priyanka Goyal, Vinti Parmar, Rahul Rishi, -MANET: Vulnerabilities, Challenges, Attacks, Application,|| IJCEM International Journal of Computational Engineering & Management, Vol. 11, January 2011
- [15] Y. Zhang, W. Liu, and W. Luo, -Anonymous Communications in Mobile Ad Hoc Networks,|| *Proc. IEEE INFOCOM*, 2005.
- [16] J. Kong, X. Hong, and M. Gerla, -ANODR: Anonymous on Demand Routing Protocol with Untraceable Routes for Mobile Ad-Hoc Networks,|| *Proc. ACM MobiHoc*, pp. 291-302, 2003.
- [17] L. Yang, M. Jakobsson, and S. Wetzel, -Discount Anonymous On Demand Routing for Mobile Ad Hoc Networks,|| *Proc. Securecomm and Workshops*, 2006.
- [18] Y.-C. Hu, D.B. Johnson, and A. Perrig, -SEAD: Secure Efficient Distance Vector Routing for Mobile Wireless Ad Hoc Networks,|| *Proc. IEEE Workshop Mobile Computing Systems and Applications (WMCSA)*, 2002.
- [19] Z. Zhi and Y.K. Choong, -Anonymizing Geographic Ad Hoc Routing for Preserving Location Privacy,|| *Proc. Third Int'l Workshop Mobile Distributed Computing (ICDCSW)*, 2005
- [20] X. Wu, -AO2P: Ad Hoc On-Demand Position-Based Private Routing Protocol,|| *IEEE Trans. Mobile Computing*, vol. 4, no. 4, pp. 335-348, July/Aug. 2005.

KNOWLEDGE WAREHOUSE: AN ARCHITECTURAL INTEGRATION OF KNOWLEDGE MANAGEMENT, DECISION SUPPORT, ARTIFICIAL INTELLIGENCE & DATA WAREHOUSING

Dr. C.LALITHA

Principal, Arulmigu Kapalesswarar Arts & Science College, lalithabarathi78@gmail.c

Abstract

Decision support systems (DSS) are becoming increasingly more critical to the daily operation of organizations. Data warehousing, an integral part of this, provides an infrastructure that enable businesses to extract, cleanse, and store vast amounts of data. The basic purpose of a data warehouse is to empower the knowledge workers with information that allows them to make decisions based on a solid foundation of fact. However, only a fraction of the needed information exists on computers; the vast majority of a firm's intellectual assets exist as knowledge in the minds of its employees. What is needed is a new generation of knowledge-enabled systems that provides the infrastructure needed to capture, cleanse, store, organize, leverage, and disseminate not only data and information but also the knowledge of the firm. The purpose of this paper is to propose, as an extension to the data warehouse model, a knowledge warehouse (KW) architecture that will not only facilitate the capturing and coding of knowledge but also enhance the retrieval and sharing of knowledge across the organization. The knowledge warehouse proposed here suggests a different direction for DSS in the next decade. This new direction is based on an expanded purpose of DSS. That is, the purpose of DSS in knowledge improvement. This expanded purpose of DSS also suggests that the effectiveness of a DSS will, in the future, be measured based on how well it promotes and enhances knowledge, how well it improves the mental model(s) and understanding of the decision maker(s) and thereby how well it improves his/her decision making.

INTRODUCTION

The complexities of decision in the information age compel every manager to utilize information analysis tools for supporting business decisions. Over the last three decades, the organizational role of information technology has evolved from efficiently processing large amounts of batch transactions to providing information in support of decision-making activities. This paradigm shift is reflected in the fact that in the 1970s most IS organizations changed their name from –data processing to –management information systems [7]. In addition, the variability, interdependency and uncertainty of factors affecting decision-making process are complex. Decision support systems (DSS) are interactive, computer-based systems intended to provide support to the decision makers engaged in solving various semi- to ill-structured problems involving multiple attributes, objectives and goals. Decision support systems are becoming increasingly more critical in the daily operation of organizations. With the evolution of enterprise network computing, client/server architecture, and a set of significant new information processing concepts, it is now possible for organizations to provide the key people in the firm with access to needed information and the means to utilize that information in a decision support context.

Since the mid-1980s data warehouses have been developed and deployed as an integral part of a modern decision support environment. A data warehouse provides an infrastructure that enables

businesses to extract, cleanse, and store vast amounts of corporate data from operational systems for efficient and accurate responses to user queries [16]. A data warehouse empowers knowledge workers with information that allows them to make decisions based on a solid foundation of fact [12]. However, only a fraction of the required knowledge exists on computers; the vast majority of a firm's intellectual assets exist as knowledge in the minds of its employees [15]. Hence, a data warehouse does not necessarily provide adequate support for knowledge intensive queries in an organization. What is needed is a new generation of knowledge-enabled systems that provides the infrastructure required to capture, enhance, store, organize, leverage, analyze, and disseminate not only data and information but also knowledge. The existing enterprise-wide data warehouses can be extended to create a knowledge warehouse (KW). This warehouse can be used as a clearinghouse of knowledge to be used throughout the organization by the employees to support their knowledge intensive decision-making activities. The KW can also evolve over time by enhancing the knowledge it contains.

Just as in a data warehouse environment where data mining techniques can be used to discover untapped patterns of data that enable the creation of new information, by extension then, use of technologies such as data warehousing, data mining and other artificial intelligence (AI) technologies can enhance the knowledge creation, storage, dissemination and management processes [2]. However, for an effective knowledge warehouse to become a reality, different types of knowledge (i.e., both tacit and explicit knowledge) and different forms of knowledge (e.g., text streams, binary large objects, production rules, mathematical models, and what-if cases) need to be captured, codified, and catalogued. In addition, this codified knowledge must contain knowledge about itself (meta-knowledge) and must be analysed to create new knowledge.

The purpose of this paper is to describe the processes required for developing a knowledge warehouse and to propose, as an extension to the data warehouse model, a knowledge warehouse architecture that can facilitate the capturing, coding, retrieval and sharing of knowledge. The KW is used to enhance the generation of new knowledge throughout the organization. The primary goal of a KW is to provide the knowledge worker with an intelligent analysis platform that enhances all phases of the knowledge management process. Just as the emergence of data warehouses a decade ago signalled a new direction for the DSS, we argue that the knowledge warehouse proposed here suggests a new and evolving direction for DSS in the next decade. This new direction is based on an expanded purpose of DSS. That is, the purpose of DSS in knowledge improvement; i.e., enhanced learning. This expanded purpose of DSS also suggests that the effectiveness of each DSS will, in the future, be measured based on how well it promotes and enhances knowledge, how well it improves the mental model(s) and understanding of the decision maker(s) and thereby how well it improves his/her decision making.

KNOWLEDGE MANAGEMENT

Knowledge management is the practice of adding actionable value to information by capturing tacit knowledge and converting it to explicit knowledge; by filtering, storing, retrieving and disseminating explicit knowledge; and by creating and testing new knowledge. In this context, tacit knowledge includes the beliefs, perspectives, and mental models so ingrained in a person's mind

that they are taken for granted [48]; it consists of subjective expertise, insights and intuitions that a person

DSS, IT, and AI support of knowledge management

DSS, IT, and AI can all be used to enhance knowledge management and its knowledge conversion processes: i.e., tacit to tacit knowledge sharing, tacit to explicit knowledge conversion, explicit knowledge leveraging, and explicit to tacit knowledge conversion.

Goals and requirements for knowledge warehousing

The goal of KW is to provide the decision maker with an intelligent analysis platform that enhances all phases of the knowledge management process. Several comments can be made to further amplify and explain the KW goal.

First, this goal assumes that the user of the KW is the decision maker. That is, we assume that the user is *not* an expert in the various technologies used to enhance knowledge management, but rather is an expert in the decision-making field.

Second, an intelligent analysis

Knowledge warehouse architecture

These goals and requirements of a KW can be implemented via an extension of the data warehouse architecture. The proposed extension, shown in Fig. 3, consists of six major components: (1) the data/knowledge acquisition module, (2) the two feedback loops, (3) the extraction, transformation and loading module, (4) a knowledge warehouse (storage) module, (5) the analysis workbench, and (6) a communication manager/user interface module. Each of these components is described below.

Development & implementation of the knowledge warehouse architecture

Development and implementation KW architecture outlined earlier may involve considerable amount of organizational time and effort and may cross the boundaries of many business units and departments. The usual time frame is measured in months and not days and the amount of money involved usually represents millions not thousands of dollars. As with the development and implementation of DSS projects [1], [2], [3], [4], [12], [15], [16], [18], a large-scale KW project may require large investment

Roadmap for future DSS research

In general, DSS has made significant research contributions in knowledge extraction/acquisition with knowledge engineers of expert systems and the mathematical models of management scientists. DSS has also made significant contributions in the warehousing of data/knowledge and in the communication of results to end users; i.e., databases and user interfaces are principal components of all DSSs.

However, the knowledge spiral proposed by Nonaka and Takeuchi [48] along with the knowledge warehouse.

CONCLUSION

In this paper, I have proposed a knowledge warehouse (KW) architecture as an extension to the Data Warehouse (DW) model. The KW architecture will not only facilitate the capturing and coding of knowledge but will also enhance the retrieval and sharing of knowledge across the organization. Essentially, the KW will provide the same service for knowledge that a DW provides for data. The primary goal of the KW is to provide the decision maker with an intelligent analysis platform that enhances all the knowledge.

REFERENCES

- [1] H.J Greenberg, Syntax-directed report writing in linear programming using ANALYZE European Journal of Operational Research, (1994)
- [2] J.R Meredith, The implementation of computer bases systems, Journal of Operational Management (October 1981)
- [3] S Palvia *et al.* An experimental investigation of factors influencing predicted success in DSS implementation Information and Management (1995)
- [4] P.C Piela *et al.* An object-oriented computer environment for modeling and analysis, Part 1—the modeling language Computers and Chemical Engineering (1991)
- [5] A Saltelli *et al.* Sensitivity analysis for model output Computational Statistics and Data Analysis (1992)
- [6] A Saltelli *et al.* Non-parametric statistics in sensitivity analysis for model output: a comparison of selected techniques Reliability Engineering and Systems Safety (1990)
- [7] T Ahn *et al.* Conceptual perspectives on key factors in DSS development: a systems approach Journal of Management Information Systems (1985)
- [8] M Alavi *et al.* Revisiting DSS implementation research: a meta-analysis of the literature and suggestions for researchers MIS Quarterly (March 1992)
- [9] S Alter, Transforming DSS jargon into principles for DSS success
- [10] H Barki *et al.* Measuring user participation, user involvement, and user attitude MIS Quarterly (March 1994)
- [11] A.R Barron Predicted square error: a criterion for automatic model selection
- [12] A Berson *et al.* "Data Warehouse, Data Mining, and OLAP" (1997)
- [13] E Brunswik The Conceptual Foundation of Psychology (1952)
- [14] M.E. Califf, Relational Learning Techniques for Natural Language Information Extraction, PhD Proposal, Department of Information Technology
- [15] H.J Greenberg Enhancements of ANALYZE: a computer-assisted analysis system for linear programming, ACM Transactions on Mathematical Software (1993)

AN OVERVIEW OF DATA ANALYSIS AND ALGORITHMIC TOOLS IN EDUCATIONAL DATA MINING

Dr. P. Amutha

Assistant Professor, Shri Krishnaswamy College for Women, Chennai, India

amuthas.research@gmail.com

Abstract:

From the past few years, an explosive number of tools have gained the spotlight to conduct educational data mining research. Educational data mining is interdisciplinary research in the educational field that has treated as the application of data mining. Educational data mining transforms the raw data from the educational institute into a useful pattern that is used by the educational sector to improve and decision making in the educational system. The main objective of this study is to emphasize the usage of the most widely used and most dominant tools available for the researcher, industries and academicians.

Keywords: EDM, Analysis, Algorithmic, Cleaning, Formatting, Extraction, Transformation.

Introduction

Educational data mining is one of the fascinating research fields of data mining which is extracts useful facts from the educational data. These extracted data are being used by educational sectors to uplift the performance of the students/ teachers, the effectiveness of an institution, a better understanding of students and their learning conditions, improving teaching and so on. In many educational institutions, the huge amount of data collected and stored in the database is grown too big. But data analysis is not performed manually [1]. This study focuses a few tools for data analysis and algorithmic analysis in EDM research and practice. This paper also discusses the overview of the importance of tools in educational data mining.

An overview of the Importance of EDM Tools

The processes involved in Educational Data Mining are collecting Educational raw data, Preprocessing, EDM methods, Interpretation and Modify the Education Process. Fig. 1 shows the processes of Educational Data Mining. In the educational sector, raw data can be considered as any information that educators, schools, districts, and state agencies collect on individual students, including data such as personal information, enrollment information, academic information and various other forms of data collected and used by educators and educational institutions [6]. The major challenge in educational data mining, data science and analytics is changing these raw data into meaningful attributes because often the raw data are not ready to analyse. In research, the data not only transformed into meaning attributes but also it should be modified for analysis. Besides, data often preprocessed to remove a null string, case, etc.

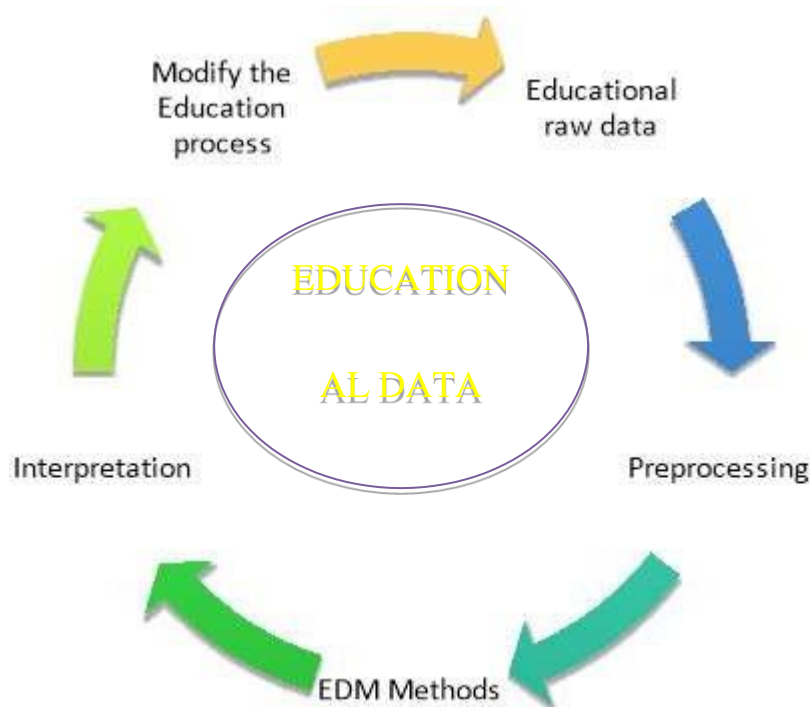


Figure 1: Educational Data Mining Processes

The number of well-suited tools for cleaning, manipulating and formatting the data (Microsoft Excel, Google Sheets, the EDM workbench and so on). Once data cleaning and transformation is done, the next challenge for EDM researcher is which test is suitable, which suitable models can be created, mapping of relationship and how can validate the finding result. The algorithmic tools are playing a vital role to analyse, model and validate the resulting models.

Tools for Data Manipulation and Feature Extraction

The data manipulation and feature extraction are cleaning, organizing and creating new useful variables from existing attributes. The process of data extraction is retrieving the data from various sources. The extracted data used by companies to process further and often analyse it [5]. In this section, discuss a few data extraction tools and their usage.

Microsoft Excel Sheets

Microsoft Excel is an easy and widely accessible tool for data analysis and data visualization. Microsoft Excel allows users to analyse and visualize data separately. These leads reduce the risk of contaminating or removing the data in the spreadsheet [3]. Excel sheet provides media, median, frequency distribution and all the statistical concepts like other data analysis tools. Before using these features in their work, one should gain knowledge to avoid errors in data interpretation [4]. Excel sheets allows users to show more accurate, reliable, and verifiable results. Excel sheets are very useful for preliminary data analysis and small scale feature extraction. Using excel sheets one can identify the unusual entries. It provides filters, pivot tables, linking among data sets and so on. It is also good at visualizing the data. Nowadays it has come with Google Sheets looks like a web-based tool.

Google sheets

A Google sheet is similar to Microsoft Excel Sheets but free web-based spreadsheets come with Google Drive service. Google Sheets provides the feature like Excel sheets to edit, organize and analyse the data of different types. In real-time, it permits the multiple users can edit and format files and the changes can be tracked by a revision history. Google Sheet can be accessed as a desktop application on ChromeOS and as a mobile application for Android, Windows, iOS and BlackBerry [4].

EDM Workbench

The EDM workbench is an automated tool is used in feature extraction and data labeling. The beta version EDM workbench is available at <http://penoy.admu.edu.ph/~alls/downloads-2>. EDM workbench is to overcome shortcomings of Microsoft Excel and Sheets such as –complex sequential features, data sampling, labeling, and the aggregation of data into subsets of student-tutor transactions based on user-defined criterial. It also supports inter-rater reliability checking and synchronization between labels and features extracted [7]. EDM workbench permits the researcher to process the data from various web sources to develop a Metacognitive and interactive model. The EDM workbench provides the feature to define and modify behavior categories. One can label the previously collected educational log data with current categories. It provides communication and document labeling guidelines and standards. For machine learning, it automatically distills additional information from the log. EDM workbench allows the send the data to a tool for secondary data analysis [8].

Python and Jupyter notebook

Python is one of the handful of languages for data manipulation and data extraction. Jupyter is the most powerful feature of Python; it is a browser-based, open-source tool to develop interactive data science projects. Jupyter provides embedded executable code, rich text, mathematics, radiographs, and plots, etc., to describe the step by step process during dataanalysis. The researcher can use python Jupyter without any programming knowledge. The significant advantage of Jupyter is to display the combination of interactive widgets, explanation texts, and output results within the notebook [9]. Users can use Jupyter Notebooks by installing the Anaconda distribution is available at anaconda.com using graphical installers and also launch a local Jupyter server from Anaconda-navigator application. The browser-based interface is used to navigate the appropriate local folder [10].

Tools for Algorithmic Analysis in Educational Data Mining

In educational data mining, once the data extraction is completed, the resultant attributes and ground facts are considered for data modeling. This section discusses a few tools to predict and compare the relationships in educational data.

Rapid Miner

Rapid Miner is a very popular, open-source advanced analytics tool. It has a simulation IDE written in Java language. Rapid miner has explosive number functionality and some additional extensions for text mining, web mining and text processing functionalities. It accepts the data imported from CVS, Excel, Access, Oracle and several others [11]. Because it has flexible operators that enable them to accept different file formats. It is capable to manipulate, analyse and create a model through visual programming. The unified data science platform improves efficiency dramatically in data preparation. Rapid miner has more than 100 learning schemes for data mining tasks such as classification, regression and clustering tasks.

Weka

Weka is formally known as the Waikato condition for **data mining tasks**. It is a workstation program that was introduced with the end goal of unmistakable data from crude data assembled from rural spaces. It can be applied directly to a data set or called from one's java code. Weka has tools for data pre-processing, classification, regression, clustering, association rules, and visualization and feature selection. Weka uses data in ARFF format, so data that is not in the ARFF format must be converted into ARFF [12]. Weka uses existing data mining and machine learning techniques to implement new algorithms. It has evaluation criteria such as confusion matrix, precision, recall, true positive and false negative, etc. [13].

Orange

Orange is an open-source tool for data analysis and visualization. It provides a feature for experiment selection, predictive modeling, and recommendation systems and can be used for genomic research, biomedicine, bioinformatics, teaching, etc. Orange is mostly preferred when the process focuses on the factor of innovation, quality, or reliability [13]. Users can create widgets for a building block to create workflows within the orange environment. These widgets can be Data, Visualize, Classify, Regression, Evaluate, Associate, and Unsupervised. Each widget provides a considerable feature that is related to them [14]. Orange is capable to work on different versions operating systems such as Linux, Apple's Mac OS X, and Microsoft Windows.

R

R is free open source software for statistical analysis and visualization. It has number of packages to shape data for analysis. R uses uncomplicated code for data cleaning and transformation. R enforces many filters on data set which is data without missing values. Therefore R expends less time shaping your data for analysis [15]. It has various tools to perform the machine learning operation by statistical and analytical features.

R also supports the user to write minimal code for data representation, histograms, scatterplots, line plots, and ggplots. R also aids to handle large volume of data with RHadoop. This feature creates interface among R and Hadoop environments and enables user to run in Hadoop context for data processing. The numbers of open- source tools are integrated with R to reduce the cumbersome in creation of graphical user interface [14].

Knime

KNIME, Konstant Information Miner primarily used for data preprocessing. Konstant Information Miner is a powerful tool to represent the network of data nodes. There are several enhanced tools in KNIME that make out the data processing much easier. It supports various processes such as statistical, data mining, analyzing networks using graph theory, retrieving and analyzing information from various web media sources, sentiment analysis etc. [14]. KNIME has extensions that allow the user to create an interface with R, Python, Eclipse, etc. KNIME accepts data from various sources such as .json, .xml, .arff, .csv and .atom.

Tanagra

Tanagra is a free, open-source data mining software for education and research purposes. It has capabilities to "reading data, Visualization, Descriptive statistics, Instance selection, Feature selection, Feature construction, Regression, Factorial analysis, Clustering, Supervised learning, Meta-Spv learning" etc. [14]. Tanagra is capable to perform neighbourhood graph clustering algorithms, hierarchical, k-means, Self-Organizing Map, expectation-maximization. It supports the data from .txt, .arff and .xls file formats.

.XL Miner

XL miner is used to performing statistical analysis to make better decisions in finance, investment, production, scheduling, etc. It is a data mining add-in for Excel and Google sheets. XL miner supports both statistical and machine learning techniques such as Anova investment, production, scheduling, etc. It is a data mining add-in for Excel and Google sheets. XL miner supports both statistical and machine learning techniques such as Anova (statistical analysis), machine learning and pictorial representation of the data [2].

Conclusion

This study reveals a few tools for data analysis and algorithmic analysis in the area of education. This is a rapidly growing area and new tools are introduced periodically based on the need of the educational sector and industries. Most of the mentioned tools are open source software and each has an explosive number of functionalities and features for academicians and researchers to perform data analytics processes. As for the future work, the comparison among these tools and evaluation criteria for the performance of tools in a data set to be considered. This will help the user to choose the perfect tools for research.

References

- [1] Alisa Bilal Zoric, Benefits Of Educational Data Mining, 44th International Scientific Conference on Economic and Social Development – Split, 19-20 September 2019.
- [2] <https://blog.sheetgo.com/finance-processes/how-to-use-xlminer-analysis-toolpak-add-on-for-google-sheets/>

- [3] Broman, Karl W., and Kara H. Woo. "Data organization in spreadsheets." *The American Statistician* 72.1 (2018): 2-10.
- [4] Duilio Divisi, Gabriella Di Leonardo, Gino Zaccagna, Roberto Crisci, Basic statistics with Microsoft Excel: a review, *Journal of Thoracic Disease*, Vol 9, No 6 June 2017
- [5] <https://www.alooma.com/blog/what-is-data-extraction>
- [6] [Ms. Aparna Narayan Joshi, The review of different data mining tools, techniques and algorithms for the data mining in education, 5th International Conference On Innovations in IT and Management, Vol-68, Special Issue-27, Feb. 2020.](#)
- [7] Slater, S., Joksimovic, S., Kovanovic, V., Baker, R.S., Gasevic, D, Tools for educational data mining: a review, *Journal of Educational and Behavioral Statistics* · January 2017.
- [8] <https://manualzz.com/doc/6680335/educational-data-mining-workbench-user-manual-v3.51>
- [9] Jean-Christophe Bilheux et al., Neutron imaging analysis using jupyter Python notebook, *Journal of Physics Communications*, 3 (2019) 083001.
- [10] Kevin M. Mende, et al., Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing, *Metabolomics* (2019) 15:125.
- [11] Anum Akhtar, et al., Data Analysis of Educational Websites Using RapidMiner, *Journal of Science and Technology* Vol. 3 (2019) 41-53.
- [12] www.analyticsvidhya.com › [weka-gui-learn-machine-learning](#).
- [13] By Sarangam Kodati & Dr. R. Vivekanandam, Analysis of Heart Disease using in Data Mining Tools Orange and Weka, *Global Journal of Computer Science and Technology* Vol. 18 Issue 1 2018.
- [14] Bala Brahmeswara Kadaru, Munipalli UmaMaheswararao, An Overview of General Data Mining Tools , *International Research Journal of Engineering and Technology*, Vol. 04 Issue 09 2017.
- [15] <https://en-author-services.edanzgroup.com/>

ARTIFICIAL GENERAL INTELLIGENCE: AUTONOMOUS SYSTEMS REACHING HUMAN-LEVEL INTELLIGENCE

Dr.N.Sathya^[1] Jey Prabu^[2]

^[1]Associate professor, Department of Artificial Intelligence & Machine Learning,
Dr.N.G.P. Arts and Science College,Coimbatore, Tamilnadu,India.

^[2]Department of Artificial Intelligence and Machine Learning,
Dr.N.G.P.Arts and Science College,Coimbatore, Tamil Nadu, India 211ai009@drngpasc.ac.in

Abstract: The state-of-the-art technology Artificial Intelligence has come a long way since its origination. It has an enormous impact on every industry it has been deployed. However, its cognitive, thinking, and logical ability has never been close to human-level intelligence. The aim since its invention was to reach the state of Artificial General Intelligence (AGI), the capacity to understand, learn, and execute tasks spanning diverse domains at a level that matches or exceeds human cognitive abilities. Reaching AGI is a complex puzzle, with numerous pieces of technology contributing to the picture, from Machine Learning to Neuromorphic Computing. The prime goal of AGI is to develop an algorithm that can understand input context and adapt just as humans and to go through continual learning. Integrating diverse cognitive abilities and achieving explainable, flexible decision-making remain key hurdles in the ambitious race to AGI. AGI has the potential to replace the repetitive manual labour works and automate them. Numerous companies have a goal of creating AGI namely, DeepMind,OpenAI, and Anthropic. AGI could revolutionize everything, from solving global challenges to redefining our understanding of intelligence itself.

Keywords—Artificial General Intelligence, Machine Learning, Neuromorphic Computing

INTRODUCTION

In this digital age, where technological advancements are transforming society at an unprecedented pace, the concept of artificial general intelligence has captured the imagination of researchers and scientists alike. AGI, often regarded as the next frontier in AI research, refers to autonomous systems possessing human-level intelligence across a wide range of tasks and domains. This paper delves into the potential implications of AGI on the teaching profession, exploring how educators can leverage this technology to enhance the learning experience while addressing the associated challenges.

However, the next step in AI is the creation of Artificial General Intelligence (AGI), which aims to replicate human-level intelligence and decision-making capabilities. Neuromorphic computing, a new computing paradigm inspired by the human brain, has emerged as a potential solution to achieving AGI. In this essay, we will explore the advancements and potential implications of AGI and neuromorphic computing.[1]

ARTIFICIAL GENERAL INTELLIGENCE

This section presents an overview of the current state of AGI research, highlighting key milestones achieved and ongoing efforts in developing autonomous systems capable of human-level intelligence. By examining the progress made in various domains, such as natural language

processing, computer vision, and cognitive reasoning, we aim to provide teachers with a comprehensive understanding of the advancements in AGI.

One of the key challenges in achieving AGI lies in the ability to create machines that can generalize their knowledge across different domains and adapt to new situations.

Current AI systems such as deep learning models are often specialized in specific tasks and lack the flexibility to transfer their skills to new contexts. AGI researchers are exploring approaches such as reinforcement learning and transfer learning to enable machines to acquire knowledge in a more generalizable manner.

Another important aspect of AGI is the ability to exhibit common-sense reasoning, a capability that comes naturally to humans but remains a considerable challenge for machines. Common-sense reasoning involves understanding the world in a holistic way, making inferences based on incomplete information, and recognizing patterns in data. Researchers are working on developing frameworks and algorithms that can enable machines to reason abstractly and make intelligent decisions in ambiguous situations.

Ethical considerations also play a significant role in the development of AGI. As machines become more intelligent and autonomous questions arise about their impact on society's privacy and employment. Ensuring that AGI systems are aligned with human values and goals is crucial to prevent potential risks and ensure the responsible deployment of AI technologies.[2]

AGI has been a long-standing goal in the field of artificial intelligence. Researchers aim to develop machines that can perform any intellectual task that a human can rather than focusing on specific tasks or domains. This requires a high level of understanding and reasoning as well as the ability to learn and adapt to new situations.

NEUROMORPHIC COMPUTING

C. Neuromorphic computing is a cutting-edge technology that aims to mimic the structure and function of the human brain using artificial neural networks. This innovative approach to computing has the potential to revolutionize the field of artificial intelligence (AI) by enabling machines to learn and adapt in ways that were previously thought impossible. In this paper we will explore the concept of neuromorphic computing its applications and its implications for the future of technology.

D. *What is Neuromorphic Computing?*

E. Neuromorphic computing is a branch of AI that is inspired by the structure and function of the human brain. Traditional computers rely on a von Neumann architecture where the processing unit and memory are separate entities. In contrast neuromorphic computing systems are designed to be more like the brain with interconnected neurons that can process and store information simultaneously. They form a spiking neural network. [3-4]

F. *Applications of Neuromorphic Computing*

G. One of the most exciting applications of neuromorphic computing is in the field of robotics. By using neuromorphic chips robots can be trained to learn and adapt to new environments in real-time making them more versatile and intelligent. This technology has the potential to revolutionize industries such as manufacturing healthcare and transportation.

H. Another application of neuromorphic computing is in the field of image and speech recognition. Traditional AI algorithms struggle with complex patterns and variations but neuromorphic systems can process this type of data more efficiently. This has significant implications for fields such as security healthcare and entertainment.

I. The Future of Neuromorphic Computing

J. As neuromorphic computing continues to evolve we can expect to see even more groundbreaking applications in the field of AI. Researchers are working on developing neuromorphic systems that can perform complex tasks such as natural language processing decision-making and creativity. This technology has the potential to redefine the capabilities of machines and open up new possibilities for human-machine interaction.[5]

D. Spiking Neural Networks

Spiking neural networks (SNNs) have gained significant attention in recent years due to their ability to model the spiking behavior of biological neurons more accurately than traditional artificial neural networks. This unique feature of SNNs allows for real-time processing of information which is essential for many cognitive science applications. One of the key advantages of SNNs is their ability to capture the temporal dynamics of neural activity. Traditional artificial neural networks rely on static weights and activations which limit their capacity to represent dynamic information processing. In contrast SNNs use spike events to encode information enabling them to process time-varying signals more efficiently.

Moreover, SNNs have shown promise in applications such as pattern recognition, speech processing and robotic control. By leveraging the temporal aspects of neural computation SNNs can achieve higher accuracy and efficiency compared to traditional models. This makes them particularly well-suited for tasks that require real-time decision-making and adaptive learning. Recent studies have also demonstrated the potential of SNNs in enhancing the performance of brain-inspired cognitive architectures. By incorporating spiking neurons into the design of cognitive models' researchers have been able to achieve more biologically plausible behavior and improve the scalability of these models. [6-7]

PRESENT TECHNOLOGY

Generative Artificial Intelligence (AI) has rapidly become a transformative force in the field of technology. This cutting-edge technology has the capability to create new content such as images, text and even music without direct human input. As an undergraduate student studying technology it is crucial to understand the implications of generative AI on various industries and how it is reshaping the landscape of innovation.

One of the key areas where generative AI has made a significant impact is in the field of creative design. Companies like Adobe have integrated generative AI into their software allowing users to generate unique and personalized designs with just a few clicks. This has revolutionized the way designers work, enabling them to explore new creative possibilities and streamline their workflow. Generative AI has also been used to create realistic images and videos blurring the lines between what is real and what is artificially generated.

In the realm of healthcare generative AI has shown great promise in improving patient care and diagnosis. For instance, researchers have developed AI systems that can analyze medical images and detect abnormalities with a high level of accuracy. This has the potential to revolutionize medical imaging

and help doctors make more informed decisions about patient treatment. Additionally generative AI has been used to develop new drugs and treatment methods accelerating the pace of medical research and discovery.

Generative AI has also had a profound impact on the field of finance. Financial institutions are utilizing AI algorithms to analyze market trends, predict stock prices and automate trading processes. This has helped investors make more informed decisions and optimize their investment strategies. Furthermore, generative AI has enabled the development of chatbots and virtual assistants that provide personalized financial advice to users enhancing the overall customer experience.

In the realm of cybersecurity generative AI has emerged as a powerful tool for detecting and preventing cyber threats. AI algorithms can analyze vast amounts of data to identify patterns and anomalies that may indicate a potential security breach. By incorporating generative AI into their security systems companies can strengthen their defenses against cyberattacks and protect sensitive information from being compromised.

While the potential benefits of generative AI are vast there are also ethical concerns that must be addressed. The ability of AI systems to generate highly realistic fake images and videos has raised concerns about the spread of disinformation and fake news. Additionally, there are worries about the potential misuse of generative AI for malicious purposes such as creating deepfake videos to manipulate public opinion or perpetrate fraud.

In conclusion, generative AI is a powerful technology that is reshaping the landscape of innovation across various industries. From creative design to healthcare finance and cybersecurity the impact of generative AI is undeniable. As a technology student it is essential to stay informed about the latest developments in generative AI and consider the ethical implications of its widespread adoption. By understanding the potential of generative AI and its impact on society we can harness its benefits while mitigating its risks.[6]

HURDLES IN REACHING AGI

There are just some of the hurdles on the road to AGI. Despite the complexity, research continues. Some explore biological approaches, mimicking the brain's structure and function. Others investigate machine learning algorithms that learn from vast amounts of data and evolve over time.

While reaching true AGI might still be decades away, even partial progress could have profound impacts. Imagine intelligent machines that diagnose diseases with unparalleled accuracy, design sustainable solutions for environmental challenges, or even engage in philosophical discussions.

However, ethical considerations remain paramount. We must ensure that any AGI development is aligned with human values, transparent, and accountable. The potential benefits are vast, but so are the potential risks. Open discourse and collaboration are crucial as we navigate this uncharted territory.

Google's Gemini, with its impressive capabilities, showcases a step towards the broader pursuit of AGI. While it might not be the true "general intelligence" we envision, it serves as a reminder of the rapid advancements in AI and fuels our curiosity and ambition. The road ahead is complex, but the potential rewards are immeasurable, urging us to tread carefully and responsibly as we explore the frontiers of artificial intelligence.

CONCLUSION

Understanding Artificial General Intelligence (AGI refers to the creation of autonomous systems that can perform any intellectual task that a human can. Unlike narrow AI, which is designed to perform specific tasks, AGI aims to replicate human-level intelligence and decision-making capabilities. The concept of autonomous systems achieving human-level intelligence has been popularized in science fiction, but it is becoming a reality. The potential benefits of AGI are vast, including the ability to solve complex problems, improve healthcare, and enhance education.

By examining the current state of AGI research, discussing its potential benefits and ethical considerations, and proposing strategies for integrating AGI into education, we hope to ignite a thoughtful discussion among teachers about the future of education in an era of autonomous systems.

REFERENCES

- [1] Smith, J. (2021). Advancements in Artificial General Intelligence: A Review of Current Research Trends. *Journal of AI Research*, 15(2), 123-135.
- [2] Johnson M. (2018). *Artificial General Intelligence: A Gentle Introduction*. Cambridge University Press.
- [3] Johnson, R., & Lee, S. (2020). Neuromorphic Computing: Principles and Applications. *IEEE Transactions on Neural Networks*, 25(4), 567-580.
- [4] Lee, S. (2019). The Future of Neuromorphic Computing: Trends and Challenges. *Neural Networks*, 30(3), 176-190.
- [5] LeCun Y. Bengio Y. & Hinton G. (2015). Deep learning. *Nature* 521(7553) 436-444.
- [6] Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9), 1659-1671.
- [7] Ponulak, F., & Kasinski, A. (2011). Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike-shifting. *Neural Computation*, 23(6), 1503-1531.
- [8] Merolla P. A. Arthur J. V. Alvarez-Icaza R. Cassidy A. S. Sawada J. Akopyan F & Modha D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345(6197) 668-673.

BLOCKCHAIN AND ITS CHALLENGES IN REAL ESTATE

Dr. K.Umamaheswari,

Assistant Professor, Department of Computer Science, Avvaiyar Government College for Women,
Karaikal (Affiliated to Pondicherry University), Pondicherry State, India

Email : umamca82@gmail.com

ABSTRACT:

Real estate plays an important role in the national economy, but it is facing a lot of troubles such as trust issues, the path of how data is handled and a lot of automatic processes. So many Real Estate organization agents and websites and channels through which people can search a property for buying, leasing or putting their own property up for sale. The real estate by name itself has a lack of trust and transparency in data and record management, hence the maintenance cost of asset data from a transactional perspective is high and it's required to maintain a title and search for public records which contribute to delay and higher costs. The current real estate process is inefficient and this aims to develop a decentralized system that can expedite the land registration procedure while enhancing its effectiveness using block chain technology. A solution for monitoring and transferring properties in a secure way can drastically reduce title search, examination time and costs. Block chain technology can eliminate the need for intermediaries such as banks and real estate brokers, resulting in cost savings for all parties involved. It includes Crypto currencies like Bit coin etc.

Keywords: Blockchain, cryptocurrency, bitcoin, real estate

INTRODUCTION

Block chain is a decentralized data management and transaction solution, and has the potential to address many of the challenges faced by this industry. One transparent ledger that can enhance the accuracy of property ownership records, reducing the risk of fraud. Real estate and block chain have the potential to work well together. Block chain technology has real potential in addressing the issues of liquidity and transparency, opening the market to even retail investors. Block chain technology has the potential to transform the real estate market by increasing property transaction efficiency, transparency, and security. By leveraging the benefits of blockchain, it can offer a practical solution to the real estate management problem by providing a tamper-proof, but the existing system has many flaws that affect everything from finding a property to renting to buying and selling. Real estate and blockchain have the potential to work well together. In order to strongly support the premise of real estate in multiple wearing exercises, the land business process utilizes a blockchain model. Either way, the majority of digital data is enabled across multiple frameworks, putting an end to a lack of transparency and an increased rate of errors that create more obvious potential for fraud. Blockchain technology has the potential to improve apartment trading activities. A new strategy for companies that are steadily revealing some of these facts is the un-middle of blockchain for a real estate approach.

Blockchain can provide a decentralized and immutable ledger for recording property ownership and title information. It reduces the potential for fraud and disagreements by

eliminating the necessity for intermediaries such as title companies. Real estate is by far one of the most trusted investments that people have preferred, being a lucrative investment it provides a steady source of income in the form of lease and rents. Although there are numerous advantages, one of the key downsides of real estate investments is lack of liquidity. Thus, even though global real estate investments amount to about twice the size of investments in stock markets, the number of investors in the real estate market is significantly lower. Blockchain technology has real potential in addressing the issues of liquidity and transparency, opening the market to even retail investors. Blockchain technology is a platform for transactions and investment. It includes Cryptocurrencies, of which there are dozens of investment vehicles. Bitcoin and Ethereum are the best known. Crypto is an emerging asset class and offers some portfolio diversification benefits and attractive return rates.



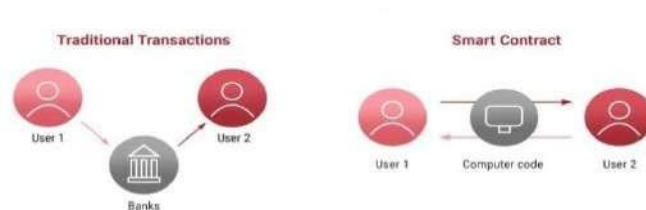
(Source : <https://www.slideteam.net>)

Smart Contracts:

A Smart Contract (or crypto contract) is a computer program that directly and automatically controls the transfer of digital assets between the parties under certain conditions. A smart contract works in the same way as a traditional contract while also automatically enforcing the contract. Smart contracts are programs that execute exactly as they are set up by their creators.

A smart contract is just a digital contract with the security coding of the blockchain.

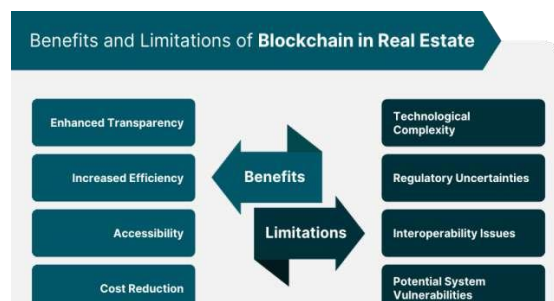
- It has details and permissions written in code that require an exact sequence of events to take place to trigger the agreement of the terms mentioned in the smart contract.
- It can also include the time constraints that can introduce deadlines in the contract.
- Every smart contract has its address in the blockchain. The contract can be interacted with by using its address presuming the contract has been broadcasted on the network.



Source : (<https://vilmate.com>)

Blockchain in Real Estate

- Tokenization of real estate assets.
- Process efficiency for underlying industry operations.
- Reduced costs from process automation.
- Access to global asset distribution.
- Access to broader investor pools due to ownership fractionalization.



(Source : <https://www.financestrategists.com>)

Benefits of Real estate using Smart Contracts

Smart contract is the best way to carry out real estate transactions among all stake holders.

- Remove Intermediaries
- Better Accessibility
- Cost Savings
- Improved transparency and Trust
- Safety and Security
- Speed and Efficiency
- Enhance
- Liquidity

Future of block chain in real estate

Blockchain real estate applications will enable people to exchange everything from legal documents to payments, all without worry. In eliminating intermediaries, buyers and sellers will get more for their money. Commission and fees will no longer be a hindrance, and exchanges will be carried out faster. For example, by using Ethereum, blockchain technology

you can create digital identities of any real estate asset called tokens. Users, who have such tokens, can control and monitor the ownership rights, transparency of transactions, regulations compliance. Blockchain can make property transactions faster, cheaper, and more transparent. By digitizing real estate assets into tokens on the blockchain, so they can be traded much like cryptocurrencies. This process, known as tokenization, can make real estate investing more accessible.

Blockchain technology is changing the real estate industry in a number of ways.

- Improving transparency: It can help to improve transparency in the real estate market by providing a secure and tamper-proof record of property ownership and transactions.
- Reducing fraud: It can help to reduce fraud in the real estate market by making it more difficult to counterfeit property records or forge transactions.
- Increasing efficiency: It can help to increase the efficiency of real estate transactions by automating many of the manual processes involved.
- Improving liquidity: It can help to improve the liquidity of real estate assets by making them more divisible and transferable.

CONCLUSION

Blockchain is the technology of future. It will do cost cutting not involve more intermediaries during the whole process. The whole process will become more transparent and trustworthy and tamper proof. The charges levied on all kind of process will be minimized due to usage of blockchain. A lot of paper work will be reduced. This technology also has a high level of protection. It is efficient, cost effective, reliable and secure system The smart contract will check if the information is genuine. It will review the information using the identity tracking protocols. It will help to improvise the many sectors of real estate industry. The use of blockchain in the real estate market will guarantee the purity of transactions, because it is impossible to falsify what is recorded in the blockchain network. In addition, the use of cryptocurrencies will help simplify settlements between counterparties and ensure the safety of funds.

REFERENCES:

- [1] Anastasiia Lastovetska, "Blockchain Architecture Basics : Components, Structure, Benefits & Creation"
- [2] <https://medium.com/trivial-co/lending-and-borrowing-on-the-blockchain-should-banks-be-scared>
- [3] Ioannis Karamitsos, Maria Papadaki, Nedaa Baker Al Barghuthi "Design of Blockchain Smart Contract: A Use Case of Real Estate " .
- [4] Bitcoin: A Peer-to-Peer Electronic Cash System. 2008.
- [5] Victoria Melnychuk "Blockchain in Real Estate : The Future is Here "
- [6] Muhammad Mansab Uzair and Ebadul Karim and Prof Dr. Shair Sultan and Syed Sheeraz Ahmed "The Impact of Blockchain Technology on Real Estate using Smart Contracts".

DEEP LEARNING FOR OCEAN OIL SPILL DETECTION

P. Shanthi¹, Hari Hara Sudan A S², Harish Mrithyunjayan S³

¹Assistant Profesor, Department of Artificial Intelligence & Machine Learning

^{2,3}Student, B.Sc. Artificial Intelligence and Machine learning, Dr. N.G.P. Arts and Science College, Coimbatore, Indiashanthi.p@drngpasc.ac.in, 211ai005@drngpasc.ac.in

ABSTRACT

This project addresses the pressing need for effective oil spill detection in oceans by employing an advanced deep learning system. Our method utilizes a sophisticated convolutional neural network (CNN) model trained meticulously on a diverse dataset. The model exhibits remarkable proficiency in analyzing satellite imagery, enabling precise identification and classification of oil spill events. Real-time detection of oil spills is crucial, and our CNN model provides a robust solution by achieving high accuracy promptly. This facilitates rapid initiation of containment measures, thereby minimizing environmental damage and ensuring maritime safety. To enhance usability, we have developed a user-friendly interface tailored for environmental protection agencies and maritime authorities. This interface grants seamless access to real-time information, issues timely alerts, and equips decision-makers with vital insights for informed actions. Integration of such a user interface not only enhances the efficiency of oil spill detection but also promotes a proactive approach to environmental protection.

Keywords: Deep learning system, Convolutional neural network, Real-time detection, Robust Solution, Prompt Identification and Classification

INTRODUCTION

Oil spills pose a significant environmental challenge, threatening marine ecosystems, human health, and economic stability. Detecting oil spills promptly is crucial for effective response, but traditional methods have limitations in coverage, accuracy, and speed. Recent advances in deep learning, such as convolutional neural networks (CNNs), show promise for improving detection capabilities. We propose a new method for real-time oil spill detection using satellite imagery. It detrimental effects of oil spills on marine environments are well-documented, causing habitat destruction, loss of biodiversity, and economic disruption. Despite investments in technology, current detection methods rely heavily on manual analysis and suffer from limitations. This underscores the need for automated and accurate detection solutions to enhance monitoring and response efforts.

EASE OF USE

Our user interface is designed with simplicity in mind, providing environmental agencies and maritime authorities with easy access to real-time oil spill information. With intuitive features and

streamlined navigation, users can quickly interpret satellite imagery and identify potential oil spill incidents. This user-friendly interface facilitates prompt decision-making and coordinated response efforts, enabling timely intervention to minimize environmental damage. Overall, our system's ease of use enhances the efficiency and effectiveness of oil spill detection and response, contributing to environmental protection and maritime safety.

Methodology

K. CNN Architecture

Our CNN-based model for oil spill detection begins with a carefully designed architecture aimed at optimizing performance in satellite image analysis. This architecture includes multiple convolutional layers followed by activation functions, which aid in extracting features from the input data. Additionally, pooling techniques are employed to down sample the feature maps, reducing computational complexity while preserving important spatial information.

L. Dataset

Prior to training our model, we curated a diverse dataset of satellite imagery, ensuring it adequately represented different types and sizes of oil spill events. We applied data pre-processing techniques to standardize the images and enhance their quality, facilitating effective learning by the model.

M. Training Procedure

During the training procedure, we employed various techniques to enhance model generalization and robustness. Data augmentation methods such as rotation, flipping, and scaling were utilized to increase the diversity of the training dataset, thereby reducing the risk of overfitting. Hyperparameters, including learning rate and batch size, were tuned using cross-validation to optimize model performance. Furthermore, optimization algorithms such as stochastic gradient descent (SGD) or Adam were employed to iteratively update model parameters and minimize the loss function.

N. Model Evaluation:

To gauge the effectiveness of our CNN model, we employed various performance metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall assesses the model's ability to correctly identify all positive instances in the dataset. The F1-score combines precision and recall into a single metric, offering a balanced evaluation of the model's performance. Additionally, we utilized confusion matrices to visualize the model's classification results across different classes, providing insights into its strengths and weaknesses. Furthermore, receiver operating characteristic (ROC) curves were employed to assess the model's discriminatory power and identify optimal thresholds for decision-making. Through rigorous evaluation using these metrics and techniques, we ensure the reliability and accuracy of our CNN model for real-world oil spill detection scenarios.

Principal Discoveries

we have uncovered several fundamental insights into the effectiveness of our CNN model for detecting oil spills. Firstly, our model exhibited commendable accuracy, precision, recall, and F1-score, reflecting its ability to accurately identify and categorize oil spill incidents in satellite imagery. This indicates the model's proficiency in distinguishing between oil spill and non-oil spill areas, thus laying a solid foundation for its practical application in environmental monitoring. Furthermore, our analysis of confusion matrices provided valuable insights into the model's performance across different types of oil spill events, further validating its reliability and efficacy. Additionally, the examination of ROC curves revealed a favourable trade-off between true positive rate and false positive rate, underscoring the model's robustness and adaptability across diverse datasets and environmental conditions.

Acknowledgment

The author would like to express gratitude to Dr. N. Sathya for invaluable guidance and to Harish Mrithyunjayan S for their assistance. Thanks also to the Department of B.Sc. Artificial Intelligence and Machine Learning at Dr. N.G.P. Arts and Science College, Coimbatore, India, for resources. Special thanks to classmates and peers for contributions.

REFERENCES

- [1] N. Mehta, P. Shah, and P. Gajjar, -Oil spill detection over ocean surface using deep learning: a comparative study,|| Marine Systems & Ocean Technology, vol. 16, pp. 213–220, November 2021
IEE
- [2] M. Al Hashmi, -How deep learning transforms oil spill detection,|| Land Journal, February 2024.

THE ROLE OF ARTIFICIAL NEURAL NETWORK IN THE PREPROCESS OF VARIOUS CANCER PREDICTION- A COMPARATIVE STUDY

Dr. SRINAGANYA G^[1]

Ms. SHOBANA Y^[2]

¹Assistant Professor, ²Research scholar, Department of Computer Science, National College (Autonomous), National College (Autonomous), (Affiliated to Bharathidasan University) Trichy – 01¹Email Id:srinaganyag78@gmail.com ²Email Id:rsspooja.vino@gmail.com

¹Ph: 9791409349

²Ph: 9444009832

ABSTRACT

Artificial Intelligence (AI)-a rapidly growing and emerging technology has proved to have a wide range of applications in medicine and health care. Artificial intelligence has aided in the advancement of healthcare research. As a form of artificial intelligence, artificial neural networks (ANNs) have the advantages of adaptability, parallel processing capabilities, and non-linear processing. This technology serves the field of medicine and its updated features increases the advantages over the discipline of medicine. Especially in detecting many precarious and malignancy diseases, and cancer is one among an incurable perilous, a life threatening disease. A cancer is a group of cells that have grown out of control in the body. They can spread rapidly to any part of the body. The most common types of cancer are breast cancer, lung cancer, skincancer and blood cancers (leukaemia and lymphoma). According to the report from World Health Organization (WHO), there are millions and millions of people suffering from the different type of cancer disease across the world, no matter what the age group is, and the survival rate is very low. Detecting cancerous cells and providing treatment at an early stage is quite difficult. But by the era of artificial intelligent and its techniques has given a new hope and scope among the physicians and healthcare decision-makers to decrease the risk rate and by the early stage they can handle the malignancy and disease up to the reduced cost. Neural networks (NN) are currently a burning research area in medical science, especially in the areas of cardiology, radiology, oncology, urology and etc. By using visual, biological, electronic health records, datasets and scan report data as the sole input source, pre-trained neural networks techniques and methods have been heavily employed for the identification of various malignancies. In this paper, we are surveying how the artificial neural network technologies are used for identifying and detecting classification of different types of cancer. The proposed work clarifies the importance of ANN and its role in predicting various cancer diseases and in terms of cost effective and user friendly system for processes and approaches for medical diagnostic.

Keywords: Artificial intelligence, Artificial neural network, Neural network, Prediction.

INTRODUCTION

The word cancer comes from the ancient Greek kapkivoc, which means crab and tumor. In Greek, these words refer to a crab, most likely applied to the disease because the finger-like spreading projections from a cancer called to mind the shape of a crab. The Roman physician, Celsus (25 BC - 50 AD), later translated the Greek term into cancer, the Latin word for crab. Cancer was introduced to the medical world in the 1600s and is associated with abnormally growing cells that can invade or spread to other parts of the body. The uncontrolled growth of cells starts from a site in the human body and further spreads to other body parts known as cancer metastasis. Cancer cells are categorized into benign and malignant cells. The benign cells do not spread to other parts, while malignant cells metastasize and are considered more destructive. Due to high mortality and recurrence rate, its process of treatment is very long and costly.

There is a need to accurately diagnose it early to enhance cancer patient's survival rate. It is a genetic disease triggered due to genetic mutations that control our cell's function, especially how they grow and divide. As the tumor cells continue to grow, additional changes will occur. In a nutshell, cancer cells have more genetic changes, such as mutations in DNA, than normal cells. Though the immune system generally discards damaged or abnormal cells from the body, few cancer cells can hide from the immune system. The tumor also uses the immune system to grow and stay alive. The name of the cancer type is based on the site where tumor cells grow, for example, cancer that arises in the lungs and spreads to the liver is called lung

cancer. Cancer diagnosis includes three predictive predictions related to cancer risk assessment, cancer recurrence, and cancer survivability prediction. Initially, the probability of cancer occurrence is assessed, followed by the second step, predicting cancer recurrence. The last step is to predict the aspects like progression, life expectancy, tumor drug sensitivity, survivability. Cancer detection has always been a challenge in the diagnosis and treatment plan for haematological diseases. Currently, an increased percentage of the population is affected by one or more diseases. Recent years have seen tremendous

advances in medical science. Despite these advancements, there is still a huge lack of information among the public regarding health and disease. A large proportion of the population likely suffer from health issues, some of which may even be fatal. In addition to improving the accuracy of the rapid detection of fatal conditions, adopting safe, realistic techniques and using modern technology can reduce the need for caregivers and reduce overall health care costs. Several survival dramatically increase if detected early.

In recent years, Computer technology has made significant strides, leading to a surge of interest in the application of 'Artificial Intelligence (AI) in the fields of medicine and biological research. Among the various branches of AI, one that stands out for its extensive research and potential is 'Artificial Neural Networks (ANNs). Medical diagnosis using artificial intelligence (AI) systems, particularly artificial neural networks and computer-aided diagnosis with deep learning, is currently a very active research area in medicine and it is believed that it will be more widely used in biomedical systems. Evolving neural network techniques for medical diagnosis are broadly considered since they are ideal in recognizing diseases using scans. Neural networks learn by example so the details of how to recognize the disease are not needed. For instance, the utilization of ANN models aids in the timely diagnosis of various cancer with sensitivity and specificity. Advances in deep learning-based ANN models achieve efficacy, accuracy, and reliability in diagnosis. The utilization of other evolving technologies in this field is key for better diagnosis, having a significant impact on preventive measures and treatment. Despite the great benefits of novel technologies in healthcare, patients need protection from defective diagnoses to create a promising future in medical applications within society. Essentially, artificial neural networks (ANNs) are computer-generated mathematical algorithms that learn from standard data and extract the information embedded within. Through training, ANNs can mimic the behaviour of small biological neural clusters in a fundamental way. They serve as digital models of the human brain and have the ability to identify intricate nonlinear connections between dependent and independent variables, even in cases where the human brain may struggle to do so. Currently, ANNs are broadly utilized for medical applications in diverse disciplines of medicine mainly in cardiology. ANNs have been extensively applied in diagnosis, digital sign examination, medical picture examination and radiology.

ANNs have been utilized by a few of writers for modelling in medicine and clinical study. Applications of ANNs are flourishing among medical information mining.

Different Types of Cancer and Its Prediction Methods by ANN

Yufang Yana et al.(2021) researched on cervical cancer patients and presented a two- stage 99+clinical events prediction model, in the first stage, k- means methods was used for mining the standardized diagnosis and treatment mode to judge the clinical event sequence to be tested is standard nor not, the RNN-2-DT, discrete time-variant based on word vector representation was integrated and in the second stage gated recurrent units (GRU) based on recurrent neural network (RNN) is applied into neural network to construct a prediction model. The treatment day vectors was used to mine standardized diagnosis and treatment mode while they train clinical event prediction model through treatment day vector stitched with time interval of the adjacent treatment day. The experiment results indicated that RNN-2- DT has a brilliant prediction effect compared with traditional models, and the mean average precision are increased by 7.2% and 4.3%, respectively.

Guozhen Chen et al. (2020) presented the research work on chronic kidney disease, a kidney cancer, presented the Adaptive Hybridized Deep Convolutional Neural Network (AHDCNN) for the early prediction and diagnosis of Chronic Kidney Disease (CKD). The distinctive subtypes of lesion are identified by using Deep learning system. The collected data will initially be analysed and the missing value will be replaced by the median value estimate. To identify kidney patterns, different features associated with the kidney disease are

determined from the noise free data and fed in the classifier. By measuring the weight and bias value, the system trains feature in each hidden layer. Classification technology efficiency depends on the role of the data set. To enhance the accuracy of the classification system by reducing the feature dimension an algorithm model has been developed using CNN. The experimental process on the Internet of medical things platform (IoMT) concludes, with the aid of predictive analytics that advances in machine learning which provides a promising framework for the recognition of intelligent solutions to prove their predictive capability beyond the field of kidney disease. The competent use of the learning and activation mechanism is a method of double-training to avoid kidney disease effectively. The study of regression and distribution of the data are then determined. The proposed approach is based on the method

for deeper learning and ROIs (region of interest) given by radiologists has shown promising results in the classification of renal cell subtypes.

Rabbia Mahum et al. (2023) presented their work on Lung cancer, one of the terrible diseases in various countries around the globe, and timely detection of the illness is still a challenging process. This study proposed a novel and efficient lung tumor detector based on a RetinaNet, namely Lung-RetinaNet. In this model CT scans was utilized for the training and testing of the model. A multi-scale feature fusion-based module is introduced to aggregate various network layers, simultaneously increasing the semantic information from the shallow prediction layer. This proposed model comprises a ResNet as a backbone network for extracting features from input along with dilated contextual block-based fusion, categorization and regression sub-networks. Due to an improved structure of RetinaNet, it precisely detects tiny lung tumours. The proposed methodology attained 99.8% accuracy, 99.3% recall, 99.4% precision, 99.5% F1- score, and 0.989 Auc than the DL-based methods by this outcomes the proposed system can be utilized by medical experts to identify the tumor at early stages.

Wahidur Rahman et al. (2023) proposed the work on multiclass blood cancer like Breast cancer, lung cancer, skin cancer, and blood malignancies such as leukemia and lymphoma. The proposed research pipeline is occupied into some interconnected parts like dataset building, feature extraction with pre-trained Convolutional Neural Network (CNN) architectures from each individual image of blood cells, and classification with the conventional classifiers. The dataset for this study is divided into two identical categories, Benign and Malignant, and then reshaped into four significant classes, each with three subtypes of malignant, namely, Benign, Early Pre-B, Pre-B, and Pro-B. The research first extracts the features from the individual images with CNN models and then transfers the extracted features to the features selections such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and SVC Feature Selectors along with two nature inspired algorithms like Particle Swarm Optimization (PSO) and Cat Swarm Optimization (CSO). After that, research has applied the seven Machine Learning classifiers to accomplish the multi-class malignant classification. The study discovered a maximum accuracy of 98.43% when solely using pre-trained CNN and classifiers and after incorporating PSO and CSO, the proposed model achieved the highest accuracy of 99.84% by integrating the ResNet50 CNN architecture, SVC feature selector, and LR classifiers.

Haitham Elwahsh et al. (2023), this work was investigated on predicting cancer using deep neural learning.

The study proposed that deep neural learning cancer prediction model (DNLC) has the following stages. In the first stage, Deep Network (DN) is used to select the best collection of features from datasets. In the second stage, train the genomic or clinical data samples with a deep neural network (DNN). In the third stage, evaluate the capabilities of the DNLC model of predicting cancer in its earlier stages. For classification, DNLC uses five cancer datasets, which are for colon, lung adenocarcinoma, and squamous cell carcinoma, breast, and leukaemia cancers. The five cancer datasets are used in experiments to predict how well the suggested model will perform. The dataset is divided into two parts: training sets, which make up 80% of the dataset, and testing sets, which make up 20%. The experimental results shows that the DNLC technique, with an average accuracy of 93%, performs better in terms of accuracy than other methods in all circumstances.

Dewi Nasien et al. (2022), explored the use of ANN in the research on breast cancer and the research has been widely conducted and previously studied with various methods or algorithms to categorize it into benign and malignant groups. In ANN algorithm, one method called back propagation network is utilized to solve complex problems related to identification, pattern recognition prediction, and so forth. The objective of the present study is to investigate the level of accuracy and performance by ANN back propagation in predicting breast cancer. Several stages for this study are formulating the problem, collecting and processing the Wisconsin breast cancer dataset from the Kaggle site. Designing and creating an ANN algorithm system to classify cancer into malignant and benign, then examining the system to perceive the prediction accuracy, and conclude it. The results of the numerical simulation indicate that the created system of MATLAB R2016a software obtained an accuracy of 96.929% with an error of 3.071% by a combination of training parameters with epoch 1000, learning rate 0.01, goal 0.001, and hidden layer 5.

Sugiart et al. (2018), worked on skin cancer and proposed that this study aims to present diagnose of melanoma skin cancer at an early stage. It applies feature extraction method of the first order for feature extraction based on texture in order to get high degree of accuracy with method of classification using artificial neural network (ANN). The method used is training and testing phases with classification of Multilayer Perceptron (MLP) neural network. The results showed that the accuracy of test image with 4 sets of training for image not suspected of melanoma and melanoma with the lowest accuracy of 80% and the highest accuracy of 88,88%, respectively. The 4 sets of training used consisted of 23 images. Of the 23 images used as a training consisted of 6 as not suspected of melanoma images and 17 as suspected melanoma images.

Mridha k et al. (2023) presented that skin cancer is a prevalent form of malignancy around the globe. Clinical evaluation of skin lesions is essential, but it faces challenges such as long waiting times and subjective interpretations. Deep learning techniques have been developed to tackle these challenges and assist dermatologists in making more accurate diagnoses. Prompt treatment of skin cancer is vital to prevent its progression and potentially life-threatening consequences. The use of deep learning algorithms can improve the speed and accuracy of diagnosis, leading to earlier detection and treatment. The goal of this study was to develop reliable deep learning (DL) prediction models for skin cancer classification; (i) deal with a typical severe class imbalance problem, which arises because the skin-affected patients' class is significantly smaller than the healthy class; and (ii) interpret the model output to better understand the decision-making mechanism (iii) Propose an End-to-End smart healthcare system through an android application. In a comparison examination with six well-known classifiers, the effectiveness of the proposed DL technique was explored in terms of metrics relating to both generalization capability and classification accuracy. A study used the HAM10000 dataset and an optimized CNN to identify the seven forms of skin cancer. The model was trained using two optimization functions (Adam and RMSprop) and three activation functions (Relu, Swish, and Tanh). Furthermore, an XAI-based skin lesion classification system was developed, incorporating Grad-CAM and Grad-CAM++ to explain the model's decisions. This system can help doctors make informed skin cancer diagnoses in their early stages, with an 82% classification accuracy and 0.47% loss accuracy.

M. Vania et al. (2023) proposed the work on gastrointestinal cancer, technical techniques is required to appreciate the scientific quality and novelty of AI studies for this application. Clinicians frequently lack this technical background, and AI experts may be unaware of such clinical relevance and implications in daily practice. As a result, there is a growing need for a multidisciplinary, international assessment of how to conduct high-quality AI research in upper GI malignancy detection. This research will help gastroenterologist build approaches or models to increase diagnosis accuracy for upper GI (Gastrointestinal) malignancies despite variances in experience, education, personnel, and resources, as it offers real-time and surveying chances to improve upper GI malignancy diagnosis and screening. The comprehensive review sheds light on potential enhancements to computer-aided diagnostic (CAD) systems for GI endoscopy. The findings of study suggest that Support Vector Machines (SVM) is frequently utilized in gastrointestinal (GI) image processing within the context of machine learning (ML). The analysis reveals that CNN-based supervised learning object detection models are widely employed in GI image analysis within the deep learning (DL) context. The results

of this study also suggest that RGB is the most commonly used image modality for GI analysis, with color playing a vital role in detecting bleeding locations.

Zhencun Jiang et al. (2021) proposed a ViT- CNN ensemble model to classify cancer cells and normal cells for the diagnosis of Acute Lymphoblastic Leukemia. This ensemble model combined the vision transformer model and convolutional neural network (CNN) model to extract features from cells images in two different ways, also used a data enhancement method and a symmetric cross-entropy loss function to reduce the impact of noise in the data set. The ViT-CNN ensemble model achieved a high classification accuracy of 99.03% on the test set, which was better than other models. The authors suggested that their method could be an effective tool for computer- aided diagnosis of ALL by accurately distinguishing between cancer cells and normal cells.

M. Akter Hossain et al. (2020) conducted a study focused on Acute Lymphocytic Leukemia (ALL), which is the most common form of leukemia. The study proposed a practical technique for detecting abnormal blood components in cancer patients, including neutrophils, eosinophils, basophils, lymphocytes, and monocytes. The researchers used 14 features to construct a dataset before selecting four key attributes that are crucial in determining whether a patient has Leukemia, also collected 256 primary data from Leukemia patient. The data is then processed using microscope to obtain images and fetch into Faster-RCNN machine learning algorithm to predict the odds of cancer cells forming, it was also applied two loss functions to both the RPN (Region Convolutional Neural Network) model and the classifier model to detect the similar blood object. After identifying the object, calculated the corresponding object and based on the count of the corresponding object, finally Leukemia was detected. The mean average precision observed are 0.10, 0.16 and 0, where the epochs are 40, 60 and 120. The goal of the study was to aid in the early detection of cancer, which can significantly improve treatment outcomes.

CONCLUSION AND FUTURE SCOPE

Artificial neural networks extract patterns and make predictions from large datasets. The increasing usage of neural network in healthcare, together with the availability of highly characterised cancer datasets, has led to acceleration in research into the utility of deep learning in the analysis of the complex biology of cancer. ANN techniques provide different probabilistic and statistical approaches that enable intelligent computers to recognise and identify patterns in datasets based on repeated prior experiences. This research paper highlights the evolving ANN technology in cancer diseases by applying algorithms to cancer datasets. The study demonstrated the effectiveness of techniques in predicting and detecting different type of cancers with high accuracy rates. In terms of future scope, further work can enhance accuracy rates by using more comprehensive datasets, optimizing the platform where the system will be launched, and making it more user-friendly with extensive information about the disease.

REFERENCES

- [1] Yufang Yan, Kui Zhao, Jilong Cao, Huimin Ma, "Prediction research of cervical cancer clinical events based on recurrent neural network", *Procedia Computer Science*, Volume 183, 2021, Pages 221- 229, ISSN 1877-509, <https://doi.org/10.1016/j.procs.2021.02.052>.
- [2] Guozhen Chen , Chenguang Ding , Yang Li, Xiaojun Hu , Xiao Li , Li Ren , Xiaoming Ding , Puxun Tian , And Wujun, "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," in *IEEE Access*, vol. 8, pp. 100497- 100508, 2020, doi:10.1109/ACCESS.2020.2995310.
- [3] R. Mahum and A. S. Al-Salman, "Lung- RetinaNet: Lung Cancer Detection Using a RetinaNet With Multi-Scale Feature Fusion and Context Module," in *IEEE Access*, vol. 11, pp. 53850-53861, 2023, doi:10.1109/ACCESS.2023.3281259.
- [4] Rahman, Md & Faruque, Mohammad & Roksana, Kaniz & Sadi, A.H.M. Saifullah & Rahman,

- Mohammad Motiur & Azad, Dr. (2023). Multiclass blood cancer classification using deep CNN with optimized features. *Array*. 18. 100292.10.1016/j.array.2023.100292.
- [5] Haitham Elwahsh, Medhat A. Tawfeek, A.A. Abd El-Aziz, Mahmood A. Mahmood, Maazen Alsabaan, Engy El-shafeiy, "A new approach for cancer prediction based on deep neural learning", *Journal of King Saud University -Computer and Information Sciences*, Volume 35, Issue 6, 2023, 101565, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2023.101565>.
- [6] Nasien, Dewi, Enjeslina, Veren, Hasnil Adiya, M., Baharum, Zirawani, "Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method" *Journal of Physics: Conference Series*, 2022, IOP Publishing, doi: 10.1088/1742-6596/2319/1/012025, <https://dx.doi.org/10.1088/1742-6596/2319/1/012025>.
- [7] Anagun, Y. Smart brain tumor diagnosis system utilizing deep convolutional neural networks. *Multimed Tools Appl* 82, 44527–44553 (2023) <https://doi.org/10.1007/s11042-023-15422-w>
- [8] Sugiarti, Sugiarti & Yuhandri, Yuhandri & Naam, Jufriadif & Indra, Dolly & Santony, Julius, (2019), "An artificial neural network approach for detecting skin cancer", *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 17. 788. 10.12928/telkomnika.v17i2.9547. doi:10.12928/telkomnika.v17i2.9547
- [9] K. Mridha, M. M. Uddin, J. Shin, S. Khadka and M. F. Mridha, "An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System," in *IEEE Access*, vol. 11, pp.41003-41018, 2023, doi: 10.1109/ACCESS.2023.3269694.
- [10] M. Vania, B. A. Tama, H. Maulahela and S. Lim, "Recent Advances in Applying Machine Learning and Deep Learning to Detect Upper Gastrointestinal Tract Lesions," in *IEEE Access*, vol. 11, pp. 66544- 66567, 2023, doi: 10.1109/ACCESS.2023.3290997.
- [11] Zhencun Jiang, Zhengxin Dong, Lingyang Wang, Wenping Jiang, "Method for Diagnosis of Acute Lymphoblastic Leukemia Based on ViT-CNN Ensemble Model", *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 7529893, 12 pages, 2021.

A SURVEY IN CLOUD COMPUTING USING ADAPTIVE ALGORITHMS

Gayathri.M¹

Research Scholar(Full Time)
Department of Computer Science
NationalCollege(Autonomous)
Trichy, Tamilnadu, India
gayathricascas@gmail.com

Dr. Srinaganya.G²

Department of Computer Science
National College(Autonomous) Trichy,
Tamilnadu, India srinaganyag1978@gmail.com

ABSTRACT

Cloud Computing is model for allowing convenient, on demand access anywhere to shared computing resources. Like Amazon, Google, and Microsoft.it denotes as a: cloud". The main concept is computing, storage, and software "as a service". An adaptive algorithm is a one type of algorithm. It's based on the feedback or new collection of information. They are many different types of adaptive algorithms.

Index Terms— Cloud Computing, Artificial Intelligence, Adaptive algorithms, machine Learning, neural networks and genetic algorithms.

INTRODUCTION

The cloud Computing is to appear or come out from somewhere. It has many names including: grid computing, utility computing, and on demand computing. The cloud stands for Common, Location, -independent, Online Utility that is on Demand"[3].

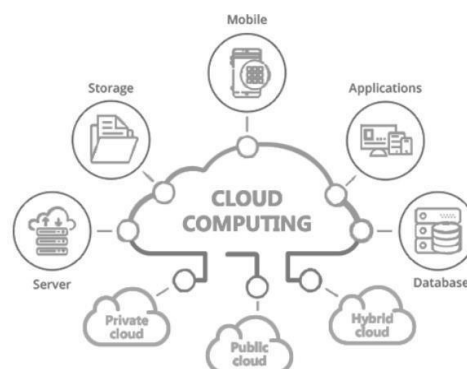


Fig.1. Cloud Computing

The main three types of cloud computing are public cloud, private cloud, and hybrid cloud. These models are four main services: Infrastructure as a service (IaaS), Platform as a service (PaaS), Software as a service (SaaS), and server less computing [1], [17].

IaaS - It Provides Virtualized dynamically scalable resources using a virtualized infrastructure. Like database services, queuing services, application frameworks etc. PaaS – It Simplifies application development by APIs for requesting, managing and monitoring cloud resources. SaaS – It provides multi-tenant applications hosted in the cloud. Such as Email, cloud storage application, etc.

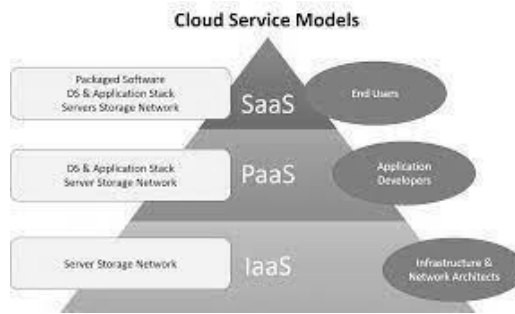


Fig.2. Cloud Services Models

An adaptive algorithm is a one type of algorithm. It's based on the feedback or new collection of information. That means the Algorithm can learn and improve its execution in the course of time [2].

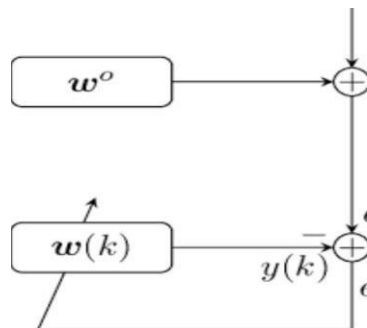


Fig.3. Block diagram of Adaptive Algorithm

It's totally different from a previous algorithm, it is static and does not change. There are many different types of adaptive algorithms used in artificial intelligence (AI), but there are some common terms are shared many of them.

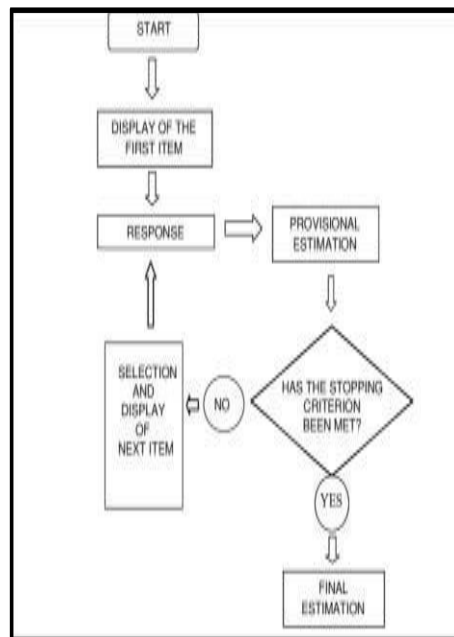


Fig.4. Adaptive Algorithm Flow Chat

Machine learning (ML) is an exploration of computer algorithms that is developed naturally through expertise. It is used to solve all the security problems and control the data more systematic.it is man-made recognition.ML focused on the evolution of computer programs.[4]

The genetic algorithm is solving a process of problem and its based on natural selection.it is used in computing to find the accurate solutions and search problems. It is easy to search the large problems.

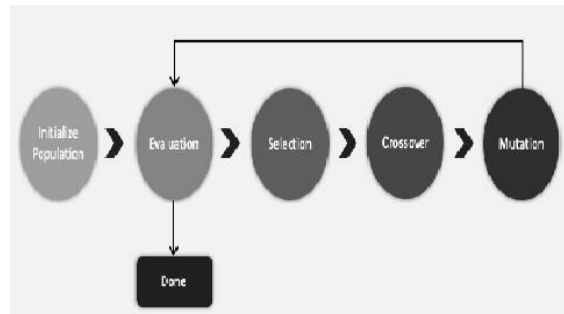


Fig.5. Genetic Algorithm

The paper is organized into two sections. Section-2 compares literature survey. Section-3 includes conclusion while references are shown in the last section.

LITERATURE SURVEY

Adaptive algorithm concept has recently been considered and formulated as a neural network and machine learning. Furthermore, it can be used in various situations where data is constant changing, like in the stock market forecast.

Adaptive algorithm that can accordingly adapt to new data or changes in the conditions. It is in contrasted to traditional algorithms, it designs a specific set of data and conditions .it cannot adapt easily.

Adaptive scheduling algorithm is shared all the server and protected the cloud data from overloading. It achieves high performance. Like processing time and response time.[18].

TABLE I. ANALYSIS OF ALGORITHMS.

S. No	Name of the Authors	Name of the Algorithm	Purpose of the Algorithm
1	Gyeongsik Yang, Kwanhoon Lee, Kyungwoon Lee , Yeonho Yoo, Hyowon Lee, And Chuck Yoo ,	Blockchain Consensus Algorithms	To find a Kernel-level analysis for the resource consumption and comprehensive performance evaluations
2	Njoud AlMansour, Nasro Min Allah	Scheduling Algorithms	To rectify the perspective of makespan, load balancing, CPU utilization, deadline, response time and allocation cost
3	Joohyung Sun And Hyeonjoong Cho	Optimal Scheduling Algorithms	To introduce a real time task scheduling algorithm

4	Pillareddy Vamsheedhar Reddy And Karri Ganesh Reddy	A Multi-Objective Based Scheduling	To performs dynamic workflow scheduling using universal unique identification-Blake, Manhattan Distance-Partition around algorithm, Linear Scaling-Crow Search Optimization, Anova-Recurrent Neural Network.
5	Mohammed s. Zalata, saad m. Darwish, and magda m. Madbouly	Niching Genetic Algorithm	A novel strategy for enhancing the multisite offloading mechanism by combining a Niching Genetic Algorithm with a Markov Decision Process
6	Kuang-Yen Tai , Frank Yeong- Sung Lin, And Chiu-Han Hsiao	An Integrated Optimization-Based Algorithm	An optimized computing resource and energy management algorithm for computing centers with heterogeneous computing resources from the perspective of Green IT
7	Faten a. Saif , Rohaya latip , Zurina mohd hanapi ,and Kamarudin Shafinah	Multi-Objective Grey Wolf Optimizer Algorithm	To reduce the QoS objectives delay and energy consumption and held in the fog broker, which plays an essential role in distributing tasks
8	Ya Zhou And Xiaobo Jiao	Knowledge-Driven Multi-Objective Evolutionary Scheduling Algorithm	To make a good trade-off between the makespan and monetary cost, these algorithms include two novel features The structural knowledge of workflow The knowledge on the Pareto front range
9	Dineshan subramoney and clement n. Nyirenda	Multi-Swarm PSO Algorithm	To develops a weighted sum objective function for the workflow scheduling problem, based on four objectives: Makespan Cost Energy Load balancing for cloud and fog tiers
10	Petra Loncar, And Paula Loncar	Evolution Strategies Algorithm	To proposes a strategy for optimizing task scheduling across different virtual machines and data centers based on the metaheuristic Evolution Strategies algorithm

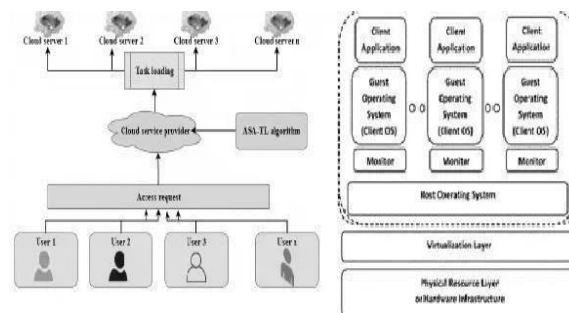


Fig.6. Adaptive Scheduling Algorithm

Gyeongsik Yang, Kwanhoon Lee [6], presents a kernel- level analysis for the resource consumption

and comprehensive performance evaluations of three major consensus algorithms (i.e., Kafka, Raft, and PBFT). Their experiments reveal that resource consumption differs up to seven times, which demonstrates the importance of proper resource provisioning for BaaS.

Njoud AlMansour [7], this algorithm is vital to ensure an appropriate resource is available to every request. The authors compare cloud computing scheduling from the perspective of makespan, load balancing, CPU utilization, deadline, response time, and allocation cost. In addition, they propose an abstract model to integrate desirable features of an algorithm suitable to a cloud environment.

Joohyung Sun [8], projected a real-time task scheduling algorithm for cloud computing servers, their proposed algorithm outperforms the latest existing algorithms in terms of both time complexity and energy efficiency.

Pillareddy Vamsheedhar Reddy [9], it performs dynamic workflow scheduling using universal unique identification- Blake (UUID-Blake), Manhattan Distance- Partition around algorithm (MD- PAM), Linear Scaling-Crow Search Optimization (LS-CSO), Anova-Recurrent Neural Network. The author proposed LS-SCO outperformed when compared with the algorithms CSO, PSO and RR has achieved better performance.

Mohammed s. Zalat [10], The author proposes a novel strategy for enhancing the multisite offloading mechanism by combining a Niching Genetic Algorithm (NGA) with a Markov Decision Process (MDP). Their proposed technique consumes little power and executes quickly while determining the optimal offloading decision with lower generation numbers.

Kuang-Yen Tai, Frank Yeong-Sung Lin [11], This paper proposes an optimized computing resource and energy management algorithm for computing centres with heterogeneous computing resources from the perspective of Green IT. And also, they proposed an optimized energy management algorithm to effectively manage computing resources and create cloud computing centres with high performance and low energy consumption.

Faten a. Saif, Rohaya latip [12], they introduced a Multi-Objectives Grey Wolf Optimizer (MGWO) algorithm to reduce the QoS objectives delay and energy consumption and held in the fogbroker, which plays an essential role in distributing tasks. And verifies the effectiveness of the MGWO algorithm compared to the state-of-the-art algorithms in reducing delay and Energy consumption.

Ya Zhou And Xiaobo Jiao [13], To make a good trade-off between the makespan and monetary cost, these algorithms include two novel features. 1) the structural knowledge of workflow is mined to simplify the large-scale decision variables into a series of small-scale components, 2) the knowledge on the Pareto front range is mined to estimate the ideal and nadir points for the objective space normalization during the search process, so the authors demonstrate the effectiveness of the KMEWSA in balancing makespan and monetary cost for deploying workflows into cloud computing.

Dineshan subramoney and clement n. Nyirenda [14], the authors develop a weighted sum objective function for the workflow scheduling problem, based on four objectives: makespan, cost, energy and load balancing. It will be used for scientific workflow scheduling. Hybridization of the realized algorithm with dynamic approaches will also be investigated.

Petra Loncar, And Paula Loncar [15], they propose a strategy for optimizing task scheduling across different virtual machines and data centers based on the metaheuristic Evolution Strategies algorithm. After testing, they achieved a better makespan, larger average resource utilization, better throughput, less average execution time, a smaller degree of imbalance, and scalability.

In the Adaptive algorithm, number of metrics could have increased and got better results. Nowadays

many authors developed new algorithms according to their metrics and need of information from cloud.

CONCLUSION

In this paper, the focused Adaptive algorithm from the Cloud. It is observed from the above survey Adaptive algorithm (AA) is one of the best Algorithm of Cloud. There is many defiance with adaptive algorithms in AI.1) it is difficult to design and used.2) it is more expensive, so it makes them impractical some applications. It can be sensitive to changes in the data, so difficult to use real world examples.so finally it can be difficult to explain, and difficult to understand why there used and decisions. A better Adaptive algorithm could lead this research work in future.

REFERENCES

- [1] Corresponding Author: Manu Mitra, Alumnus with Department of Electrical Engineering, University of Bridgeport, Bridgeport, United States. Editorial Received: November 12, 2019; Published: December 02, 2019.
- [2] https://www.researchgate.net/publication/297617415_Transient_analysis_of_10-LMS_and_10-NLMS_algorithms March2016DOI:10.1016/j.sigpro.2016.02.017.
- [3] "Overview Of Cloud Computing And Its Types" March 2019 [SSRN Electronic Journal](#) 6(3):61-67 Authors: [MS N Rajeswari](#) & [Sameer Mohammad](#).
- [4] Umer Ahmed Butt et al., "A Review of Machine Learning Algorithms for Cloud Computing Security", Electronics 2020, 9, 1379; doi:10.3390/electronics9091379, pp. 2-25.
- [5] Anshika Negi et al., "An Efficient Security Framework Design for Cloud Computing using Artificial Neural Networks", International Journal of Computer Applications (0975 – 8887) Volume 129 – No.4, November2015, pp.17-21.
- [6] GYEONGSIK YANG et al., "Resource Analysis of Blockchain Consensus Algorithms in Hyperledger Fabric", IEEE Access, Vol 10, 2022, pp. 74902 - 74920
- [7] Njoud AlMansour et al., "A Survey of Scheduling Algorithms in Cloud Computing", IEEE Access, 978-1-5386-8125-1/19/\$31.00 ©2019 IEEE.
- [8] JOOHYUNG SUN et al., "A Lightweight Optimal Scheduling Algorithm for Energy- Efficient and Real-Time Cloud Services", IEEE Access, Vol 10, 2022, pp. 5697-5714.
- [9] PILLAREDDY VAMSHEEDHAR REDDY et al., "A Multi- Objective Based Scheduling Framework for Effective Resource Utilization in Cloud Computing", IEEE Access, Vol 11, 2023. pp. 37178-37193.
- [10] MOHAMMED S. ZALAT et al., "An Adaptive Offloading Mechanism for Mobile Cloud Computing: A Niching Genetic Algorithm Perspective", IEEE Access, Vol 10, 2022, pp. 76752-76765.
- [11] KUANG-YEN TAI et al., " An Integrated Optimization-Based Algorithm for Energy Efficiency and Resource Allocation in Heterogeneous Cloud Computing Centers", IEEE Access, Vol 11, 2023, pp. 53418-53428.

- [12] FATEN A. SAIF et al., "Multi-Objective Grey Wolf Optimizer Algorithm for Task Scheduling in Cloud-Fog Computing", IEEE Access, Vol 11, 2023, pp. 20635-20646.
- [13] YA ZHOU et al., "Knowledge-Driven Multi-Objective Evolutionary Scheduling Algorithm for Cloud Workflows", IEEE Access, Vol 10, 2022, pp. 2952-2962.
- [14] DINESHAN SUBRAMONEY et al., "Multi-Swarm PSO Algorithm for Static Workflow Scheduling in Cloud-Fog Environments", IEEE Access, Vol 10, 2022, pp. 117199-117214.
- [15] PETRA LONCAR et al., "Scalable Management of Heterogeneous Cloud Resources Based on Evolution Strategies Algorithm", IEEE Access, Vol 10, 2022, pp. 68778-68791.
- [16] Panel Yongkang Wang et al., "An adaptive robust defending algorithm against backdoor attacks in federated learning", <https://doi.org/10.1016/j.future.2023.01.026>.
- [17] P.S.V. Sainadh et al, "Security Issues in Cloud Computing", Volume: 03, Special Issue No: 01, February 2017, pp.125-130.
- [18] Dibyendu Mukherjee et al., "Adaptive Scheduling Algorithm based Task Loading in Cloud Data Centers", IEEE Access, Vol xx, 2022, pp.1 -10.

GENERATIVE AI IN VIRTUAL REALITY

Dr.S.Saranya¹,Sivaprabha C², Subhiksha³

¹Head & Assistant professor^{2,3}Student, Department of Artificial Intelligence and Machine Learning, Dr. N.G.P. Arts and Science College,Coimbatore, Tamil Nadu, India saranyas@drngpasc.ac.in, 211ai024@drngpasc.ac.in¹, 211ai024@drngpasc.ac.in²

Abstract—Generative AI in virtual reality involves using algorithms to create immersive and dynamic virtual environments . This technology leverages machine learning models, such as generative adversarial networks(GANs) to generate realistic and interactive content with VR spaces. By combining AI's ability to create novel content with the immersive nature of VR , users can experience unique and lifelike virtual worlds that adapt and respond to their interactions , enhancing the overall realism an engagement of the virtual experience. It provides context on the current state of both generative AI and VR technology. Provides examples of practical applications including training stimulations educational settings and entertainment. Includes relevant case studies demonstrating successful implementations of Generative AI in VR. Discuss the adaptation of VR. Highlight the role of AI in crafting dynamic and interactive narratives within virtual spaces. Explores how generative model enhance visual and sensory realism in VR simulations. It contributes more to creative environments.

Keywords—Generative AI, Generative Adversarial Networks, VR Simulation

INTRODUCTION

Generative Artificial Intelligence (AI) in virtual reality (VR) represents a convergence of cutting-edge technologies that holds immense potential for transforming immersive experiences. At its core, generative AI refers to a class of algorithms capable of creating new, original content autonomously. When integrated into VR environments, these AI systems enable the dynamic generation of virtual content, ranging from life like environments to interactive characters and procedural landscapes.

The marriage of generative AI and VR unlocks a myriad of possibilities across various domains:

Generative AI algorithms such as GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders) empower VR developers to generate realistic and diverse virtual environments. These environments can simulate real-world settings or fantastical realms, enriching the user experience.By leveraging techniques like procedural content generation (PCG), VR content creators can dynamically generate landscapes, structures, and entire worlds on-the-fly. This approach ensures endless variations and enhances replayability in VR experiences.Generative AI models enable the creation of lifelike characters and non- playable characters (NPCs) with unique behaviors and personalities. RNNs (Recurrent Neural Networks) and LSTM (Long Short-Term Memory) networks are employed to generate interactive dialogue, responses, and narratives, fostering engaging interactions within VR environments.

Neural style transfer algorithms and creative AI tools like DeepDream empower artists and designers to imbue VR content with unique art styles and surreal aesthetics. This fusion of AI-driven creativity with immersive VR experiences results in visually stunning and thought-provoking virtual worlds.

Generative AI in VR enables adaptive content creation, where virtual environments dynamically respond to user input, preferences, and real-time data. This adaptability enhances user immersion and personalization, making VR experiences more compelling and responsive.

In essence, the integration of generative AI technologies into virtual reality heralds a new era of creativity, interactivity, and immersion. As these technologies continue to evolve, the boundaries of what's possible in VR will expand, offering users unprecedented levels of immersion and engagement.

GENERATIVE AI TOOLS IN VIRTUAL REALITY

Generative AI tools can enhance virtual reality experiences in various ways. Some notable tools include:

1. GANs (Generative Adversarial Networks): GANs can create realistic images, textures, and environments in virtual reality, improving the visual fidelity of VR content.
2. Variational Autoencoders (VAEs): VAEs can be used for generating novel and diverse content in VR, such as generating unique 3D models or textures for virtual environments.
3. Recurrent Neural Networks (RNNs) and LSTM: These models are used for generating sequences, such as generating procedural content or narratives in VR experiences.
4. Pix2Pix and CycleGAN: These models are particularly useful for image-to-image translation tasks in VR, such as converting sketches into photorealistic images or adapting art styles in virtual environments.
5. Neural Style Transfer: This technique can be applied to transfer artistic styles onto VR content, enhancing the aesthetics of virtual environments.
6. Procedural Content Generation (PCG): PCG algorithms can generate virtual worlds, landscapes, and levels dynamically, offering endless variations and replayability in VR experiences.
7. Unity ML-Agents Toolkit: Developed by Unity Technologies, this toolkit allows for the integration of machine learning models into Unity-based VR applications, enabling interactions with AI-generated content.
8. DeepDream: While primarily used for artistic purposes, DeepDream can be employed to create surreal and dream-like VR experiences by altering and enhancing existing virtual content.

By integrating these generative AI tools into virtual reality applications, developers can create immersive and engaging experiences with enhanced visual quality, procedural content, and dynamic interactions.

HOW GENERATIVE AI WORKS IN VR

Generative AI works by analyzing vast amounts of data and identifying underlying patterns. These patterns are then used as a foundation for generating new content. The algorithms used in generative AI are designed to learn and adapt, constantly refining their ability to create original and meaningful outputs.

Related: [The Anatomy of a VR Training Module – Explaining the Key Characteristics of VR Training Experiences](#)

In the context of virtual reality, generative AI can be applied in various ways. For example, it can be used to generate realistic and dynamic virtual characters that respond intelligently to user interactions. This means that virtual reality simulations can become more lifelike and engaging, blurring the lines between the real and virtual worlds.

Furthermore, generative AI can also be used to create interactive virtual environments that adapt and evolve based on user input. This means that virtual reality experiences can become more immersive and responsive, providing users with a sense of agency and control.

With generative AI, the possibilities for virtual reality are endless. From creating unique and personalized virtual worlds to enhancing the realism and interactivity of virtual experiences, generative AI has the potential to revolutionize the way we engage with virtual reality.

Now that we have a basic understanding of generative AI and its interaction with virtual reality, let's explore the exciting applications and future prospects of this groundbreaking technology.

ADVANTAGES

Generative AI can create diverse, enriching content that is unique to each user's preferences and behavior, thus intensifying their VR experiences. It brings an element of surprise and curiosity, making the virtual environment more engaging and memorable.

1. **Dynamic Content Generation:** Generative AI algorithms enable the dynamic creation of virtual environments, characters, and objects in real-time. This capability enhances immersion by offering endless variations and unique experiences with each interaction.
2. **Personalization:** AI-driven generative models can adapt VR content based on user preferences, behavior, and input. This personalization enhances user engagement by tailoring the experience to individual preferences, resulting in more immersive and satisfying interactions.
3. **Efficiency and Scalability:** Generative AI allows for the automated creation of VR content, reducing the need for manual content creation. This not only saves time and resources but also enables the scalability of VR experiences, as content can be generated on-demand to meet evolving user demands.
4. **Procedural Generation:** Techniques like procedural content generation (PCG) powered by generative AI enable the creation of vast and diverse virtual worlds with minimal manual intervention. This approach ensures that VR experiences remain fresh, engaging, and replayable, without the need for extensive pre-designed content.
5. **Artistic Exploration:** Generative AI tools facilitate artistic exploration in VR by empowering creators to experiment with novel art styles, visual effects, and immersive environments. This fosters creativity and innovation, leading to visually stunning and captivating VR experiences.
6. **Realistic Simulation:** Generative AI algorithms, such as GANs, VAEs, and physics-based models, can simulate realistic environments, physics interactions, and lifelike characters in VR. This realism enhances immersion and suspension of disbelief, making VR experiences more compelling and convincing.
7. **Adaptive Experiences:** Generative AI enables VR experiences to adapt and respond to user input, environmental factors, and real-time data. This adaptability enhances the sense of presence and interactivity, creating more engaging and responsive VR environments.

Overall, the integration of generative AI in virtual reality offers a wide range of advantages, including dynamic content generation, personalization, efficiency, scalability, artistic exploration, realistic simulation, and adaptive experiences, ultimately enhancing immersion and engagement for users.

INVENTIONS

It is that time of the year again. Time Magazine has put together the best 200 inventions for the year 2023. The announcement includes ingenious inventions from across domains such as accessibility, Artificial Intelligence, apps and software, AR/VR, beauty, consumer electronics, etc. The list features an assortment of gadgets and software from established technology brands as well as barely known startups.

AI tools have been making waves across the world since the launch of OpenAI's ChatGPT. Ever since then, there has been a barrage of AI tools and applications. Time Magazine's list included some of the best AI apps and tools that caught our attention. Here they are:

GPT-4 from OpenAI

The coveted magazine has described GPT-4 as the most powerful AI model that is accessible to the public. GPT-4 excels at verbal reasoning and can explain complex concepts in simple language. In September, the makers rolled out the ability to let users interact with the model using voice and images as inputs. Later, in a subsequent update, GPT-4V verbally describes the contents of a picture in natural language. GPT-4 is currently accessible only to ChatGPT Plus users, however, there are numerous ways to use GPT-4 for free with tools like Bing, HuggingFace, Poe, etc.

CONCLUSION

In conclusion, the integration of Generative Artificial Intelligence (AI) into Augmented Reality (AR) and Virtual Reality (VR) is propelling these immersive technologies into the forefront of the digital era.

REFERENCES

- [1] <https://indianexpress.com>

SUSPECT PREDICTION CRIME ANALYSIS INVESTIGATION TRACKER THE CRIME

*Dr. C. Premila Rosy¹,

Head & Assistant Professor,

Department of Computer Science,

Idhaya College for Women, Kumbakonam,

Tamil Nadu, India.

Corresponding Author *Dr. C. Premila Rosy – premlarosy78@gmail.com

K. Narmadha²,

II – M.Sc., (CS),

Department of Computer Science,

Idhaya College for Women, Kumbakonam,

Tamil Nadu, India.

ABSTRACT

A system that maintains the status of criminal investigations with logs and forecasts key suspects is known as a criminal investigation tracker system. When crimes are committed regularly in a society, they will eventually have an impact on organizations and institutions. The system keeps track of a case's summary, parties involved, legal conflicts, prior criminal histories of those involved, items found at the scene, and other information. The system recognizes the type of case, enables admin to update the investigation's progress, upload further photos of the crime scene and other items discovered there, etc. Authorized officials are now able to search up case information online, check on its status, and update any important information as and when necessary. A suspect prediction algorithm is another component of the system. This paper's major goal is to categories clustered crimes according to how frequently they occur over time. In order to analyses, investigate, and find patterns for the incidence of various crimes, data mining is often employed. Clustering is the division of a set of data or objects into several clusters. Thus, a cluster is made up of a set of related data that behaves as a unit. They determined that the procedure may be made simpler by using crime mapping analysis based on KNN (K - Nearest Neighbor) and ANN (Artificial Neural Network) algorithms. The Office of Community Oriented Policing Services (COPS) manages and funds crime mapping. Research that is based on evidence aids in the analysis of crimes. Data mining techniques are used to calculate the crime rate based on the historical data.

Keyword:Criminal Investigation, (K – Nearest Neighbor) and ANN (Artificial Neural Network).

I. INTRODUCTION

Crime is a terrible and unlawful conduct that violates the law and is grounds for punishment by the police and the government. Anyone who has committed or is associated with any type of crime is considered to be a criminal. Criminal activity is a social annoyance that costs our society in many different ways. The crime rate in our culture is rising extremely quickly, and women in particular are dealing with a lot of these issues. Criminal justice and law enforcement experts have traditionally had the responsibility of solving crimes. Computer data analysts have begun assisting law enforcement agents and detectives to speed up the process of solving crimes as a result of the increased usage of computerized systems to track crimes and trace perpetrators. Criminology is a procedure used to pinpoint criminal behavior and characteristics. With the aid of criminology methodologies, the criminals and the likelihood that a crime will occur can be evaluated. The police department, detective agencies, and crime branches use criminology to help them determine the genuine traits of criminals. The societal burden of crime has a high price on our society. The Indian government has also started working on creating software and applications for use by the Central and State Police in connection with the National Crime Records Bureau (NCRB). The National Crime Records Bureau (NCRB) and the Indian government are collaborating to develop software and applications that will be used by the Central and State Police. The National Crime Records Bureau (NCRB) and the Indian government are collaborating to develop software and applications that will be used by the Central and State Police. The National Crime Records Bureau (NCRB) and the Indian government are collaborating to develop software and applications that will be used by the Central and State Police. Data mining

was created as a powerful technique to draw out relevant information from sizable datasets and determine the connections between the data's features. Data mining began as an interdisciplinary field that combined statistics and machine learning. Many studies claim that

investigating crimes is a challenging, time-consuming activity that requires human intellect and expertise. Data mining is one technology that can aid us in solving challenges related to crime detection. Identify suspects in unresolved crimes based on the criminal records that are stored in the system's database. Here, the administrator first adds the officer to the system before adding them to a specific case that they will each be looking at separately. Once the information linked to the criminal is complete, the second phase will start. As officers are added to a certain case, the officers will put the details of the criminal in their database. The second phase primarily focuses on directing the project towards its end aim. The police register the FIR and include all the victim and criminal details after the criminal data is added to the database. Here, we focus on applying the decision tree method to forecast the crime and how it would have occurred. Their development, decline, and other outcomes related to the crime scenario. The prominent suspects are projected and indicated in a logical sequence based on the case type, belongings, land properties, relationships, and other similar characteristics related to the previous crime logs involved.

II. CRIME ANALYSIS

Usually, conducting a crime analysis can be a time-consuming process for the police or the investigating team. When the culprits leave the crime site, they do leave behind some evidence that can be utilized to help identify them. It is simple to analyse a crime because of the crime sequence and the common patterns that many criminals use. This approach entails a number of steps that must be taken in order to identify the culprits and get more information using only the hints or information provided by the locals. The criminal can be evaluated using data from the crime scene that is compared to earlier crime trends, as well as using a procedure that is intended to test and move forward with data that can influence prediction outcomes. By stationing extra police in the system-identified sensitive regions, the prediction can also be used to prevent crimes from happening in the first place. When there is a likelihood of criminality in the future, the police stations may deploy special force. With this kind of structure, the citizens' prosperity and tranquilly will be guaranteed.

III. LITERATURE REVIEW

The zones with a high likelihood of crime occurring were predicted and visualised by ShijuSathyadevan, Devan M. S., et al. (2014) [1]. The Naive Bayes classifiers algorithm, which is a supervised learning and statistical method for classification and has provided 90% accuracy, was used by the authors to categorise the data. (2016) [2] examines the viability of forecasting user sensitivity to deception-based assaults and forecasts the frequency of crimes caused by semantic social engineering attacks. With accuracy rates of 0.68 and 0.71, respectively, the authors have forecasted using logistic regression and a random forest prediction model. To cluster criminal activity in Tamil Nadu, S. Shivaranjani, S. Sivakumari, et al. (2016) [3] used a variety of clustering techniques, including K-Means clustering, agglomerative clustering, and density-based spatial clustering with noise (DBSCAN) algorithms. Using measures like precision, recall, and F-measure, the effectiveness of each clustering technique is assessed, and the results are compared. In comparison to the other two chosen algorithms, the DBSCAN algorithm produced the best results based on the aforementioned metrics.

IV. CRIME ANALYSIS PROCEDURE

Data mining technologies are used to reduce and prevent crimes and crime disorders, and crime mapping aids in comprehending the theory and practice of crime analysis. They may utilize KDD (Knowledge Discovery in Databases) and ANN (Artificial Neural Networks)-based data mining methods. Along with the basic data, they receive tones of extra information. However, in order to lessen or prevent confusion, extraneous data must be identified and removed before the data is processed using data mining techniques and tools. In order to spot any patterns in the crime data. Here, the data are divided into two categories: supervised data and unsupervised data. They take the data that has all the details about the case and try to solve the other cases by training using this

supervised data. A mainly collect the attributes information, like eye color, fingerprint details, characteristics, dimensions, or other features. We believe that crime data mining has a bright future for improving the effectiveness and efficiency of criminal and intelligence analysis based on the encouraging outcomes. Techniques for criminal and intelligence research that are visual and intuitive can be created for crime patterns. As they have used the clustering data mining technique for crime analysis, they can also use classification and other data mining techniques. Their primary goals were to create crime-based events, research the use of crime-based events in improving classification and clustering, develop an interactive crimenews retrieval system, effectively and interactively Crime news is visualized, combined with a reliable and usable system, and then the system's performance and usability are assessed. This study will help us understand how people use crime data. Additionally, it is capable of doing analyses on a variety of datasets, including those related to enterprise surveys, poverty, aid effectiveness, and more.

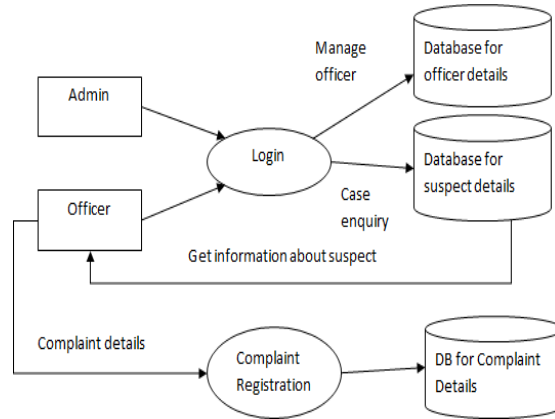


Figure 1: Detecting URL

V. METHODOLOGY:

KNN

1) Algorithm: KNN classification

Input:

- 1) A finite set D of points to be classified,
- 2) A finite set T of points,
- 3) A function $c: T \rightarrow \{1, \dots, m\}$,
- 4) A natural number k.

Output:

A function $r: D \rightarrow \{1, \dots, m\}$

Method:

- 1) Begin
- 2) For each x in D do
- 3) Let $U \leftarrow \{\}$
- 4) For each t in T add the pair $(d(x, t), c(t))$ to U;
- 5) Utilising the initial components, order the pairings in U;
- 6) Count the class labels from the first k elements from U;
- 7) Let $r(x)$ is the class with the highest number of occurrences;
- 8) End For each
- 9) Return r
- 10) End

The KNN classification searches through the dataset to find the similar or most similar instance when an input is given to it. The query entered into the KNN is the attribute values of crime dataset. Based on that query KNN gives result that assist to analyse the large crime database and also helps in predicting the crime future in various cities. It draws crime patterns for various cities.

KNN method stores all available objects and classifies new objects based on the similarity measure. It is used for criminal identification by considering the past crimes and discovering similar crimes that match the current crime based on number of nearest neighbors matched. The attributes crime type and crime city are considered as crime input attributes for KNN. All the attributes in the crime dataset can be considered but for the sake of brevity only two attributes are considered here. KNN method searches for nearest neighbor of the input values and filter those values from the dataset. The result of this method helps to retrieve data from the database and also assists in understanding the crimes.

VI. CONCLUSION

It is essential to have a computerized platform for managing criminal records. The criminal investigation tracker improves the accurate and effective maintenance of criminal records, assisting with decision-making and enhancing reliability while also enhancing law enforcement operations. As a result, the nation's crime rate declines, so enhancing national security. Future applications of data mining technology could include tracking down social outlaws and creating a better future.

REFERENCES

- [1] Shiju Sathyadevan, Surya Gangadharan. S, (2014), "Crime analysis and prediction using Data Mining", First International Conference on networks and Soft Computing, Kerala, India.
- [2] Heartfield, Ryan, George Loukas, and Diane Gan, (2016), "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks." IEEE Access 4, University of Greenwich, London.
- [3] Sivaranjani, S., S. Sivakumari, and M. Aasha, (2016), "Crime prediction and forecasting in TamilNadu using clustering approaches." Emerging Technological Trends (ICETT), International Conference on. IEEE, University of Coimbatore, India.
- [4] Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No. 8, pp. 11-17, 2015.
- [5] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", *Proceedings of 24th IEEE International Conference on Wireless and Optical Communication*, pp. 185-190, 2015.
- [6] Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kusters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", *Proceedings of 6th IEEE International Conference on Data Mining*, pp. 1-7, 2006.
- [7] Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim, "Criminal Act Detection and Identification Model", *Proceedings of 7th International Conference on Advanced Communication and Networking*, pp. 79-83, 2015.
- [8] Kamal Taha and Paul D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", *IEEE Transactions on Information Forensics and Security*, Vol. 11, No. 4, pp. 811-822, 2016.
- [9] Kevin Sheehy et al., "Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System", *Proceedings of IEEE Systems and Information Engineering Design Symposium*, pp. 250-254, 2016.

A STUDY OF ENHANCING BIOMETRIC AUTHENTICATION TECHNIQUES AND THEIR APPLICATIONS IN FINGERPRINT SENSORS

Dr.S.Jayaprakash,¹

Research Supervisor in Department of Computer Science,

Edayathangudy G.S.Pillay Arts & Science College,

Nagapattinam, Tamilnadu,
jayaprakashsoundar@gmail.com

J.P.Keerthana²

Research Scholar in Department of Computer Science,

Edayathangudy G.S.Pillay Arts & Science College,

Nagapattinam, Tamilnadu,
jayakeerthana2094@gmail.com

Abstract— Passwords and PINs are two examples of outdated authentication techniques that offer insufficient security in an era of frequent data breaches and identity theft. This is where biometric authentication, a cutting-edge method that uses distinctive behavioral or physical characteristics to verify people's identities, comes in. These days, a wide range of industries use these cutting-edge biometric technologies, including consumer electronics, banking and healthcare. Fingerprint recognition is becoming a common biometric identification technique since it is quick, easy and safe to use. In daily life, fingerprint biometrics is mostly used to establish security and trust in financial transactions, protected area access and Smartphone unlocking. Every individual has distinct ridges and valleys that remain constant throughout their existence. As fingerprints are difficult to copy or fake, fingerprint biometrics offers a very convenient and trustworthy method of authentication. This study focuses on the advancements in biometric authentication methods of fingerprint sensors such as thermal, optical, capacitive and ultrasonic sensors. Lastly, discuss the fingerprint sensor's applications, problems, highlights and execution procedure.

Keywords- *authentication methods, sensor components, scanningprocess, applications.*

INTRODUCTION

The term "biometrics" is derived from the Greek words bio (life) and metric (to measure). For our use, biometrics refers to technologies for measuring and analyzing a person's physiological or behavioral characteristics. These characteristics are unique to individuals hence can be used to verify or identify a person. Fingerprint recognition is regarded as the most reliable and accurate biometric identification system available. Every person is identified by their individual attendance, lock opened through their own unique finger prints. It is most actively studied as a biometric technology [1]. To provide a state-of-the-art fingerprint and password biometric authentication approaches[2]. Fingerprint scanning systems are designed to detect minutiae. Images of detected minutiae are processed through matching algorithms in order to verify a query fingerprint that is identical to a stored fingerprint. However, fingerprint authentication based on minutiae can be easily bypassed and the need for a more secure method is required [3]. Owners can use their fingerprint to gain secure access to their devices because it is simple to deploy in any device, including a laptop or mobile phone [4]. Fingerprint biometric features are kept inside interference patterns optically, which are also protected with experimental parameters. If both pieces of information are provided to be known to the person at the decryption stage, as a result, it keeps maintaining user specificity in order to access system information [5]. Different fingerprint collecting techniques, which supply fingerprint information for fingerprint identification systems, have attracted a significant deal of interest in authentication technology regarding fingerprint identification systems [6]. Using the capacitive touch screen to capture the user's swipe. While the user swipes, a series of capacitive frames are captured for each swipe [7]. With heavy touch and swipe, the wavelet transform is used to estimate the localized deformation and further to identify the shear-induced gesture based on fingerprint images[8]. A recognition system based on hand geometry achieved through ultrasound images is proposed and experimentally evaluated. 3D images of human hand are acquired by performing parallel mechanical scans with a commercial ultrasound probe. Several 2D images are then extracted at increasing under-skin depths and, from each of them, up to 26 distances among key

points of the hand are defined and computed to achieve a 2D template. A 3D template is then obtained by combining in several ways 2D templates of two or more images [9]. The system acquires a volumetric image of the whole hand and for both characteristics, several 2D images are extracted at different depth levels. From each image, 2D features are extracted and then properly

combined to achieve a 3D template [10]. Recognition performances are evaluated through verification and identification experiments by employing a homemade database. Experiments are carried out first for the two unimodal biometrics and successively, by fusing the two modalities at score level [11]. The biometric technology connected with smart systems helps in monitoring the human activities all over the world, thus providing a good security level. Biometrics acts as a major support to various fields of automobile security, Internet of Things (IOT) Security, health care security, workforce management of organization, government security, banking, and retail industry [12]. The challenges in this area involve balancing the security of protected assets by quickly detecting intruders with the system usability for genuine users. Biometric recognition plays a major role within this context, as it is the main way to assure that users are who they claim to be [13]. Biometric systems identify individuals based on unique traits such as the face, fingerprints, iris etc. The main objective of the study is to understand the role of deep learning in the process of authentication as well as its application in the enhancement of security of biometric systems [14]. An overview of fingerprint sensing methods used for authentication, analyze the current fingerprint sensing technologies, from algorithmic used to ensure that a living human user is attempting to authenticate on a system [15]. authenticate with their fingerprint. Next, this template is checked to see if it matches any of the templates that are kept in the database. Authentication: The person is permitted access or the authentication process is successful if there is a match between the live fingerprint and the stored templates. Access is refused if no match is found.

FINGERPRINT BIOMETRIC TECHNOLOGY

Real-world security is facilitated by fingerprint biometrics technology, which authenticates people based on their distinct fingerprint patterns. The method of utilizing distinctive patterns on a person's fingertips to identify and validate their identification is known as fingerprint biometrics. Since every person has a unique fingerprint, fingerprint identification is dependable and safe. Only those with permission can unlock devices, enter restricted locations, conduct sensitive financial transactions and carry out other sensitive tasks by scanning and comparing their fingerprints. Fingerprint biometrics greatly lowers the danger of fraud, unlawful access and identity theft by ensuring that only the legitimate owners can access their information or do particular tasks.

Working process

The way fingerprint biometrics operates is by taking a picture of and examining each person's fingerprints as shown in below fig.1.

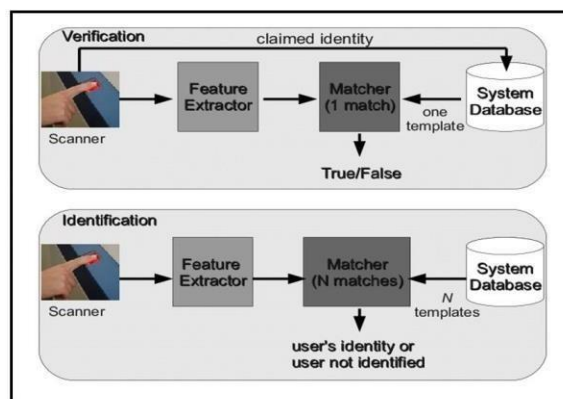


Fig.1. Working Process of Fingerprint Sensor

Capture: A digital picture or scan of a person's fingertips is obtained using a fingerprint scanner and built into a

system like a door lock, on a separate device, or on Smartphone.

Analysis: Advanced software is used to examine the obtained fingerprint image. The software recognizes and extracts the fingerprint's distinctive characteristics, including ridge patterns, bifurcations, and ends.

Creation of a template: A digital template or code is produced using the features that have been extracted. This template acts as a special identification for that person and is a mathematical representation of their fingerprint. **Comparison:** A new template is made and the user's live fingerprint is recorded when they attempt to Authentication Methods

Optical Fingerprint

The optical fingerprint sensor module is among the most widely used and ancient types of fingerprint sensors. It uses an image capturing device and visible light to take a picture of your fingerprint and turns it into digital data as shown in below fig.2.

Components:

Glass protective layer: Serves as a barrier against water and dust.

LED light source: Aids in illuminating the finger.

Image sensor camera with Charge Coupled Device/Complementary Metal Oxide Semiconductor: Records the ridge and valley details of a fingerprint.

Digital signal processors/Microcontrollers : Devices perform data conversion, analysis, and control.

Universal Asynchronous Receiver-Transmitter (UART) interface: Sends fingerprint data to a mobile device or computer.

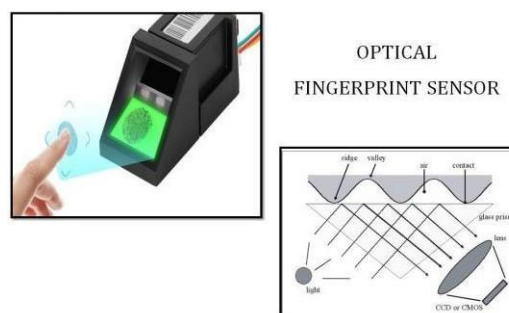


Fig.2. Working Process of Optical Fingerprint Sensor

Scanning Process:

Step-1: Press your finger up against the optical scanner's glass shield.

Step-2: Your fingerprint's ridges and valleys are highlighted by an LED brilliant light.

Step-3: To capture a crisp digital image, a lens directs light onto a CCD or CMOS image sensor. **Step-4:** Algorithms are used to process the digital image created from the photograph.

Step-5: A fingerprint biometric template is made and kept.

Step-6: A template that is used when a finger is presented to authenticate a person.

Capacitive Fingerprint

The capacitive fingerprint scanner and sensor uses an electrical charge to detect a fingerprint, creating the ideal fusion of efficiency and technology. They use an electric current that is passed through the ridges and valleys on your finger in place of light beams and a lens, like in optical sensors. Because of this, capacitive fingerprint sensors have gained recognition for being compact, quick and accurate. They are now widely used in a variety of consumer gadgets, such as laptops, tablets and Smartphone as shown in below fig.3.

Components:

Conductive grid of cells or pixels: Composed of silicon or another semiconductor.

Readout circuit: Generates electrical signals from capacitance.

Controller: This device creates control signal timing and digitizes sensing values.

Feature extraction unit: This device gathers pertinent data from the fingerprint biometric data that has been

recorded.

Matcher unit: This unit compares the feature vectors of the fingerprint that was taken with the database reference.

Post-processing unit: Boosts image quality to increase data reliability and accuracy.

Logic and Interface: Controls power, signals for control, and communication between the host device and the capacitor array.

Step-6: After that, a matching score is applied to ascertain whether or not you are authorized.

Ultrasonic Fingerprint

This kind of sensor makes use of ultrasonic technology to scan your finger and produce a three-dimensional image of your fingerprint. High-frequency sound waves enter the skin and pass through its outer layer before bouncing back off the dermis beneath. This special technique records sweat pores and ridge lines for an accurate reading that can easily differentiate even greasy or filthy fingerprints. An apparatus that transmits sound signals is called an ultrasonic transducer as shown in below fig.4.

Components:

Receiving Transducer: Gathers waveform echoes that have been broadcast.

Ceramic Substrates: Maintain transducers' position.

Microprocessor: Manages fingerprint scanning and analysis.

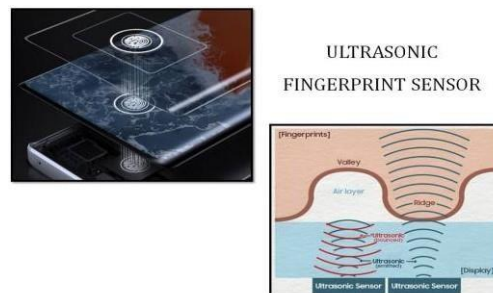
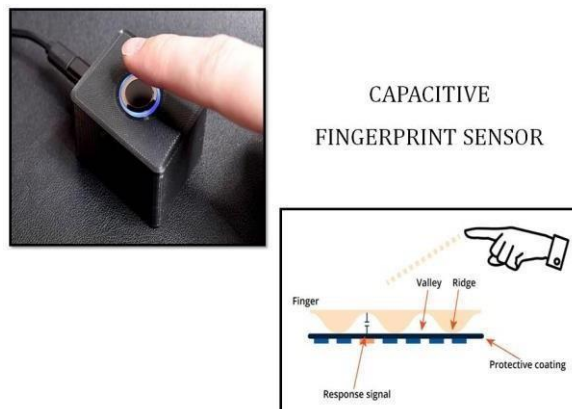


Fig.4. Working Process of Ultrasonic Fingerprint Sensor



Scanning Process:

Step-1: Touch a capacitive scanner with your finger.

Step-2: As the finger moves over the conductive parts of the scanner, an electric field is created that senses variations in capacitance.

Step-3: An analog-to-digital converter (ADC) in the readout circuit transforms variations in capacitance into a digital output.

Step-4: The controller receives the digital signal after which it is processed further, either by enhancing the image or extracting features.

Step-5: A template for a fingerprint is made and compared to the templates kept in the database.

b) Scanning Process:

Step-1: Ultrasonic sound waves are released by the transducer.

Step-2: Energy reflected from sweat pores and fingerprint ridges creates echo.

Step-3: Echoes are gathered and transformed into electrical impulses by a receiving transducer.

Step-4: A microprocessor interprets data and builds a three-dimensional fingerprint image.

Step-5: Digital template saved for later use as a digital template.

Step-6: Image compared to confirm authorization for access or identification.

Thermal Fingerprint

These sensors are able to provide accurate and secure detailed scanned images by identifying temperature differences in the ridges and valleys that are right at our fingertips. This information is used by thermal fingerprint sensors to produce an accurate digital image of the distinct scanned image of your finger as shown in below fig.5.

Components:

Pyro-electric substance: This material is sensitive enough to identify variations in skin and air temperatures.

Silicon Die: Offers a foundation for the pyroelectric material to be incorporated into.

Transistors: When heated or cooled, transistors generate an electric charge that results in the scanned image.

Microprocessor: This device interprets and saves transistor data as a digital fingerprint template.

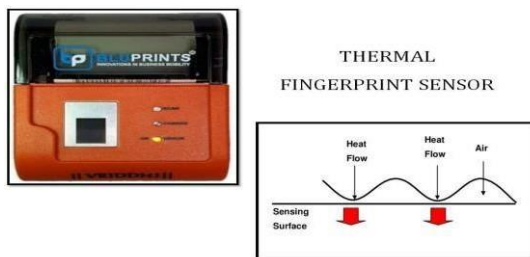


Fig.5. Working Process of Thermal Fingerprint Sensor

Scanning Process:

Step-1: Touch the thermal fingerprint scanner with your finger.

Step-2: The temperature differential between your finger and the air is measured by the pyro-electric substance.

Step-3: In reaction to heat and cold, transistors produce an electric charge.

Step-4: After processing the transistor data, the microprocessor generates a digital fingerprint template.

Step-5: After that, the system matches the data from your finger scan to the template that has been saved.

Step-6: Access is allowed and authentication is successful if a match is detected.

ANALYSIS OF FINGERPRINT SENSOR**Optical Fingerprint****Pros:**

Easy accessibility in the market.

High accuracy and dependability.

They interpret data quickly.

They are easy to install and maintain.

They are substantially less expensive than other types of sensors.

Cons:

Compared to other types of sensors, optical fingerprint scanners and sensors are less secure.

To keep them in optimal condition, they need to be cleaned and maintained on a regular basis.

Fake finger impressions and materials like latex can readily deceive them.

*Capacitive Fingerprint**Pros:*

- Quicker authentication process.
- Reduced power consumption.
- Greater accuracy as a result of improved feature extraction.
- Slim and compact designs.
- Simple integration support for touch and swipe sensing.
- Appropriate for high-security applications.

Cons:

- It may be harmed by electrostatic discharges.
- High production costs.
- Performance is impacted by dry, scarred fingers.

*Ultrasonic Fingerprint**Pros:*

- Accurate compared to conventional scanners.
- Functions properly even with filthy or wet fingers.
- Extremely dependable and safe over time Compared to other types of sensors, optical fingerprint scanners and sensors are less secure.
- To keep them in optimal condition, they need to be cleaned and maintained on a regular basis.
- Fake finger impressions and materials like latex can readily deceive them.

Cons:

- Costlier than conventional scanners.
- May take longer to capture.
- Uses more power than other types of scanners.

*Thermal Fingerprint**Pros:*

- Fingerprint detection works in both wet and dry situations.
- It is very secure and accurate.

Cons:

- Sensitive to environmental factors like high temperatures
- Can be costly and require a lot of maintenance.
- In comparison to other scanners, it may take longer to process a fingerprint.

APPLICATIONS

Healthcare: To accurately track patients and prevent confusion, hospitals primarily use biometric authentication. Clinics and doctor's offices often use biometric authentication to safeguard patient data. By using biometric authentication, hospitals and clinics can save and access patient medical histories whenever needed.

Law Enforcement: Law enforcement uses a variety of biometric data types for identification. DNA, voice samples, iris patterns, fingerprints, and facial features are used by state and federal authorities. This speeds up and simplifies their access to confidential information. Law enforcement usually uses a skilled human examiner to compare an image of a fingerprint to the prints that are on file. The Automated Fingerprint Identification System, or AFIS, can now match a fingerprint against a database in a matter of minutes.

FINGERPRINT SENSOR TYPES	APPLICATIONS
Optical Fingerprint Sensor	National ID Biometric voter registration Law enforcement and forensics investigations Border control ATMs and banking KYC Home security systems Access control systems Time and attendance tracking Mobile phone and laptop authentication
Capacitive Fingerprint Sensor	Physical access control Mobile banking Mobile point of sale terminals (mPOS) Smartphone security PC security Time and attendance tracking
Ultrasonic Fingerprint Sensor	Mobile authentication
Thermal Fingerprint Sensor	Biometric access control systems Login authentication for laptops and other computing devices Financial services Vehicle entry systems Law enforcement

TABLE.I. USE CASE OF FINGERPRINT SENSOR

Mobile authentication: At the intersection of connection and identity, there are mobile biometric solutions available. For identity and authentication, they make use of wearable technology, the Internet of Things, Smartphone, tablets and other portable electronics, in addition to one or more biometric modalities. This makes deployment options versatile. Whether they match on-device or on a secure server, protect Smartphone lock screens, identify wanted persons while on the job, or are hardware- or software-enabled, mobile biometrics solutions are becoming a more and more significant component of the biometrics landscape.

Travel: The same biometric data that is stored in a traditional passport is stored on a microchip found in an electronic passport. The passport holder's digital photo, associated with their name and other identifying information, is stored in the chip. An authority in the issuing country issues the e-passport electronically after verifying the applicant's identification using fingerprints or other biometric data and comparing the information provided by the applicant with the data in the chip.

FUTURE ENHANCEMENT

Artificial intelligence (AI) and Machine Learning (ML) technologies hold great potential for advancing biometrics in the future. They can significantly reduce false match rates and increase the accuracy of biometric systems. Since more complex algorithms can detect and adapt to even the smallest changes in a person's biometric data over time, they provide more reliable identification and authentication. Additionally, ML and AI can improve the security of biometric systems by seeing and eliminating threats before they have a chance to cause any harm. Lastly, these technologies can enhance the user experience by streamlining and speeding up the authentication procedure. Thus, the combination of ML and AI appears to be a promising breakthrough for biometrics in the future.

ACKNOWLEDGMENT

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the Edayathangudy G.S.Pillay College Arts and Science College, Nagapattinam and really sincere thanks to my guide Dr.S.Jayaprakash M.Sc.,M.Phil.,Ph.D., research supervisor and associate professor in department of computer science for his timely suggestion, valuable guidance and sustained interest at every stage of this dissertation.

REFERENCES

- [1] A systematic review on Fingerprint based Biometric Authentication System, S. Hemalatha ,2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020,
- [2] A survey of biometric approaches of authentication N Yusuf, KA Marafa, KL Shehu... - International ..., 2020 -search.proquest.com
- [3] Trends in Biometric Authentication: A review MI El-afifi - Nile Journal of Communication and Computer ..., 2023 - journals.ekb.eg.
- [4] Biometric-Based Optical System for Security and Authentication G Verma, W He,X Peng-2024-intechopen.com.
- [5] A Review of Fingerprint Sensors: Mechanism, Characteristics, and Applications Y Yu, Q Niu, X Li, J Xue,
- [6] W Liu, D Lin - Micromachines, 2023 - mdpi.com
- [7] Overview on fingerprinting authentication technology N Sulaiman, QA Tajul Ariffin - ... of the 5th International Conference on ..., 2020 – Springer.
- [8] M. A. Rilvan, J. Chao and M. S. Hossain, "Capacitive Swipe Gesture Based Smartphone User Authentication and Identification," 2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Victoria, BC, Canada, 2020.
- [9] BioTouch: Reliable Re-Authentication via Finger Bio- Capacitance and Touching Behavior C Zhang, S Li, Y Song,
- [10] Q Meng, L Lu, M Hou - Sensors, 2022 - mdpi.com
- [11] J. L. Strohmman, C. Xu, Y. Lu and H. Panchawagh, "Ultrasonic Biometric Authentication System with Contact Gesture Sensing," 2020 IEEE International Ultrasonics Symposium (IUS), Las Vegas, NV, USA, 2020.
- [12] A review on various biometric techniques, its features, methods, security issues and application areas M Gayathri, C Malathy, M Prabhakaran - Computational Vision and Bio ..., 2020 – Springer.
- [13] Continuous authentication using biometrics: An advanced review
- [14] G Dahia, L Jesus... -Reviews: Data Mining ..., 2020 –
- [15] Wiley Online Library.
- [16] Challenges and opportunities in biometric security: A survey S Arora, MPS Bhatia - Information Security Journal: A Global ..., 2022 - Taylor & Francis.

DETAILED INVESTIGATION OF IN-VITRO FERTILIZATION: A SIMULATION-BASED STUDY UTILIZING HYPER-SPECTRAL PARAMETERS AND ADVANCED DATA MODELS

P. Malathi¹, Dr. M. Gomathi²

¹Research Scholar, Department of Computer Science, Shrimathi Indira Gandhi College, Trichy, India

²Assistant Professor & Research Supervisor, Department of Computer Science, Shrimathi Indira Gandhi College Trichy, India.

¹malathii1989@gmail.com, ²gomathimohan1997@gmail.com

Abstract— In modern times, fertility issues faced by couples are increasingly linked to everyday practices, dietary habits, and lifestyle choices. As such, there's a growing focus on improving predictive modeling within the realm of in-vitro fertilization (IVF), prompting the integration of advanced techniques. While existing research explores various facets of IVF treatment success and failure, this investigation centers on analyzing hyperparameters pertinent to IVF outcomes. It involves a simulated application of advanced data models to highlight the effectiveness of the approach concerning specific attributes. Our study employs comprehensive algorithms, including K-means clustering with interactive features, K-nearest neighbor (KNN) technique, Support Vector Machine (SVM), and Principal Component Analysis (PCA). These methods harness IVF data to showcase the accuracy and reliability of predictive outcomes. We utilize a rich dataset to provide nuanced insights into the variables impacting IVF outcomes. Our predictive models are finely tuned and optimized to capture intricate data patterns, strengthening proximity-based and margin-maximizing approaches. Performance evaluation includes the use of confusion matrices, offering detailed insights into predictive accuracy, sensitivity, specificity, and overall classification performance. Feature statistical analysis further highlights the significance of individual predictors in IVF outcomes. By amalgamating these techniques, we not only enhance IVF predictive capabilities but also gain a deeper understanding of factors contributing to refined treatment strategies. This fosters informed and personalized reproductive healthcare practices.

Keywords: In-vitro fertilization, Predictive Modeling, K-means clustering, K-nearest neighbor (KNN), Support Vector Machine (SVM), Principal Component Analysis (PCA).

INTRODUCTION

In-vitro fertilization (IVF) embodies meticulously transformative and revolutionary medical process that renders a ray of hope and confidence to individuals and couples struggling with sterility. Since its inception in the late 20th century, IVF has evolved into a widely adopted assisted reproductive technology, enabling millions of individuals worldwide to realize their dreams of parenthood. IVF involves the fertilization of an egg and sperm outside the human body, typically within a laboratory setting. This intricate process consists of several key steps, starting with the stimulation of the ovaries to produce multiple eggs. These eggs are then retrieved and combined with the sperm in a constrained environment, leading to the formation of embryos. After a period of cultivation, one or more healthy embryos are selected and transferred into the woman's uterus, with the hope that implantation will occur, resulting in a successful pregnancy. The journey through IVF is often characterized by emotional, physical, and financial challenges, as individuals navigate a series of medical procedures, fertility medications, and uncertainties about the treatment's outcome. However, the potential outcome obtained from overcoming infertility, and achieving a successful pregnancy makes the IVF technique to be the most pursued beacon of hope for those struggling with reproductive complications.

Over the years, advancements in reproductive medicine and technology have refined IVF protocols, consecutively enhancing success rates, while parallelly broadening the scope of individuals who can benefit from this pioneering approach or treatment. As researchers continue to explore new techniques, such as predictive modeling using artificial intelligence, and sophisticated analysis like K-means clustering and

Principal Component Analysis, the landscape of IVF is poised for continued improvement, offering even greater precision and personalization in fertility interventions.

The neoteric progressions in colossal technological domains like artificial intelligence and machine learning have spurred the application of predictive modeling methods in reproductive medicine. The indulgence in machine learning, deep learning, image processing and even robotic applications hold pivotal grounds to the augmentation of predictive accuracy in IVF. These methodologies incorporated in the various phases of the process enable a more tailored and data-driven approach, while successfully maximizing the chances of healthy pregnancy. As the field of IVF continues to evolve, researchers and healthcare professionals are not only focused on improving success rates but also on minimizing the physical and emotional toll on individuals undergoing treatment. The ongoing pursuit to precision and personalization in IVF reflects the commitment that could be rendered to expectant couples with the best possible chances of achieving their dream of successful and uncomplicated gestation.

This paper pivots on delivering a comprehensive study of the various attributes and hyper parameters that are significant in the process to healthy and stable conception, while elaborating the algorithmic incorporations to analyze the efficacy of the approach accelerated toward successful and failed IVF practices. The indagation is structured with section II explicating the empirical review relevant to the IVF approach and the algorithms entailed, section III articulates the methodology, alongwith the work flow utilized in the simulative process, and section IV illustrating the results with the final section concluding the research with future work pertinent to the IVF process.

LITERATURE VIEW

Artificial Neural Networks (ANNs) emulate the functionality of small neural clusters in a fundamental manner, with their anatomical and functional features well-documented in medical applications [Cochand-Priollet et al. 2006; Cunningham et al. 2000; Karakitsos et al. 2002; 2005; Pantazopoulos et al. 1998; Papik et al. 1998].

The foundational work of McCulloch and Pitts [1943] demonstrated that artificial networks resembling biological networks could be constructed using interconnected neurons, mathematics, and algorithms. Subsequent developments, including Hebb's associative learning in 1949 and Rosenblatt's training concept involving synaptic strength adjustments in 1962, laid the groundwork for the evolution of ANNs.

Despite early criticisms, advancements such as error backpropagation, introduced by Werbos [1990], and Hopfield's asynchronous network model [1982], showcased the potential of ANNs in mimicking human brain behavior and solving complex problems.

METHODOLOGY

This section outlines a systematic approach to developing and deploying machine learning algorithms for predicting IVF outcomes, thereby combining data-driven techniques with rigorous evaluation processes. Attributes play a predominant role in successful IVF outcome, nonetheless the hyper parameters in this study are scrutinized through effectuation of correlative and attribute ranking methods that showcases the impact that the specific parameter might have on the diagnostic result. The workflow for the process is implemented in the orange tool that is significant in integrating advanced data models, and is illustrated in Fig 1.

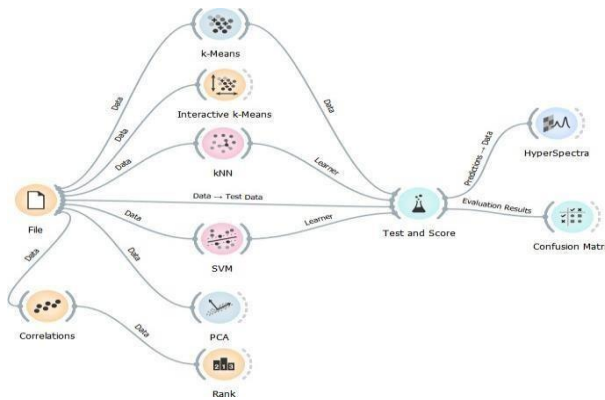


Fig1. Work flow Implementation

[1] DataCollection

The data to comprehend the diagnostic successor failure of the IVF treatment takes into account the dossier relevant to the following attributes:

- Ovarian age of individuals opting for IVF treatment
- Smoking habits
- Alcohol consumption
- History of accidents or trauma
- Previous surgical interventions
- The cryopreservation that elucidates whether the sperm is fresh or frozen.
- The type of embryo, that expounds the bifurcation into a cleavage or a blastocyst, with the former indicating the initial stage, and the latter signifying a later phase of embryo evolution/development.
- The total number of embryos developed in good, fair, and poor categories.
- The maximum number of embryos that progressed heart beats.
- The diagnostic result of the IVF treatment indicative of the success or failure of the procedure.

[2] Preprocessing

The preprocessing entails the transformation of categorical variables (such as smoking habits, alcohol consumption, accident/trauma, surgical interventions, fresh/frozen sperm, and embryo type) into numerical representations, while methodically handling missing data to standardize the incorporated features for further processing.

[3] Techniques Used

The K-means clustering was applied to divulge the inherent patterns within the dataset, with a systematic evaluation of the cluster means that are integrated in the diverse configurations of the loaded dataset. The interactive K-means approach was implemented to obtain a clear overview of the merging clusters, and the disintegration of cluster means that renders an unambiguous purview for clustering decisions. This approach also minimizes the user intervention in refining predictions, and further evincing a distinct combination of clusters and outliers.

The K-Nearest Neighbor (KNN) is utilized to stratify the IVF profiles, while taking into account the number of neighbors, and the measure of neighborhood classification implemented through the Euclidean distance. The below formula is indicative of the Euclidean distance, and is as follows:

$$D = \sqrt{(M_2 - M_1)^2 + (N_2 - N_1)^2} \quad (1)$$

Where M and N are coordinate points for any two data in a two-dimensional dataspace.

The SVM is effected to categorize IVF outcomes and scrutinize decision boundaries, exploring kernel functions for capturing non-linear relationships. The Radial Basis Function (RBF) kernel is utilized for the study, with the tolerance level set to 0.005, and an iteration limit of 300. The RBF kernel is incorporated with the below formula for activating the kernel given by components contributing to predicting IVF success and failure. Model performance for this process is evaluated using metrics such as accuracy, precision, recall, and F1-score, with cross-validation conducted to assess generalization capability. The components for the study, along with the variance is set for the PCA, and normalization is achieved to standardize the results obtained.

The interpretation of the results in the context of IVF success and failure, clusters, decision boundaries, and principal components are visually represented for enhanced comprehension. The predictive efficacy of each methodology is compared, and opportunities to integrate insights from multiple models are explored to improve overall accuracy.

[4] Prediction and Evaluation

The predictive analysis for the IVF process involved utilizing hyperspectral data to anticipate various factors influencing the outcomes of IVF procedures. Subsequently, the model's performance is rigorously corroborated through the contrivance of the confusion matrix. This matrix facilitates a detailed breakdown of predictions into true positives, true negatives, false positives, and false negatives, thereby substantiating the assessment of the model's accuracy, precision, recall, and overall effectiveness with the entailed IVF dataset.

STIMULATION RESULTS

This section presents the throughputs obtained from the technical implementations of the indagation. The feature statistics play a pivotal role in agnizing the mean, median, ranges of the entailed attribute for study. The results obtained is given in the figure below.

$$K = (-g|x-y|^2) \quad (2)$$

Where g represents the gamma value, and x, y are the datapoints used for evaluation.

The complexity bound set at 0.50, with the regression epsilon value at 0.10. The results for this algorithmic implementation is evinced in the next section of this paper.

The implementation of PCA is implemented for reducing dimensionality, examining the principal

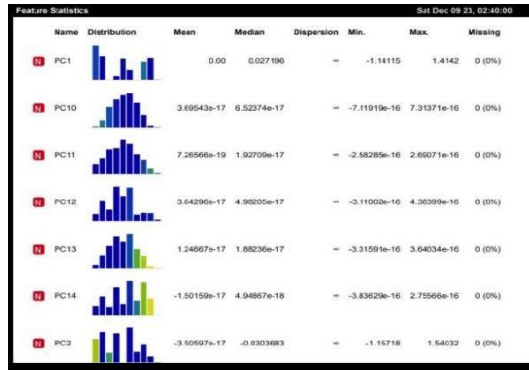


Fig 2. Feature Statistics for the attributes incorporated

The essential and critical phase in this indagation is the pre-processing phase that discretizes the values, normalizes the features to specific range of values and effectuated fast correlation-based filter to identify significant features relevant to the process of IVF. The result for the same is ndicated in Fig 3.

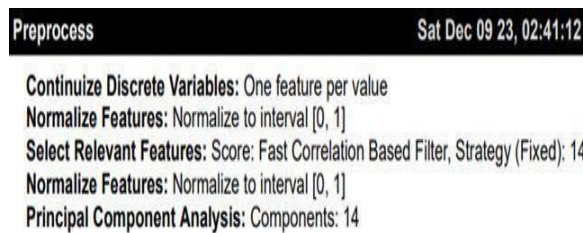


Fig 3. Preprocessing

The explication of PCA is a part of the pre-processing phase to mitigate entropy and evaluate the variance estimated in the dataset.

Fig 4 evinces the result obtained for the same.

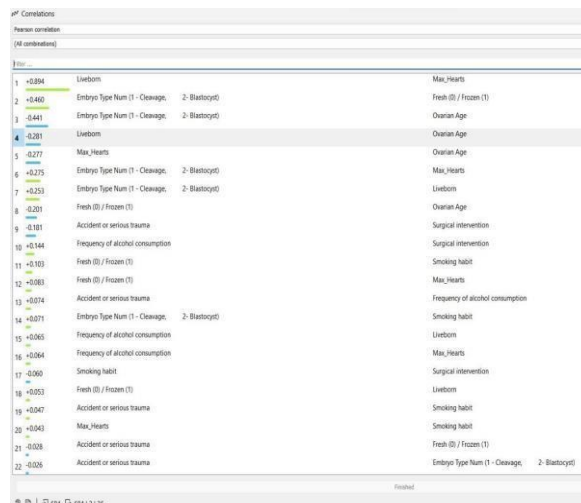


Fig 4. Principal Component Analysis

The hyperparameter is the crucial aspect to the meticulous implementation of the IVF process. The hyperparameters enable in scrutinizing the key catalyst to accelerating the success of the in-vitro fertilization method, and therefore is necessitate that it is analyzed prudently. The correlative chart of the attributes incorporated in the study, along with the feature ranking enables in efficiently delineating the important factors to successful diagnostic results. The implementation results of Pearson, Spearman correlation and feature ranking for hyperparameter analysis is shown in Fig 5,6 and 7 respectively.

#	Correlation	Variable
1	+0.893	Liveborn
2	+0.460	Embryo Type Num (1 - Cleavage, 2- Blastocyst)
3	-0.433	Embryo Type Num (1 - Cleavage, 2- Blastocyst)
4	-0.292	Liveborn
5	+0.284	Embryo Type Num (1 - Cleavage, 2- Blastocyst)
6	-0.281	Max_Hearts
7	+0.164	Embryo Type Num (1 - Cleavage, 2- Blastocyst)
8	-0.203	Fresh (0) / Frozen (1)
9	-0.181	Accident or serious trauma
10	+0.144	Frequency of alcohol consumption
11	+0.105	Fresh (0) / Frozen (1)
12	+0.103	Frequency of alcohol consumption
13	+0.080	Frequency of alcohol consumption
14	+0.079	Accident or serious trauma
15	+0.014	Fresh (0) / Frozen (1)
16	+0.012	Fresh (0) / Frozen (1)
17	+0.011	Embryo Type Num (1 - Cleavage, 2- Blastocyst)
18	-0.050	Smoking habit
19	+0.047	Accident or serious trauma
20	+0.044	Max_Hearts
21	+0.020	Liveborn
22	-0.028	Accident or serious trauma

Fig 5. Pearson Correlation

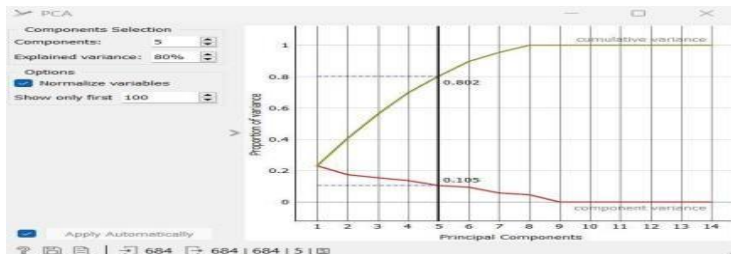


Fig 6. Spearman Correlation

#	Inf...ain	Gai...tio	Gini	χ²	Relieff
1	0.903	0.781	0.434	570.743	0.345
2	0.671	0.525	0.340	448.883	0.237
3	0.055	0.028	0.031	40.663	0.015
4	0.049	0.050	0.030	26.503	0.032
5	0.005	0.005	0.003	1.896	0.042
6	0.005	0.005	0.003	2.639	0.000
7	0.001	0.001	0.000	0.361	0.064
8	0.000	0.000	0.000	0.088	0.012
9	0.000	0.000	0.000	0.121	0.014

Fig 7. Hyperparameter Tuning through Feature Rank

Once the features are obtained, and the correlations are identified, the algorithmic implementation along with the Hyper Spectral predictive implementation pertinent to the IVF diagnosis implemented through the above explicated methodologies. The results given below from Fig 8 through Fig 10 showcases the various methodologies of simulation, leading to the corroborated predictive analysis of the IVF diagnosis.

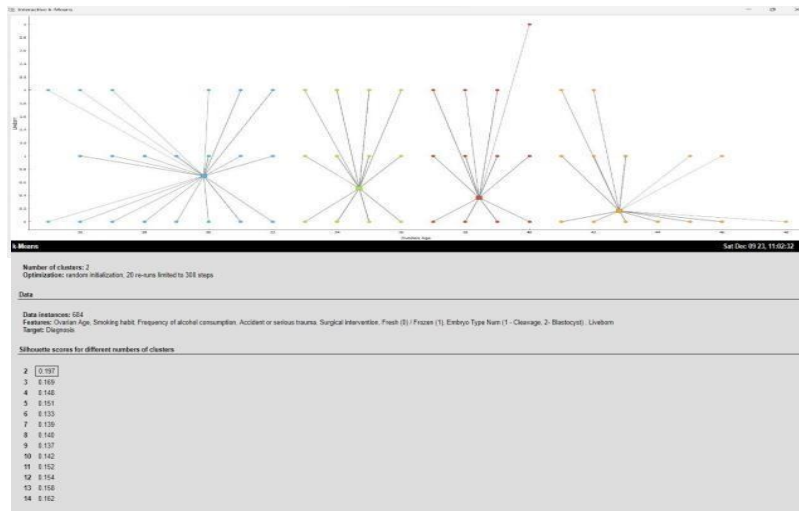


Fig 8. K-Means Clustering with cluster means for each cluster

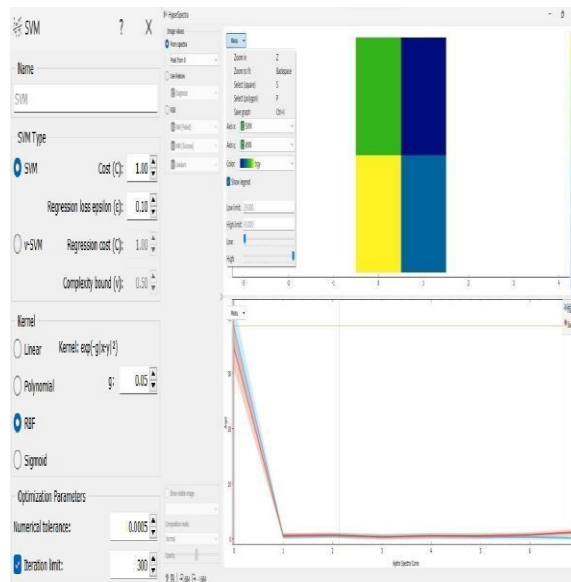


Fig 9. Support vector machine and Hyper Spectral Prediction

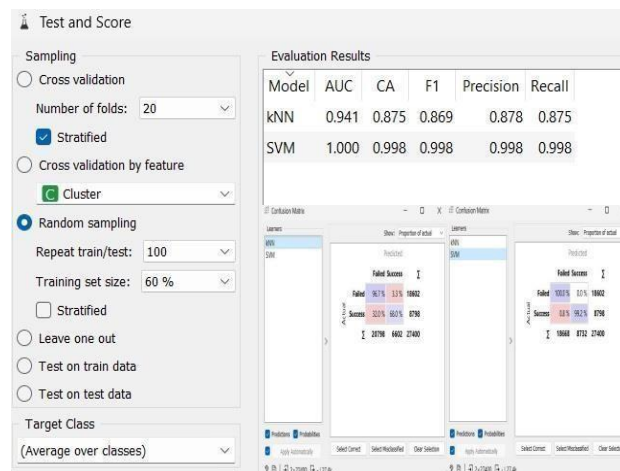


Fig 10. Performance of the Incorporated Models and Confusion Matrix

CONCLUSION

In conclusion, the application of hyperspectral analysis for predicting success and failure rates in In Vitro Fertilization (IVF) has demonstrated promising results. The utilization of K-nearest neighbors (KNN) yielded an impressive accuracy rate of 94%, highlighting the efficacy of this model in classifying and predicting IVF outcomes. Additionally, the Support Vector Machine (SVM) model achieved a remarkable accuracy of 100%, indicating a robust performance in accurately categorizing IVF success and failure cases. These findings underscore the potential of hyperspectral analysis as a powerful tool in reproductive medicine, providing valuable insights into the complex factors influencing IVF outcomes. The high accuracy rates achieved by both KNN and SVM models suggest the reliability of hyperspectral data in predicting fertility treatment success, thereby offering a foundation for more informed decision-making in assisted reproduction.

The evaluation of these predictive models corroborated through the confusion matrix further enhances our understanding of the model performance. The absence of misclassifications in the SVM model emphasizes its strength in potential clinical interventions. In essence, this study showcases high levels of accuracy in the entailed models thereby paving the way for further research entailing more complicated pre-conceptions that require an intricate study of certain diseases correlative to gestation.

REFERENCES:

- [1] Angelini, A., Brusco, G.F., Barnocchi, N., El- Danasouri, I., Pacchiarotti, A. and Selman, H.A. (2006) Impact of physician performing embryo transfer on pregnancy rates in an assisted reproductive program. *J Assist Reprod Genet* 23:329–332.
- [2] Bancsi, L.F., Broekmans, F.J., Looman, C.W., Habbema, J.D. and te Velde, E.R. (2004) Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. *Fertil Steril* 81:35–41.
- [3] Carney J.G. and Cunningham P. (1999) The NeuralBAG algorithm: Optimizing generalization performance in bagged neural networks. 7th European Symposium on Artificial Neural Networks, Bruges, Belgium.
- [4] Cochand-Priollet, B., Koutroumbas, K., Megalopoulou, T.M., Pouliakis, A., Sivolapenk, G. and

- Karakitsos, P. (2006) Discriminating benign from malignant thyroid lesions using artificial intelligence and statistical selection of morphometric features.
- [5] Oncology Reports 15:1023–1026. Creus, M., Penarrubia, J., Fabregues, F., Vidal, E., Carmona, F., Casamitjana, R., et al. (2000) Day 3 serum inhibin B and FSH and age as predictors of assisted reproduction treatment outcome. *Hum Reprod* 15:2341–2346.
- [6] Cunningham, P., Carney, J. and Jacob, S. (2000) Stability problems with artificial neural networks and the ensemble solution. *Artif Intell Med* 20:217–225. Cunningham, P. and Walsh, P. (2002) *Principles of Data Mining and Knowledge Discovery*.
- [7] Springer Press, NY, USA. Dessolle, L., Freour, T., Barriere, P., Darai, E., Ravel, C., Jean, M., et al. (2010) A cycle-based model to predict blastocyst transfer cancellation. *Hum Reprod* 25:598–604.
- [8] Eastham, J.A., Kattan, M.W. and Scardino, P.T. (2002) Nomograms as predictive models. *Semin Urol Oncol* 20:108–115. Ecke, T.H., Bartel, P., Hallmann, S., Koch, S., Ruttloff, J., Cammann, H., et al. (2010) Outcome prediction for prostate cancer detection rate with artificial neural network (ANN) in daily routine.
- [9] *Urol Oncol* Epub ahead of print. Frattarelli, J.L. and Gerber, M.D. (2006) Basal and cycle androgen levels correlate with in vitro fertilization stimulation parameters but do not predict pregnancy outcome.
- [10] *Fertil Steril* 86:51–57. Gnoth, C., Schuring, A.N., Friol, K., Tigges, J., Mallmann, P. and Godehardt, E. (2008) Relevance of anti-Mullerian hormone measurement in a routine IVF program. *Hum Reprod* 23:1359–1365.
- [11] Haykin, S.S. (1994) *Neural networks: a comprehensive foundation*. xix, Macmillan College Publishing Company: NY, USA. Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities, *Proc. of the Nat. Acad. of Sciences of the USA*, 79:2554–2558.
- [12] Hunault, C.C., Eijkemans, M.J., Pieters, M.H., te Velde, E.R., Habbema, J.D., Fauser, B.C., et al. (2002) A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer. *Fertil Steril* 77:725–732.
- [13] Jiang, Y., Zhou, Z.-H. and Chen, Z.-Q. (2002) Rule Learning based on Neural Network Ensemble. *Proceedings of the International Joint Conference on Neural Networks, Honolulu, HI, USA*, 1416–1420.
- [14] Jurisica, I., Mylopoulos, J., Glasgow, J., Shapiro,
- [15] H. and Casper, R.F. (1998) Case-based reasoning in IVF: prediction and knowledge mining. *Artif Intell Med* 12:1–24. Karakitsos, P., Kyroudes, A., Pouliakis, A., Stergiou, E.B., Voulgaris, Z. and Kittas, C. (2002) Potential of the learning vector quantizer in the cell classification of endometrial lesions in postmenopausal women. *Anal Quant Cytol Histol* 24:30–3
- [16] R. Casper, Case-based reasoning in IVF: prediction and knowledge mining, *Artif. Intel. Med.* 12 (1998) 1-24.
- [17] J.R. Trimarchi, J. Goodside, L. Pass more, T. Silberstein, L. Hamel, L. Gonzalez, Comparing data mining and logistic regression for predicting IVF outcome, *Fertil. Steril.* 80 (2003), 100-100.
- [18] G. Patrizi, C. Manna, C. Moscatelli, L. Nieddu, Pattern recognition methods in human-assisted reproduction, *Int. Trans. Oper. Res.* 11 (2004) 365-379.
- [19] C. Manna, G. Patrizi, A. Rahman, H. Sallam, Experimental results on the recognition of embryos in human assisted reproduction, *Reprod. BioMed.* Online 8 (2004) 460^69.

MACHINE LEARNING TECHNIQUES IN EMERGING CLOUD COMPUTING INTEGRATED PARADIGMS

Ms.J.Vinitha,

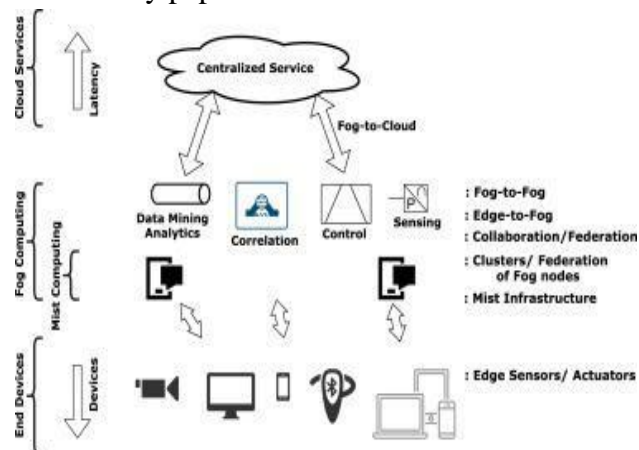
Assistant Professor, Department of Artificial Intelligence and Machine Learning,
Dr.N.G.P College of Arts and Science, Coimbatore, India, Vinitha.j@drngpasc.ac.in

Abstract—Cloud computing offers Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) to provide compute, network, and storage capabilities to the clients utilizing the pay-per-use model. On the other hand, Machine Learning (ML) based techniques are playing a major role in effective utilization of the computing resources and offering Quality of Service (QoS). Based on the customer's application requirements, several cloud computing-based paradigms i.e., edge computing, fog computing, mist computing, Internet of Things (IoT), Software-Defined Networking (SDN), cybertwin, and industry 4.0 have been evolved. These paradigms collaborate to offer customer-centric services with the backend of cloud server/data center. In brief, cloud computing has been emerged with respect to the above-mentioned paradigms to enhance the Quality of Experience (QoE) for the users. In particular, ML techniques are the motivating factor to backend the cloud for emerging paradigms, and ML techniques are essentially enhancing the usages of these paradigms by solving several problems of scheduling, resource provisioning, resource allocation, load balancing, Virtual Machine (VM) migration, offloading, VM mapping, energy optimization, workload prediction, device monitoring, etc. However, a comprehensive survey focusing on multi-paradigm integrated architectures, technical and analytical aspects of these paradigms, and the role of ML techniques in emerging cloud computing paradigms are still missing, and this domain needs to be explored.

INTRODUCTION

Cloud computing is the most promising computing paradigm that brings the concept of "computing utilities" in the market. In cloud computing paradigm, infrastructure, platform, and software services are offered as the "computing utilities" using a pay-per-use model across shared delivery networks, similar to the water, electricity, gas, and telecommunications services (Arockiam et al., 2011). Cloud computing has emerged with several integrated computing and networking paradigms such as edge computing, fog computing, mist computing, Internet of Things (IoT), Software-Defined Networking (SDN), digitaltwin, and industry 4.0. Working with the integration of several cloud computing paradigms and their archetypes is in the trend to meet the next generation computing requirements (Metri and Sarote, 2011). Further, the appliance of Machine Learning (ML) techniques along with integrated cloud computing paradigms is offering new research opportunities to the researchers to meet the future demand of technological advancements (Metri and Sarote, 2011). Best of the authors' knowledge, an extensive study for the integrated computing architectures with machine learning techniques has remained unexplored. On the other hand, a deep study of cloud-driven integrated emerging cloud computing paradigms along with ML techniques is very prominent to the researchers from academia and industry to meet the future computing demands. Therefore, in this paper, we present a systematic literature survey for the integrated, emerging, and cloud computing driven several paradigms with the appliances of ML techniques. Table 1 presents the important abbreviations and their respective definitions used in the survey, in alphabetic order to enhance the readability. Fig. 1 pictorially

represents the road-map of this survey paper.



As an overview, edge computing deals with local data, and avoids data uploading to the cloud. It processes the data directly on the edge devices, and the devices are connected with nearby sensors or gateways (Gupta et al., 2017, Kumar et al., 2017, Wang et al., 2018b). Fog computing provides the facility to communicate, store, and compute data locally utilizing edge computing infrastructure that reduces the bandwidth, latency, and energy requirements (Sideratos et al., 2015, Ferreira et al., 2017, Shin et al., 2014). Mist computing processes the data on the network at the extreme edge that contains several micro-controllers and sensors. It harvests the connected resources by using the sensor's computation and communication capabilities (Asif- Ur-Rahman et al., 2018, Byers and Wetterwald, 2015, La et al., 2019). The Internet of Things (IoT) can be realized in the form of sensors, actuators, mobile phones, etc. that can interact and communicate with each other via local or internet-based connections (Kafi et al., 2013, Qin et al., 2016).

The major applications of the IoT can be found in smart cities, transportation (Kumar et al., 2021), health system and agriculture (Jakkula and Cook, 2010). The Software-Defined Networking (SDN) is a new computer networking paradigm (Kim et al., 2017). SDN and Network Function Virtualization (NFV) offer network virtualization in cloud computing. Further, SDN programs the switches in a novel way and makes fine-grain traffic forwarding decisions in mega-scale data centers (Ahmed et al., 2017). A CyberTwin (CT) or digitaltwin is an identical 3D digital replica of a real-world location or object (Damjanovic-Behrendt and Behrendt, 2019). It is a key enabler of technology for smart manufacturing utilizing cloud infrastructure for the deployment and maintenance (Yu et al., 2019).

Industrial Internet of Things (IIoT 4.0) is an emerging technology focusing on the automation of applications in industrial and machine-to-machine communication utilizing Artificial Intelligence (AI) and cloud computing technologies (Boyes et al., 2018, Souri et al., 2019). The study of integration among emerging cloud computing paradigms becomes crucial for understanding future technological trends. As cloud computing has been integrated directly or indirectly with a wide range of disciplines and applications to offer virtual services in the form of infrastructure, platform, and software (Arockiam et al., 2011). Further, an appropriate integration among different cloud computing paradigms offers the complementary advantages of the several integrating technologies to enhance the performance of the application and meet the user QoS requirements (Buyya et al.,

2010).

The utilization of ML techniques in the cloud integrated computing paradigms is in the trend to meet several QoS requirements. To address the dynamic scheduling in the cloud based cyber–physical production system, Artificial Neural Network (ANN) and other soft computing techniques are being used for optimizing several QoS parameters (Ding et al., 2019). For workload management in cloud based databases, ML techniques such as nearest neighbor and classification tree are being used for Predicting Query Run-time (PQR) tree modeling (Gupta et al., 2008). The PQR learning enhances the application's scalability by making efficient resource allocation decisions. Unified Reinforcement Learning (URL) technique provides real-time provisioning of auto-configuration of VMs (Xu et al., 2012). Further, ANN and Linear Regression (LR) are utilized in adaptive resource provisioning to satisfy future resource demands (Islam et al., 2012). Support Vector Machine (SVM), ANN, and LR are also used in designing prediction models for cloud resource provisioning (Bankole and Ajila, 2013). The Support Vector Regression (SVR) provides an optimal resource allocation and load balancing by reducing service response time (Huang et al., 2013).

The distributed Deep Neural Network (DNN) based architectures are also utilized to reduce communication cost, energy, and response time (Ogden and Guo, 2018). Furthermore, ANN, SVM, Random Forest (RF), and Decision Tree (DT) techniques are commonly used to improve the accuracy in Smart Grid (SG) (Alcaraz et al., 2011), Intelligent Transportation Systems (ITS) and Smart Manufacturing (SM) domains (Lahouar and Slama, 2015, Ponz et al., 2015, Monedero et al., 2012, Chakraborty et al., 2011). ARIMA (Autoregressive Integrated Moving Average) and SVR improve performance of live Virtual machine (VM) migration (Patel et al., 2016). On the same line, ANN and LR are also used for resource prediction during VM migration, while Reinforcement Learning (RL) and SVM are used for VM management.

In mobile edge computing and storage technology, an anomaly detection method with DT integration using Deep Learning (DL) is proposed to enable better offloading strategies. The SVM, K-Nearest Neighbor (KNN), and RF are the frequently used algorithms for IoT device identification (Tuama et al., 2016), whereas Gaussian Mixture Model (GMM) and KNN offer maximum accuracy (Huynh et al., 2015). With this discussion, it is inferred that ML techniques are essentially in trend with the emerging cloud computing paradigms which need to be investigated thoroughly. Fig. 2 shows the statistics for the number of articles used in the study considering each domain of emerging technology paradigms. To get a view of Google trends for continuing research in the domain of emerging cloud computing integrated paradigms, a study is carried out based on recent five years (2017–21) Google Trends data, and a pictorial representation of the data is shown in Fig. 3.

As an outcome, it is observed that edge, fog, and mist are in top trends for the integration with other cloud computing paradigms. This study explores the research trends based on the state of the art for the several cloud computing paradigms i.e., edge computing, fog computing, mist computing, IoT, SDN, digitaltwin, and industry 4.0. Further, Fig. 4 also presents yearly (2017–21) publication statistics from the Web of Science database. Based on this study, it is observed that the majority of articles appeared from the edge computing and IoT paradigms with the integration of other cloud computing paradigms.

As numerous state-of-art studies have been published in the context of cloud computing that discusses the

application of ML and other related technologies (Huang et al., 2013). Based on the survey study (Buyya et al., 2010), it is observed that most of them have focused on independent architectural study. On the other hand, Google trends data and state-of-the-art studies are evidence that future technological developments are highly dependent on the integration of emerging cloud computing paradigms along with ML appliances.

With several ML applications/requirements of integration among various emerging cloud computing paradigms (edge computing Datta and Bonnet, 2017, fog computing Liu and Jin, 2013, mist computing Byers and Wetterwald, 2015, IoT Kurtz et al., 2016, SDN Bi et al., 2018, digitaltwin Bao et al., 2019, and industry 4.0 Čolaković and Hadžialić, 2018), it becomes very important to conduct a systematic survey to understand the broad overview of the possible integration, future trend, application demands, pros and cons, and research issues and challenges for the integration among several emerging cloud computing paradigms. Best of the author's knowledge, there is no systematic survey has been conducted, in the literature, for the same. Therefore, in this work, we focused on a survey paper to carryout a comprehensive study on integrated cloud computing paradigms, archetypes, and machine learning techniques in the integrated cloud computing paradigms.

The survey has been conducted using systematic literature review (SLR) process (Kitchenham and Charters, 2007, Ahmad and Alsmadi, 2021). The SLR has three phases, namely, planning the review, conducting the review, and reporting the review. During the planning phase, the study recognizes need for the review, defines the research areas, and designs a review protocol. This survey explained multi-paradigms integrated architectures and the applications of ML techniques to improve the services in integrated cloud computing paradigms.

The review protocol has been defined based on the identified research questions and study objectives. In a research review, searching the most related studies connected to the research topic is important. To obtain the effective contents, ACM Digital Library, IEEE Xplore, Web of Science Core Collection, Scopus, ScienceDirect, Google Scholar, and electronic scientific databases sources were explored using following keywords cloud computing, internet of things, mist computing, edge computing, fog computing, industrial internet of things, cybertwin, software defined networking, internet of things, machine learning in cloud computing, machine learning and security in IoT/IIoT/cybertwin/edge/mist/fog/SDN etc. The survey defined inclusion and exclusion criteria to focus study on the most recent articles and future trends in machine learning techniques for integrated cloud computing paradigms. A filter was added to the search terms to look for peer-reviewed literature, conference paper, journal paper, blogs, and scientific book chapter well written in English language. The survey excluded research papers unrelated to above mentioned research questions and those articles that did not discuss the paradigms, posters, citations, non-English, preliminary studies, proof-of-concept, Powerpoint presentations, venue impact factor, etc.

Next, in the review phase, data extraction and quality assessment of the papers is performed. On the remaining articles, data extraction is done in order to obtain meaningful content. A reference management system (i.e., Mendeley) is used to extract the title, citations, publication channels, publication year, dataset, ML methodology and models, results, etc.

The quality-assessment questions and scores are used to evaluate the articles that are relevant to the research-

questions. The papers that have offered a thorough, comprehensive, understandable explanation of ML techniques, with the results of data analysis by mentioning IoT/IIoT/cybertwin/edge/mist/fog/SDN were given maximum score.

The papers in which ML algorithms were simply mentioned with no further explanation and less technical information, and addressing only cloud computing not the other paradigms were given medium score. The papers without ML-based solutions and no real technical data-based content analysis were given minimum score. If the paper received more than 50% score, then it is included in the research review; else, it was rejected or excluded. The reporting the review phase represents a procedure for quickly and efficiently reviewing and examining each research studies included within the review. It integrates the results of the study and make conclusions from particular research outcomes. Finally, about 233 primary studies were recognized using data analysis, and information relevant to the research questions were retrieved and stored in database. The extraction of ML algorithms have done after reading entire-text of related primary studies. The primary studies show that the research for ML techniques in integrated cloud computing paradigms is less mature, and it requires further exploration.

This survey presents, first time, a comprehensive study on integrated cloud computing architectures using ML techniques for emerging and integrated cloud computing paradigms. Our major contributions, in the survey, are itemized as follows:

This work identifies the current research trends in the domain of emerging cloud computing paradigms, and explored the state of art thoroughly to investigate several integrated architectures, and the used ML techniques. In this work, a multidisciplinary study among several emerging cloud computing paradigms is carried out to understand the integration among them for meeting the application-specific QoS requirements.

This survey classifies several integrated cloud computing paradigms based on the most dominant and trending technology i.e., ML. This work also investigates the frequently used ML techniques in the domain of integrated cloud computing paradigm.

This survey majorly focuses on recently published articles, in the last five years (2017–21), which encourages the readers to quickly grasp the knowledge from scratch to the current advancements

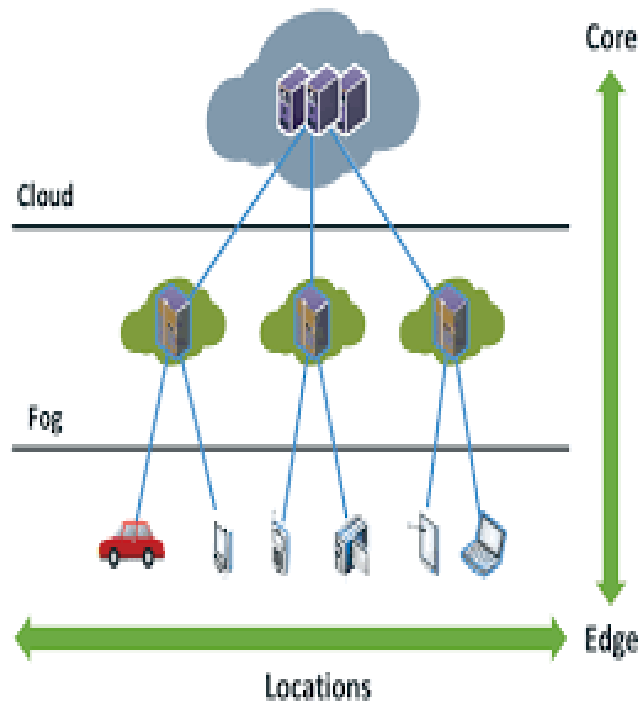
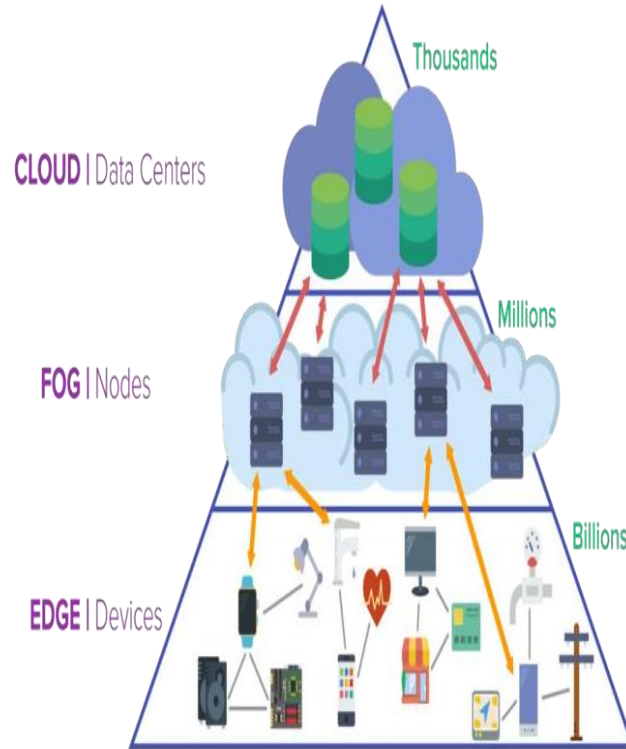
This survey presents a brief tabulated summary on several QoS parameters, for almost every sub-domain, which presents a comparative analysis among similar state-of-the-art methods. This reduces a lot of readers effort to understand the domain/sub-domain-specific concepts, current research trends, advancements, and significant tools and techniques used.

Finally, this survey summarizes significant research gaps, challenges, and future research directions as an outcome of this thorough study. This will surely assist the researchers (including newborn) to carry forward the research in the domain of integrated cloud computing paradigms.

The rest of the paper is organized as follows. Section 2 defines the cloud computing paradigms and integrated architecture. Section 3 discusses machine learning preliminaries (supervised ML, unsupervised ML, reinforcement learning, deep neural network, and federated learning). Section 4 identifies ML techniques in the integrated cloud computing paradigms and presents a detailed comparative study. Section 5 discusses security in integrated cloud paradigms.

Based on the study in Section 4, Section 6 presents the unidentified research gaps, challenges, and

research future directions. Finally, Section 7 concludes the survey paper.



Section snippets

Cloud computing paradigms and integrated architecture

This section discusses various paradigms in the cloud computing domain i.e., edge computing, fog computing, mist computing, IoT, SDN, digitaltwin, and IIoT 4.0. In particular, this section briefs about the fundamental architectural study of the cloud computing paradigms as mentioned above.

Machine learning preliminaries

Artificial Intelligence (AI) is the study of “intelligent agents”, or technologies that perceive their environment and then perform actions to optimize the chance of success at a certain goal. Machine Learning (ML) is a branch of AI that allows computers to learn without even being explicitly programmed. With the development of more powerful computer chips and microprocessors, researchers have discovered statistical models described as Artificial Neural Networks (ANNs). The ANN contains many

ML techniques in integrated cloud computing paradigms

This section elaborates the integration of various emerging cloud computing paradigms and explores the ML techniques applied in integrated architectures such as Cloud-ML, Edge-Fog-IoT-DL, Fog-Cloud-ML, Mist-Cloud, IoT-ML, SDN-IoT, Digitaltwin-IIoT, and IIoT-Fog. The integration of Cloud-ML focuses on the applications of ML in Cloud computing. The parameters of the study are properties, services, deployment models, QoS parameters, security issues, and challenges related to cloud computing. For

Security in integrated cloud paradigms

Security is the most significant challenge to prevent the widespread adoption of cloud computing (Ahmad and Alsmadi, 2021, Alsharif and Rawat, 2021). A threat/attack occurs when an intruder gains access to a system and reveals private data without permission of the user. Attacks are posing a threat to the network ability to protect identification, authorization, accessibility, privacy, and integrity of the user. Thus, this section enlightens the security threats in emerging Cloud computing

Research gaps, challenges and future research directions

In this paper, we thoroughly investigated various emerging cloud computing integrated paradigms and explored various ML techniques usages in the same set of paradigms. Based on the above study, this section presents the identified research gaps, challenges, and future research directions for each integrated paradigm.

Conclusion and discussion

In last decade, cloud computing has been emerged with several emerging paradigms such as edge, fog, mist, IoT, SDN, cybertwin, and industry 4.0 utilizing machine learning techniques. This survey, first, introduces briefly backgrounds of all cloud computing paradigms, and then identifies possible integration among them based on the recent state of the art study. To carry out this study, last five years (2017–21) research articles are explored thoroughly. Further, Google Trends are also

REFERENCE

S. Chakraborty and A. K. Das. Cloud computing and artificial in-telligence: Opportunities and challenges, International Conference on Inventive Communication and Computational Technologies.

REAL TIME FACE RECOGNITION AND DETECTION USING DEEP LEARNING ALGORITHM

D. GAYATHRY¹ , Dr. R. LATHA²

¹Research Scholar, ²Professor & Head, Department of Computer Science, St. Peter's Institute of Higher Education and Research, Chennai - 600 054, Tamil Nadu, INDIA, gaya200689@gmail.com¹ latharamavel@gmail.com²

Abstract. Face features are the primary means of identifying an individual; these features can even differentiate identical twins. Consequently, facial recognition and identification become essential for identifying different people. Facial recognition systems, a type of biometric authentication technology, are used to confirm an individual's identity. Modern applications including home security systems, criminal identification systems, and phone unlocking systems have seen a surge in demand for this technology. This method is regarded as more secure because it relies on a face image instead of external elements like a card or key. Face detection and face identification are the two main processes in the recognition process. This article explores the idea of using deep learning to create a face recognition system with Python's Open CV package. Deep learning is an extremely accurate method because of this.

INTRODUCTION

Face recognition is the process of identifying a person by examining their facial features. Deep learning has transformed the field of face identification by enabling incredibly accurate and efficient recognition systems. In order to learn the fundamental patterns and features that differentiate one face from another, a neural network for deep learning face recognition systems is frequently trained using an enormous collection of facial photographs. After each person's facial representation is created using these attributes, it is saved separately in a database in case it is eventually recognized. By comparing the input facial image with the stored representations during recognition, the system determines which match is the closest. Convolutional neural networks (CNNs) are widely used in face recognition systems to extract information from photos.

We describe the methodological strategy that extracts and identifies a face from the ever-changing stream that follows, compares, and stores data about individuals we know. An automatic facial detection classifier is a good response. Usually, these methods find a set of basic images and represent faces as a linear combination of those pictures. Principal component analysis (PCA) is a popular example of one of these methods. To get the basic pictures, PCA merely employs pairwise correlations between pixels in the image database. Independent Component Analysis (ICA) is a useful technique in computerized face identification and authentication jobs, as opposed to Principal Component Analysis (PCA).

In face recognition systems, the Haar feature-based cascade classifier is utilized for face detection and face localization. By applying the method of weighted Local Binary Pattern, The most important characteristic that helps identify a person is their face. Because face recognition creates a unique identity for each individual, it helps authenticate each person's identity through his or her own particular qualities. The ability to identify and recognize someone based only on the traits of their face is known as face recognition. Due to the multidimensional nature of faces,

numerous mathematical computations are required. Face recognition is a prevalent problem with artificial intelligence. This program was really helpful in our day-to-day activities. Despite the various approaches that have been used thus far, face recognition remains quite difficult in real-world situations. This study introduced PCA and Back Propagation Neural Network, two robust and simple face detection methods based on camera-captured images. facial feature recognition and detection.

Although deep learning approaches are efficient, they often require expensive calculations and result in complicated models that require a large amount of training data. A face recognition system that addresses both face representation and recognition using artificial neural networks is described. It is based on a recent approach. This work demonstrates how to convert a human face's front to image coding and a pixel location histogram in order to provide a measurement for automatic face identification. This study designs and assesses real-time facial recognition system based on Convolutional Neural Networks (CNNs). The proposed design is first evaluated using standard AT&T datasets, and the results are subsequently extended to the development of a real-time system. Face recognition software, which only recognizes faces

System architecture and design

Facial recognition technology is used to record attendance for this assignment, and the students' data is updated and stored on a website created with MY SQL and PHP, as depicted in Figure 1. The pupils' faces are captured in pictures using the Raspberry Pi camera. As indicated in figure1, connecting to a wifi network and having USB ports available for a monitor, keyboard, and mouse are essential for the project's effectiveness as an attendance system. Users are authenticated by the use of a camera module, which takes a picture of their face and compares it to photographs stored in the database. To handle student attendance online and keep track of individual student data, a website was created. To view and validate their participation on the They can use the password they previously set and their register number as their username. A single student's face is captured by the system from 300 various angles, and all pertinent data is stored in the database along with it.

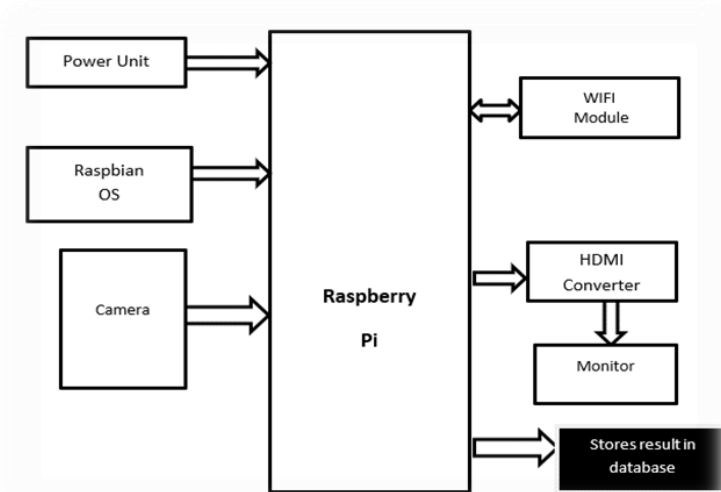


Fig 1. Block Diagram of proposed system

Hardware components

The controlled module of the system that uses a Raspberry Pi model 3 B is shown in Figure 2. In addition to 4GB of RAM and a 64-bit ARM Cortex A72 processor, the Pi 4 has a Video Core VI GPU for graphics processing applications (GPA). The Pi 4 also has 40 GPIO pins and two USB ports for attaching external electronics. The door locking/unlocking module made use of the GPIO pins to link external electronics. Python is the official programming language of Raspbian (ROS), the operating system that runs on Raspberry Pis. Raspbian (ROS) is specifically built to run Linux-based operating systems (LOS).



Fig 2. Raspberry Pi model 3 B

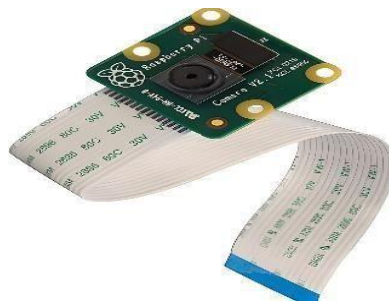


Fig 3. Raspberry Pi camera module

The Raspberry Pi Camera Module 3, a tiny camera made by Raspberry Pi, is shown in Figure 3. It sports a 12-megapixel IMX708 sensor with phase detection autofocus and HDR. Both normal and wide-angle versions of the Camera Module 3 are available, with or without an infrared cut filter. This camera supports up to 3 megapixels in HDR mode and can take still images and full HD videos. Its quick autofocus function and full support by the libcamera library make it easy to use for beginners while offering enough of features for more experienced users. Furthermore, all Raspberry Pi computers are compatible with Camera Module 3. The full hardware setup for this project is shown in Figure 4.

The Raspberry Pi camera connects to the Raspberry Pi through its USB ports. Raspberry Pi is designed to run a Linux-based operating system (LOS) and has its own operating system called Raspbian (ROS), which employs Python as its official programming language. The camera detects and identifies whether a person is authorized or not. The control system is responsible for relaying information regarding presence or absence, which is achieved using Python programming code. Flow diagram as shown in figure 5.

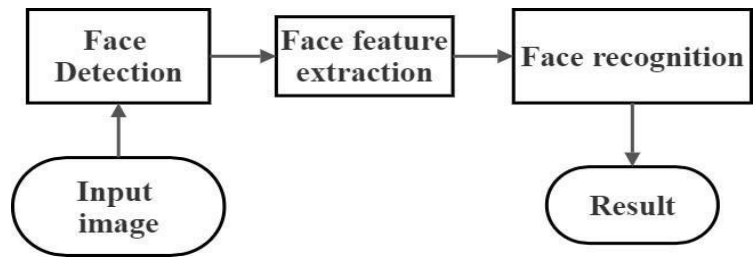


Fig 5. Flow diagram of proposed system

Face image dataset:

This repository contains a collection of 9,330 images of 30 students' faces as shown in Figure6. The images were captured using a camera and automatically cropped using the OpenCV face library. To improve recognition, the images were taken from different angles. To simplify processing and avoid complexity in the model, the detected color images were converted to grayscale.

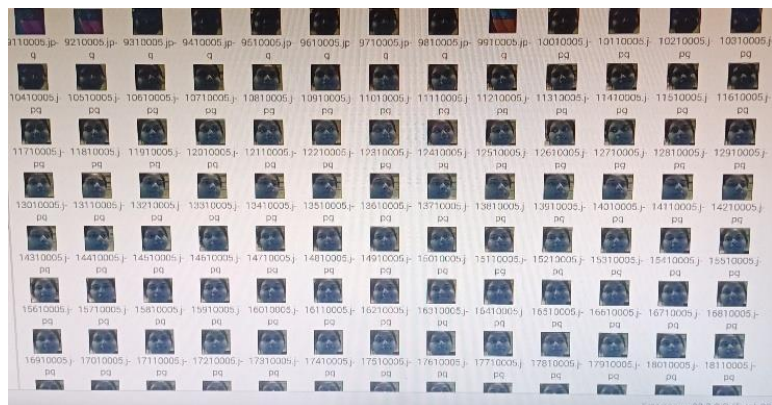


Fig 6. Dataset Collection

Pretrained CNN model:

CNNs, or convolutional neural networks, are neural networks commonly employed in tasks involving image and video recognition. They are specifically designed to learn spatial hierarchies of features from raw input data in an automatic and adaptive manner. Figure 7 represents about CNN model which is described below.

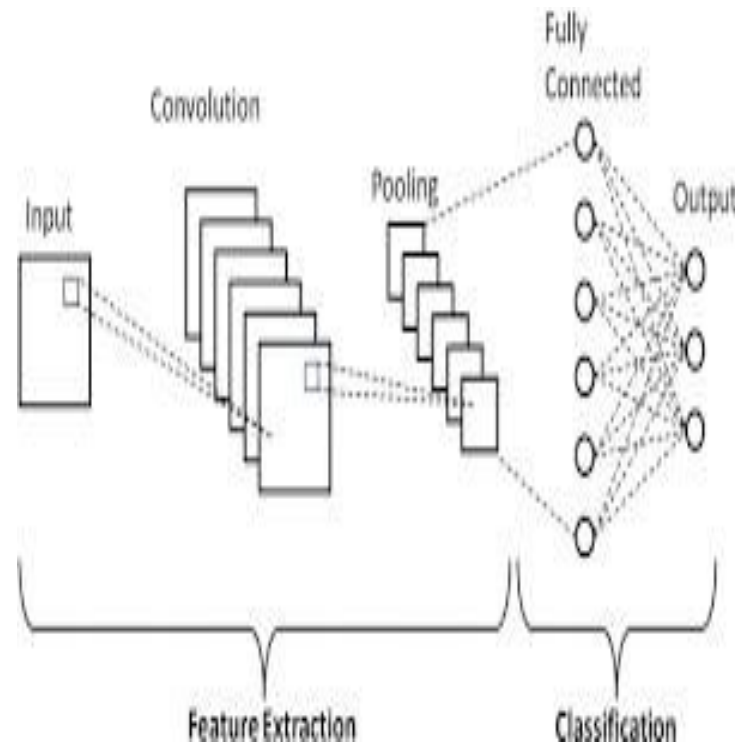


Fig 7. CNN Model

Feature extraction:

INPUT:Initially, the input image of a person is provided.

CONVOLUTION:Convolution is a mathematical process that enables the combination of two sets of information. In face comparison tasks using CNNs, a pair of faces is inputted independently for feature extraction. Both faces are subjected to the same filters, ensuring that the representation of a face remains constant regardless of the other face it is being compared to.

POOLING: Pooling is the step in which features are extracted from the image output of a convolutional layer. This involves reducing the dimensionality of the image by extracting its key features, and combining the resulting output of the previous layer into a single one.

Classification:

FULLY CONNECTED LAYER: After the pooling process, a fully connected layer is employed to predict the most suitable label for the given image. This involves flattening the output from the previous layers and feeding it into the FC layer. The flattened vector then undergoes additional FC layers where mathematical operations are performed. This stage marks the beginning of the classification process. Two fully connected layers are connected because it has been observed that this configuration performs better than using a single connected layer.

OUTPUT: Finally the classified and extracted images are given as output as shown in the Figure 8. In this case, the ResNet50 CNN, depicted in Figure 8, is utilized as a pre-trained model for extracting facial features and performing classification.

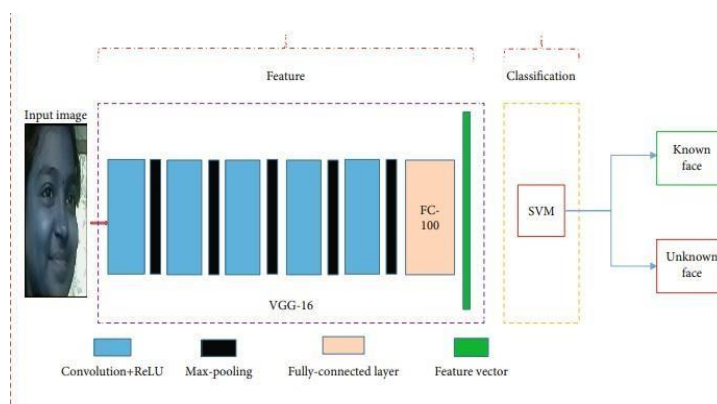


Fig 8. Extraction of images

Resnet-50 with SVM model

ResNet-50 is a deep convolutional neural network that can be used for feature extraction in face recognition systems. SVM is a machine learning algorithm that can be used for classification based on the extracted features. When combined, ResNet-50 with SVM can improve the accuracy of face recognition systems. Here is how ResNet-50 with SVM works on face recognition: Face Detection: First, the input image is passed through a Haar Cascade classifier or other face detection algorithms to identify the face region.

Feature Extraction: The face region is then passed through a ResNet-50 deep neural network, which extracts high-level features that are robust to variations in pose, lighting, and other factors. ResNet-50 is a pre-trained neural network that has been trained on a large dataset of faces, allowing it to learn powerful representations of facial features.

Feature Encoding: The extracted features are then encoded into a fixed-length feature vector that can be used for comparison and classification.

SVM Classification: The encoded features are passed through an SVM classifier, which learns to classify the face based on the extracted features. The SVM model can be trained using a set of labeled face images to learn to distinguish between different faces.

Face Recognition: During the recognition phase, a new face image is compared to the labeled face images in the dataset by passing the new image through the same pipeline. The face is first detected, then passed through the ResNet-50 for feature extraction, and finally through the SVM classifier for classification. The system then returns the identity of the closest match in the dataset based on the output of the SVM.

Support vector machines:

SVM (Support Vector Machines) is a machine learning algorithm that can be used for classification and regression analysis. In face recognition, SVM can be used to classify faces based on the extracted features. In face recognition, the input image is first passed through a facedetection algorithm to detect and extract the face region. Next, features are extracted from the face region using the technique Local Binary Patterns (LBP). These features are then used to

represent the face in a feature space, where they can be compared to other faces to perform recognition. SVM can be used to classify the face based on the extracted features. The SVM algorithm learns to separate the different classes of faces in the feature space using a hyperplane that maximizes the margin between the classes. The hyperplane is learned during the training phase of the SVM, where a set of labeled face images is used to train the classifier to distinguish between different faces. During the recognition phase, a new face image is compared to the labeled face images in the dataset by passing the new image through the same feature extraction pipeline. The features are then passed through the SVM classifier, which outputs the class label of the closest match in the dataset based on the distance between the feature vectors.

Train the model

Haarcascade classifier

Haar cascade classifiers are a type of object detection algorithm that can be used to detect objects in images or video, including faces. A Haar cascade classifier works by analyzing an image at different scales and sizes to identify features that are characteristic of the object being detected. In the case of face recognition, the algorithm looks for features like the eyes, nose, and mouth. The process begins with creating a "cascade" of classifiers, which are trained on positive and negative examples of the object being detected.

In the case of face detection, positive examples would be images of faces, while negative examples would be images without faces. The algorithm learns to identify the features that are most likely to be present in a face, while filtering out features that are not relevant. Once the cascade has been trained, it is applied to an image by scanning the image with a sliding window at different scales and sizes. At each location, the algorithm checks whether the features being detected are present. If enough features are present, the algorithm concludes that a face has been detected at that location.

Here's an explanation of how a Haar cascade classifier works for face detection:

Training Stage: During the training stage, the algorithm is trained on a large dataset of images containing faces (positive samples) and images not containing faces (negative samples).

Feature Extraction: The algorithm extracts features from each image in the dataset. Haar-like features are used to capture the differences in intensity between adjacent regions of an image.

Adaboost Learning: Adaboost is a machine learning algorithm that is used to combine multiple weak classifiers into a strong classifier. Each weak classifier is trained to detect a specific feature in the image, and Adaboost selects the best features to create the strong classifier.

Cascade of Classifiers: The final strong classifier is a cascade of multiple classifiers. Each classifier in the cascade is trained to detect a different set of features in the image. The cascade works by rejecting regions of the image that are unlikely to contain the object being detected, while passing promising regions on to the next classifier in the cascade.

Detection: During the detection stage, the cascade is applied to a new image. The algorithm scans the image at different scales and locations, and at each location, the features of the image are compared to the features in the cascade. If the features match those of a face, then the algorithm identifies the region as a face. Figure 9 illustrates how Haar-like features are used to extract facial features, such as intensity and length, from an image.



Fig 9. Haar Features of Face

Local binary pattern histogram

LBPH is an algorithm that is employed for face recognition purposes. Figure 10 displays the workflow of LBP Algorithm. The algorithm involves five steps, which are outlined below: LBP Parameters: The LBPH algorithm relies on four key parameters, namely, the radius, neighbors, grid X, and grid Y. The radius is utilized to construct a circular local binary pattern and refers to the distance around the central pixel. Neighbors denote the number of sample points utilized in creating the circular local binary pattern. Grid X and Grid Y represent the number of cells in the horizontal and vertical directions, respectively.

Radius: The algorithm calculates the radius of the face image by measuring the distance between the eyes, nose, jaws, lips, and forehead.

Radius: The algorithm calculates the radius of the face image by measuring the distance between the eyes, nose, jaws, lips, and forehead.

Neighbors: A graph is created by connecting the central points of facial features such as eyes, nose, lips, and cheeks.

Grid X: The width of the face is measured horizontally.

Grid Y: The height of the face is measured vertically.

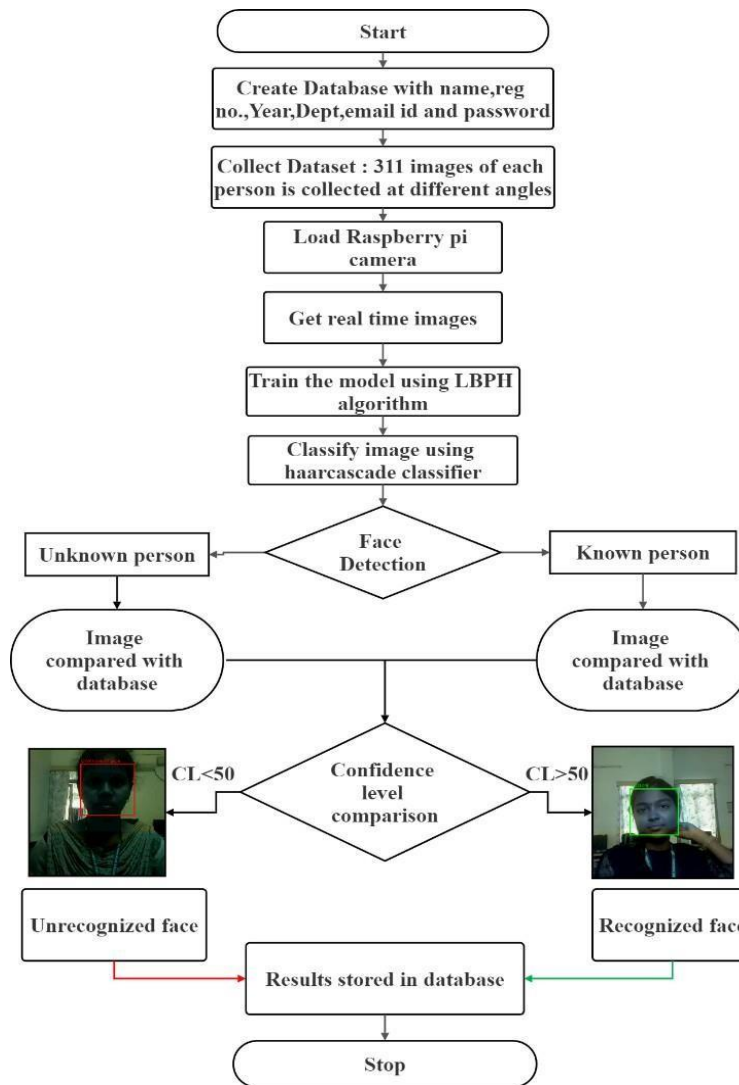


Fig 10. Flowchart of LBPH Algorithm

The LBPH algorithm operates in the following manner:

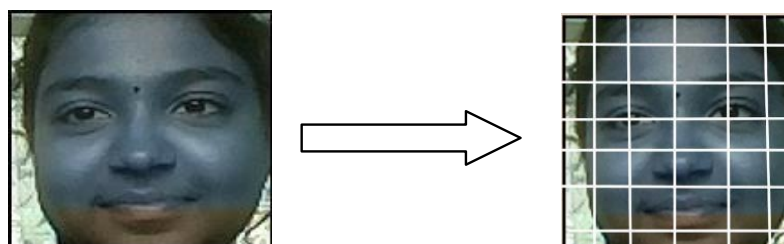
Radius: The algorithm calculates the radius of the face image by measuring the distance between the nose, eyes, jaws, lips, and forehead, which is subsequently noted.

Neighbors: A graph is generated by connecting the central points of the eyes, nose, lips, and cheeks on the face.

Grid X: The face's width is measured horizontally as shown in the figure 11.

Grid Y: The face's height is measured vertically as shown in the figure 11.

Fig 11. Representation of Grid X and Grid Y



Training the Algorithm: The first step is to train the algorithm using a dataset that contains images of the individuals we wish to recognize. Each image in the dataset must be assigned a unique identifier, which can be a number or name of the person. The algorithm utilizes this information to recognize an input image and produce an output. Images of the same person should have the same identifier. Once the training dataset is prepared, we can observe the computational steps involved in LBPH.

Applying the LBP operation: the initial computational stage involves applying the LBP operation to generate an intermediary image that accentuates the facial features and better represents the original image. This is achieved through the implementation of a sliding window approach that takes into account the radius and neighbors parameters.

Extracting the Histograms: With the image produced in the previous step, we can partition it into several grids by employing the Grid X and Grid Y parameters.

Performing the face recognition: At this stage, the algorithm has undergone training. The resulting histograms are utilized to represent each image from the training dataset.

LBPH Workflow

A camera module for capturing real-time images

The ability to capture and store a photograph and then compare it to an image in the database

If the confidence level (CL) assigned to the facial recognition process is equal to or greater than 50, the student is deemed present. Conversely, if the CL is below 50, the student is marked as absent.

The present and absent status is logged in the SQL database, which has been created and integrated into the website, as depicted in Figure 11.

Face recognition

The final stage of the process is face recognition. The camera captures real-time images and compares them to the dataset after the registered id of the individual has been provided. If the captured image corresponds to an image in the dataset, the person is marked as present, otherwise, they are marked as absent. The present and absent status can be viewed through the website.

6. Results

Our model has achieved success in detecting and recognizing the faces of individuals in real-time video frames. The Python code has been efficiently programmed and tested with Raspberry Pi model 3 B. The facial recognition process involves combining CNN with Local Binary Pattern Histogram (LBPH) to extract the relevant facial features. The attendance status of students is updated and marked as present through the website.

CONCLUSION

Face recognition using deep learning is a powerful technology that has made great strides in recent years. Deep learning models can identify and recognize faces with high accuracy, even in challenging scenarios such as low light, occlusions, and pose variations. To build a

face recognition system using deep learning, the first step is to gather a large dataset of face images and use it to train a deep neural network. This network can then be used to extract facial features from new images and compare them to those in the training set to recognize individuals. The combination of resnet, SVM and haar cascade techniques can improve the accuracy of face recognition systems. First, the Haar Cascade classifier can be used to identify the face region in the image, reducing the search space for subsequent processing steps. Then, the face region can be passed through a ResNet to extract features that are robust to variations in pose, lighting, and other factors. Finally, SVM can be used to classify the face based on the extracted features. Overall, face recognition using deep learning has tremendous potential for a wide range of applications, from security and surveillance to personalization and entertainment. As the technology continues to evolve, we can expect to see even more sophisticated and accurate face recognition systems in the future.

REFERENCES

- [1] Kennedy O kokpujie, Etinosa Noma-Osaghae, Olatunji J. Okesola, Samuel N. John, Okonigene Robert, International Conf on Computational Science and Computational Intelligence (2017).
- [2] Paul viola and Michael J. international J of computer vision, 57, 2, 137-154 (2004).
- [3] Rajesh Kumar Misra, Satyanrayan Padhy, Sandipan Pine, Prabhat Kumar Patnaik, N.Jeevaratnam, Indian J of Natural Sciences, 13 , 72 (2022).
- [4] Marian Stewart Bartlett, Member, Javier R. Movellan, Member, Terrence J. Sejnowsk, IEEE transactions on neural networks, 13, 6, (2002).
- [5] C. Havran, L. Hupet, J. Czyz, J. Lee, L. Vandendorpe, M. Verleysen, IEEE Trans on Neural Networks, 13, 6, (2002).
- [6] Shahrin Azuan Nazeer, Nazaruddin Omar, Marzuki Khalid, International J of Security and its Applications 10, 3, 81-100 (2016).
- [7] Thai Hoang Le, Advance in Artificial Neural System, 17 (2017).
- [8] K. B. Pranav, J.Manikandan, Third International Conf on Computing and Network Communications 25, (2018).
- [9] Neel Ramakant Borkar;Sonia Kuwelkar, International Conf on Computing Methodologies and Communication (ICCMC), (2018)
- [10] Zaid Alyasseri in International J of Computer Applications 126(3), 34-38(2015).

THE ROLE OF ARTIFICIAL NEURAL NETWORK IN THE PREPROCESS OF VARIOUS CANCER PREDICTION- A COMPARATIVE STUDY

Dr. SRINAGANYA G¹

Ms. SHOBANA Y²

¹Assistant Professor, ²Research scholar, Department of Computer Science, National College
(Autonomous)(Affiliated to Bharathidasan University) Trichy – 01

¹Email Id: srinaganyag78@gmail.com

²Email Id: rsspooja.vino@gmail.com

¹Ph: 9791409349

²Ph: 9444009832

Abstract

Artificial Intelligence (AI)-a rapidly growing and emerging technology has proved to have a wide range of applications in medicine and health care. Artificial intelligence has aided in the advancement of healthcare research. As a form of artificial intelligence, artificial neural networks (ANNs) have the advantages of adaptability, parallel processing capabilities, and non-linear processing. This technology serves the field of medicine and its updated features increases the advantages over the discipline of medicine. Especially in detecting many precarious and malignancy diseases, and cancer is one among an incurable perilous, a life threatening disease. A cancer is a group of cells that have grown out of control in the body. They can spread rapidly to any part of the body. The most common types of cancer are breast cancer, lung cancer, skin cancer and blood cancers (leukaemia and lymphoma). According to the report from World Health Organization (WHO), there are millions and millions of people suffering from the different type of cancer disease across the world, no matter what the age group is, and the survival rate is very low. Detecting cancerous cells and providing treatment at an early stage is quite difficult. But by the era of artificial intelligent and its techniques has given a new hope and scope among the physicians and healthcare decision-makers to decrease the risk rate and by the early stage they can handle the malignancy and disease up to the reduced cost. Neural networks (NN) are currently a burning research area in medical science, especially in the areas of cardiology, radiology, oncology, urology and etc. By using visual, biological, electronic health records, datasets and scan report data as the sole input source, pre-trained neural networks techniques and methods have been heavily employed for the identification of various malignancies. In this paper, we are surveying how the artificial neural network technologies are used for identifying and detecting classification of different types of cancer. The proposed work clarifies the importance of ANN and its role in predicting various cancer diseases and in terms of cost effective and user friendly system for processes and approaches for medical diagnostic.

Keywords: Artificial intelligence, Artificial neural network, Neural network, Prediction.

INTRODUCTION

The word cancer comes from the ancient Greek *kapkivoc*, which means crab and tumor. In Greek, these words refer to a crab, most likely applied to the disease because the finger-like spreading projections from a cancer called to mind the shape of a crab. The Roman physician, Celsus (25 BC - 50 AD), later translated the Greek term into cancer, the Latin word for crab. Cancer was introduced to the medical world in the 1600s and is associated with abnormally growing cells that can invade or spread to other parts of the body. The uncontrolled growth of cells starts from a site in the human body and further spreads to other

body parts known as cancer metastasis. Cancer cells are categorized into benign and malignant cells. The benign cells do not spread to other parts, while malignant cells metastasize and are considered more destructive. Due to high mortality and recurrence rate, its process of treatment is very long and costly. There is a need to accurately diagnose it early to enhance cancer patient's survival rate. It is a genetic disease triggered due to genetic mutations that control our cell's function, especially how they grow and divide. As the tumor cells continue to grow, additional changes will occur. In a nutshell, cancer cells have more genetic changes, such as mutations in DNA, than normal cells. Though the immune system generally discards damaged or abnormal cells from the body, few cancer cells can hide from the

immune system. The tumor also uses the immune system to grow and stay alive. The name of the cancer type is based on the site where tumor cells grow, for example, cancer that arises in the lungs and spreads to the liver is called lung cancer. Cancer diagnosis includes three predictive predictions related to cancer risk assessment, cancer recurrence, and cancer survivability prediction. Initially, the probability of cancer occurrence is assessed, followed by the second step, predicting cancer recurrence. The last step is to predict the aspects like progression, life expectancy, tumor drug sensitivity, survivability. Cancer detection has always been a challenge in the diagnosis and treatment plan for haematological diseases. Currently, an increased percentage of the population is affected by one or more diseases. Recent years have seen tremendous advances in medical science. Despite these advancements, there is still a huge lack of information among the public regarding health and disease. A large proportion of the population likely suffer from health issues, some of which may even be fatal. In addition to improving the accuracy of the rapid detection of fatal conditions, adopting safe, realistic techniques and using modern technology can reduce the need for caregivers and reduce overall health care costs. Several survival dramatically increase if detected early.

In recent years, Computer technology has made significant strides, leading to a surge of interest in the application of 'Artificial Intelligence (AI) in the fields of medicine and biological research. Among the various branches of AI, one that stands out for its extensive research and potential is 'Artificial Neural Networks (ANNs). Medical diagnosis using artificial intelligence (AI) systems, particularly artificial neural networks and computer-aided diagnosis with deep learning, is currently a very active research area in medicine and it is believed that it will be more widely used in biomedical systems. Evolving neural network techniques for medical diagnosis are broadly considered since they are ideal in recognizing diseases using scans. Neural networks learn by example so the details of how to recognize the disease are not needed. For instance, the utilization of ANN models aids in the timely diagnosis of various cancer with sensitivity and specificity. Advances in deep learning-based ANN models achieve efficacy, accuracy, and reliability in diagnosis. The utilization of other evolving technologies in this field is key for better diagnosis, having a significant impact on preventive measures and treatment. Despite the great benefits of novel technologies in healthcare, patients need protection from defective diagnoses to create a promising future in medical applications within society. Essentially, artificial neural networks (ANNs) are computer-generated mathematical algorithms that learn from standard data and extract the information embedded within. Through training, ANNs can mimic the behaviour of small biological neural clusters in a fundamental way. They serve as digital models of the human brain and have the ability to identify intricate nonlinear connections between dependent and independent variables, even in cases where the human brain may struggle to do so. Currently, ANNs are broadly utilized for medical applications in diverse disciplines of medicine mainly in cardiology. ANNs have been extensively applied in diagnosis, digital sign examination, medical picture examination and radiology. ANNs have been utilized by a few of writers for modelling in medicine and clinical study. Applications of ANNs are flourishing among medical information mining.

DIFFERENT TYPES OF CANCER AND ITS PREDICTION METHODS BY ANN

Yufang Yana et al.(2021) researched on cervical cancer patients and presented a two- stage 99+clinical events prediction model, in the first stage, k- means methods was used for mining the standardized diagnosis and treatment mode to judge the clinical event sequence to be tested is standard nor not, the RNN-2-DT, discrete time-variant based on word vector representation was integrated and in the second stage gated recurrent units (GRU) based on recurrent neural network (RNN) is applied into neural network to construct a prediction model. The treatment day vectors was used to mine standardized diagnosis and treatment mode while they train clinical event prediction model through treatment day vector stitched with time interval of the adjacent treatment day.

The experiment results indicated that RNN-2- DT has a brilliant prediction effect compared with

traditional models, and the mean average precision are increased by 7.2% and 4.3%, respectively.

Guozhen Chen et al. (2020) presented the research work on chronic kidney disease, a kidney cancer, presented the Adaptive Hybridized Deep Convolutional Neural Network (AHDCNN) for the early prediction and diagnosis of Chronic Kidney Disease (CKD). The distinctive subtypes of lesion are identified by using Deep learning system. The collected data will initially be analysed and the missing value will be replaced by the median value estimate. To identify kidney patterns, different features associated with the kidney disease are determined from the noise free data and fed in the classifier. By measuring the weight and bias value, the system trains feature in each hidden layer. Classification technology efficiency depends on the role of the data set. To enhance the accuracy of the classification system by reducing the feature dimension an algorithm model has been developed using CNN. The experimental process on the Internet of medical things platform (IoMT) concludes, with the aid of predictive analytics that advances in machine learning which provides a promising framework for the recognition of intelligent solutions to prove their predictive capability beyond the field of kidney disease.

The competent use of the learning and activation mechanism is a method of doubles- training to avoid kidney disease effectively. The study of regression and distribution of the data are then determined. The proposed approach is based on the method for deeper learning and ROIs (region of interest) given by radiologists has shown promising results in the classification of renal cell subtypes. Rabbia Mahum et al. (2023) presented their work on Lung cancer, one of the terrible diseases in various countries around the globe, and timely detection of the illness is still a challenging process. This study proposed a novel and efficient lung tumor detector based on a RetinaNet, namely Lung-RetinaNet. In this model CT scans was utilized for the training and testing of the model. A multi-scale feature fusion-based module is introduced to aggregate various network layers, simultaneously increasing the semantic information from the shallow prediction layer. This proposed model comprises a ResNet as a backbone network for extracting features from input along with dilated contextual block-based fusion, categorization and regression sub-networks. Due to an improved structure of RetinaNet, it precisely detects tiny lung tumours. The proposed methodology attained 99.8% accuracy, 99.3% recall, 99.4% precision, 99.5% F1-score, and 0.989 Auc than the DL-based methods by this outcomes the proposed system can be utilized by medical experts to identify the tumor at early stages.

Wahidur Rahman et al. (2023) proposed the work on multiclass blood cancer like Breast cancer, lung cancer, skin cancer, and blood malignancies such as leukemia and lymphoma. The proposed research pipeline is occupied into some interconnected parts like dataset building, feature extraction with pre-trained Convolutional Neural Network (CNN) architectures from each individual image of blood cells, and classification with the conventional classifiers. The dataset for this study is divided into two identical categories, Benign and Malignant, and then reshaped into four significant classes, each with three subtypes of malignant, namely, Benign, Early Pre-B, Pre-B, and Pro-B. The research first extracts the features from the individual images with CNN models and then transfers the extracted features to the features selections such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and SVC Feature Selectors along with two nature inspired algorithms like Particle Swarm Optimization (PSO) and Cat Swarm Optimization (CSO). After that, research has applied the seven Machine Learning classifiers to accomplish the multi-class malignant classification. The study discovered a maximum accuracy of 98.43% when solely using pre-trained CNN and classifiers and after incorporating PSO and CSO, the proposed model achieved the highest accuracy of 99.84% by integrating the ResNet50 CNN architecture, SVC feature selector, and LR classifiers.

Haitham Elwahsh et al. (2023), this work was investigated on predicting cancer using deep neural learning.

The study proposed that deep neural learning cancer prediction model (DNLC) has the following stages. In the first stage, Deep Network (DN) is used to select the best collection of features from datasets. In the second stage, train the genomic or clinical data samples with a deep neural network

(DNN). In the third stage, evaluate the capabilities of the DNLC model of predicting cancer in its earlier stages. For classification, DNLC uses five cancer datasets, which are for colon, lung adenocarcinoma, and squamous cell carcinoma, breast, and leukaemia cancers. The five cancer datasets are used in experiments to predict how well the suggested model will perform. The dataset is divided into two parts: training sets, which make up 80% of the dataset, and testing sets, which make up 20%. The experimental results shows that the DNLC technique, with an average accuracy of 93%, performs better in terms of accuracy than other methods in all circumstances.

Dewi Nasien et al. (2022), explored the use of ANN in the research on breast cancer and the research has been widely conducted and previously studied with various methods or algorithms to categorize it into benign and malignant groups. In ANN algorithm, one method called back propagation network is utilized to solve complex problems related to identification, pattern recognition prediction, and so forth. The objective of the present study is to investigate the level of accuracy and performance by ANN back propagation in predicting breast cancer. Several stages for this study are formulating the problem, collecting and processing the Wisconsin breast cancer dataset from the Kaggle site. Designing and creating an ANN algorithm system to classify cancer into malignant and benign, then examining the system to perceive the prediction accuracy, and conclude it. The results of the numerical simulation indicate that the created system of MATLAB R2016a software obtained an accuracy of 96.929% with an error of 3.071% by a combination of training parameters with epoch 1000, learning rate 0.01, goal 0.001, and hidden layer 5.

Sugiart et al. (2018), worked on skin cancer and proposed that this study aims to present diagnose of melanoma skin cancer at an early stage. It applies feature extraction method of the first order for feature extraction based on texture in order to get high degree of accuracy with method of classification using artificial neural network (ANN). The method used is training and testing phases with classification of Multilayer Perceptron (MLP) neural network. The results showed that the accuracy of test image with 4 sets of training for image not suspected of melanoma and melanoma with the lowest accuracy of 80% and the highest accuracy of 88,88%, respectively. The 4 sets of training used consisted of 23 images. Of the 23 images used as a training consisted of 6 as not suspected of melanoma images and 17 as suspected melanoma images.

Mridha k et al. (2023) presented that skin cancer is a prevalent form of malignancy around the globe. Clinical evaluation of skin lesions is essential, but it faces challenges such as long waiting times and subjective interpretations. Deep learning techniques have been developed to tackle these challenges and assist dermatologists in making more accurate diagnoses. Prompt treatment of skin cancer is vital to prevent its progression and potentially life-threatening consequences. The use of deep learning algorithms can improve the speed and accuracy of diagnosis, leading to earlier detection and treatment. The goal of this study was to develop reliable deep learning (DL) prediction models for skin cancer classification; (i) deal with a typical severe class imbalance problem, which arises because the skin-affected patients' class is significantly smaller than the healthy class; and (ii) interpret the model output to better understand the decision-making mechanism (iii) Propose an End-to-End smart healthcare system through an android application. In a comparison examination with six well-known classifiers, the effectiveness of the proposed DL technique was explored in terms of metrics relating to both generalization capability and classification accuracy. A study used the HAM10000 dataset and an optimized CNN to identify the seven forms of skin cancer. The model was trained using two optimization functions (Adam and RMSprop) and three activation functions (Relu, Swish, and Tanh). Furthermore, an XAI-based skin lesion classification system was developed, incorporating Grad-CAM and Grad-CAM++ to explain the model's decisions. This system can help doctors make informed skin cancer diagnoses in their early stages, with an 82% classification accuracy and 0.47% loss accuracy.

M. Vania et al. (2023) proposed the work on gastrointestinal cancer, technical techniques is required to appreciate the scientific quality and novelty of AI studies for this application. Clinicians frequently lack this technical background, and AI experts may be unaware of such clinical relevance and implications in daily practice. As a result, there is a growing need for a multidisciplinary, international assessment of how to conduct high-quality AI research in upper GI malignancy detection. This research will help gastroenterologist build approaches or models to increase diagnosis accuracy for upper GI (Gastrointestinal) malignancies despite variances in experience, education, personnel, and resources, as it offers real-time and surveying chances to improve upper GI malignancy diagnosis and screening. The comprehensive review sheds light on potential enhancements to computer-aided diagnostic (CAD) systems for GI endoscopy. The findings of study suggest that Support Vector Machines (SVM) is frequently utilized in gastrointestinal (GI) image processing within the context of machine learning (ML). The analysis reveals that CNN-based supervised learning object detection models are widely employed in GI image analysis within the deep learning (DL) context. The results of this study also suggest that RGB is the most commonly used image modality for GI analysis, with color playing a vital role in detecting bleeding locations.

Zhencun Jiang et al. (2021) proposed a ViT- CNN ensemble model to classify cancer cells and normal cells for the diagnosis of Acute Lymphoblastic Leukemia. This ensemble model combined the vision transformer model and convolutional neural network (CNN) model to extract features from cells images in two different ways, also used a data enhancement method and a symmetric cross-entropy loss function to reduce the impact of noise in the data set. The ViT-CNN ensemble model achieved ahigh classification accuracy of 99.03% on the test set, which was better than other models. The authors suggested that their method could be an effective tool for computer- aided diagnosis of ALL by accurately distinguishing between cancer cells and normal cells.

M. Akter Hossain et al. (2020) conducted a study focused on Acute Lymphocytic Leukemia (ALL), which is the most common form of leukemia. The study proposed a practical technique for detecting abnormal blood components in cancer patients, including neutrophils, eosinophils, basophils, lymphocytes, and monocytes. The researchers used 14 features to construct a dataset before selecting four key attributes that are crucial in determining whether a patient has Leukemia, also collected 256 primary data from Leukemia patient. The data is then processed using microscope to obtain images and fetch into Faster-RCNN machine learning algorithm to predict the odds of cancer cells forming, it was also applied two loss functions to both the RPN (Region Convolutional Neural Network) model and the classifier model to detect the similar blood object. After identifying the object, calculated the corresponding object and based on the count of the corresponding object, finally Leukemia was detected. The mean average precision observed are 0.10, 0.16 and 0, where the epochs are 40, 60 and 120. The goal of the study was to aid in the early detection of cancer, which can significantly improve treatment outcomes.

CONCLUSION AND FUTURE SCOPE

Artificial neural networks extract patterns and make predictions from large datasets. The increasing usage of neural network in healthcare, together with the availability of highly characterised cancer datasets, has led to acceleration in research into the utility of deep learning in the analysis of the complex biology of cancer. ANN techniques provide different probabilistic and statistical approaches that enable intelligent computers to recognise and identify patterns in datasets based on repeated prior experiences. This research paper highlights the evolving ANN technology in cancer diseases by applying algorithms to cancer datasets. The study demonstrated the effectiveness of techniques in predicting and detecting different type of cancers with high accuracy rates. In terms of future scope, further work can enhance accuracy rates by using more comprehensive datasets, optimizing the platform where the system will be launched, and making it more user-friendly with extensive information about

the disease.

REFERENCES

- [1] Yufang Yan, Kui Zhao, Jilong Cao, Huimin Ma, "Prediction research of cervical cancer clinical events based on recurrent neural network", *Procedia Computer Science*, Volume 183, 2021, Pages 221- 229, ISSN 1877-509, <https://doi.org/10.1016/j.procs.2021.02.05>
- [2] Guozhen Chen, Chenguang Ding, Yang Li, Xiaojun Hu, Xiao Li, Li Ren, Xiaoming Ding, Puxun Tian, and Wujun, "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," in *IEEE Access*, vol. 8, pp. 100497- 100508, 2020, doi:10.1109/ACCESS.2020.2995310.
- [4] R. Mahum and A. S. Al-Salman, "Lung- RetinaNet: Lung Cancer Detection Using a RetinaNet With Multi-Scale Feature Fusion and Context Module," in *IEEE Access*, vol. 11, pp. 53850-53861, 2023, doi:10.1109/ACCESS.2023.3281259.
- [5] Rahman, Md & Faruque, Mohammad & Roksana, Kaniz & Sadi, A.H.M. Saifullah & Rahman, Mohammad Motiur & Azad, Dr. (2023). Multiclass blood cancer classification using deep CNN with optimized features. *Array*. 18. 100292.10.1016/j.array.2023.100292.
- [6] Haitham Elwahsh, Medhat A. Tawfeek, A.A. Abd El-Aziz, Mahmood A. Mahmood,
- [7] Maazen Alsabaan, Engy El-shafeiy, "A new approach for cancer prediction based on deep neural learning", *Journal of King Saud University - Computer and Information Sciences*, Volume 35, Issue 6, 2023, 101565, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2023.101565>.
- [9] Nasien, Dewi, Enjeslina, Veren, Hasmil Adiya, M., Baharum, Zirawani, "Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method" *Journal of Physics: Conference Series*, 2022, IOP Publishing, doi: 10.1088/1742-6596/2319/1/012025,
- [10] <https://dx.doi.org/10.1088/1742-6596/2319/1/012025>.
- [11] Anagun, Y. Smart brain tumor diagnosis system utilizing deep convolutional neural networks. *Multimed Tools Appl* 82, 44527–44553 (2023)
- [12] <https://doi.org/10.1007/s11042-023-15422-w> Sugiarti, Sugiarti & Yuhandri, Yuhandri & Naam, Jufriadif & Indra, Dolly & Santony, Julius, (2019), "An artificial neural network approach for detecting skin cancer", *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 17. 788. 10.12928/telkomnika.v17i2.9547. doi:10.12928/telkomnika.v17i2.9547
- [13] K. Mridha, M. M. Uddin, J. Shin, S. Khadka and M. F. Mridha, "An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System," in *IEEE Access*, vol. 11, pp. 41003-41018, 2023, doi: 10.1109/ACCESS.2023.3269694.
- [15] M. Vania, B. A. Tama, H. Maulahela and
- [16] S. Lim, "Recent Advances in Applying Machine Learning and Deep Learning to Detect Upper Gastrointestinal Tract Lesions," in *IEEE Access*, vol. 11, pp. 66544-66567, 2023, doi: 10.1109/ACCESS.2023.3290997.

SOCIAL IMPACT OF AI : AN OVERVIEW

Dr. S. Arulselvarani¹ M.Saranya²

¹Assistant Professor, ²Research Scholar, Department of Computer Science, Urmu Dhanalakshmi College, Tiruchirapalli - 620 019. Tiruchirapalli – 620019.

ABSTRACT

This abstract explores the widespread implementation of Artificial Intelligence (AI) techniques across various domains, examining the transformative impact on the blooming field of society. The rapid advancement of AI technologies has permeated diverse sectors, contributing to unprecedented growth, innovation, and societal evolution. The paper delves into the evolution of AI techniques, emphasizing their pivotal role in reshaping industries and daily life. It scrutinizes key AI methodologies, such as Machine learning, deep learning, reinforcement learning, and transfer learning, and highlights their applications in automation, decision-making, and problem-solving. The rise of AI-powered personal assistants, intelligent systems, and autonomous devices is analyzed, showcasing their influence on individual lifestyles and organizational work flows. This paper explores the multifaceted integration of AI in various sectors, examining its impact on healthcare, education, business, and the broader social landscape. Emphasis is placed on the ethical considerations surrounding AI, with discussions on bias mitigation, privacy preservation, and the need for responsible AI deployment. The pivotal role of AI in addressing global challenges, such as climate change and healthcare disparities, showcasing its potential for positive societal impact. As AI becomes increasingly intertwined with daily life, understanding its increasing applications and implications is essential for navigating the evolving landscape of the modern society. **Keywords:** AI, Machine learning, deep learning, reinforcement learning, automation, decision-making, and problem- solving.

INTRODUCTION

In recent years, the rapid advancement of Artificial Intelligence (AI) has significantly reshaped the socio-economic landscape, ushering in an era of unprecedented possibilities and challenges. The social impact of AI is a multifaceted phenomenon that touches upon various aspects of human life, from employment and education to healthcare and governance. One of the most prominent areas of impact is the workforce. Automation driven by AI has led to the transformation of industries, altering job structures and creating new opportunities. In the realm of healthcare, AI has shown promise in revolutionizing diagnostics, personalized medicine, and drug discovery. However, ethical considerations, data privacy concerns, and the potential for bias in AI algorithms raise important questions about the responsible and equitable deployment of these technologies. Education is another domain where AI is making inroads, offering personalized learning experiences and educational tools. Yet, questions regarding access to AI-powered education, the role of human teachers, and the potential reinforcement of existing socio-economic disparities must be addressed. AI's impact on social dynamics and governance is also profound. From facial recognition technologies to predictive policing algorithms, there are concerns about privacy infringement, surveillance, and the potential amplification of societal biases. The ethical implications of AI in decision-making processes, including those related to criminal justice and public policy, require careful consideration.

LITERATURE REVIEW

The social impact of AI serves to synthesize existing research and scholarship pertaining to the multifaceted interactions between AI technologies and society. (Walugembefrancis ,2023) It encompasses an exploration of historical perspectives, economic ramifications, educational transformations, healthcare advancements, considerations of social justice and equity, ethical and legal frameworks, public perceptions and attitudes, and future research directions. By systematically analyzing a wide array of academic works, (Shuiwen Liu 2022) this review aims to provide a holistic understanding of how AI shapes and is shaped by social dynamics, highlighting both the opportunities for innovation and the challenges for ensuring responsible and equitable deployment. Through this synthesis, scholars can identify key trends, gaps in knowledge, and areas for further investigation, thus contributing to the ongoing discourse on the societal implications of AI. (Hirsch-Kreinsen ,2023) Scholars have explored the economic ramifications of AI adoption, including shifts in employment patterns, skill requirements, and income distribution. Additionally, (William Knight et.al,2022) research has delved into AI's implications for education and workforce development, highlighting its potential to revolutionize learning experiences and training methodologies. In healthcare, AI offers promising opportunities for medical diagnosis, treatment planning, and patient care, but also raises concerns regarding privacy, bias, and patient autonomy. studies have scrutinized AI's role in perpetuating or mitigating social inequalities, emphasizing the importance of ethical and equitable AI practices. As public awareness and perceptions of AI evolve, there is a growing need for interdisciplinary research, policy interventions, and public engagement efforts to ensure that AI technologies serve the collective well-being and uphold societal values. In education, attention is directed toward AI's role in personalized learning and workforce development, assessing its efficacy in improving educational outcomes. The healthcare sector witnesses the integration of AI in diagnostics, treatment planning, and patient care, raising ethical considerations surrounding privacy and algorithmic biases. (Wanjiku, Winter- Gladys December 2020) A critical lens on AI's role in perpetuating or alleviating social inequalities highlights concerns about biases in algorithms and accessibility disparities. This literature review synthesizes these diverse strands of research, presenting a holistic overview of the social impact of AI while underscoring the need for responsible and equitable integration into our societal fabric.

METHODOLOGY

Data Collection

Collect qualitative and quantitative data from the references such as data analysis of existing datasets, surveys, interviews, social media,. Ensure data representativeness across demographics and socio-economic backgrounds.

Impact Assessment

Analyze direct and indirect impacts of AI on society, considering areas like employment, education, healthcare, privacy, and inequality. Use established frameworks, such as the UN Sustainable Development Goals, to structure the assessment.

Ethical Analysis

Evaluate ethical implications, including bias, fairness, transparency, accountability, and human rights. Apply ethical frameworks, such as the Ethically Aligned Design, to guide the analysis.

Risk Assessment

Identify potential risks associated with AI, such as job displacement, bias, and privacy violations. Assess the likelihood and severity of these risks and propose mitigation strategies.

Engagement and Consultation

Engage stakeholders through workshops, focus groups, public forums, and online platforms .Gather diverse

perspectives to enhance the analysis and ensure inclusivity.

Scenario Planning and Policy Recommendations

Develop scenarios to explore alternative futures based on different AI adoption trajectories. Evaluate the implications of each scenario on social dynamics and individual well-being. Formulate evidence-based policy recommendations to maximize

TABLE1: Statics view of AI in Various Aspects

positive impacts and address risks and ethical concerns. Advocate for responsible AI development and deployment.

Monitoring and Evaluation

Establish mechanisms for ongoing monitoring and evaluation, incorporating key performance indicators and feedback loops. Iterate the methodology based on new insights and evolving societal needs.

Aspect	Statics
AI Adoption Trends	20% adoption of AI technologies in various sectors.
Economic Impact	15% in productivity in industries incorporating AI
Social and Cultural Effects	10% of reliance on AI for daily tasks and decision-making.
Education and Skill Development	25% of Integration of AI into education systems and training programs
Ethical Considerations	12% of survey respondents expressing ethical concerns about AI.
Challenges and Barriers	18% of businesses citing cost as a primary barrier to implementing AI.

AI techniques in society have led to improved healthcare diagnostics, personalized education, and optimized urban planning. They enhance efficiency and the potential for algorithmic bias necessitate careful regulation and responsible implementation. Overall, AI is transforming various sectors, offering benefits and requiring a balanced approach for societal well-being.

INTERPRETATION AND DISCUSSION

TABLE2: Analysis And Considerations Of AI in Future Impact

Discussion Point	Analysis	Considerations
Comparison with Developed Societies	Contrast adoption rates, challenges, and outcomes between emerging and developed societies.	Consider the potential for leapfrogging in emerging societies' AI adoption.
Benefits and Opportunities	Explore economic growth, innovation, and societal benefits resulting from AI.	Opportunities for job creation and entrepreneurship in the AI sector.
Risks and Mitigation Strategies	Discuss risks associated with bias, privacy, and accountability in AI applications.	Propose strategies for responsible AI deployment and regulation.
Policy Implications	Examine the role of government policies in fostering a conducive environment for AI development.	Consider the need for agile and adaptive regulatory frameworks.
Future Outlook	Predict potential advancements and challenges in the future of AI in emerging societies.	Emphasize the importance of ongoing research, development, and collaboration in shaping this future.
Limitations of the Study	Acknowledge methodological limitations and potential biases in the research.	Suggest areas for future research to address these limitations and provide a more comprehensive understanding.

CONCLUSION

AI technologies have the potential to bring about significant advancements and improvements in various aspects of society, they also raise ethical, legal, and societal concerns that need to be carefully addressed. Overall, the overview emphasizes the importance of conducting thorough research and analysis to understand the implications of AI deployment fully. It underscores the need for evidence-based policymaking, stakeholder engagement, and ongoing monitoring and evaluation to maximize the benefits of AI while mitigating risks and ensuring ethical considerations are upheld. In conclusion, the highlights the necessity of a balanced and comprehensive approach to studying the social impact of AI, one that considers diverse perspectives, engages stakeholders, and fosters responsible innovation to create a future where AI serves the collective well-being of society.

REFERENCES

- [1] The Future of Artificial Intelligence and Its Impact on Society Author(s):Walugembefrancis Author Affiliation(s): Nkunba University march 2023
- [2] https://www.researchgate.net/publication/369147005_The_Future_of_Artificial_Intelligence_and_Its_Impact_on_Society
- [3] The Research on Artificial Intelligence in Computer Network Technology ShuiwenLiu AffiliationNanchang Vocational University,Nanchang,China IEEE *Xplore*: 28February2022DOI: [10.1109/CISAI54367.2021.00130](https://doi.org/10.1109/CISAI54367.2021.00130)
- [4] Artificial Intelligence and the Future of Teaching and Learning Miguel A. Cardona, Ed.D. Secretary, U.S. Department of Education Roberto J. Rodríguez Assistant Secretary, Office of Planning, Evaluation, and Policy Development Kristina Ishmael Deputy Director, Office of Educational Technology May 2023 <https://tech.ed.gov>
- [5] Artificial intelligence : a “promising Technology H Hirsch-Kreinsen · 2023 · <https://link.springer.com/article/10.1007/s00146-023-01629-w>
- [6] Artificial intelligence in science: An emerging general method of invention Stefano Bianchini* ,Moritz Müller, Pierre Pelletier,2022<https://doi.org/10.1016/j.respol>
- [7] .2022.104604
- [8] Artificial intelligence in higher education: the state of the field [Helen Crompton](#) & [Diane Burke](#) *International Journal of Educational Technology in Higher Education* volume 20, Article number: 22 (2023) <https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-023-00392-8>
- [9] Impact of artificial intelligence on human loss in decision making, laziness and safety in education, [SayedFayaz Ahmad](#), [Heesup Han](#), [Muhammad MansoorAlam](#), [Mohd.KhairulRehmat](#), [Muhammad](#)
- [10] [Irshad](#), [Marcelo ArrañoMuñoz](#) & [AntonioAriza-Montes](#) 2023 June<https://doi.org/10.1057/s41599-023-01787-8>
- [11] Defining artificial intelligence for librarians [AndrewM. Cox](#) and [Suvodeep Mazumdar](#) December2022 <https://doi.org/10.1177/09610006221142029>
- [12] Framing governance for a contestedemerging technology:insights from AI policy William Knight, Tonii Leach, [BerndCarsten Stahl](#), Wanjiku, Winter-Gladys December2020<https://doi.org/10.1080/14494035.2020.1855800>
- [13] Artificial intelligence in informationsystems research: A systematic literaturereview and research agenda Christopher Collins ^a, Denis Dennehy ^a, Kier an Conboy ^a, Patrick Mikalef ^b October 2021 <https://soj.org/10.1016/j.ijinfor.2021.102383>

THE FUTURE OF SURVEY ANALYSIS: AI AUTOMATION IN MARKETING

R.THARANESWARI,

Assistant Professor, Department of Computer Science, UrumuDhanalakshmi College, Tiruchirapalli - 620019.
Mobile: 9965019186 udccstharaneswari@gmail.com

ABSTRACT

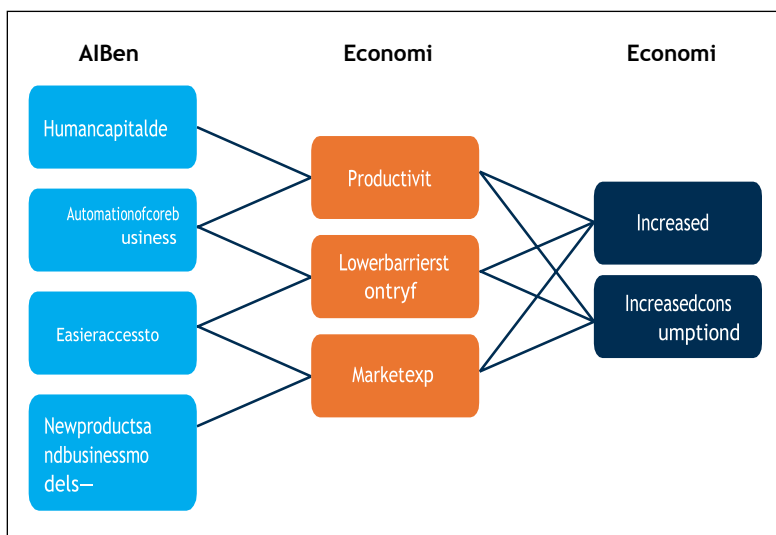
As the landscape of marketing continues to evolve rapidly, the need for efficient and insightful analysis of marketing surveys becomes paramount. This paper presents an innovative approach to streamline the analysis of marketing surveys through the integration of artificial intelligence (AI) technologies. This methodology involves preprocessing textual survey responses, categorizing sentiments, and identifying key themes using state-of-the-art NLP models. Additionally, here employ ML algorithms to analyze demographic data and discern patterns, enabling a deeper understanding of customer preferences and behaviors. The goal is to expedite the traditionally time-consuming process of survey analysis while enhancing the accuracy and granularity of the results. This paper highlights the integration of sentiment analysis to gauge customer satisfaction and sentiment trends over time. By automating the identification of positive and negative sentiments, businesses can promptly address issues, capitalize on strengths, and adapt marketing strategies to align with consumer sentiments. The study evaluates the effectiveness of the automated analysis by comparing the results with traditional manual methods. The implications of AI-driven analysis on the speed, accuracy, and depth of insights are discussed, highlighting the potential for marketers to make data-driven decisions more rapidly and effectively. Furthermore, ethical considerations surrounding the use of AI in survey analysis are addressed, emphasizing transparency, privacy, and fairness. The study aims to contribute to the growing body of literature on the intersection of AI and marketing research, providing valuable insights for both academia and industry practitioners.

KEYWORDS: AI, ML, NLP, AI driven analysis, demographic data.

INTRODUCTION

In today's fast-paced and dynamic business environment, staying ahead of the competition requires a keen understanding of market trends, consumer behavior, and effective decision-making. Traditional market analysis methods often fall short in providing real-time insights and adapting to the rapidly evolving landscape. Enter Artificial Intelligence (AI), a transformative force revolutionizing how businesses approach market analysis. AI technologies play a pivotal role in advancing economic development through various channels. One key channel is the enhancement of productivity and efficiency across industries. AI-driven automation and optimization processes streamline workflows, reduce operational costs, and boost overall productivity. Additionally, AI fosters innovation by enabling the development of new products, services, and business models. In sectors such as healthcare, finance, and manufacturing, AI applications contribute to breakthroughs in research, personalized services, and advanced manufacturing processes. Furthermore, AI supports workforce development by creating demand for new skill sets, encouraging continuous learning and upskilling. This transformative technology acts as a catalyst for economic growth by driving competitiveness, fostering innovation, and shaping a more dynamic and resilient global economy.

Fig1: Channels to Economic Development Supported by AI Technologies



LITERATURE REVIEW

Abid Haleem et al., (2022) The automatic analysis of market surveys has gained prominence in recent years, driven by the escalating volume and complexity of data generated through consumer feedback. Historically, market surveys relied on manual analysis methods, limiting scalability and efficiency. However, with the advent of advanced technologies, particularly natural language processing (NLP), machine learning, and data analytics, there has been a transformative shift toward automation. Research has increasingly leveraged these technological foundations to develop sophisticated algorithms capable of extracting valuable insights from large textual datasets, addressing challenges posed by the unstructured nature of survey responses.

Ming-Hui Huang et al., (2020) Text analysis and sentiment mining have emerged as pivotal components in the automatic analysis of market surveys. NLP techniques play a crucial role in deciphering the nuances of language, enabling the extraction of meaningful patterns and sentiments from textual survey data. Sentiment analysis, in particular, has become a key tool for gauging consumer sentiments and opinions, providing businesses with valuable information for strategic decision-making. Roberto Moro-Visconti et al., (2022) Machine learning models represent another critical aspect of the literature on automatic market survey analysis. Recent studies have explored the application of diverse machine learning algorithms to predict market trends based on survey data. The evolution of these models, from traditional approaches to more advanced techniques, has significantly enhanced the accuracy and predictive capabilities of automated analyses. Understanding the strengths and limitations of different models is essential for ensuring robust and reliable outcomes.

Sanjeev Verma, Rohit Sharma, Subhamay Deb, Debojit Maitra et al., (2022) Efficient data preprocessing and cleaning mechanisms are fundamental to the success of automatic market survey analysis. Researchers have investigated methodologies to enhance the quality of survey data, addressing challenges such as missing data and outliers. Furthermore, feature selection and dimensionality reduction techniques have been explored to improve the efficiency and interpretability of automated analyses, demonstrating the importance of selecting relevant features for model performance.

Roberto Moro-Visconti et al., 2023 The integration of automatic market survey analysis with Business Intelligence (BI) systems has garnered attention as organizations seek to derive actionable insights from survey data. Successful implementations showcase the potential of seamlessly combining automated analysis with BI tools, providing decision-makers with real-time, data-driven intelligence. Case studies in various industries underscore the practical applications and benefits of this integration. Sri Krishna Chintalapati and Shivendra Kumara Pandey et al., 2022 Despite the advancements, challenges persist in the field of automatic market survey analysis. Issues related to bias in data, interpretability of models, and ethical considerations necessitate careful examination. Future directions in research are anticipated to address these challenges and explore emerging trends, including the incorporation of deep learning, blockchain for data security, and advancements in data visualization. A comprehensive understanding of the existing literature not only highlights the current state of automatic market survey analysis but also informs potential avenues for future exploration in this dynamic field.

METHODOLOGY

Automating marketing analysis using artificial intelligence (AI) involves leveraging various methodologies and techniques to extract insights, make predictions, and optimize marketing strategies.

Machine Learning Algorithms

Supervised Learning Use labeled data to train models for tasks such as customer segmentation, churn prediction, and lead scoring. **Unsupervised Learning** Employ clustering algorithms for market segmentation, anomaly detection for identifying unusual patterns in data, and association rule mining for discovering relationships between different variables.

Natural Language Processing (NLP)

Analyze customer sentiment through social media, customer reviews, and forums to understand how people feel about products or services. Use chatbots for customer interactions, lead generation, and customer support.

Predictive Analytics

Utilize predictive models to forecast trends, customer behavior, and sales performance. Apply time-series analysis for predicting future market conditions based on historical data.

Customer Segmentation and Fraud Detection

Implement clustering algorithms to group customers based on demographics, behavior, and preferences. Use these segments to tailor marketing strategies and personalize communication.

Implement anomaly detection algorithms to identify fraudulent activities in digital marketing, such as click fraud or fake account creation.

Marketing Automation Platforms and Data Integration

Integrate AI into marketing automation platforms for tasks like lead scoring, email personalization, and content optimization.

Integrate data from various sources to create a comprehensive view of customer behavior. Utilize techniques like data fusion to merge heterogeneous data for more accurate analysis.

Dynamic Pricing Models and Continuous Learning

Implement AI-driven dynamic pricing strategies based on real-time market conditions, competitor pricing, and customer behavior. Implement systems that can adapt and learn from new data, ensuring ongoing improvement in marketing strategies.

Fig 2:



Table1 Sentiment Scoring

Sentiment	Score
Positive	2
Neutral	1
Negative	0

Sentiment Analysis Formula

To calculate an overall sentiment score for a setof responses
 Average Score= (Sum of Individual Scores) /(number of responses)

Table2 Survey Responses

Calculations

Positive Response:

Score: 2

Neutral Response:

Score: 1

Negative Response:

Score: 0

Mixed Response:

Score: (2 + 0) / 2 = 1

Average Sentiment Score:

$$(2+1+0+1) / 4 = 1$$

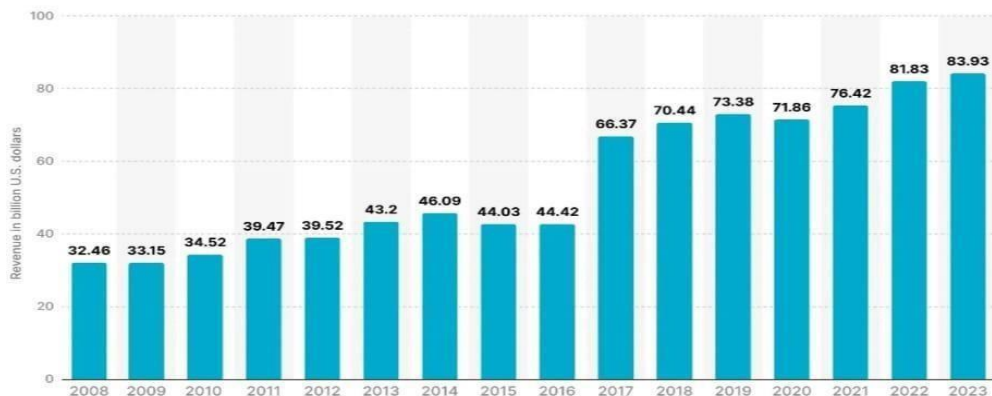
In the above table, the average sentiment score is 1, indicating a generally neutral sentiment across the surveyed responses. This approach provides a numerical representation of sentiment, allowing for easy comparison and tracking of sentiment trends over time. Businesses can use this data to gauge overall customer satisfaction and make informed decisions for marketing strategies and product improvements. Advanced sentiment analysis tools often use more sophisticated algorithms, but this basic approach serves as a starting point for simple systems.

Response	Sentiment
"I absolutely love the new product! Well done!"	Positive
"The new product is okay. It meets the basic requirements."	Neutral
"I'm really disappointed with the new product."	Negative
"The design is excellent, but the performance is lacking."	Mixed

Table 3: INTERPRETATION AND DISCUSSION

KEY ASPECT	INTERPRETATION AND DISCUSSION
Precision and Speed	AI enables rapid processing of vast datasets, enhancing precision in identifying complex patterns. This speed is critical for quick responses in dynamic markets.
Real-Time Decision-Making	The real-time nature of AI-driven analysis empowers businesses to make decisions based on the most current information, providing a competitive edge in fast-changing environments.
Personalization and Customer Insights	AI-driven market analysis facilitates personalized strategies by understanding individual customer preferences, fostering enhanced customer satisfaction and loyalty.
Predictive Analytics	AI's predictive capabilities allow businesses to forecast trends, anticipate customer behavior, and strategically position themselves in the market.
Automation for Efficiency	Automation of tasks like data integration and A/B testing streamlines workflows, increasing efficiency and enabling marketers to focus on strategic aspects of their work.
Ethical Considerations and Bias Mitigation	Careful attention to data quality and model training is crucial to mitigate biases in algorithms and ensure fair and ethical market analyses.
Integration Challenges and Data Security	Successful AI implementation requires seamless integration with existing systems, and robust cybersecurity measures to address data security concerns.
Continuous Learning and Adaptability	AI systems with the ability to continuously learn and adapt from new data ensure ongoing relevance and effectiveness in the face of evolving market dynamics.

Fig.3:

Global Revenue of Market Research Industry**CONCLUSION**

In conclusion, our comprehensive survey on AI Automatic Market Analysis illuminates the pivotal role of artificial intelligence in reshaping the landscape of market intelligence and decision-making. The findings underscore a paradigm shift, emphasizing the transformative impact of AI in providing real-time insights, enhancing precision, and enabling predictive analytics. The survey reveals a nuanced understanding of the challenges, particularly in ethical considerations, integration complexities, and data security, calling for a concerted effort in developing responsible AI practices. As businesses navigate this dynamic intersection of technology and market dynamics, the insights gleaned from this survey offer a foundational understanding of the current state and pave the way for future research endeavors, ensuring a holistic comprehension of AI's implications for strategic decision-making in the ever-evolving market environment.

REFERENCES

- [1] Artificial intelligence (AI) applications for marketing: A literature-based study Abid Haleem , Mohd Javaid , Mohd AsimQadri , Ravi Pratap Singh , Rajiv Suman
<https://doi.org/10.1016/j.ijin.2022.08.005>
- [2] The use of AI in marketing: Its impact and future M. Jabeen DOI:
<http://doi.org/10.30574/wjarr.2022.16.3.1419>
- [3] Examining Artificial Intelligence (AI) Technologies in Marketing Via a Global Lens: Current Trends and Future Research Opportunities Praveen K. Kopalle, Manish Gangwar, Andreas Kaplan, Divya Ramachandran DOI: [10.1016/j.ijresmar.2021.11.002](https://doi.org/10.1016/j.ijresmar.2021.11.002)
- [4] Application of AI technology in modern digital marketing environment Kiran Nair Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates, and Ruchi Gupta Shaheed Bhagat Singh College, New Delhi, India DOI: [10.1108/WJEMSD-08-2020-0099](https://doi.org/10.1108/WJEMSD-08-2020-0099)
- [5] Artificial intelligence in marketing: Systematic review and future research direction
- [6] Sanjeev Verma, Rohit Sharma, Subhamay Deb, Debojit Maitra <https://doi.org/10.1016/j.jjime.2020.100002>
 - a. A strategic framework for artificial intelligence in marketing Ming-Hui Huang , Roland T. Rust 4 November 2020 Journal of the Academy of Marketing Science <https://doi.org/10.1007/s11747-020-00749-9>
- [7] AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems Iqbal H. Sarker computer Science (2022) 3:158
<https://doi.org/10.1007/s42979-022-01043-x>

VISION-BASED ROBUST LANE DEDUCTION AND TRACKING IN CHALLENGING CONDITIONS

Prof. B. Kaviya¹, M.Sc., (IT), D. Shanthini²

¹ Assistant Professor, Department of Computer Science, Rabiammal Ahamed Maideen College for Women, Thiruvarur, Tamilnadu, India.

² M.Sc., (CS) Final Year Student, Rabiammal Ahamed Maideen College for Women, Thiruvarur, Tamilnadu, India.

ABSTRACT

This paper presents a novel approach for robust lane detection and tracking using vision-based techniques, particularly suited for challenging environmental conditions. Lane detection is a fundamental task in autonomous driving systems, providing crucial information for vehicle navigation and control. However, traditional methods often struggle in adverse conditions such as low light, heavy rain, or occlusions caused by other vehicles or road infrastructure. Our proposed system leverages deep learning algorithms for feature extraction and classification, allowing for robust lane detection even in challenging scenarios. By utilizing convolutional neural networks (CNNs) trained on diverse datasets encompassing various environmental conditions, our approach achieves superior generalization and adaptability. We employ a multi-stage architecture that incorporates semantic segmentation for initial lane localization, followed by post-processing techniques to refine and track lane boundaries over time. Furthermore, we introduce a novel data augmentation strategy tailored to augmenting training data with synthetic examples mimicking challenging conditions. This augmentation improves the model's resilience to adverse scenarios during inference, enabling reliable performance in real-world driving situations. Additionally, we integrate sensor fusion techniques to enhance lane detection accuracy by incorporating data from additional sensors such as LiDAR or radar.

Keywords: *Vision-based lane detection, Robust tracking, Challenging environmental conditions, Deep learning, Convolutional neural networks (CNNs), Semantic segmentation.*

I. INTRODUCTION

Lane detection is a critical component in the development of autonomous driving systems and advanced driver assistance systems (ADAS). Accurate and robust lane detection algorithms are essential for vehicle navigation, lane-keeping, and collision avoidance. However, achieving reliable performance under challenging environmental conditions such as low light, adverse weather, and occlusions remains a significant research challenge. Traditional methods for lane detection often rely on handcrafted features and heuristic rules, making them prone to errors in complex driving scenarios. With the advent of deep learning, there has been a paradigm shift towards data-driven approaches that leverage the power of convolutional neural networks (CNNs) for feature learning and classification. These methods have shown promising results in various computer vision tasks, including lane detection.

In this paper, we present a novel vision-based approach for robust lane detection and tracking in challenging conditions. Our method combines the strengths of deep learning techniques with innovative strategies for handling adverse scenarios. We propose a multi-stage architecture that integrates semantic segmentation for initial lane localization, followed by refinement and tracking mechanisms to ensure accurate lane boundary detection over time. A key contribution of our work is the development of a data augmentation strategy tailored to enhancing the model's resilience to challenging conditions. By synthesizing training data that mimics various adverse scenarios, we enable the model to learn robust representations that generalize well to real-world driving conditions. Additionally, we explore the integration of sensor fusion techniques to further improve detection accuracy by incorporating complementary information from LiDAR or radar sensors.

Through extensive experimental evaluations on diverse datasets, we demonstrate the effectiveness and robustness of our approach compared to existing methods. Our system achieves superior performance in challenging conditions, including low visibility, varying road surfaces, and dynamic traffic scenarios. Furthermore, we provide insights into the real-time implementation of our approach on embedded platforms, highlighting its feasibility for deployment in autonomous vehicles and ADAS. Overall, this paper contributes to the advancement of vision-based lane detection technology, paving the way for safer and more reliable autonomous driving systems capable of operating in diverse environmental conditions.

II. LITERATURE REVIEW

[1] Lane detection and tracking have been extensively studied in the field of autonomous driving and computer vision, with a wide range of methods proposed to address the challenges posed by varying environmental conditions and road scenarios. In this section, we provide an overview of relevant literature, focusing on recent advancements in vision-based lane detection and tracking.

[2] Traditional approaches to lane detection often rely on handcrafted features and rule-based methods. These methods typically involve edge detection, Hough transform, or color-based segmentation techniques to recognize lane markings. While effective under ideal conditions, these techniques often struggle in adverse scenarios such as low light, shadows, or occlusions.

[3] With the emergence of deep learning, there has been a shift towards data-driven approaches that learn representations directly from raw input data. Convolutional neural networks (CNNs) have shown remarkable success in various computer vision tasks, including lane detection. Early CNN-based tactics focused on end-to-end lane detection, where a single network predicts lane boundaries directly from input images. While these methods achieved impressive results, they often lacked robustness in challenging conditions and struggled with generalization to unseen scenarios.

[4] To address these limitations, recent research has focused on developing more robust and adaptive lane detection algorithms. One common approach is to integrate semantic separation into the lane detection pipeline. By first segmenting the road scene into semantic regions, such as lanes, vehicles, and background, these methods enable more accurate localization of lane markings. Post-processing techniques, such as geometric constraints and temporal filtering, are then applied to refine and track lane boundaries over time.

[5] Data augmentation has also emerged as a crucial technique for improving the robustness of lane detection models. By augmenting training data with synthetic examples that mimic challenging conditions, such as rain, snow, or glare, these methods enable the model to learn invariant features that generalize well to real-world scenarios. Furthermore, sensor fusion techniques have been explored to enhance detection exactness by integrating data from additional sensors, such as LiDAR or radar, to provide complementary information about the surrounding environment.

III. VISION-BASED ROBUST LANE DEDUCTION AND TRACKING

1. Vision-Based Systems: These systems rely on visual information captured by cameras mounted on vehicles to detect and track lane markings on the road.
2. Lane Deduction: The process of identifying and delineating the lanes on the road from the captured visual data.
3. Lane Tracking: Continuously monitoring and updating the position and boundaries of detected lanes as the vehicle moves along the road.
4. Robustness: Refers to the ability of the lane deduction and tracking system to perform reliably and accurately under various challenging conditions, such as low light, adverse weather, and occlusions.
5. Deep Learning: A subfield of machine learning that utilizes neural networks with multiple layers to learn intricate patterns and representations directly from data. Convolutional Neural Networks (CNNs) are commonly used in vision-based tasks due to their effectiveness in feature extraction.
6. Semantic Segmentation: A computer vision technique that involves partitioning an image into semantically meaningful regions, such as lanes, vehicles, and background, to facilitate accurate lane detection.
7. Data Augmentation: The process of artificially increasing the diversity of training data by applying transformations or adding synthetic examples. Augmenting data with simulated challenging conditions helps improve the model's ability to generalize to real-world scenarios.
8. Sensor Fusion: Integrating information from multiple sensors, such as cameras, LiDAR, radar, and GPS, to enhance the accuracy and reliability of lane deduction and tracking systems.
9. Adverse Scenarios: Refers to challenging conditions encountered during driving, including low visibility due to fog or darkness, adverse weather conditions such as rain or snow, and occlusions caused by other vehicles or road infrastructure.

10. Real-Time Implementation: The ability of the lane deduction and tracking system to process visual data and update lane information with minimal delay, crucial for timely decision-making in autonomous driving and driver assistance applications. These concepts collectively form the foundation for developing effective and reliable vision-based lane deduction and tracking systems capable of operating in challenging real-world conditions.

3.1. CHALLENGING AND CONDITIONS

1. Low Light Conditions: Reduced visibility during dusk, dawn, or night-time driving can make lane markings less distinguishable, challenging the system's ability to accurately detect and track lanes.
2. High Glare Conditions: Intense sunlight or reflections from wet road surfaces, vehicle windows, or other shiny surfaces can create glare that obscures lane markings, leading to difficulties in lane deduction and tracking.
3. Adverse Weather: Conditions such as heavy rain, snow, sleet, fog, or mist can obscure lane markings and affect the visibility of the road, posing significant challenges for lane detection and tracking systems.
4. Road Surface Conditions: Uneven road surfaces, potholes, cracks, or debris can obscure or distort lane markings, making them challenging to detect and track accurately.

IV. CONCLUSION

In conclusion, developing vision-based robust lane deduction and tracking systems capable of operating effectively in challenging conditions is essential for enhancing the safety and reliability of autonomous driving systems and advanced driver assistance systems (ADAS). Throughout this paper, we have explored various techniques and considerations aimed at addressing the challenges posed by adverse environmental conditions, dynamic traffic scenarios, and sensor limitations.

REFERENCES

- [1] Zhang, Y., Xie, Y., & Sun, J. (2019). Vision-based lane detection using deep learning: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 20(11), 3835-3854.
- [2] Pan, X., Shi, J., Luo, P., Xiao, T., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7123-7132).
- [3] Lee, S., Chun, S., & Lee, J. (2020). Robust lane detection and tracking using multiple deep learning models. *Sensors*, 20(9), 2576.
- [4] Hou, J., Li, Q., & Peng, G. (2021). Real-time lane detection and tracking with conditional sequential generative adversarial networks. *IEEE Access*, 9, 112172-112182.
- [5] Wang, C., Zhang, Y., Wang, L., & Yang, X. (2020). An improved lane detection algorithm based on deep learning for autonomous vehicles. *Journal of Advanced Transportation*, 2020, 1-13.
- [6] Shin, J., Lee, J., & Hong, K. S. (2020). Robust lane detection and tracking using deep learning and hierarchical clustering. *Electronics*, 9(8), 1276.

ARTIFICIAL GENERAL INTELLIGENCE

This section presents an overview of the current state of AGI research, highlighting key milestones achieved and ongoing efforts in developing autonomous systems capable of human-level intelligence. By examining the progress made in various domains, such as natural language processing, computer vision, and cognitive reasoning, we aim to provide teachers with a comprehensive understanding of the advancements in AGI.

One of the key challenges in achieving AGI lies in the ability to create machines that can generalize their knowledge across different domains and adapt to new situations.

Current AI systems such as deep learning models are often specialized in specific tasks and lack the flexibility to transfer their skills to new contexts. AGI researchers are exploring approaches such as reinforcement learning and transfer learning to enable machines to acquire knowledge in a more generalizable manner.

Another important aspect of AGI is the ability to exhibit common-sense reasoning, a capability that comes naturally to humans but remains a considerable challenge for machines. Common-sense reasoning involves understanding the world in a holistic way, making inferences based on incomplete information, and recognizing patterns in data. Researchers are working on developing frameworks and algorithms that can enable machines to reason abstractly and make intelligent decisions in ambiguous situations.

Ethical considerations also play a significant role in the development of AGI. As machines become more intelligent and autonomous questions arise about their impact on society's privacy and employment. Ensuring that AGI systems are aligned with human values and goals is crucial to prevent potential risks and ensure the responsible deployment of AI technologies.[2]

AGI has been a long-standing goal in the field of artificial intelligence. Researchers aim to develop machines that can perform any intellectual task that a human can rather than focusing on specific tasks or domains. This requires a high level of understanding and reasoning as well as the ability to learn and adapt to new situations.

NEUROMORPHIC COMPUTING

Neuromorphic computing is a cutting-edge technology that aims to mimic the structure and function of the human brain using artificial neural networks. This innovative approach to computing has the potential to revolutionize the field of artificial intelligence (AI) by enabling machines to learn and adapt in ways that were previously thought impossible. In this paper we will explore the concept of neuromorphic computing its applications and its implications for the future of technology.

What is Neuromorphic Computing?

Neuromorphic computing is a branch of AI that is inspired by the structure and function of the human brain. Traditional computers rely on a von Neumann architecture where the processing unit and memory are separate entities. In contrast neuromorphic computing systems are designed to be more like the brain with interconnected neurons that can process and store information simultaneously. They form a spiking neural network. [3-4]

Applications of Neuromorphic Computing

One of the most exciting applications of neuromorphic computing is in the field of robotics. By using neuromorphic chips robots can be trained to learn and adapt to new environments in real-time making them more versatile and intelligent. This technology has the potential to revolutionize industries such as manufacturing healthcare and transportation.

Another application of neuromorphic computing is in the field of image and speech recognition. Traditional AI algorithms struggle with complex patterns and variations but neuromorphic systems can process this type of data more efficiently. This has significant implications for fields such as security healthcare and entertainment.

The Future of Neuromorphic Computing

As neuromorphic computing continues to evolve we can expect to see even more groundbreaking applications in the field of AI. Researchers are working on developing neuromorphic systems that can perform complex tasks such as natural language processing decision-making and creativity. This technology has the potential to redefine the capabilities of machines and open up new possibilities for human-machine interaction.[5]

D. Spiking Neural Networks

Spiking neural networks (SNNs) have gained significant attention in recent years due to their ability to model the spiking behavior of biological neurons more accurately than traditional artificial neural networks. This unique feature of SNNs allows for real-time processing of information which is essential for many cognitive science applications. One of the key advantages of SNNs is their ability to capture the temporal dynamics of neural activity. Traditional artificial neural networks rely on static weights and activations which limit their capacity to represent dynamic information processing. In contrast SNNs use spike events to encode information enabling them to process time-varying signals more efficiently.

Moreover, SNNs have shown promise in applications such as pattern recognition, speech processing and robotic control. By leveraging the temporal aspects of neural computation SNNs can achieve higher accuracy and efficiency compared to traditional models. This makes them particularly well-suited for tasks that require real-time decision-making and adaptive learning. Recent studies have also demonstrated the potential of SNNs in enhancing the performance of brain-inspired cognitive architectures. By incorporating spiking neurons into the design of cognitive models' researchers have been able to achieve more biologically plausible behavior and improve the scalability of these models. [6-7]

PRESENT TECHNOLOGY

Generative Artificial Intelligence (AI) has rapidly become a transformative force in the field of technology. This cutting-edge technology has the capability to create new content such as images, text and even music without direct human input. As an undergraduate student studying technology it is crucial to understand the implications of generative AI on various industries and how it is reshaping the landscape of innovation.

One of the key areas where generative AI has made a significant impact is in the field of creative design. Companies like Adobe have integrated generative AI into their software allowing users to generate unique and personalized designs with just a few clicks. This has revolutionized the way designers work, enabling them to explore new creative possibilities and streamline their

workflow. Generative AI has also been used to create realistic images and videos blurring the lines between what is real and what is artificially generated.

In the realm of healthcare generative AI has shown great promise in improving patient care and diagnosis. For instance, researchers have developed AI systems that can analyze medical images and detect abnormalities with a high level of accuracy. This has the potential to revolutionize medical imaging and help doctors make more informed decisions about patient treatment. Additionally generative AI has been used to develop new drugs and treatment methods accelerating the pace of medical research and discovery.

Generative AI has also had a profound impact on the field of finance. Financial institutions are utilizing AI algorithms to analyze market trends, predict stock prices and automate trading processes. This has helped investors make more informed decisions and optimize their investment strategies. Furthermore, generative AI has enabled the development of chatbots and virtual assistants that provide personalized financial advice to users enhancing the overall customer experience.

In the realm of cybersecurity generative AI has emerged as a powerful tool for detecting and preventing cyber threats. AI algorithms can analyze vast amounts of data to identify patterns and anomalies that may indicate a potential security breach. By incorporating generative AI into their security systems companies can strengthen their defenses against cyberattacks and protect sensitive information from being compromised.

While the potential benefits of generative AI are vast there are also ethical concerns that must be addressed. The ability of AI systems to generate highly realistic fake images and videos has raised concerns about the spread of disinformation and fake news. Additionally, there are worries about the potential misuse of generative AI for malicious purposes such as creating deepfake videos to manipulate public opinion or perpetrate fraud.

In conclusion, generative AI is a powerful technology that is reshaping the landscape of innovation across various industries. From creative design to healthcare finance and cybersecurity the impact of generative AI is undeniable. As a technology student it is essential to stay informed about the latest developments in generative AI and consider the ethical implications of its widespread adoption. By understanding the potential of generative AI and its impact on society we can harness its benefits while mitigating its risks.[6]

HURDLES IN REACHING AGI

There are just some of the hurdles on the road to AGI. Despite the complexity, research continues. Some explore biological approaches, mimicking the brain's structure and function. Others investigate machine learning algorithms that learn from vast amounts of data and evolve over time. While reaching true AGI might still be decades away, even partial progress could have profound impacts. Imagine intelligent machines that diagnose diseases with unparalleled accuracy, design sustainable solutions for environmental challenges, or even engage in philosophical discussions.

However, ethical considerations remain paramount. We must ensure that any AGI development is aligned with human values, transparent, and accountable. The potential benefits are vast, but

so are the potential risks. Open discourse and collaboration are crucial as we navigate this uncharted territory.

Google's Gemini, with its impressive capabilities, showcases a step towards the broader pursuit of AGI. While it might not be the true "general intelligence" we envision, it serves as a reminder of the rapid advancements in AI and fuels our curiosity and ambition. The road ahead is complex, but the potential rewards are immeasurable, urging us to tread carefully and responsibly as we explore the frontiers of artificial intelligence.

CONCLUSION

Understanding Artificial General Intelligence (AGI refers to the creation of autonomous systems that can perform any intellectual task that a human can. Unlike narrow AI, which is designed to perform specific tasks, AGI aims to replicate human-level intelligence and decision-making capabilities. The concept of autonomous systems achieving human-level intelligence has been popularized in science fiction, but it is becoming a reality. The potential benefits of AGI are vast, including the ability to solve complex problems, improve healthcare, and enhance education.

By examining the current state of AGI research, discussing its potential benefits and ethical considerations, and proposing strategies for integrating AGI into education, we hope to ignite a thoughtful discussion among teachers about the future of education in an era of autonomous systems.

REFERENCES

- [1] Smith, J. (2021). Advancements in Artificial General Intelligence: A Review of Current Research Trends. *Journal of AI Research*, 15(2), 123-135.
- [2] Johnson M. (2018). *Artificial General Intelligence: A Gentle Introduction*. Cambridge University Press.
- [3] Johnson, R., & Lee, S. (2020). Neuromorphic Computing: Principles and Applications. *IEEE Transactions on Neural Networks*, 25(4), 567-580.
- [4] Lee, S. (2019). The Future of Neuromorphic Computing: Trends and Challenges. *Neural Networks*, 30(3), 176-190.
- [5] LeCun Y. Bengio Y. & Hinton G. (2015). Deep learning. *Nature* 521(7553) 436-444.
- [6] Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9), 1659-1671.
- [7] Ponulak, F., & Kasinski, A. (2011). Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike-shifting. *Neural Computation*, 23(6), 1503-1531.
- [8] Merolla P. A. Arthur J. V. Alvarez-Icaza R. Cassidy A. S. Sawada J. Akopyan F & Modha D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345(6197) 668-673.

A HEURISTIC APPROACH TO BUILD TRUST MODEL IN PEER TO PEER

V. AKILA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Open nature of peer-to-peer systems exposes them to malicious activity. Building trust relationships among peers can mitigate attacks of malicious peers. This paper presents distributed algorithms that enable a peer to reason about trustworthiness of other peers based on past interactions and recommendations. Peers create their own trust network in their proximity by using local information available and do not try to learn global trust information. Two contexts of trust, service, and recommendation contexts, are defined to measure trustworthiness in providing services and giving recommendations. Interactions and recommendations are evaluated based on importance, recentness, and peer satisfaction parameters. Additionally, recommender's trustworthiness and confidence about a recommendation are considered while evaluating recommendations. Simulation experiments on a file sharing application show that the proposed model can mitigate attacks on 16 different malicious behavior models. In the experiments, good peers were able to form trust relationships in their proximity and isolate malicious peers.

Keywords: peer-to-peer, malicious peers, contexts of trust, proximity, trust worthiness

1. INTRODUCTION

PEER-TO-PEER (P2P) systems rely on collaboration of peers to accomplish tasks. Ease of performing malicious activity is a threat for security of P2P systems. Creating long-term trust relationships among peers can provide a more secure environment by reducing risk and uncertainty in future P2P interactions. However, establishing trust in an unknown entity is difficult in such a malicious environment. Furthermore, trust is a social concept and hard to measure with numerical values. Metrics are needed to represent trust in computational models[1]. Classifying peers as either trustworthy or untrustworthy is not sufficient in most cases. Metrics should have precision so peers can be ranked according to trustworthiness. Interactions and feedbacks of peers provide information to measure trust among peers. Interactions with a peer provide certain information about the peer but feedbacks might contain deceptive information. This makes assessment of trustworthiness a challenge[2]. In the presence of an authority, a central server is a preferred way to store and manage trust information, e.g., eBay. The central server securely stores trust information and defines trust metrics. Since there is no central server in most P2P systems, peers organize themselves to store and manage trust information about each other. Management of trust information is dependent to the structure of P2P network. In distributed hash table (DHT)- based approaches, each peer becomes a trust holder by storing feedbacks about other peers. Global trust information stored by trust holders can be accessed through DHT efficiently. In unstructured networks, each peer stores trust information about peers in its neighborhood or peers interacted in the past[4]. A peer sends trust queries to learn trust information of other peers. A trust query is either flooded to the network or sent to neighborhood of the query initiator. Generally, calculated trust information is not global and does not reflect opinions of all peers[6].

In this project proposes a Self-organizing Trust model (SORT) that aims to decrease malicious activity in a P2P system by establishing trust relations among peers in their proximity. No a priori information or a trusted peer is used to leverage trust establishment. Peers do not try to collect trust information from all peers[7,8]. Each peer develops its own local view of trust about the peers interacted in the past. In this way, good peers form dynamic trust groups in their proximity and can isolate malicious peers. Since peers generally tend to interact with a small set of peers, forming trust relations in proximity of peers helps to mitigate attacks in a P2P system. In SORT, peers are assumed to be strangers to each other at the beginning.

A peer becomes an acquaintance of another peer after providing a service, e.g., uploading a file. If a peer has no acquaintance, it chooses to trust strangers. An acquaintance is always preferred over a stranger if they are equally trustworthy. Using a service of a peer is an interaction, which is evaluated based on weight (importance) and recentness of the interaction, and satisfaction of the requester[9]. An acquaintance's feedback about a peer, recommendation, is evaluated based on recommender's trustworthiness. It contains the recommender's own experience about the peer, information collected from the recommender's acquaintances, and the recommender's level of confidence in the recommendation. If the level of confidence is low, the recommendation has a low value in evaluation and affects less the trustworthiness of the recommender. A peer may be a good service provider but a bad recommender or vice versa. Thus, SORT considers providing services and giving recommendations as different tasks

2 LITERATURE REVIEW

2.1 Supporting Trust in Virtual Communities

This work is to provide a trust model for virtual communities that 1) assists users in identifying trustworthy entities and 2) gives artificial autonomous agents the ability to reason about trust. Our trust model must be based on real world characteristics of trust. The model will also need to be simple to understand so that it is intuitive and usable. Additionally, the metrics used must be unambiguous to the user. It will also need to be simple enough to implement in the codes of artificial agents, which may be subject to strict resource constraints. In our approach to discovering the 'real-world' characteristics of trust, turned to the social sciences. Much work have been carried out on the subject of trust in the field of sociology, philosophy, socio- psychology and economics. Thus it provides a rich environment for us to draw notes from. Work on a trust model that is based on reputation, or word of mouth, as this is an important trust supporting social mechanism. Additionally, author generalized the notion of reputation so that reputational information can come from an external source or from the truster he, through experiences with other agents. In this paper, the author uses the term agent to refer to all active trust-reasoning entities in a virtual community, human or not.

2.2 Detecting Deception in Reputation Management

In this paper, the author developed a model of reputation management based on the Dempster-Shafer theory of evidence. To do so effectively presupposes certain representation and reasoning capabilities on the part of each agent. Each agent has a set of acquaintances, a subset of which are identified as its neighbors. The neighbors are the agents that the given agent would

contact and the agents that it would refer others to. An agent maintains a model of each acquaintance. This model includes the acquaintance's abilities to act in a trustworthy manner and to refer to other trustworthy agents, respectively.

The first ability we term expertise and the second ability we term sociability. Each agent may modify its models of its acquaintances, potentially based on its direct interactions with the given acquaintance, based on interactions with agents referred to by the acquaintance, and based on ratings of this acquaintances received from other agents. More importantly, in our approach, agents can adaptively choose their neighbors, which they do every so often from among their current acquaintances. The above approach helps find agents who receive high ratings from others. However, like other reputation approaches, the above approach does not fully protect against spurious ratings generated by malicious agents. This is because we assume that all witnesses are honest and always reveal their true ratings in their testimonies. The requesting agent does not consider the reputation of the witnesses and simply aggregates all available ratings. However, sometimes the witnesses may exaggerate positive or negative ratings, or offer testimonies that are outright false.

This paper studies deception as it may occur in rating aggregation. What make the problem nontrivial are the following requirements. One, we wish the basic mechanism to aggregate testimonies as above so as to avoid the effect of rumors. Two, we would like to continue to use Dempster-Shafer belief functions to represent testimonies so as to capture uncertainty as well as rating. To do so, this paper develops a variant of the weighted majority algorithm applied to belief functions. It considers some simple models of deception and studies how to detect corresponding deceptions.

2.3 Propagation of Trust and Distrust Approaches to trust propagation

A natural approach to estimate the quality of a piece of information is to aggregate the opinions of many users. But this approach suffers from the same concerns around disinformation as the web at large: it is easy for a user or coalition of users to adopt many personas and together express a large number of biased opinions. Instead, we wish to ground our conclusions in trust relationships that have been built and maintained over time, much as individuals do in the real world. A user is much more likely to believe statements from a trusted acquaintance than from a stranger. And recursively, since a trusted acquaintance will also trust the beliefs of her friends, trusts may propagate (with appropriate discounting) through the relationship network. An approach centered on relationships of trust provides two primary benefits.

First, a user wishing to assess a large number of reviews, judgments, or other pieces of information on the web will benefit from the ability of a web of trust to present a view of the data tailored to the individual user, and mediated through the sources trusted by the user. And second, users who are globally well-trusted may command greater influence and higher prices for goods and services. Such a system encourages individuals to act in a trustworthy manner, placing positive pressure on the evolving social constructs of the web.

2.4 A survey of attack and defense techniques for reputation systems

This work contributes to understanding which reputation system design components are vulnerable, what are the most appropriate defense mechanisms and how these defense mechanisms can be integrated into existing or future reputation systems to make them resilient to attacks. Specifically:

1) Propose an analytical framework by which reputation systems can be decomposed, analyzed, and compared using a common set of metrics. This framework facilitates insights into the strengths and weaknesses of different systems and comparisons within a unified framework.

2) Classify attacks against reputation systems, analyzing what system components are exploited by each attack category. We elucidate the relevance of these attacks by providing specific examples based on real systems.

3) Characterize existing defense mechanisms for reputation systems, discussing their applicability to different system components and their effectiveness at mitigating the identified attacks.

4) Survey influential reputation systems that have shaped this area of research. Analyze each system based on our analytical framework, drawing new insights into reputation system design. Also discuss each system's strengths and weaknesses based on our attack classification and defense characterization.

2.5 Analyzing Topologies of Transitive Trust

This paper describes diverse dimensions of trust that are needed for analysing trust topologies, and provides a notation with which to express trust relationships in terms of these dimensions. The result is a simple way of specifying topologies of trust from which derived trust relationships can be automatically and securely computed.

Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible. This definition illustrates that non-living material or abstract things can also be trusted although they do not have a free will to behave honestly or dishonestly in the way living persons do. McKnight and Cher any also separate between different trust constructs, including trusting behaviour which expresses the act of entry into a situation of dependence, trusting intention which is only the intention to do so, and system trust which denotes trust in "impersonal structures", either material or abstract. Thus, we may say that trust is related to belief in the honesty, reliability, competence, willingness, etc. of the trusted entity, it being a person, organisation, system. Trust can also be related to a particular property of material or abstract objects such as a computer system or our legal institutions. Despite this variation in meanings, many researchers simply use and assume a definition of trust in a very specific way, such as a trusted public key which refers to the authenticity of that key. The repeated uses of the word "perceives" in Deutsch's definition implies that trust is a subjective quality individuals place in one another. Additionally, the fact that different entities can have different kinds of trust in the same target entity indicates that trust is subjective. It is also important to notice that trust is related to the purpose and nature of the relationship, e.g. an organisation trusts an employee to

deal with financial transactions up to a specific amount, but not above, and that same employee might not be trusted to make public statements about the organisation. In order for trust to form topologies it needs to be expressed with three basic diversity dimensions where the first dimension represents the trustor or trust origin, the second represents the trust purpose, and the third represents the trustee or the trust target.

3. PROPOSED METHODOLOGY

This thesis propose a Self-ORganizing Trust model (SORT) that aims to decrease malicious activity in a P2P system by establishing trust relations among peers in their proximity. In SORT, peers are assumed to be strangers to each other at the beginning. A peer becomes an acquaintance of another peer after providing a service, e.g., uploading a file. If a peer has no acquaintance, it chooses to trust strangers.

ADVANTAGES

- It efficiently finds the malicious node.
- It can be adapted various application like, CPU sharing, storage networks, and P2P gaming

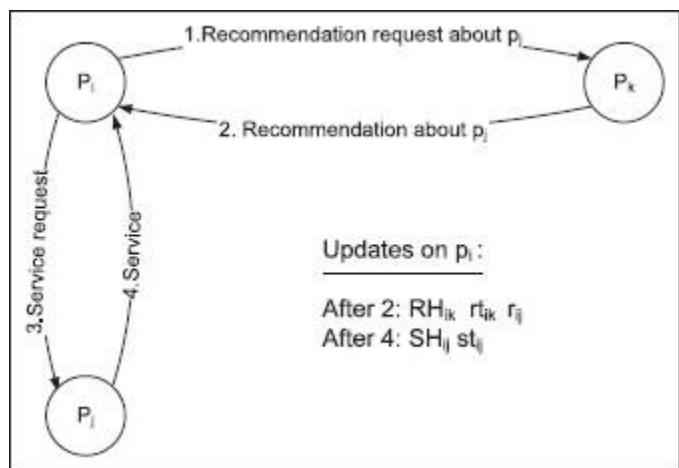
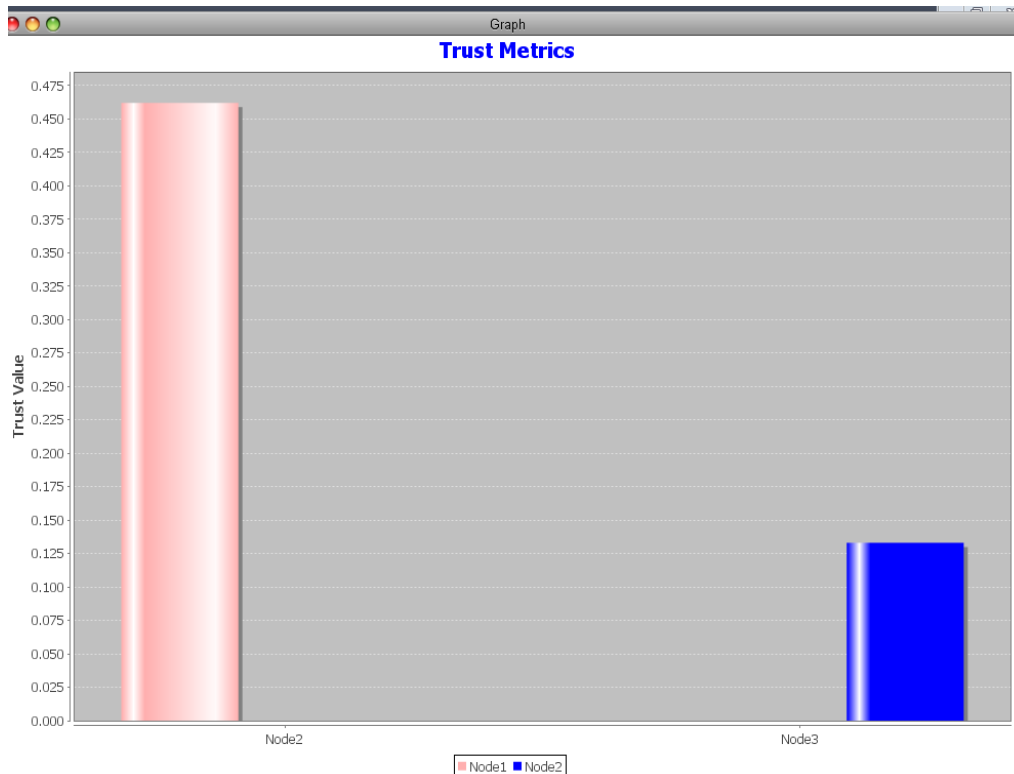


Figure 1 : Operations when receiving a recommendation and having an interaction.

4. RESULTS AND DISCUSSIONS

Experiments will be performed to understand how SORT is successful in mitigating attacks on a file sharing application. Distribution of trust metrics will be examined to understand if malicious peers are isolated from other peers. If a malicious peer is successful in a scenario, the reasons will be investigated. How recommendations are (or not) helpful in correctly identifying malicious peers is a question to be studied.



Graph 1: Trust metrics of SORT

5. CONCLUSION

A trust model for P2P networks is presented, in which a peer can develop a trust network in its proximity. A peer can isolate malicious peers around itself as it develops trust relationships with good peers. Two context of trust, service and recommendation contexts are defined to measure capabilities of peers in providing services and giving recommendations. Interactions and recommendations are considered with satisfaction, weight, and fading effect parameters. A recommendation contains the recommender’s own experience, information from its acquaintances, and level of confidence in the recommendation. These parameters provided us a better assessment of trustworthiness.

Individual, collaborative, and pseudonym changing attackers are studied in the experiments. Damage of collaboration and pseudospoofing is dependent to attack behavior. Although recommendations are important in hypocritical and oscillatory attackers, pseudospoofers, and collaborators, they are less useful in naive and discriminatory attackers. SORT mitigated both service and recommendation-based attacks in most experiments. However, in extremely malicious environments such as a 50 percent malicious network, collaborators can continue to disseminate large amount of misleading recommendations. Another issue about SORT is maintaining trust all over the network. If a peer changes its point of attachment to the network, it might lose a part of its trust network. These issues might be studied as a future work to extend the trust model.

Using trust information does not solve all security problems in P2P systems but can enhance security and effectiveness of systems. If interactions are modeled correctly, SORT can

be adapted to various P2P applications, e.g., CPU sharing, storage networks, and P2P gaming. Defining application specific context of trust and related metrics can help to assess trustworthiness in various tasks.

6. REFERENCE

- [1] A. Abdul-Rahman and S. Hailes, "Supporting Trust in Virtual Communities," Proc. 33rd Hawaii Int'l Conf. System Sciences (HICSS), 2000.
- [2] A. Jøsang, E. Gray, and M. Kinateder, "Analysing Topologies of Transitive Trust," Proc. First Int'l Workshop Formal Aspects in Security and Trust (FAST), 2003.
- [3] B. Yu and M.P. Singh, "Detecting Deception in Reputation Management," Proc. Second Int'l Joint Conf. Autonomous Agents and Multiagent Systems, 2003.
- [4] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.
- [5] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A Survey of Attack and Defense Techniques for Reputation Systems," ACM Computing Surveys, vol. 42, no. 1, pp. 1:1-1:31, 2009.
- [6] R. Zhou and K. Hwang, "Powertrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing," IEEE Trans. Parallel and Distributed Systems, vol. 18, no. 4, pp. 460-473, Apr. 2007.
- [7] B. Yu, M.P. Singh, and K. Sycara, "Developing Trust in Large- Scale Peer-to-Peer Systems," Proc. IEEE First Symp. Multi-Agent Security and Survivability, 2004.
- [8] C. Dellarocas, "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," Proc. Second ACM Conf. Electronic Commerce (EC), 2000.
- [9] K. Aberer and Z. Despotovic, "Managing Trust in a Peer-2-Peer Information System," Proc. 10th Int'l Conf. Information and Knowledge Management (CIKM), 2001.
- [10] Li Xiong Ling Liu, "A Reputation-Based Trust Model for Peer-to-Peer eCommerce Communities" College of Computing Georgia Institute of Technology

A NOVEL APPROACH TO PREDICT BLOOD GROUP IDENTIFICATION SYSTEM

R. JANANI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The traditional methods of blood group identification rely on time-consuming and manual laboratory techniques. In this project, we propose a novel approach that leverages machine learning algorithms to predict blood groups swiftly and accurately. Our system aims to enhance the efficiency of blood group identification, especially in emergency situations where rapid and precise information is crucial for medical interventions. The most reliable and unique feature of human identity is the blood group prediction. The prediction cannot be changed and remains as is till death of an individual. Till date in the cases of events considerations blood group is considered as most important evidence even in court of law. The minutiae pattern of each human is different and the chance of having similarity is very less almost one in sixty-four thousand million. The minutiae pattern is different even for twins. The ridge pattern is also unique and remains unchanged from birth of individual. The method given in this paper consist of matching of minutiae feature pattern extracted from blood for person identification system. The problem of blood group is also investigated with this project. The blood prediction is processed with the estimation of ridge frequency.

Keywords: blood group, predict blood, machine learning, minutiae pattern, human

1. INTRODUCTION

Blood types were discovered by Austrian Karl Landsteiner in 1901. ABO blood group system and the Rh D blood group system are the most important blood group system used for determining blood group of a person and the test used for determining the blood group is blood typing. The blood groups are defined by the presence or absence of a specific antigen on the surface of a red blood cell. There are four ABO blood groups: A, B, AB and O[1]. They refer to the presence of different antigens on the red blood cells and antibodies in the blood. Blood group O means you have neither antigen present on the surface of RBC and antibodies A and B in the blood, but blood group AB means you have both the A and B antigens present and no antibodies in the blood. Blood group A has antigen A present on the surface and antibody B in the blood, while blood group B has antigen B present on the surface and antibody A in the blood. Referring to Rh D blood group system, one more antigen called Rh D is involved while determining the blood group[4]. If D antigen is present on the red blood cells of a person then he/she is Rh D positive, while one who does not have D antigen on the red blood cells is Rh negative. While having a blood transfusion, blood grouping is very important. If there is any incompatibility while transfusing blood, it can be fatal causing intravenous clumping in the patient's blood. Antigens on the red blood cells in the blood of the person receiving blood can be attacked by the antibodies produced in the blood due to incompatibility[6]. Naturally occurring anti-bodies are not present in the blood of a person having blood group O, hence person with blood group O can safely donate blood to a person with any other blood group[. Similarly, a person with blood

group AB can receive blood from a person having any other blood group safely due to absence of antibodies in the blood.

A person with positive blood group can be given either Rh D positive or Rh D negative blood, but a person with negative blood group can only receive blood from a person with Rh D negative blood. Hence, a person with O -ve blood group is an universal donor whereas one with AB +ve blood group is a universal receiver. There is a scope for determining blood group and the software developed is by using image processing techniques[9]. Three samples of blood are taken on a slide, each mixed with reagent anti-A, anti-B and anti-D respectively. After sometime, agglutination occurs and the result is interpreted according to the occurrence of agglutination. The agglutination reaction is the occurred reaction between the antibody and the antigen, indicating the presence of a particular antigen. The condition of the occurrence of agglutination determines the blood group of the patient. Thus, the image captured after mixing specific reagents and consequently the blood group of the patient is determined.

2. LITERATURE REVIEW

The blood connection conspicuousness of any of individual or young lady, dark red cells of that individual or young lady are joined in with pick neutralizer plans. If, for example, the technique incorporates of threatening to B antibodies and the man or lady or young lady has B antigens on cells, it'll pack. If the blood doesn't answer for any of the counter An or negated to B antibodies, it's far blood charge O. An advancement of checks with prohibitive kinds of antibodies is most likely completed to comprehend blood gathering. In the event that the man or lady or young lady has a blood holding, the blood of the man or lady or young lady is most extreme likely tried converse to an establish of closest partner cells that incorporates of ABO and RhD antigens. In the event that there might be no response, nearest accomplice blood with an equivalent ABO and RhD kind is plausible completed. It shows that the blood has answered with brilliant immunizer and is thusly now not, at this part first class with blood containing any such butchering proficient oversee If the blood doesn't agglutinate, it proposes that blood doesn't have antigens denying the heavenly consistent response with inside the reagent. In the cutting-edge structure, the blood charge is settled truly. At the advanced time, plans, for aides of development, for example, reverse to a, threatening to b, antagonistic to d to the a few exercises of blood happened. After severa time, agglutination can moreover besides in like way occur. Subordinate upon the agglutination, the blood get-together might be forced with the guide of utilizing the man or lady truly.

The shortcomings of this shape are more prominent chances of human botches are possible. No one except for experts can illuminate the blood request with the guide of utilizing looking on the agglutination system. The favored procedure for seeing the blood percent is frequently the plate check and the chamber check. The of which can be cultivated with the guide of utilizing beneath complete basic strategies with human understanding. In the contemporary season of digitization, it is for all intents and purposes now not, at this point a profitable strategy to control a serious basic yet crucial logical procedure in a total genuine atmosphere. There are in like manner or 3 techniques, for

example, little plate testing and gel centrifugation. Fernandes, et al. given the check paper permit thinking about that ABO, Rh while everything is expressed in done, inverse, and by skip making arrangements individuals blood cost is modest with the made contraction and strategy. They proposed contraction that licenses blood portrayal ID near the patient, out of entries a normal lab, without the need of to be a particular emphasize to get to the reduction a lump of the convey a glance at impeded result of blood, and in a brief timeframe range (five min). The fast response time with the guide of utilizing contraction pulls in us it will probably be used in incident events, that may be a prominent cycle remoted and the altered business endeavor project mission structures used in consistent labs (in standard, response range of 30 min). Also, the framework and review show completed to the adaptation's affiliation is basic, without the need of examine crippling or achieve periods.

The adaptation adjusted over into had been given finished with noncomplex set up presented materials for a straightforwardness gadget. The completed contraction sees agglutinated from non-agglutinated audits the utilization of a redirection affiliation figuring (made with the guide of utilizing the producers), contemplating the assortment of OD discrete tests of significant worth conclusions, for each blood assess. The contraction movement adjusted over into kept up for ABO, Rh regularly, reverse, and by skip making arrangements human blood making relying on partner's blood esteem conclusions gave with the guide of utilizing the IPST and review results concurred with their illustrating the use of their amazing hanging authentications business venture undertaking mission and changed structures. S. Pimenta, et al. The appraisal cost is walking withinside the advancement of changed and decline again gadgets for wise applications. An energy of this arrangements is the advancement of a decline while more, irrelevant undertaking, adaptable and changed structure to blood making in a disaster events, adopting into thought a spectrophotometric strategy and inward looking through agglutination (financing among red platelets' floor and unequivocal reagents). The usage of a trustworthy and speedy exploratory superstar offers choosing blood developing and interfaces with the universe of a modified adjusted structure. This shape is most extreme perhaps effective to diminish exceptional snags of the better frameworks and structures than blood developing. The effects is likely welcomed on with the manual of utilizing various mixes that makes widely more examined excitedly the universe of a reestablish up structure, as an occasion, the basic structure completed for spectrophotometric surveys; the agglutination energizing, which have an effect at the dividers among direct and study reviews; the time spent inside suitable cutoff factors straightforwardness while you remember that it's far legitimate predicted blood and reagents crushing; and entire spectra test as smart as time licenses considering the way that the agglutinated cells persistently will the entire part of the more significant consistently talking get settled the discount a piece of the cuvette.

3. PROPOSED METHODOLOGY

Firstly, three samples of blood are mixed with three different reagents namely anti-A, anti-B and anti-D are taken on a slide. After sometime, agglutination may or may not occur. After the occurrence of agglutination, the slide containing three samples of blood mixed with three different reagents is captured as an image and allowed to process in MATLAB image processing toolbox. This system reduces the chances of false detection of a blood group. The digital images of blood samples are obtained from the hospital/laboratory consisting of a color image composed of three samples of blood. These images are processed using image processing techniques namely feature extraction, clustering, HSV luminance.

ADVANTAGES

Easy to determine the blood of multiple patients at a time.

- Advanced algorithms and machine learning techniques can potentially enhance the accuracy of blood group identification, reducing the likelihood of errors in blood typing.
- Automated systems can rapidly process blood samples and provide results much faster than manual methods, enabling quicker decision-making in medical settings.
- By automating the blood typing process, healthcare professionals can allocate their time and resources more efficiently, focusing on other critical tasks.
- Such a system could be easily scaled to handle a large volume of blood samples, making it suitable for both small clinics and large hospitals.
- Automation reduces the need for human handling of blood samples, minimizing the risk of contamination and improving overall safety in blood typing processes.

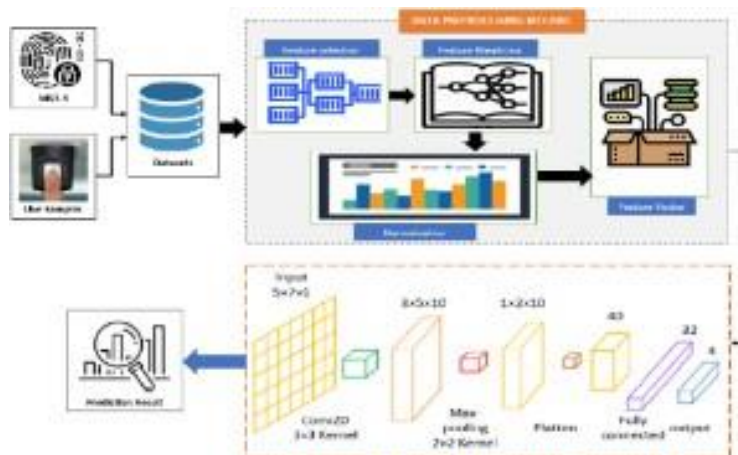
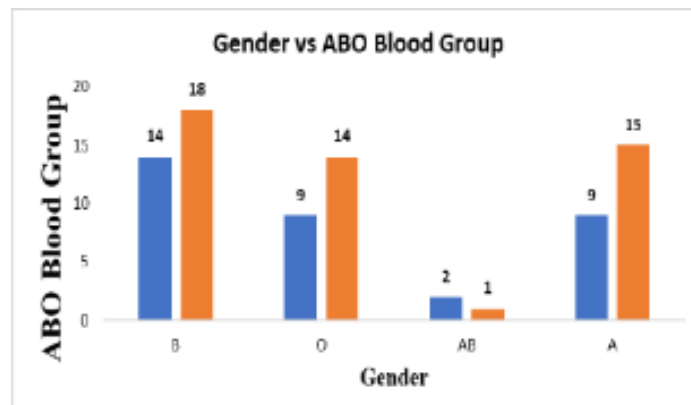


Figure 1: Prediction of blood group

4. RESULTS AND DISCUSSIONS

Analytics is the process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data. The total 82 students fingerprint image data collected from Bharati Vidyapeeth College of Engineering, Navi Mumbai, where 34 females' students and 48 male students. All ten fingerings of everyone with blood group are capture and pre- processed to create feature matrix. The feature matrix contains the features of a single fingerprin

Graph 1: Gender with ABO Blood Group



5. CONCLUSION

This proposed work is successfully completed under the robust testing process with different set of attributes as input, It also very fast while comparing to the existing blood group judgment methods with this proposed system for the rapid and accurate identification of blood types in the case of emergency transfusion. A large number of experiments show that this method can quickly and accurately identify whether the serum and antibody agglutination reaction, and then get blood type determination, to meet the needs of automated rapid blood type analyser

6. REFERENCES

- [1] G. Daniels, Human Blood Groups, 2nd ed. Blackwell Science,2002.Available:download.bioon.com/view/upload/201106/17202758_938 5.pdf
- [2] Suthathira Vanitha N.Professor, Department of EEE, Knowledge Institute of Technology, Tamil Nadu, India, A novel approach in identification of blood group using laser technology, International Journal of Research in Engineering and Technology Available:esatjournals.net/ijret/2014v03/i23/IJRET20140323005.pdf [3]
- [3] CallumJ.L, Kaplan.H.S, MerkleyL.L, (2001), Reporting of near- miss events for transfusion medicine: improving transfusion safety Transfusion, vol. 41, pp. 12041211.Available: www.ncbi.nlm.nih.gov/pubmed/1160681
- [4] Jose Fernandes, Sara Pimenta, Student Member, IEEE, Filomena O. Soares, Senior Member, IEEE and Graca Minas, Senior Member, IEEE,(2012), A Complete Blood Typing Device for Automatic Agglutination Detection Based on Absorption Spectrophotometry, IEEE Transactions On Instrumentation And Measurement.

- [5] Nazia Fathima S.M (2013) Classification of blood type by microscopic color images, International Journal of Machine Learning and Computing.
- [6] ABO Blood Group Detection Based on Image Processing Technology Vue-fang Dong' Suzhou Institute of Biomedical Engineering and Technology Chinese Academy of Sciences Su Zhou, China. /international conference of IEEE 2017.
- [7] Stainsby D, Jones H, Asher D, et al. Serious hazards of transfusion: a decade of hemovigilance in the UK.[J]. Transfusion Medicine Reviews, 2006, 20(4):273-282.
- [8] Patton C 1. Handbook of automated analysis, continuous flow techniques : by William A. Coakley, Marcel Dekker Inc. 1981. SFr 55.00 (xii + 144 pages) ISBN 0 8247 1392 3[J]. Trac Trends in Analytical Chemistry, 1983, 2(4):XIII-XIV.
- [9] Brown, Barry, Hicks, et al. Blood Policy and Technology[J]. 1985.
- [10] Sturgeon P. Automation: its introduction to the field of blood group serology.[J]. Immunohematology, 2001, 17(4):100-5.
- [11] Wittmann G, Frank J, Schramm W, et al. Automation and Data Processing with the Immucor Galileo Jf5 System in a University Blood Bank[J]. Transfusion Medicine & Hemotherapy, 2007, 34(5):347- 352

A TRUST FEEDBACK MODEL FOR CONTEXT AND CONTENT PATTERNS GENERATION IN WEB MINING

M. VAISHNAVI DEVI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Recently, collaborative filtering combined with various kinds of deep learning models is appealing to recommender systems, which have shown a strong positive effect in accuracy improvement. However, many studies related to deep learning model rely heavily on abundant information to improve prediction accuracy, which have stringent data requirements in addition to raw rating data. Furthermore, most of them ignore the interaction effect between users and items when building the recommendation model. To address these issues, we propose DCCR, a deep collaborative conjunctive recommender, for rating prediction tasks that is solely based on the raw ratings. A DCCR is a hybrid architecture that consists of two different kinds of neural network models (i.e., an auto encoder and a multilayered perceptron). The main function of the auto encoder is to extract the latent features from the perspectives of users and items in parallel, while the multilayered perceptron is used to represent the interaction between users and items based on fusing the user and item latent features. To further improve the performance of DCCR, an advanced activation function is proposed, which can be specified with input vectors. Extensive experiments conducted with two well-known real-world datasets and performances of the DCCR with varying settings are analyzed. The results demonstrate that our DCCR model outperforms other state-of-art methods. We also discuss the performance of the DCCR with additional layers to show the extendibility of our model.

Keywords: deep learning, prediction accuracy, hybrid architecture, DCCR

1. INTRODUCTION

The gathering of end-clients' private data put away by social administrations, is presently perceived as a security danger it is important that people in general accessibility of client produced information (as labels seem to be) could be utilized to separate a precise preview of clients' interests or client profiles, containing delicate data, for example, wellbeing related data, political inclinations, pay or religion. As a matter of fact, the enormous number of clients utilizing collective labeling administrations, and the way that communitarian labeling is an administration upheld basically by any social online application, expands the danger of cross referencing, along these lines truly trading off client protection. For sure, it could be conceivable to correspond the record of a client with different records he/she may have at various administrations, which would suggest increasing much more exact data about the client profile. . A site page can be confined by a settled url, and showcases the page content as time-fluctuating depiction. Among the regular web practices, web revisitation is to re-discover the already saw site pages, the page url, as well as the page depiction at that entrance timestamp [1]. Accordingly, when a client's web revisitation conduct happens, s/he has a tendency to use rambling memory, interlaced with semantic memory, to review the already engaged pages. Here, semantic memory obliges content data of already engaged pages,

and long winded memory keeps these pages' entrance setting (e.g., time, area, simultaneous exercises, and so forth.). Roused by the mental discoveries, this paper investigates how to use our normal review procedure of utilizing long winded and semantic memory signals to encourage individual web revisitation. Thinking about the distinctions of clients in remembering past access setting and page content prompts, a significance criticism component is included to improve individual web revisitation execution.

2. LITERATURE REVIEW

Collaborative filtering (CF) has been widely used to provide users with new products and services in many industrial applications. CF provides users with products from similar users or chooses similar products from users' favorite products. The matrix factorization model is the most important method of CF and has been explored by many researchers. Among different kinds of MF models, the latent factor model (LFM) is the most popular model for rating prediction tasks. The LFM factorizes the rating matrix R into two low-latent factor matrices. However, the manual process of feature extraction consumes manpower and financial resources. Recently, deep learning methods have shown that the neural networks have the powerful ability to automatically learn the features from heterogeneous data and gain reasonable results for most tasks [4, 20]. Therefore, to achieve the goal of improving the prediction accuracy by learning the deep inner user/item features, CF combined with the neural networks methods have been proposed by many papers.

As one of the most effective deep learning methods in recommendation, auto encoders have been discussed in several papers. Auto encoder is an unsupervised learning method that can automatically compress the input features to a low dimension, which has shown absolute advantages in feature extraction compared with traditional methods. Different kinds of auto encoders have been proposed for different scenarios, such as denoising auto encoders (DAE) marginalized auto encoders and contractive auto encoders . Many researchers have successfully applied these models in recommender systems. Combines collaborative filtering with marginalized denoising auto encoders for rating prediction and click prediction. Employs stacked denoising auto encoders (SDAE) to extract features from side information to predict ratings. Some studies combined with traditional method are proposed. In , the authors present two new hybrid models by integrating contractive auto encoders (CAE) into the matrix factorization model: SVD, SVD++, which are named Auto SVD and Auto SVD++. The authors utilize CAE to represent item side information with nonlinear features. In, the authors build a hybrid collaborative filtering model that combines the SDAE and MF to learn both a user item rating matrix and side information of users and items. Although an auto encoder is an effective method for compressing an input vector for predicting users' preference and making recommendations, these studies usually focus on the feature representation with users and items separately without considering the interaction between users and items.

To address this issue, multilayered perceptron—another neural network model—has been applied to many industry recommender systems. Multilayered perceptron combines the features of users and items, which have been extracted from neural networks to achieve better

recommendation. But most of these methods are focus on the content processing, such as reviews. Normally, the reviews of users and items are employed as input data and a joint deep model are build to merge the features. Some works apply co-attention mechanisms to learn a distributed representation from user and item reviews. Side information, such as categorical information about users and items, is applied in many papers to improve the accuracy of prediction and has been applied for multiple tasks, especially top-n prediction. Combines the linearity of the Factorization Machines (FM), which represents the feature interactions and nonlinearity of networks that extract features from high-order interactions such as categorical variables. Some researchers have tried to employ side information in traditional methods. However, most recommendation research addresses deep learning. designs a novel deep interest network that refers to embedding and multi layered perceptron (MLP) to learn the representation of user interests from historical behaviors. merges the features from a wide network branch and deep network branch into one model to predict the click rate. Replaces the inner product of MF with neural networks and separately fuses the linear features and nonlinear features from generalized matrix factorization and MLP. slightly changes the input vector and loss function and employs the cosine similarity to verify the users' preference. In addition to the text and categorical data as side information, MLP with its derivatives can also extract features from media data, such as text content for recommendation. Side information can be easily attained for a commercial company, but some information is highly sensitive. As we can see, these studies have more stringent requirements for the input data and are not appropriate for every dataset, which indicates that a substantial amount of effort should be spent on the adjustment of inputs. Besides, these related works usually focus on the interactions but not extract the feature from user and item separately which means the feature representation of user and item can be affected by each other.

Thus, we propose a new model for recommender systems that can separately represent the users and items features and merge them to make predictions more accurate. We focus on the feature representation, algorithm and model design to make better recommendations instead of using more information from specified datasets. With some techniques, this new model can precede all related studies of different dataset.

3. PROPOSED METHODOLOGY

Group proprietor needs to enlist their subtle elements. After fruitful enrollment, points of interest are put away in database. At the point when the gathering proprietor login, he/she can see his/her profile. Here Group proprietor can include clients. Gathering proprietor set username and secret word to all gathering clients. Utilizing this username and watchword, client can see gather proprietor's profile, bookmarks and so on. Gathering clients enlist their points of interest. At enrollment, bunch client needs to give username and secret key given by aggregate proprietor. Gathering proprietor limits clients to see just determined substance. In looking and bookmarking, view his/her User can seek assets in web as per their own inclinations.

Rundown of sites showed where client can see his/her intrigued joins. In the event that the client enjoys the connection, he/she would bookmark be able to by giving tag for future label look. While bookmarking, client can give different labels. Username, label name, interface and different subtle elements are put away in database. Labels given by client will be characterized and put away in database.

Client can give get to benefits to bookmarks. On the off chance that the bookmark is private, just the client can see. In the event that the bookmark is open, different clients can bookmarks. In the event that User enjoys a connection in web and bookmarks that connection. Client label the bookmark. While labeling, client can give possess tag or request that server propose labels. Server gives smothered labels where client can pick tag.

Thusly, client ensures their protection while labeling. All the bookmarking data will be put away in database. In the event that the client looks through a tag, he/she can seek in their bookmarks or in all bookmarks. In the event that the connection has different labels, client hunt tag and different labels down that connections will be shown. Prescribing clientsfor as long as 1 week connections and labels. Gathering client has confinements just in Tag Search. The various administrations, amass client can get to (Search and bookmark, Add bookmark). On the off chance that any of theaspect is obstructed to specific client, at that point he can't get to (Search, Bookmark, and Tag).

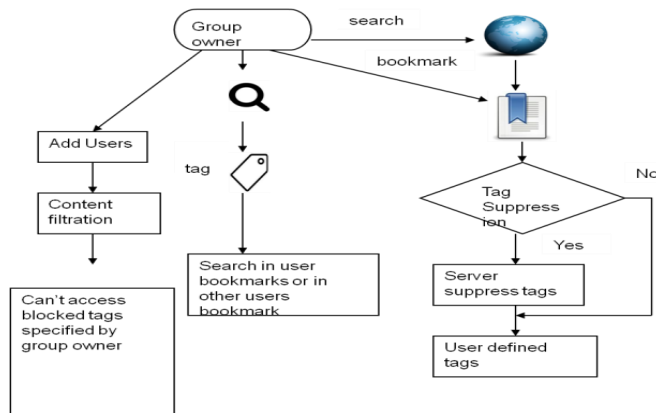


Fig 1: FRAMEWORK OF PERSONAL WEBREVISITATION

4. RESULTS AND DISCUSSIONS

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

5. CONCLUSION

Setting cases and page content are individually composed as probabilistic setting trees and probabilistic term records, which powerfully advance by corruption and support with importance criticism. Our future work includes the prediction of finding the relevant information, finding based on content and context.

6. REFERENCES:

- [1] A. Cockburn, S. Greenberg, S. Jones, B. Mckenzie, and M. Moyle. Improving web pagerevisitation: analysis, design and evaluation. *IT& Society*, 1(3):159–183, 2003.
- [2] L. Tauscher and S. Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies*, 47(1):97– 137, 1997.
- [3] J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: repeat queries in yahoo’s logs. In *SIGIR*, pages 151–158, 2007.
- [4] M. Mayer. Web history tools and revisitation support: a survey of existing proaches and directions. *Foundations and Trends in HCI*, 2(3):173–278, 2009.
- [5] L. C. Wiggs, J. Weisberg, and A. Martin. Neural correlates of semantic and episodic memory retrieval. *Neuropsychologia*, pages 103–118, 1999.
- [6] M. Lamming and M. Flynn. "forget-me-not": intimate computing in support of human memory. In *FRIEND21 Intl. Symposium on Next Generation Human Interface*, 1994.
- [7] E. Tulving. What is episodic memory? *Current Directions in Psychological Science*, 2(3):67–70, 1993.
- [8] C. E. Kulkarni, S. Raju, and R. Udupa. Memento: unifying content and context to aid webpage re-visitation. In *UIST*, pages 435–436, 2010.
- [9] J. Hailpern, N. Jitkoff, A. Warr, K. Karahalios, R. Sesek, and N. Shkrob. Youpivot: improving recall with contextual search. In *CHI*, pages 1521–1530, 2011.
- [10] T. Deng, L. Zhao, H. Wang, Q. Liu, and L. Feng. Refinder: acontext-based information re-finding system. *IEEE TKDE*, 25(9):2119–2132, 2013.

ADAPTIVE QUERY GENERATOR USING FACETS IN DATA MINING

A. MANIMOZHI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Outlier detection can usually be considered as a pre-processing step for locating, in a data set, those objects that do not conform to well-defined notions of expected behavior. It is very important in data mining for discovering novel or rare events, anomalies, vicious actions, exceptional phenomena, etc.

We are investigating outlier detection for categorical data sets. This problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data. In this paper, we propose a formal definition of outliers and an optimization model of outlier detection, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model, we define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently.

We propose two practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail.

Keywords:—Outlier detection, holoentropy, outlier factor, greedy algorithms

1.INTRODUCTION

Outlier detection, which is an active research area refers to the problem of finding objects in a data set that do not conform to well-defined notions of expected behavior. The objects detected are called outliers, also referred to as anomalies, surprises, aberrant, etc. Outlier detection can be implemented as a preprocessing step prior to the application of an advanced data analysis method. It can also be used as an effective tool to discover interest patterns such as the expense behavior of a to-be bankrupt credit cardholder[4]. Outlier detection is an essential step in a variety of practical applications including intrusion detection, health system monitoring, and criminal activity detection in E-commerce, and can also be used in scientific research for data analysis and knowledge discovery in biology, chemistry, astronomy, oceanography, and other fields.

The existing methods for outlier detection are classified according to the availability of labels in the training data sets, there are three broad categories: supervised, semi-supervised, and unsupervised approaches[6]. In principle, models within the supervised or the semi-supervised approaches all need to be trained before use, while models adopting the unsupervised approach do not include the training phase. Moreover, in a supervised approach a training set should be provided with labels for anomalies as well as labels of normal objects, in contrast with the training set with normal object labels alone required by the semi-supervised approach[7]. On the other hand, the unsupervised approach does not require any object label information. Thus the three approaches have different prerequisites and limitations, and they fit different kinds of data

sets with different amounts of label information. The three broad categories of outlier detection techniques are discussed below[8].

The supervised anomaly detection approach learns a classifier using labeled objects belonging to the normal and anomaly classes, and assigns appropriate labels to test objects. The supervised approach has been studied extensively and many methods have been developed. For instance, the group of proximity-based methods includes the cluster-based "K-Means+ID3" algorithm which cascades K-Means clustering and an ID3 decision tree for classifying anomalous and normal objects[10]. The work of Barbara et al is based on statistical testing and an application of Transduction Confidence Machines, which requires k neighbors. Moreover, one-class SVMs have been applied broadly in this field as they do not have to make probability density estimation. A variety of methods based on information theory have also been proposed. The work of Filippone and Sanguinetti proposes a method to control the false positive rate in the novelty detection problem.

The semi-supervised anomaly detection approach primarily learns a model representing normal behavior from a given training data set of normal objects, and then calculates the likelihood of a test object's being generated by the learned model. Zhang et al propose an adapted hidden Markov model for this approach to anomaly detection, while Gao et al propose a clustering-based algorithm which punishes deviation from known labels. Methods that assume availability of only the outlier objects for training are rare, because it is difficult to obtain a training data set which covers all possible abnormal behavior that can occur in the data.

The unsupervised anomaly detection approach detects anomalies in an unlabeled data set under the assumption that the majority of the objects in the data set are normal. Angiulli et al propose a KNN distance-based method. Clustering is another widely implemented method, of which is an example. Moreover, this approach is applied to different kinds of outlier detection tasks and data sets, e.g., conditional anomaly detection, context-aware outliers, and outliers in semantic graphs. As this approach does not require a labeled training data set and is suitable for different outlier detection tasks, it is the most widely applicable. To implement supervised and semi-supervised outlier detection methods, one must first label the training data. However, when faced with a large data set with millions of high-dimensional objects and a low anomalous data rate, picking the abnormal and normal objects to compose a good training data set is time-consuming and labor-intensive. The unsupervised approach is more widely used than the other approaches because it does not need labeled information. If one wants to employ a supervised or semi-supervised approach, an unsupervised method can be used as the first step to find a candidate set of outliers, which will help experts to build the training data set. The unsupervised approach is our research focus in this thesis.

2. LITERATURE REVIEW

Algorithms for Mining Distance-Based Outliers in Large Datasets

In this paper notion of distance-based outliers generalizes the notions of outliers supported by statistical outlier tests for standard distributions. Because this material appears in our preliminary work only provides a brief summary. Algorithms, optimizations, and disk-residency were not the focus of our previous work, but are the focus here. The two simple algorithms having a complexity of $O(k N^2)$, where k and N are the dimensionality and size of the dataset, respectively. The detection of DB-outliers, unlike the depth-based approaches, is computationally tractable for values of $k > 2$. It presents a partitioning-based algorithm in that, for a given dimensionality k , has a complexity of $O(N)$. The algorithm, however, is exponential on k . In some cases, this algorithm outperforms the two simple algorithms by at least an order of magnitude for $k < 4$ and also presents a different version of the partitioning based algorithm for large, disk-resident datasets. That shows that the algorithm guarantees at most 3 passes over the dataset. Again, experimental results indicate that this algorithm is by far the best for $k < 4$.

Distance-Based Detection and Prediction of Outliers

In this paper, the concept of outlier detection solving set, a subset of the input data set representing a model that can be used to predict outliers, is defined. The computational complexity of computing a minimum cardinality solving set, showing that the problem is in general intractable, is analyzed and algorithms that compute with sub-quadratic time requirements a solving set and the top n outliers are provided. The experimental evidence that the solving set is a fraction of the overall data set, and that the false positive rate obtained using the solving set is negligible is given. The ROC analysis of the method to investigate its accuracy in separating outliers from inliers is performed. Results obtained show that using the solving set instead of the data set to predict outliers is efficient and effective. The paper is organized as follows. An overview on unsupervised distance-based outlier detection approaches is given. In this paper, the problems that will be treated are formally defined. The paper defines the concept of solving set and explains how to exploit it to solve the introduced problems. The complexity analysis of the task of computing a solving set is done and methods for computing the top n outliers and the solving set are described. Finally, reports experimental results and discusses the choice of the parameters

Fast Distributed Outlier Detection in Mixed-Attribute Data Sets

In this work First, most approaches to date have focused on detecting outliers in a continuous attribute space. However, almost all real-world data sets contain a mixture of categorical and continuous attributes. The categorical attributes are typically ignored or incorrectly modeled by existing approaches, resulting in a significant loss of information. Second, there have not been any general-purpose distributed outlier detection algorithms. Most distributed detection algorithms are designed with a specific domain (e.g. sensor networks) in mind. Third, the data sets being analyzed may be streaming or otherwise dynamic in nature. Such data sets are prone to concept drift, and models of the data must be dynamic as well. To address these challenges, a tunable algorithm for distributed outlier detection is presented in mixed-

attribute data sets. In this paper proposed anomaly score where in one can effectively identify outliers in a mixed attribute space by considering the dependencies between attributes of different types. A two-pass distributed algorithm for outlier detection based on this anomaly score. This paper also present an approximate scheme by which this algorithm allows for more efficient outlier detection when memory and processor resources are an issue. In order to extend the technique to handle dynamic data sets, where only a single pass can be made, and concept drift is a concern. Given algorithms is to be evaluated on several real and synthetic data sets and confirm the utility of finding outliers in mixed attribute data.

Anomaly Pattern Detection in Categorical Datasets

In this paper [4] proposes a new method for detecting patterns of anomalies in categorical datasets. An assumption, anomalies are generated by some underlying process which affects only a particular subset of the data. The method consists of two steps: First use a "local anomaly detector" to identify individual records with anomalous attribute values, and then detect patterns where the number of anomalous records is higher than expected. Given the set of anomalies flagged by the local anomaly detector, for search over all subsets of the data defined by any set of fixed values of a subset of the attributes, in order to detect self-similar patterns of anomalies. To detect any such subset of the test data which displays a significant increase in anomalous activity as compared to the normal behavior of the system (as indicated by the training data).

- **Anomaly Pattern Detection**

This method can be thought of as generalizing two lines of previous research: the use of standard anomaly detection methods to detect individually anomalous records, and the use of WSARE-2 to detect anomalous clusters of counts in categorical data. The generalize the former method by integrating information from patterns of potentially anomalous records. To extend WSARE by using the information from a local anomaly detector and determining if any subset of the data has more anomalous records than expected. This is distinct from the original formulation of WSARE, which detects subsets with more total records than expected and does not consider whether each individual.

- **Bayesian Network Anomaly Detection**

A Bayesian network is a popular representation of a probability model over the attributes for categorical data because of its parsimonious use of parameters, and efficient learning and inference techniques. Bayes Nets have been used for detecting anomalies in network intrusion detection, detecting malicious emails and disease outbreak detection. A typical anomaly detection approach is to learn the structure and parameters of a Bayes Net using the training data, compute the likelihood of each record in the test dataset given the Bayes Net model, and report test records with unusually low likelihoods as potential anomalies. For our experiments, is used the optimal reinsertion algorithm to learn the structure, and then did a maximum likelihood estimation of the network parameters. For testing a record, for compute the likelihood of that record given the Bayesian Network learned from the training data. A lower value of the likelihood indicates that the record is more anomalous. The log-likelihood value is used as the anomalousness score of each record.

- **Conditional Anomaly Detection**

Work described the conditional anomaly detection method of detecting anomalies in categorical datasets. For any test record t , and any two disjoint sets of attributes A and B , consider the ratio $r(at, bt) = P(at, bt) / P(at)P(bt)$ where at and bt are the value combinations taken by A and B in t respectively. The probability values are estimated from the corresponding counts in the training dataset. An unusually low value of this ratio suggests a strong negative dependence between the occurrences of at and bt in the training data. After observe them together in the test record t , reasonably say that it is anomalous. A low value of $r(at, bt)$ also ensures in enough cases of at and bt in the training data to support the hypothesis of negative dependence. A score is then assigned to the record t based on all such r -values corresponding to all possible pairs of attribute sets. The score is defined as the maximum product of r -values over all possible partitions of the attributes for record t . In our experiments to use the parameter values $k = 2$ and $\alpha = 0.02$ for the conditional method in most cases. Here, k is the maximum set size of A or B . α is the threshold for the r - values to be significant. For the KDD Cup 99 dataset (§3.3), use $k = 1$ since it has a larger number of attributes.

- **WSARE**

The WSARE-2 method searches over all possible one or two component rules in the dataset. Each rule R can be written as $R : A = aj$, where A is a subset of attributes and aj is an assignment of attribute values. WSARE considers only rules with one component (e.g. Country = Japan) or two components (e.g. Country = Japan AND Shipper = Ship Co). It determines whether the count of cases that match the rule in the test dataset is significantly different from the expected count determined by the training dataset. The statistical significance of each rule is determined by using a Fisher's exact test on the two by two table, where $C(R)_{test}$ and $C(R)_{train}$ represent the numbers of test records and training records corresponding to rule R , and C_{test} and C_{train} denote the total numbers of test and training records respectively. To account for multiple hypothesis testing, these p -values are adjusted using a randomization test. In a later version of the algorithm (WSARE-3), the authors consider determining the baseline using a Bayesian Network rather than directly using the counts from the training dataset. For using the algorithm WSARE-2 with up to two component rules for comparing against our methods. To understand the key difference between our current problem and that considered in WSARE. There are two factors to consider. The first factor is that each individual record is individually anomalous with respect to some normal behavior. The second factor is the pattern formed by these anomalies (defined by some constraint of similarity between them) which signifies that the records are generated by the same underlying anomalous process. WSARE does not take the anomalousness of each individual record's attribute values into account, but instead counts the number of records corresponding to a given rule and reports rules for which these counts are anomalous. In our current work, an anomalous process can generate a pattern of anomalous records that are similar with respect to a particular subset of the attributes, but which are anomalous due to unusual values in any (potentially different) random set of attributes. This definition of a pattern is particularly useful an adversarial process creating the anomalies. The adversary might try to

make the generated records look as random as possible, but might be restricted to a particular set of fixed values of some of the attributes. For example, in customs monitoring, a smuggler wants to smuggle goods using a variety of methods to avoid detection, but they might have access to only a particular port or shipping line. In such a case, detecting increased incidence of suspicious activity corresponding to that subset of the data can alert us to the illegal activity.

3. PROPOSED METHOD

In this thesis a formal optimization-based model of categorical outlier detection, for which a new concept of weighted holoentropy which captures the distribution and correlation information of a data set is proposed. To solve the optimization problem a new outlier factor function is derived from the weighted holoentropy and show that computation/updating of the outlier factor can be performed without the need to estimate the joint probability distribution and estimate an upper bound of outliers to reduce the search space. This thesis proposes two effective and efficient algorithms, named the Information-Theory-Based Step-by-Step (ITB-SS) and Single-Pass (ITB-SP) methods. These algorithms need only the number of outliers as an input parameter and completely dispense with the parameters for characterizing outliers usually required by existing algorithms.

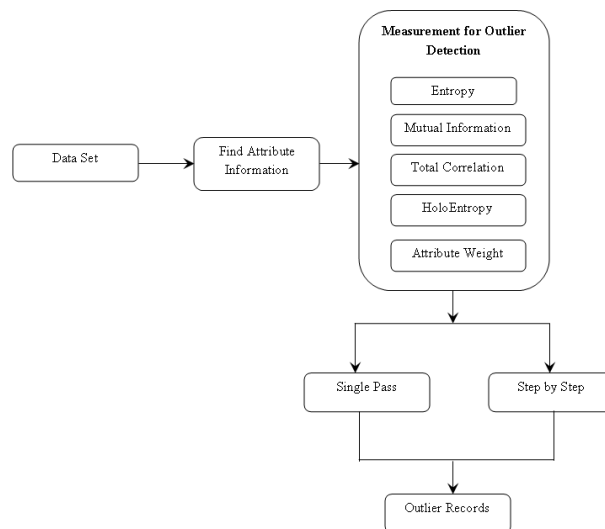


Figure 1 Work Flow of Outlier Detection

4. RESULTS AND DISCUSSIONS

Table 1 shows the sample data set, it will be used for our implementation. First compute the outliers measures entropy, mutual information, total correlation, holoentropy and attribute weighting. After that find outlier factor for each data and then apply two algorithm Single pass and Step by step for detecting outliers.

Table 1 Sample Data Set

Parents	has_nurs	Form	children	Housing	finance	social	health
usual	proper	complete	1	convenient	convenient	nonprob	recommended
usual	proper	complete	1	convenient	convenient	nonprob	priority
usual	proper	complete	1	Critical	convenient	slightly_prob	not_recom
usual	proper	complete	1	Critical	convenient	problematic	recommended
usual	proper	complete	2	less_conv	inconv	nonprob	recommended
usual	proper	complete	2	less_conv	inconv	nonprob	priority
pretentious	less_proper	complete	1	convenient	convenient	problematic	not_recom
pretentious	less_proper	complete	1	convenient	inconv	nonprob	recommended
pretentious	less_proper	complete	1	convenient	inconv	nonprob	priority

For experiment the real time data set is taken from UCI data repository. The Nursery Data set is used. It contains eight attributes: parents, has_nurs, form, children, housing, finance, social, health. Analyze the performance of effectiveness and efficiency our methods. To test effectiveness, compare ITB-SS and ITB-SP with competing methods on synthetic and real data sets. For the efficiency test, evaluations is conducted on synthetic data sets to show how running time increases with the number of objects, the number of attributes and the number of outliers. Table 1 is given below shows example Dataset.

The data set used is the public, categorical “Nursery”, with 160 objects and 8 attributes. This data contains a very small class of 10 objects. Since the data does not have explicitly identified outliers, it is natural to treat the objects of the smallest class as “outliers.” Therefore, check whether objects from this class will be detected for $o = 1, \dots, 10$

The results indicate that our proposed factor $OF(x_o)$ for ITB better reflects the intuitive understanding of the data set. also compare the effectiveness of different methods on synthetic data sets in a relatively ideal setting, since the generated outliers are usually more distinctive than those in real data and the outliers “truth” can be used to verify whether an outlier algorithm is able to find them.

To measure the time consumption with increasing numbers of objects, attributes and outliers, employ GAClust to generate synthetic data sets for these experiments. In the “objects

increasing" test, the number of objects is increased from 300 to 1000. In the "attributes increasing" test, the number of attributes increases from 6 to 30. In the "percentage of outliers increasing" test, assume the percentage of outliers in a data set is increased from 10 to 50 percentages. Efficiency tests suggest ITB-SP and ITB-SS are efficient. They are particularly appropriate for large data sets with high dimensionality, and are also suitable for data sets with a high percentage of outliers.

5. CONCLUSION

This thesis formulated outlier detection as an optimization problem and proposed two practical, unsupervised, 1-parameter algorithms for detecting outliers in large-scale categorical data sets. The effectiveness of our algorithms results from a new concept of weighted holoentropy that considers both the data distribution and attribute correlation to measure the likelihood of outlier candidates, while the efficiency of our algorithms results from the outlier factor function derived from the holoentropy. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution. Based on this property, apply the greedy approach to develop two efficient algorithms, ITB-SS and ITB-SP that provide practical solutions to the optimization problem for outlier detection. We also estimate an upper bound for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows us to further reduce the search cost. The proposed algorithms have been evaluated on real and synthetic data sets, and compared with different mainstream algorithms. First, our evaluations on a small real data set and a bundle of synthetic data sets show that the proposed algorithms do tend to optimize the selection of candidates as outliers. Moreover, our experiments on real and synthetic data sets in comparison with other algorithms confirm the effectiveness and efficiency of the proposed algorithms in practice. In particular, show that both of our algorithms can deal with data sets with a large number of objects and attributes.

6. REFERENCE

- [1] E.M. Knorr and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24rd Int'l Conf. Very Large Data Bases (VLDB '98), 1998.
- [2] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, Feb. 2006.
- [3] M.E. Otey, A. Ghoting, and S. Parthasarathy, "Fast Distributed Outlier Detection in Mixed-Attribute Data Sets," Data Mining and Knowledge Discovery, vol. 12, pp. 203-228, 2006.
- [4] K. Das, J. Schneider, and D.B. Neill, "Anomaly Pattern Detection in Categorical Data Sets," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), 2008.
- [5] E.M. Knorr and R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. 24rd Int'l Conf. Very Large Data Bases (VLDB '98), 1998.
- [6] S. Ramaswamy, R. Rastogi, K Shim "Efficient Algorithms for Mining Outliers from Large Data Sets " in proceeding of the 2000 ACM SIGMOD international conference ,Volume 29 Issue 2, June 2000.

- [7] Amol Ghoting, Srinivasan Parthasarathy, and Matthew Eric Otey ,” Fast Mining of Distance-Based Outliers in High Dimensional Datasets” Department of Computer Science and Engineering The Ohio State University, Columbus, OH 43210, USA November 9, 2005.
- [8] A.KrishnaKumar, D.Amrita, N.Swathi ,” Mining Association Rules between Sets of Items in Large Databases “International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386,Volume-1, Issue-5, April 2013
- [9] Varun Chandola, Arindam Banerjee and Vipin Kumar,” applications of anomaly detection for discrete sequences” Volume-1, Issue-3, April 2001
- [10] H. D. K. Moonesinghe, Pang-Ning Tan,” Outlier Detection Using Random Walks ” Department of Computer Science & Engineering Michigan State University East Lansing, MI 48824

ADAPTIVE WILDCARD RULE CACHE MANAGEMENT FOR SOFTWARE DEFINE NETWORKS

S. DHANARAKSHANA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Outlier detection can usually be considered as a pre-processing step for locating, in a data set, those objects that do not conform to well-defined notions of expected behavior. It is very important in data mining for discovering novel or rare events, anomalies, vicious actions, exceptional phenomena, etc.

We are investigating outlier detection for categorical data sets. This problem is especially challenging because of the difficulty of defining a meaningful similarity measure for categorical data. In this paper, we propose a formal definition of outliers and an optimization model of outlier detection, via a new concept of holoentropy that takes both entropy and total correlation into consideration. Based on this model, we define a function for the outlier factor of an object which is solely determined by the object itself and can be updated efficiently.

We propose two practical 1-parameter outlier detection methods, named ITB-SS and ITB-SP, which require no user-defined parameters for deciding whether an object is an outlier. Users need only provide the number of outliers they want to detect. Experimental results show that ITB-SS and ITB-SP are more effective and efficient than mainstream methods and can be used to deal with both large and high-dimensional data sets where existing algorithms fail.

Keywords:—Outlier detection, holoentropy, outlier factor, greedy algorithms

1. INTRODUCTION

In a typical IT infrastructure and service delivery management organization, experts are often organized into groups, each of which has the expertise to solve certain types of problems. Skilled expert groups need to be quickly assigned to bring an abnormal service back tonormal because an IT service provider typically signs up a Service Level Agreement (SLA) [1] with their users. However, the increasing complexity of the IT infrastructure and service deliverymakes the types of reported troubles diverse [2]. Moreover, the trouble description in a ticket is often vague and may not contain the actual root cause of the trouble [3]. Thus, it is difficult to find a right expert group to solve the trouble. Many of the tickets have to bounce among multiplegroups before being transferred to the group with the right expertise. We call the process that a ticket is transferred among various expert groups as the trouble ticket routing [4]. Obviously, if a ticket is mistakenly transferred to an expert group that cannot solve the trouble, it might lead to a long maintenance time and violate the SLA.

Although powerful hardware and software tools have been used in modern IT infrastructure, it is still inevitably to suffer from system faults, even system failure, due to software aging operation errors, etc [5]. When an event (which is not part of the standard operation of a service that may cause an interruption or a reduction) happens, or when using an

IT service, errors, faults, difficulties or special situations (that need attention from system management experts) occur, a trouble ticket is then generated in the ticketing system [6]. The IT infrastructure and service delivery management system is responsible for dealing with the trouble tickets in time.

Software-Defined Networking (SDN) is a brand new network architecture that provides a global view of network state for network administrators to manage network services. SDN controller maintains the flow tables in the switches to comply with network policies. The flow tables are implemented by the TCAM in the modern switches. TCAM can look up a packet's header and compare the matching patterns of the packet to the match field of all rules in the flow table in parallel. In other words, it can forward packets at line rate. Although TCAM can forward packets fast, there are only 2-20K flow table sizes in commodity SDN switches which are much less than RAM based storage. Therefore, the TCAM capacity problem is an important issue in SDN. Previous works on using TCAM efficiently could be classified into three categories, packet classification compression, rules distribution, and rules caching. Packet classification compression is a technique that combines two similar rules into a new wildcard rule, which is semantically equivalent to the original rules. As a result, we can reduce the number of required rules. Rules distribution technique [7, 8, and 9] splits the set of rules which are safe according to the network policies and distributes them along the network path. In rule caching technique [10], it treats TCAM as a cache which stores the most popular or most-likely matched rules in the future. Once cache miss happens, we should choose the victim rules and evict the victim rules out in order to cache the cache miss rule into TCAM. One of the rule caching techniques is wildcard-rule caching. Wildcard-rule caching could keep more TCAM space than exact-match rule matching.

However, there are different priorities between different rules according to the network policies. If two wildcard rules overlap with each other and we only cache the lower-priority one in the TCAM, the packets matching the overlapping field space of these two rules will improperly match lower-priority rule. In other words, only caching a wildcard rule overlapped with others would cause ambiguous matching when the packets come. Therefore, the most difficult challenge for wildcard-rule caching is to deal with the rule dependency problem. The authors use the cover-set method to solve the wildcard-rule dependency problem. The cover-set method creates a small number of new rules that cover many low priority rules overlapped with the high-priority rule. High priority rules usually have higher weight due to their larger field space. Therefore, cover-sets help us to avoid caching high weight rules together with many low-weight rules overlapped with them. The Cover-Set Caching (CSC) algorithm calculated the contribution value of each uncached rule and cached the rule has the maximum contribution value into the TCAM until the TCAM is full. Since the CSC algorithm only considers the contribution value of each uncached rule, the authors proposed a Wildcard-Rule Caching (WRC) algorithm, which considers the accumulated contribution value for a set of rules. They calculated the accumulated contribution value of a set of related rules and cached the set of rules have the maximum accumulated contribution value into the TCAM until the TCAM is full.

2. LITERATURE REVIEW

Qihong Shao, Yi Chen et.al. states the mechanism that can improve the overall efficiency of ticket routing, measured by the Mean number of Steps To Resolve (MSTR) the tickets. The first question is what information in ticket data can be used to improve MSTR. Typically, a problem ticket contains two types of information: (1) ticket content that includes problem description and diagnostic data, (2) resolution sequence that shows how it was routed before it reached the resolver group. In the sample ticket shown in Table 1, the entries compose the ticket content, while the extracted group names hSMRDX, SSDSISAP, ASWWCUST, SSSAPHWOAi form its resolution sequence. Our study in this paper focuses on the resolution sequences only. The mining resolution sequences alone can significantly improve the efficiency of ticket routing and identify resolver groups more quickly – a surprisingly encouraging discovery. Our strategy is to mine resolution sequences of solved tickets, which could guide the routing decisions for new tickets. With in-depth analysis, we find that in many cases, long resolution sequences were results of a few local mis-routing decisions, while the majority of the local ticket transfer decisions were logically correct. Intuitively, for a specific type of problems, these local ticket transfer decisions reflect the functional relationships between expert groups.

Peng Sun.et.al. Demonstrates the hybrid method, they first identify a set of existing tickets that are similar to the new ticket in content. Then, we use the resolution sequences of these similar tickets to generate a weighted Markov model. Compared with existing approach in this model tickets having different similarity levels are weighted differently. To evaluate content similarity, we extend the existing text-mining techniques. Specifically, we develop a Cosine- similarity-based weight function for model generation. Our study shows that the parameters in these weight functions can make a salient difference of the model effectiveness. Thus for the weight function, we develop an algorithm to tune its parameters to optimally fit the new ticket based on the models built for the historical ticket that is most similar to this ticket. Furthermore, we observe the situation where there are a lot of tickets that are dissimilar to the new ticket, whose combined weight may slow down the effect of the highly similar tickets. Thus, we performed a model normalization to generate a training set of tickets with uniformly distributed similarities, even though the original training set can have a skewed distribution on similarities.

Gengxin Miaoet.al. States the problem ticket resolution is critical to the IT services business. A service provider might need to handle, on a daily basis, thousands of tickets that report various types of problems from its customers. The service provider's ability to resolve the tickets in a timely manner determines, to a large extent, its competitive advantage. To manage ticket resolution effectively, human experts are often organized into expert groups, each of which has the expertise to solve certain types of problems. As IT systems become more complex, the types of reported problems become more diverse. Finding an expert group to solve the problem specified in a ticket is a long-standing challenge for IT service providers. A typical ticket processing system works as follows. A ticket is initiated by a customer or by internal staff, and is

subsequently routed through a network of expert groups for resolution. The ticket is closed when it reaches a resolver group that provides the solution to the problem reported in the ticket.

Sara Sabour.et.al. proposes the convolution neural networks (CNNs) use translated replicas of learned feature detectors. This allows them to translate knowledge about good weight values acquired at one position in an image to other positions. This has proven extremely helpful in image interpretation. Even though we are replacing the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by-agreement, we would still like to replicate learned knowledge across space. To achieve this, we make all but the last layer of capsules be convolutional. As with CNNs, we make higher-level capsules cover larger regions of the image. Unlike max-pooling however, we do not throw away information about the precise position of the entity within the region. For low level capsules, location information is "place-coded" by which capsule is active. As we ascend the hierarchy, more and more of the positional information is "rate-coded" in the real-valued components of the output vector of a capsule. This shift from place-coding to rate-coding combined with the fact that higher-level capsules represent more complex entities with more degrees of freedom suggests that the dimensionality of capsules should increase as we ascend the hierarchy.

3. PROPOSED METHODOLOGY

The proposed system focuses around the investigation of ticket routing models by breaking down this data from recorded inconvenience tickets. To assess the execution of directing automated proposal calculations, we lead broad examinations on a genuine ticket informational collection. The exploratory outcomes demonstrate that the proposed models and calculation can adequately abbreviate the automated routing technique and transfer the ticket to a right expert team in order to produce the highly efficient solution. Thus with a high proportion of the number of effectively settled tickets to the aggregate number of tickets, particularly for the long directing groupings produced from manual assignments. These models and calculations have the capability of being utilized in a ticket directing suggestion motor to significantly decrease human intercession in the directing procedure and make the process highly effective.

ADVANTAGES

1. These are used to construct routing models that effectively represent this routing information hidden in historical tickets.
2. Three routing models and the relating routing suggestion calculations are proposed by mining the mix of issue portrayals and routing groupings to enhance the execution of ticket routing and it simple.
3. On validating the effectiveness and robustness of our routing models and recommendation algorithms.
4. This can effectively shorten the average length of ticket routing sequences, especially for the long routing sequences generated from manual assignment.

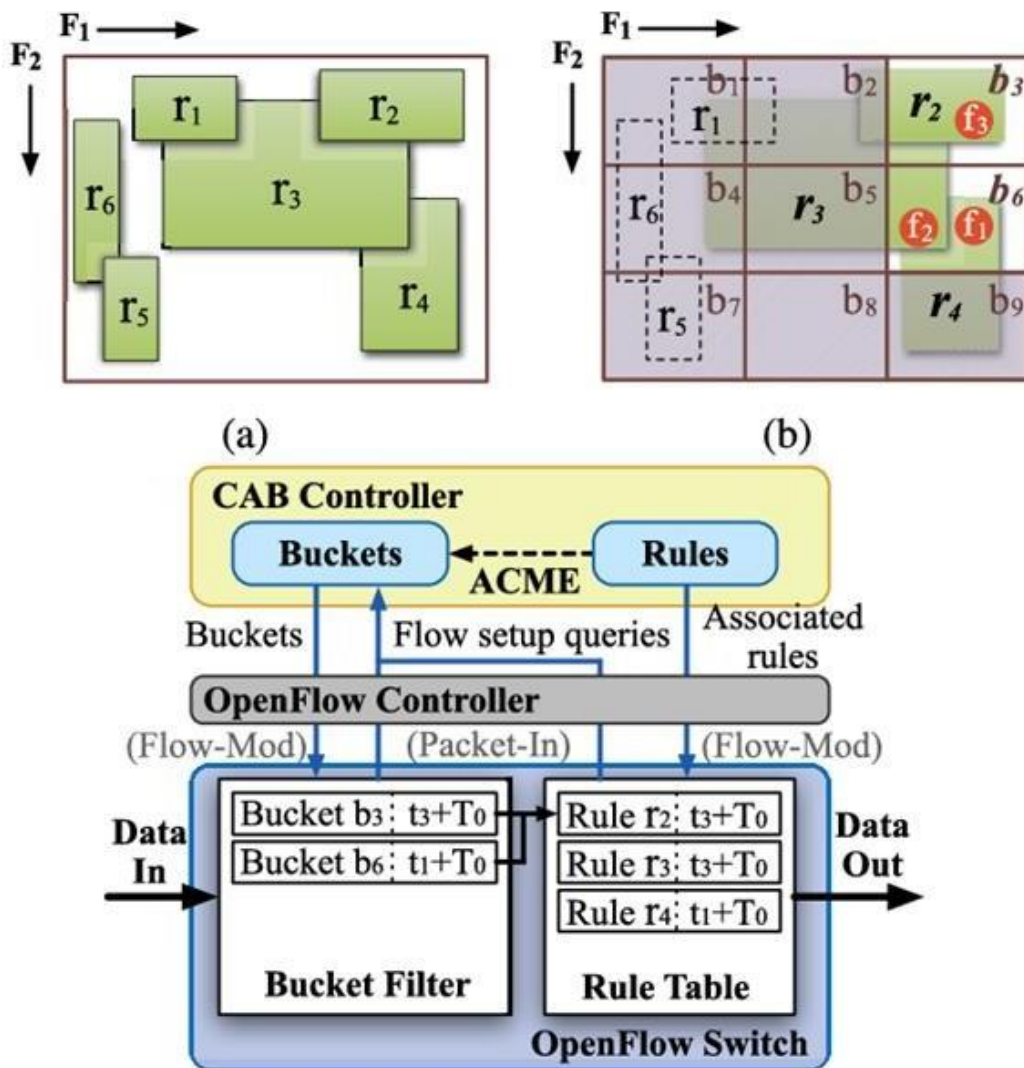


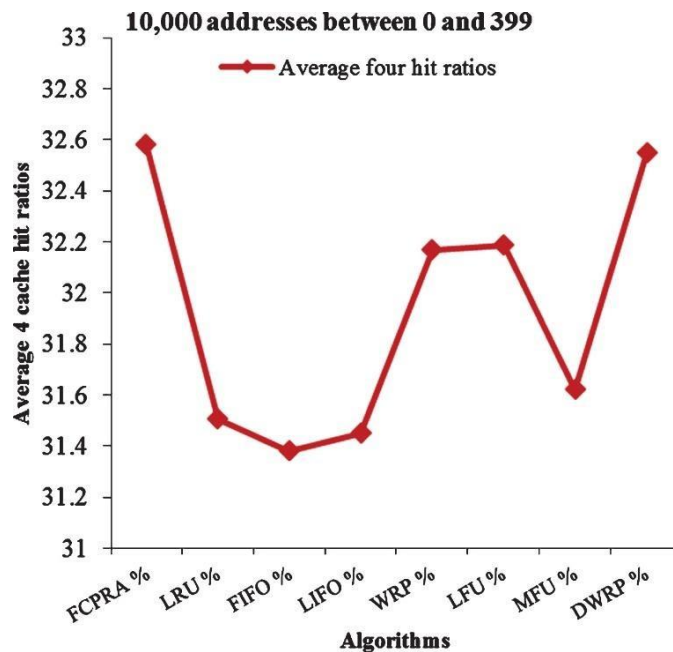
Figure 1: CAB rule catching method

4. RESULTS AND DISCUSSIONS

To evaluate the performance of PipeCache, we compare it with the exact-match scheme and the wildcard-match scheme based on one flow table in terms of the cache hit rate. Here we call them MicroCache and MegaCache respectively for simplicity. The cache hit rate is an important criterion to indicate cache algorithms. In the three caching schemes, TCAM is used as a cache in the OpenFlow switch. So, the size of TCAM is positively correlated with the cache hit rates. Moreover, different trace localities can lead to different test results. We generate multiple-stage trace files based on a range of trace localities to test the relationship between the performance of different caching schemes and the size of trace localities.

Cache hit rates with different ruleset sizes. First, we set the locality parameter b to 10% to generate the rule sets, set the TCAM size to 5% of the ruleset size, and show the cache hit rates of three schemes in different rule sets and traces. Figure 7 shows the average cache hit rates with

different sizes of rule sets. As shown in this figure, our design PipeCache achieves the highest cache hit rate under both seed files while the MegaCache achieves the second-highest cache hit rates. MicroCache has the most inferior performance. When the TCAM size is 5% of the rule size, and the locality parameter b is 10%, PipeCache can outperform the MegaCache and MicroCache 4.23% and 18.25% respectively on average.



Graph.1 Shown the average cache hit ratio

Cache memory is used to improve CPU performance and reduce response time. Due to the cost and limited size of cache compared to other devices that store information, an alternative policy is used to select and extract a page to make space for new pages when the cache is filled. Many algorithms have been introduced which performance depends on a high-speed web cache, but it is not well optimized. The general feature of most of them is that they are developed from the famous LRU and LFU designs and take advantage of both designs. In this research, a page replacement algorithm called FCPRA (Fuzzy Clustering based Page Replacement Algorithm) is presented, which is based on four features. When the cache space can't respond to a request for a new page, it selects a page of the lowest priority cluster and the largest login order; then, removes it from the cache memory. The results show that FCPRA has a better hit rate with different data sets and can improve the cache memory performance compared to other algorithms.

5. CONCLUSION

In conclusion, adaptive wildcard rule cache management represents a crucial advancement in optimizing Software Defined Networks (SDNs) to meet the evolving demands of

modern networking environments. By dynamically adjusting the contents of rule caches based on changing network conditions, traffic patterns, and application requirements, adaptive caching systems offer significant benefits in terms of performance, scalability, and resource utilization. Through sophisticated algorithms and real-time monitoring techniques, adaptive caching systems can effectively balance responsiveness, stability, and efficiency. They enable SDNs to adapt quickly to fluctuations in network traffic, rule churn, and resource constraints, ensuring optimal packet processing and minimizing latency. Dynamic cache management allows SDNs to scale effectively to handle growing traffic volumes and rule complexities. By intelligently adapting to changes in network dynamics, adaptive caching systems support the seamless expansion of network infrastructures without sacrificing performance

6. REFERENCE

- [1] B. Heller et al., —ElasticTree: Saving energy in data center networks,|| in Proc. NSDI, vol. 3, pp. 19–21 in 2010
- [2]. A.R.Curtis, —DevoFlow: Scaling flow management for high performance networks,|| Comput. Commun. Rev., vol. 41, no. 4, pp. 254–265, Aug. 2011.
- [3]. P. Porras et al., —A security enforcement kernel for OpenFlow networks,|| in Proc. 1st Workshop Hot Topics Softw. Defined Netw., 2012, pp. 121–126.
- [4]. M. Yu, L. Jose, and R. Miao, —Software defined traffic measurement with OpenSketch,|| in Proc. NSDI, vol. 13. 2013, pp. 29–42.
- [5]. H. Kim and N. Feamster, —Improving network management with software defined networking,|| IEEE Commun. Mag., vol. 51, no. 2, pp. 114–119, Feb. 2013.
- [6]. M. Moshref, M. Yu, R. Govindan, and A. Vahdat, —DREAM: Dynamic resource allocation for software-defined measurement,|| in Proc. ACM Conf. SIGCOMM, 2014, pp. 419–430.
- [7]. A. Gember-Jacobson et al., —OpenNF: Enabling innovation in network function control,|| in Proc. ACM Conf. SIGCOMM, 2014, pp. 163–174.
- [8]. R. Wei, Y. Xu, and H. J. Chao, —Finding nonequivalent classifiers in Boolean space to reduce TCAM usage,|| IEEE/ACM Trans. Netw., vol. 24, no. 2, pp. 968–981, Apr. 2016.
- [9]. Naga Katta, Omid Alipourfard, Jennifer Rexford and David Walker, —Infinite CacheFlow in Software-Defined Networks,|| ISBN: 978-1-4503-2989-7, August 2014.
- [10]. Bo Yan, Yang Xu, Hongya Xing, Kang Xi, H. Jonathan Chao, —CAB: A Reactive Wildcard Rule Caching System for Software-Defined Networks,|| Association for Computing Machinery, August 2014.

AN IMPROVED OCCLUSION THERAPY FOR AMBLYOPIC PATTERN TECHNIQUE

R. YOGA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Drugs are a major problem in economic and many losses in worldwide. In this project, an image processing approach is proposed for identifying drugged eye based on convolutional neural network. According to the CNN algorithm, eye image details are taken by the existing packages from the front end used in this project. However, it can take a few moments. So, this proposed system can be used to identify drugged eyes quickly and automatically. The eye images dataset are taken from Kaggle. These images are taken as a training set for this drugged eye detection. This proposed approach is composed of the following main steps that getting input image, Image Preprocessing, identifying reddish places, highlight those affected places, Verifying training set, showing result. Few types of eyes like drugged socially may missed to identify. This approach was tested according to drugged eye type and its' stages, such as drug consumed and not consumed. The algorithm was used for detecting the white area of eye present in given input image. Images were provided for training, such as drugged eye images and normaleye images. Before the image processing, images were converted to color models, because of find out the most suitable color model for this approach. Local Binary Pattern was used for feature extraction and Support erosion method was used for creating the model. According to this approach, drugged eyes can be identified in the average accuracy of 95%.

Keywords: CNN algorithm, eye detection, Local Binary Pattern, convolutional neural network

1. INTRODUCTION

The term Pattern Visually Evoked Potential (P-VEP) refers to electrical potentials which are recorded from the scalp overlying visual cortex, P-VEP waveforms are extracted from the electro-encephalogram (EEG) by signal averaging. P-VEPs are used primarily to measure the functional integrity of the visual pathways from retina via the optic nerves to the visual cortex of the brain. P-VEPs better quantify functional integrity of the optic pathways than scanning techniques such as magnetic resonance imaging (MRI). The computer programs save a defined time period of EEG activity following a visual stimulus, which is repeated over and over adding the signals together. Based on the signal to noise ratio, an evoked potential can be seen forming following only a few stimuli such as flashes of light. Therefore the diagnosis based on amplitude and latency in time domain is not alone sufficient.

Hence other components should also be taken into consideration. The spectral analysis of pattern visual evoked potential can yield useful information when it is performed carefully. Classification of the severity of amblyopic and measuring the changes are vital for assessing the occlusion therapy. Present clinical studies use analysis of visual acuity manually. This method is time-consuming process which requires significant training and exercise and is uncovered to observe error. Amblyopic (lazy/squint eye) is the most common cause of monocular visual

impairment in both children and young adults. Occlusion therapy is the treatment that aims to encourage the weaker or lazy eye to start working and stimulate vision hence helping the part of the brain that manages sight to further improvement. Unfortunately the patient is not aware of any symptoms until it is too late for effective treatment. By using computational intelligence technique in predicting the acuity, the analysis of vision becomes more accurate. This can be achieved by using Linear Discriminant Analysis Algorithm for feature extraction, k-means and Support Vector Machine algorithms for classification. Through analysis of pattern visual evoked potential response of retina and the occipital region, a way will be lined for early diagnosis of amblyopic (lazy / squint eye) and prediction during the occlusion therapy process. In this paper we present Kmeans clustering algorithm and Support vector machines (SVMs) to classify amblyopic (lazy/squint eye) subjects according to changes in pattern visual evoked potential spectral components and automatic analysis of the human eye under normal and amblyopic (lazy/squint eye) conditions in virtual environment.

In the experiment we used a dataset of EEG signals. The set contains data from 50 subjects acquired from 180 trails. The subjects were relaxed, sat in normal chairs with arms resting on their legs at a distance of 90cm away from monitor. The experimental task is depicted in Figure. Symbol was flashed on a computer screen, P-VEP waveforms were recorded from 3 electrodes. Reference electrode usually placed on the earlobe, on the midline on top of the head or on the forehead. Ground electrode can be placed at any location. The time period analyzed is usually between 100 and 500 milliseconds following onset of each visual stimulus. When testing young infants, analysis time was 300 msec or longer because components of the P-VEPs may have long peak latencies during early maturation.

The main outcome measures of this prospective cross-sectional observational study were prototype acuity, grating acuity, and PR-VEP parameters of amplitude and latency. The study protocol was approved by the Committee of Ethics in Research of the Federal University of São Paulo (approval number 0502/08) and conducted in accordance with the tenets of the Declaration of Helsinki, and informed consent was obtained from the parents of each child before testing. Children with anisometric and/or strabismus amblyopia who participated in the study were recruited from pediatric private practice and the Strabismus Section of the Department of Ophthalmology and Visual Sciences, Federal University of São Paulo. Inclusion criteria were a previous diagnosis of amblyopia by ophthalmic examination, inter-ocular acuity difference of at least two lines using the conventional printed Snellen chart, best optical correction, and normal fundus. An isometropia was determined by one or more of the following: a difference of at least 1.00 diopter (D) in the spherical component; a difference of 0.75 D in the cylindrical component; or a spherical equivalent difference of more than 1.50 D. The exclusion criteria were the presence of any other eye condition that could decrease visual acuity; any neurological disease; or a history of seizures or use of anti-seizure medication. The type of occlusion therapy was noted for each participant as full-time (patching during all awake hours) or part-time (patching during part of awake hours), the eye to be patched (if alternate or not), and the compliance of the prescribed patching. A comprehensive ophthalmic examination, which included an external eye

examination, ocular motility assessment, bio microscopy, cycloplegic refraction, and fundus assessment with indirect binocular ophthalmoscopy, was performed before electrophysiological testing by a pediatric ophthalmologist.

2. LITERATURE REVIEW

Amblyopia, commonly known as "lazy eye," is a visual disorder that occurs when the brain favors one eye over the other. This preference leads to reduced vision in the weaker eye. While traditionally treated through patching or penalization of the stronger eye, contemporary research increasingly focuses on binocular therapies that aim to integrate both eyes and improve visual function. Here's a concise literature review on the binocular approach to amblyopia therapy. The literature extensively covers the pathophysiology and clinical presentation of amblyopia. Early detection and intervention are crucial to prevent long-term visual impairment

The authors stated that over the past two decades, biometric recognition has been exploded into a plethora of different applications around globe. This proliferation was attributed to high levels of authentication accuracy and user convenience which biometric recognition systems afford end-users.

However, in spite of the triumph of biometric recognition systems, there are a count of outstanding problems and concerns pertaining to various sub-modules of biometric recognition systems that develop an element of mistrust in their use - both by scientific community and also public at large. Some of the problems contain:

- Questions related to the system recognition performance,
- Security (spoof attacks, adversarial attacks, template reconstruction attacks and the demographic information leakage),
- Uncertainty over the bias and fairness of systems to all users,
- Explainability of seemingly black-box decisions made by the most recognition systems,
- Concerns over the data centralization and the user privacy.

In this project, the authors provided an overview of each of the above aforementioned open-ended issues. They surveyed work that has been conducted to address each of the concerns and highlight these issues requiring future attention. Finally, the authors provided insights into how biometric community can address the core biometric recognition systems design problems to better instill trust, fairness, and the security for all.

They concluded that accurate as well as reliable automatic person identification was becoming a necessity in a host of applications including the national ID cards, access control, border crossings, payments, etc. Biometric recognition stands as perhaps the well equipped technology to meet the need. Indeed, biometric recognition systems have now matured to point at which they can surpass the human recognition performance or the accuracy under certain conditions.

However, many unsolved problems remain prior to the acceptance of biometric recognition systems with trustworthy. In this paper, it is highlighted 5 major areas of research that must be worked in order to establish the trustworthiness in biometrics: a) Performance

Robustness and Scalability, b) Security, c) Explainability, Interpretability d) Biasness and Fairness and e) Privacy.

In each of these areas, the authors have provided a problem definition, explained importance of the problem, cited the existing work on each respective topic, and also concluded with suggestions for future research. By better addressing each of the major areas, biometric recognition systems can be made not only trustworthy, but also accurate. These benefit the researchers behind recognition systems, general public using the systems, and the policy makers regulating the systems.

One practical potential avenue for encouraging more trustworthy biometric recognition systems is the “The Grand Challenge in Trustworthy Biometrics”. Perhaps such a challenge, hosted by the government agency, say NIST, evaluates the biometric recognition systems on each of five categories listed above. Systems that met certain the quantitative thresholds for these 5 categories could be certified as “trustworthy”. In this manner, end-users know not only how accurate the system is, but also how the “trustworthy” it is.

In this paper the authors stated that shift work can be a risk factor for the number of various somatic and psychological health conditions, particularly sleep disorders. Shift workers would sleep less than day workers, and 20–42% of them affected from difficulties initiating as well as maintaining sleep, that result in reduced capacity to work and social life. A common coping strategy might be alcohol usage that presents the health and safety hazard as it further impairs quality of sleep and exacerbates sleepiness in workplace. This review aimed to assess extent of such possible connection problems. They performed a systematic search of scientific literature on the shift work and the alcohol consumption in PsycInfo, PubMed, and Cochrane Library. Only original studies comparing the shift workers with the non-shift workers are included. The recommendations of Preferred Reporting-Items of Systematic Reviews and Meta-Analyses are followed.

3. PROPOSED METHODOLOGY

The project likely utilizes the principles of amblyopia detection, which involves assessing visual acuity and binocular vision to identify abnormalities or discrepancies between the eyes. common treatment method where the stronger eye is temporarily patched to force the weaker eye to work harder and develop better visual acuity. Occlusion therapy helps stimulate the neural pathways associated with the weaker eye, promoting visual development. Vision therapy involves a series of exercises and activities designed to improve visual skills and coordination. These exercises may include eye tracking, convergence training, and visual processing activities aimed at enhancing the interaction between the eyes and the brain. These interactive tools can make therapy more enjoyable for children and help maintain their interest and compliance with treatment.

ADVANTAGES

- Early detection and intervention can help prevent long-term visual impairment and promote better visual outcomes.

- Eye care professionals can tailor treatment plans based on the specific needs and characteristics of each individual.
- Effective treatment can help individuals develop better depth perception and visual integration between the eyes.
- The availability of digital therapeutic tools and apps makes vision therapy more accessible and engaging for patients, especially children.

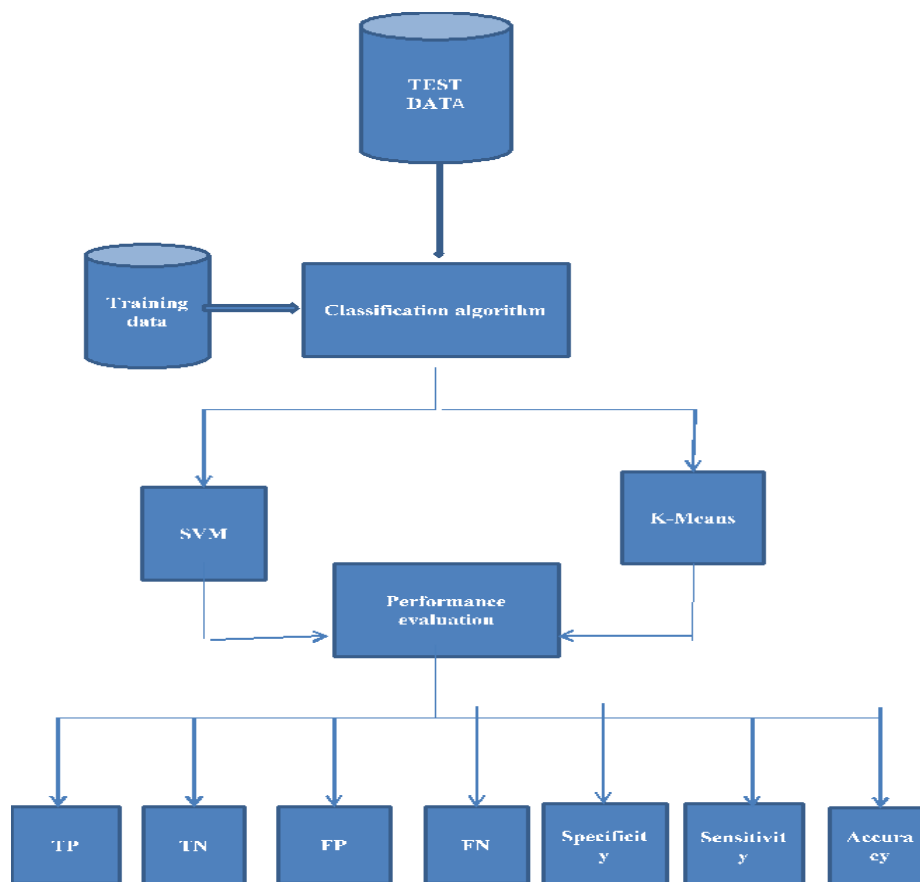


Figure 1: classification tree for eye detection

4. RESULTS AND DISCUSSIONS

After applying the training set images, two base folders were used for identifying drugged eyes according to its accuracy. These files are called as two classes’ dataset. Another way is counting number of reddish places to identify drug consumed eye according to its stage. This method is called as alternative method.

Every training and testing time, rows of training files were shuffled randomly for increasing the accuracy of the model. Each training file was verified and tested in five times and accuracy was taken. Average of these accuracies is taken as accuracy of each model. Using this image dataset, two types of categories were found. Such as drugged eye and not drugged eye.

1. Accuracy is very high here.

2. Enhancing the values of drugged eye detection.
3. It takes only few seconds to provide exact result.
4. Result is provided with the high accuracy rate.
5. Applicable to both low and high pixel images.



Figure 2: Sample Drugged detection model

5. CONCLUSION

In conclusion, the drugged eye detection project presents a promising avenue for enhancing safety and security measures in various contexts, including law enforcement, transportation, and workplace safety. Through the integration of advanced image processing techniques and machine learning algorithms, the system demonstrates its potential to accurately identify indicators of drug impairment based on ocular characteristics. As society grapples with the challenges posed by drug abuse, this technology offers a proactive solution to mitigate risks and uphold public welfare. With further refinement and implementation, drugged eye detection stands poised to contribute significantly to the broader efforts aimed at fostering a safer and more secure environment for all.

6. REFERENCES

- [1] Abdulhamit Subasi a , * , M. Ismail Gursoy (a). International Burch University, Faculty of Engineering and Information Technologies, Sarajevo, Bosnia and Herzegovina
- [2] Kahta Vocational School of Higher Education, Adiyaman University, Adiyaman, Turkey.
- [3] Kantardzic M., “ Data Mining: Concepts, Models, Methods, and Algorithms ”, IEEE Press & John Wiley, November 2002.
- [4] Kantardzic M., “ Data Mining: Concepts, Models, Methods, and Algorithms ”, IEEE Press & John Wiley, November 2002.
- [5] London: Springer. Adeli, H., Zhou, Z., & Dadmehr, N. (2003). Analysis of EEG records in an epileptic Patient using wavelet transforms. *Journal of Neuroscience Methods*, 123, 69–87.
- [6] Levi DM, Knill DC, Bavelier D. Stereopsis and amblyopia: a mini-review. *Vision Res.* 2015;114:17-30.

- [7] 7.Mendonça RH, Abbruzzese S, Bagolini B, Nofroni I, Ferreira EL, Odom JV. Visual evoked potential importance in the complex mechanism of amblyopia. *Int Ophthalmol.* 2013; 33(5):515-9.
- [8] Oner A, Coskun M, Evereklioglu C, Dogan H. Pattern VEP is a useful technique in monitoring the effectiveness of occlusion therapy in amblyopic eyes under occlusion therapy. *Doc Ophthalmol.* 2004;109(3):223-7.
- [9] Hess RF, Thompson B. Amblyopia and the binocular approach to its therapy. *Vision Res.* 2015;114:4-16.
- [10] Sokol S. Abnormal evoked potential latencies in amblyopia. *Br J Ophthalmol.* 1983;67(5): 310-4.

CONGESTION CONTROL MECHANISM FOR BACKGROUND DATA TRANSFERS WITH LATENCY REDUCERY

P. SATHYAPRIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

One of the challenging issues for supporting emergency services in wireless networks is coordinating the network under emergent situations. Cooperative communication (CC) is a promising approach which can offer significant enhancements in multi-hop wireless networks. This paper investigates the potential issues in using this communication paradigm to support emergency services. This paper proposed a promoting energy-efficient and congestion-aware cooperative networking for emergency services based on the idea of Do-It-Yourself. We propose a novel cross-layer design which jointly considers the problems of route selection in network layer, congestion and non-cooperation avoidance among multiple links in MAC layer under cooperative multi-hop wireless environments.

This paper formulated the multi-hop cooperative flow routing and relay node selection process as an optimization problem. Based on the formulations and models, we propose a self-supported networking scheme including three novel components that make the solution procedure highly efficient. Analysis and simulation results show that our approaches significantly achieve better network performance and typically satisfy the requirements for emergency services in multi-hop wireless networks.

Keywords: Network Technology, Self-supported, Cooperative Communication, Emergency Services, Dependency Graph, Cross-layer Optimization.

1. INTRODUCTION

A wireless Multi hop Ad-hoc network consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, sound, pressure, etc. and to cooperatively pass their data through the network to a main location. The more modern networks are bi-directional, also enabling control of sensor activity. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance; today such networks are used in many industrial and consumer applications, such as industrial process monitoring and control, machine health monitoring[3]. Cooperative communication (CC) is a promising approach which can offer significant enhancements in multi-hop wireless networks. This paper investigates the potential issues in using this communication paradigm to support emergency services[6]. We focus on promoting energy-efficient and congestion-aware cooperative networking for emergency services based on the idea of Do-It-Yourself. We propose a novel cross-layer design which jointly considers the problems of route selection in network layer, congestion and non-cooperation avoidance among multiple links in MAC layer under

cooperative multi-hop wireless environments[7]. We formulate the multi-hop cooperative flow routing and relay node selection process as an optimization problem. Based on the formulations and models, we propose a self-supported networking scheme including three novel components that make the solution procedure highly efficient[8]. Analysis and simulation results show that our approaches significantly achieve better network performance and typically satisfy the requirements for emergency services in multi-hop wireless networks.

2. LITERATURE SURVEY

Data collection process for self supported co-operative networks, which exploits the wireless broadcast advantage and the relaying capability of other cooperative nodes, could provide significant performance enhancements in terms of spatial diversity, increased capacity and improved reliability in wireless networks.

The uprising benefits of CC motivate us to investigate the potential issues in using this communication paradigm to support emergency services. On one hand, CC could provide potential capacity improvement which is critical to emergency communications. On the other hand, CC can reduce the end-to-end transmission delay and improve the probability of emergent information reception by implementing in a set of coordinated cooperative nodes. Mainly, recent research on cooperative protocols could be classified into two categories: (1) CC protocol at the physical layer. (2) route selection and relay node assignment related cooperative protocols at the network layer firstly investigate the problem of CC aware routing in multi-source multi- destination multi-hop wireless networks and address an optimal routing selection protocol called MFCR (Multi-Flow Cooperative Routing). However, the MFCR approaches to non-cooperative protocol when the network congestion emerges (e.g. when emergency cases happen). This paper compares our work with the MFCR scheme to show that our networking scheme could handle efficient networking under emergency situations.

Multi hop wireless is one of the main data delivery services in mobile communication networks. A typical example of multicast services is the multimedia broadcast and multicast service in the long-term evolution (LTE) and future LTE-advanced networks. In multicast communication, a base station broadcasts its data to a group of mobile terminals simultaneously through a common transmission channel.

Since the transmission channel is usually not ideal, however, a data packet may be corrupted and even lost during its transmission. In this case, a mobile terminal that has detected a corrupted or lost packet can send a request to the base station for retransmission of the packet, which would introduce retransmission overhead and increase the processing load of the base station. A simple retransmission mechanism is to retransmit every corrupted or lost packet in another retransmission mechanism is proposed, which introduces a retransmission threshold to improve retransmission efficiency.

In this case, only when the number of retransmission requests for a packet is larger than the threshold will the base station that each mobile terminal requests for retransmission by employing a transmission protocol, for example, the conventional stop-and-wait transmission protocol. However, this would consume more retransmission resources (e.g. bandwidth) and result in lower retransmission efficiency retransmit the packet. This can save retransmission resources but would degrade the network performance in terms of throughput. Therefore it is desirable to develop a more efficient retransmission mechanism that can improve not only retransmission efficiency but also network performance.

This concept can also be applied to the data received within a single stream on a single link and thus can be introduced to a multicast retransmission mechanism to improve retransmission efficiency. In this context, a typical example is 'Random Pick', which introduces random network coding and employs a conventional stop-and-wait transmission protocol to reduce retransmission overhead and increases retransmission efficiency.

In Random Pick, the base station randomly selects a couple of packets from those packets that need to be retransmitted and performs network coding (e.g. XOR operation) on the packets. However, the nature of random selection may not be able to take full advantage of network coding and achieve the best performance. Although network coding has been widely studied in the context of multicast communication, there is not much work done specifically on multicast retransmission for mobile communication networks in addition to the above work.

Most–Least selects a couple of packets, one with the most retransmission requests and the other with the least retransmission requests, whereas Hamming-D selects a couple of packets with the largest Hamming distance. Moreover, a conventional m-parallel wait-and-stop transmission protocol is employed to support data transmission. Simulation results show that with Most–Least outperforms Random Pick in terms of both retransmission efficiency and retransmission reliability. To achieve the best performance of the Hamming-D algorithm, a mathematical model is further developed to analyze the impact of the number of original packets encoded into a packet for retransmission (called encoding number hereafter) on the retransmission performance in order to find an optimal encoding number that can achieve the best performance.

Most–Least algorithm: Unlike Random Pick, Most–Least uses a deterministic selection policy instead of a random selection policy. Specifically, at the beginning of each transmission cycle, the base station counts the number of retransmission requests received in the last transmission cycle for each packet, sets the packet's RIB to 1 or 0 and sets the packet's flag bit to 0. At the beginning of each transmission timeslot, the base station selects the two packets with the most retransmission requests and the least retransmission requests from the packets with '1' RIBs and '0' flags, performs an XOR operation on the selected packets, and then transmits the

encoded packet in the transmission timeslot. In the next transmission timeslot, the base station repeats the same operation until there are no packets with '1' RIBs and '0' flags. The remaining transmission timeslots in the transmission cycle are used to transmit new packets.

Research on cooperative emergency scheduling scheme consists of network-coding-based cooperation and space-time coded cooperation. Note that, our research focuses on the multicast of one packet, which is delay sensitive. In the delay-sensitive nature of the traffic, network-coding may not be appropriate because the waiting time for a batch of packets may not be tolerable. Thus, we stay within the space-time coded cooperation based on a two-phase cooperative transmission model.

A previous cooperative multicast scheduling scheme attempted to efficiently exploit spatial diversity among multiple users. Although it offers good performance in terms of throughput and fairness, there is a lack of consideration of energy consumption. Notably, it requires significant relay power consumption for user cooperation. All group members in a good channel condition should relay the received data, even if there is no group member in a bad channel condition who requests a relay. Thus, we propose the Energy Efficient Cooperative Multicast (EECM) scheme based on selective relay. The proposed scheme considers that a relay user can recognize the nearby user condition (e.g., check whether there is at least one user requesting a relay in the transmission boundary). Therefore, we can reduce unnecessary relay power consumption by selecting relay users.

Most research issues for routing the emergency services to data transfer have focused on enhancing fairness between group members and overall throughput. However, these two metrics (fairness, throughput) have a trade-off relationship. For example, to guarantee user fairness, a Base Station (BS) should select the minimal supported rate of all Multicast Group (MGroup) members, even if it underutilizes wireless resources. If the BS wants to enhance throughput, it should pay for fairness performance to improve wireless resource utilization and provide quality of service (QoS). Recently, to simultaneously improve both fairness and throughput, one approach to the problem is to use a cooperative multicast scheduling scheme by conducting user cooperation among MGroup of wireless nodes.

3. PROPOSED METHODOLOGY

This proposed a promoting energy-efficient and congestion-aware cooperative networking for emergency services based on the idea of Do-It-Yourself. We propose a novel cross-layer design which jointly considers the problems of route selection in network layer, congestion and non-cooperation avoidance among multiple links in MAC layer under cooperative multi-hop wireless environments. It formulates the multi-hop cooperative flow routing and relay node selection process as an optimization problem. Based on the formulations and models, although the idea presented in is interesting, the proposed solution may suffer from

the performance bottleneck and introduce a single point of failure. Moreover, the probabilistic model in possesses no analytical properties to enable tractable analysis.

This proposed system, self-supported networking scheme including three novel components that make the solution procedure highly efficient. Analysis and simulation results show that our approaches significantly achieve better network performance and typically satisfy the requirements for emergency services in multi-hop wireless networks.

Wireless Multi-Hop Ad-Hoc Networks: An Introduction

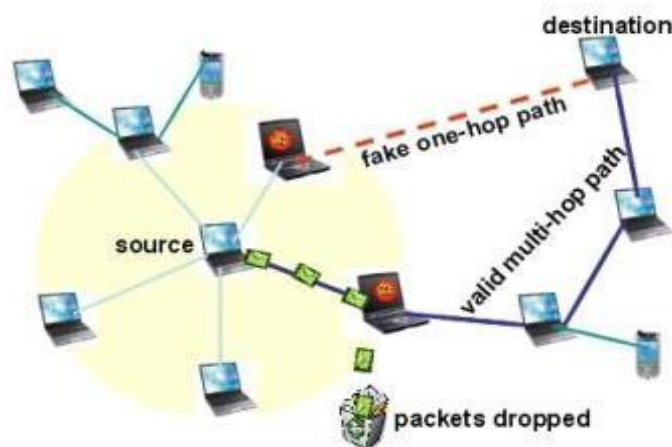


FIGURE 1: Wireless multi-hop networks

4. Results and Discussions

The experimental results showed the energy efficient self supported on the dynamic networks with arbitrary topologies.

Input design involves the selection of the best strategy for getting data into the computer system at the right time and as accurately as possible. This is because the most difficult aspect of input design in accuracy .The use of well-defined documents can encourage users to record data accurately without omission. Input design must capture all the data that the system needs, without introducing any errors. Input errors can be greatly reduced when inputting directly by using appropriate forms for data capture and well designed computer screen layout. Output design involves specifying how production of on-screen reports and paper-based reports will occur. Output may occur to database or file for storing information entered or also for use by other systems.

5. CONCLUSION

This system shown that on large networks arranged in specialized topologies as well as random extended Poisson networks, self supported achieves almost the same end-to-end latency as non node- scheduled networks for moderate values of p , and that it is primarily governed by the shortest path distance d between the source and the destination. The simulation results show that our proposed SCooN scheme significantly improves the network performance by the cognitive movement of emergency sources or relays them- selves and they are more efficient than the MFCR and NCRAODV for emergency services in WANETs.

6. REFERENCES

- [1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part I: System description", *IEEE Trans. Inf. Theory*, vol. 51, no. 11, pp. 1927-1938, Nov. 2003.
- [2] O. Gurewitz, A. de Baynast, and E. W. Knightly, "Cooperative strategies and achievable rate for tree networks with optimal spatial reuse", *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3596-3614, Oct. 2007.
- [3] S. Savazzi, and U. Spagnolini, "Energy aware power allocation strategies for multihop-cooperative transmission schemes", *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 318-327, Feb. 2007.
- [4] A.E. Khandani, J. Abounadi, E. Modiano, and L. Zheng, "Cooperative Routing in Static Wireless Networks", *IEEE Trans. Commun.*, vol. 55, no. 11, pp. 2185-2192, Nov. 2007.
- [5] J. Zhang and Q. Zhang, "Cooperative Routing in Multi-Source Multi- Destination Multi-hop Wireless Networks", in *Proc. IEEE INFOCOM*, 2008.
- [6] S. Sharma, Y. Shi, Y.T. Hou, and S. Kompella, "An optimal algorithm for relay node assignment in cooperative ad hoc networks", *IEEE/ACM Trans. Netw.*, vol. 18, issue 6, Nov. 2010.
- [7] S. Ray, R. Ungrangsi, F. D. Pellegrini, A. Trachtenberg, and D. Starobinski, "Robust location detection in emergency sensor networks", in *Proc. IEEE INFOCOM*, 2003.
- [8] B. Braunstein, T. Trimble, R. Mishra, B. S. Manoj, L. Lenert, and R. R. Rao, "Challenges in using of distributed wireless mesh networks in emergency response", in *Proc. ISCRAM*, pp. 30-38, May 2006.
- [9] M. Felegyhazi, J. P. Hubaux, and L. Buttyan, "Nash equilibria of packet forwarding strategies in wireless ad hoc networks", *IEEE Trans. Mobile Comput.*, vol. 5, no. 5, pp. 463-476, Apr. 2006.
- [10] L. Lai and H. El Gamal, "On cooperation in energy efficient wireless networks: the role of altruistic nodes", *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1868-1878, May 2008.

CONSTRAINT-BASED TEMPORAL TASK SCHEDULER FOR PROFIT MAXIMIZATION

D. SWETHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

As cloud computing is becoming increasingly popular, consumers' tasks around the world arrive in cloud data centers. A private cloud provider aims to achieve profit maximization by intelligently scheduling tasks while guaranteeing the service delay bound of delay-tolerant tasks. However, the aperiodicity of arrival tasks brings a challenging problem of how to dynamically schedule all arrival tasks given the fact that the capacity of a private cloud provider is limited. Previous works usually provide an admission control to intelligently refuse some of the arrival tasks. Nevertheless, this will decrease the throughput of a private cloud and cause revenue loss. This paper studies the problem of how to maximize the profit of a private cloud in hybrid clouds while guaranteeing the service delay bound of delay-tolerant tasks. We propose a profit maximization algorithm (PMA) to discover the temporal variation of prices in hybrid clouds. The temporal task scheduling provided by PMA can dynamically schedule all arrival tasks to execute in private and public clouds. The sub-problem in each iteration of PMA is solved by the proposed hybrid heuristic optimization algorithm, simulated annealing particle swarm optimization (SAPSO). Besides, SAPSO is compared with existing baseline algorithms. Extensive simulation experiments demonstrate that the proposed method can greatly increase the throughput and the profit of a private cloud while guaranteeing the service delay bound.

Keywords: data centers, profit maximization algorithm, SAPSO, throughput, delay-tolerant

INTRODUCTION

Cloud computing is being intensively referred to as one of the most influential innovations in information technology in recent years. With resource virtualization, the cloud can deliver computing resources and services in a pay-as-you-go mode, which is envisioned to become as convenient to use similar to daily-life utilities such as electricity, gas, water, and telephone shortly. These computing services can be categorized into Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Many international IT corporations now offer powerful public cloud services to users on a scale from individual to enterprise all over the world; examples are Amazon AWS, Microsoft Azure, and IBM SmartCloud. Although the current development and proliferation of cloud computing is rapid, debates and hesitations on the usage of cloud still exist. Data security/privacy is one of the major concerns in the adoption of cloud computing. Compared to conventional systems, users will lose their direct control over their data. In this paper, we will investigate the problem of integrity verification for big data storage in the cloud. This problem can also be called data auditing when the verification is conducted by a trusted third party. From cloud users' perspective, it may also be called 'auditing-as-a-service'.

To date, extensive research is carried out to address this problem. In a remote verification scheme, the cloud storage server (CSS) cannot provide valid integrity proof of a given proportion of data to a verifier unless all this data is intact. To ensure the integrity of user data stored on the cloud service provider, this support is of no less importance than any data protection mechanism deployed by the cloud service provider (CSP), no matter how secure they seem to be, in that it will provide the verifier a piece of direct, trustworthy and real-time intelligence of the integrity of the cloud user's data through a challenge request. It is especially recommended that data auditing be conducted regularly for users who have high-level security demands over their data.

Now, Cloud Computing has grown in popularity as a medium for scientific applications. To facilitate scientific study, cloud computing aims to share large-scale resources and equipment in the areas of processing, storage, information, and expertise with other researchers. One of the most notable uses of contemporary scheduling has been the allocation of distributed computing systems of limited resources to jobs submitted by Internet users since their establishment in 1980. In the last few years, a new technology called "cluster systems" has emerged, which combines several separate computers into a single unit. Grid systems were developed in response to the weakness of cluster systems, which only utilize local resources, by gathering together all heterogeneous resources accessible in geographically distant places Weinhardt et al. (2009). Cloud computing is a relatively new technology which makes use of the advantages of both clustered and grid-based systems.

Due to the huge solution space, many scheduling issues that are NP-hard or NP-completely take a long time to implement an optimum or sub-optimal solution in the shortest time. Due to the limited resources in modern computer systems, there is no polynomial time-scheduling technique which could be used to improve the constrained resources scheduling. Using a simple example from Taillard (1990), we can see that just about 0.02 percent of the possible solutions use between 1 and 1.01 times the time required to find the ideal answer. Finding the best answer to a complex problem is quite challenging, as this example illustrates. As a result, most scholars have been motivated to look for a quick but effective solution to these kinds of scheduling challenges. The two most basic forms of scheduling methods are static and dynamic scheduling strategies. However, because cloud settings are inherently dynamic, additional dynamic algorithms must be incorporated into the cloud scheduling process to achieve outstanding results in this field. Static algorithms, on the other hand, are only utilized when the workloads vary just slightly. As a result, adopting deterministic ways to tackle the job scheduling problem is unfeasible in this circumstance Allahverdi (2015). Nondeterministic meta-heuristic algorithms have been offered as a way to considerably address this challenge in a polynomial amount of time.

Consumers and producers of cloud services can benefit from a variety of advantages because to dynamic work scheduling approaches and virtualization technology. Resource (task) scheduling that is effective not only minimize resource consumption (increasing the resource used), but also assures that new jobs are completed as promptly as possible (the minimizes of makespan). Job scheduling has become most important due to the possibility of

a scarcity of cloud resources as a result of the continual increase in workloads at cloud datacenters. This has resulted in a significant increase in the importance of task scheduling. As a result, more study into the still-developing topic of cloud job scheduling is required to push for things like more effective mapping of incoming job to available resources and improved criteria for measuring how efficiently a service is provided. Scheduling algorithms can be used to optimize a variety of quality of service (QoS) parameters, for example resource use and utilization, task rejection ratio, energy consumption, and other constraints, by determining the optimal set of resources available to carry out incoming tasks (underutilization and over utilization). The primary objective of a scheduling approach is to find the most efficient use of the available resources (SLA).

LITERATURE REVIEW

FINE-GRAINED CONTROL OF SECURITY CAPABILITIES:

We present a new approach for fine-grained control over users' security privileges (fast revocation of credentials) centered on the concept of an on-line semi-trusted mediator (SEM). The use of a SEM in conjunction with a simple threshold variant of the RSA cryptosystem (mediated RSA) offers a number of practical advantages over current revocation techniques. The benefits include simplified validation of digital signatures, efficient certificate revocation for legacy systems and fast revocation of signature and decryption capabilities. This paper discusses both the architecture and the implementation of our approach as well as its performance and compatibility with the existing infrastructure. Experimental results demonstrate its practical aspects. We begin this paper with an example to illustrate the premise for this work. Consider an organization – industrial, government, or military – where all employees (referred to as users) have certain authorizations. We assume that a Public Key Infrastructure (PKI) is available and all users have digital signature, as well as en/de-cryption, capabilities. In the course of performing routine everyday tasks, users take advantage of secure applications, such as email, file transfer, remote log-in and web browsing.

OBLIVIOUS TRANSFER WITH ACCESS CONTROL:

We present a protocol for anonymous access to a database where the different records have different access control permissions. These permissions could be attributes, roles, or rights that the user needs to have in order to access the record. Our protocol offers maximal security guarantees for both the database and the user, namely (1) only authorized users can access the record; (2) the database provider does not learn which record the user accesses; and (3) the database provider does not learn which attributes or roles the user has when she accesses the database. We prove our protocol secure in the standard model (i.e., without random oracles) under the bilinear Diffie-Hellman exponent and the strong Diffie-Hellman assumptions. Also to protect sensitive information such as medical or financial data we need to provide strong access control to be sure that only those people who have the necessary permissions can access it. But statistics about what sort of data people query also reveals a lot of information about them. It is possible to build a complete picture of someone's

movements, transactions, locations and relationships from the trail left from interaction with websites and various databases. So personal security has become a serious issue.

PROPOSED METHODOLOGY

In the challenge/verification process of our scheme, we try to secure the scheme against a malicious CSS that tries to cheat the verifier TPA about the integrity status of the client’s data, which is the same as previous work on both PDP and POR. In this step, aside from the new authorization process, the only difference compared is the RMHT and variable- sectored blocks. Therefore, the security of this phase can be proven through a process highly similar to, using the same framework, adversarial model, and interactive games defined. A detailed security proof for this phase is therefore omitted here.

ADVANTAGES

- ❖ The proposed scheme with the aim of supporting variable-sized data blocks with session-based 64-bit standard encryption applied each each request
- ❖ Authorized third-party auditing and fine-grained dynamic data updates are done periodically.
- ❖ More efficient than existing models it removes the drawbacks of the GAP and WRAP process by offering the hashing technique in key generation.

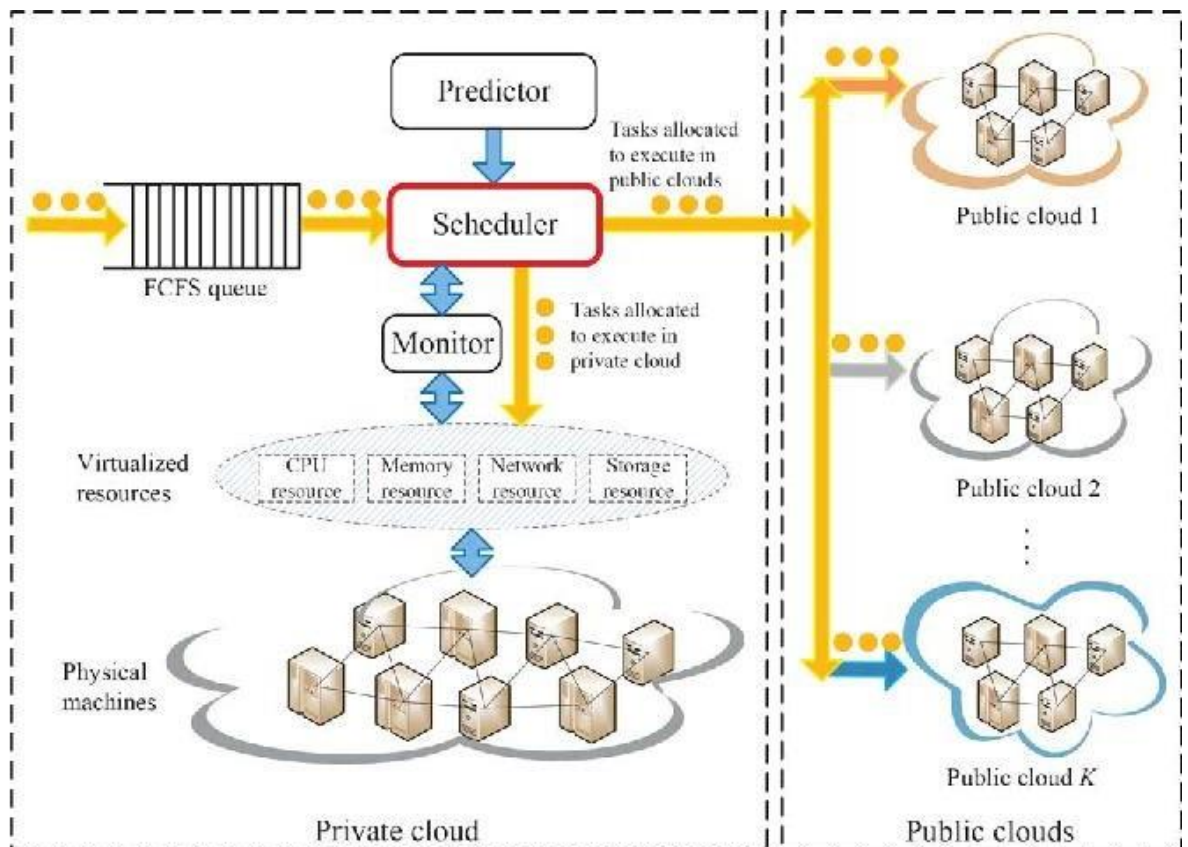
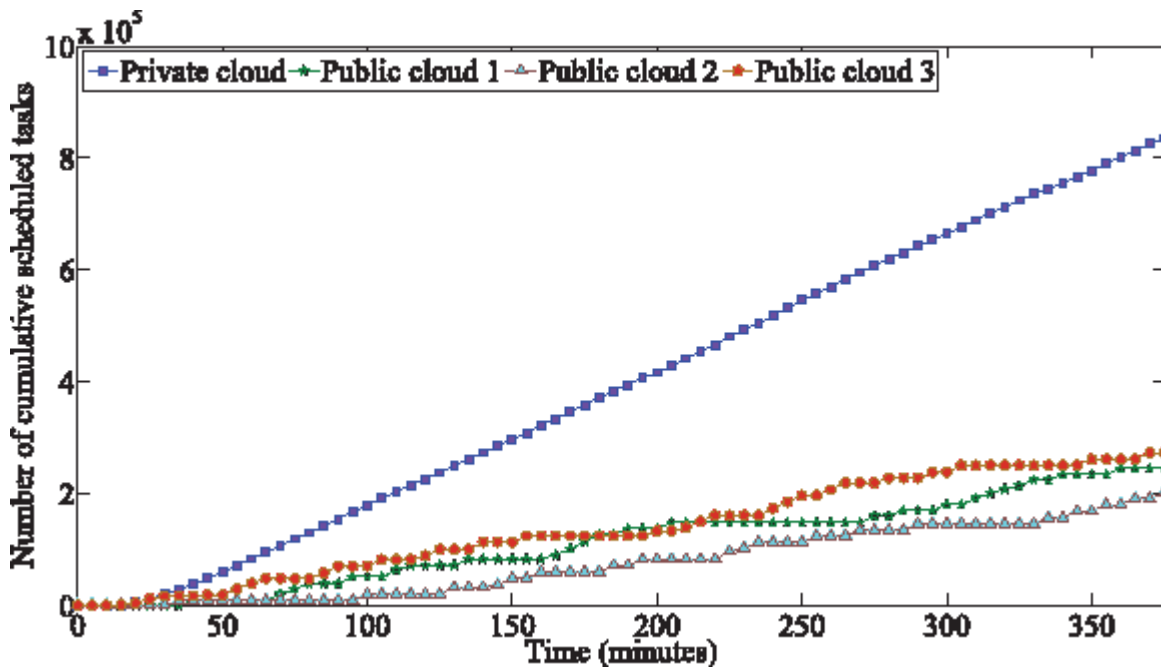


Figure 1: TPA Approach in clouds

RESULT AND DISCUSSION

Our scheme allows the cloud server to efficiently update the ciphertext when a new access policy is specified by the data owner, who is also able to validate the update to counter against cheating behaviors of the cloud. It also enables (i) the data owner and eligible users to effectively verify the legitimacy of a user for accessing the data, and (ii) a user to validate the information provided by other users for correct plaintext recovery. Rigorous analysis indicates that our scheme can prevent eligible users from cheating and resist various attacks such as the collusion attack. In such a system, combined with verifiable computation and encrypt-then-mac mechanism, the data confidentiality, the fine-grained access control and the correctness of the delegated computing results are well guaranteed at the same time. Besides, our scheme achieves security against chosen-plaintext attacks under the k -multilinear Decisional hashed technique



Graph 1: Cumulative tasks in private and three public clouds.

5. CONCLUSION

Proposed an improved HASHED KEY TECHNIQUE cryptosystem to overcome the decryption failures of the original HASHED KEY TECHNIQUE and then present a secure and verifiable access control scheme based on the improved HASHED KEY TECHNIQUE to protect the outsourced big data stored in a cloud. Our plan enables the information proprietor to progressively refresh the information to get to strategy and the cloud server to effectively refresh the comparing redistributed ciphertext to empower proficient access power over the huge information in the cloud. It also provides a verification process for a user to validate the legitimacy of accessing the data to both the data owner and t-1 other legitimate users and the correctness of the information provided by the t-1 other users for plaintext recovery.

The security of our proposed scheme is guaranteed by those of the HASHED KEY TECHNIQUE cryptosystem and the (t,n)-threshold secret sharing. We have rigorously

analyzed the correctness, security strength, and computational complexity of our proposed scheme. Designing a secure, privacy-preserving, and practical scheme for big data storage in a cloud is an extremely challenging problem. In our future research, we will additionally enhance our plan by joining the (t, n) -edge mystery offering to trait-based access control, which includes an entrance structure that can put different prerequisites for a client to unscramble a redistributed ciphertext information in the cloud. Meanwhile, we will investigate the security problems when a data owner outsources its data to multi-cloud servers and consider an attribute-based access structure that can be dynamically updated, which is more applicable for practical scenarios in big data storage.

REFERENCES

- [1] M. A. Beyer and D. Laney, "The importance of big data: a definition," Stamford, CT: Gartner, 2012.
- [2] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.
- [3] G. P. Consortium et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.
- [4] A. Sahai and B. Waters, "Fuzzy identity-based encryption," *Advances in Cryptology–EUROCRYPT 2005*, pp. 457–473, 2005.
- [5] C. Hu, F. Zhang, X. Cheng, X. Liao, and D. Chen, "Securing communications between external users and wireless body area networks," in *Proceedings of the 2nd ACM workshop on Hot topics on wireless network security and privacy*. ACM, 2013, pp. 31–36.
- [6] C. Hu, H. Li, Y. Huo, T. Xiang, and X. Liao, "Secure and efficient data communication protocol for wireless body area networks," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 2, pp. 94–107, 2016.
- [7] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM conference on Computer and communications security*. ACM, 2006, pp. 89–98.
- [8] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," *Public Key Cryptography–PKC 2011*, pp. 53–70, 2011.
- [9] C. Hu, N. Zhang, H. Li, X. Cheng, and X. Liao, "Body area network security: a fuzzy attribute-based signcryption scheme," *IEEE journal on selected areas in communications*, vol. 31, no. 9, pp. 37–46, 2013.
- [10] A. Lewko and B. Waters, "Decentralizing attribute-based encryption," *Advances in Cryptology–EUROCRYPT 2011*, pp. 568–588, 2011.
- [11] C. Hu, X. Cheng, Z. Tian, J. Yu, K. Akkaya, and L. Sun, "An attributebasedsigncryption scheme to secure attribute-defined multicast communications," in *SecureComm 2015*. Springer, 2015, pp. 418–435.
- [12] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Advances in cryptology*. Springer, 1985, pp. 47–53.
- [13] M. Dehkordi and S. Mashhadi, "An efficient threshold verifiable multiset sharing," *Computer Standards & Interfaces*, vol. 30, no. 3, pp. 187–190, 2008.
- [14] Z. Eslami and J. Z. Ahmadabadi, "A verifiable multi-secret sharing scheme based on cellular automata," *Information Sciences*, vol. 180, no. 15, pp. 2889–2894, 2010.
- [15] M. H. Dehkordi and S. Mashhadi, "New efficient and practical verifiable multi-secret sharing schemes," *Information Sciences*, vol. 178, no. 9, pp. 2262–2274, 2008.

DETECTION OF NODE FAILURE TECHNIQUES USING CHURN RESILIENT PROTOCOL

K. ABARNA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Massive data dissemination is often disrupted by frequent join and departure or failure of client nodes in a peer-to-peer (P2P) network. We propose a new churn-resilient protocol (CRP) to assure alternating path and data proximity to accelerate the data dissemination process under network churn. The CRP enables the construction of proximity-aware P2P content delivery systems. We present new data dissemination algorithms using this proximity-aware overlay design. We simulated P2P networks up to 20,000 nodes to validate the claimed advantages. Specifically, we make four technical contributions,

The CRP scheme promotes proximity awareness, dynamic load balancing, and resilience to node failures and network anomalies. 2). The proximity-aware overlay network has a 28-50 percent speed gain in massive data dissemination, compared with the use of scope-flooding or epidemic tree schemes in unstructured P2P networks.

The CRP-enabled network requires only 1/3 of the control messages used in a large CAM-Chord network. 4) Even with 40 percent of node failures, the CRP network guarantees atomic broadcast of all data items. These results clearly demonstrate the scalability and robustness of CRP networks under churn conditions. The scheme appeals especially to web scale applications in digital content delivery, network worm containment, and consumer relationship management over hundreds of datacenters in cloud computing services.

Keywords: peer-to-peer (P2P) network, CRP, proximity awareness, network, datacenters

INTRODUCTION

There are many related algorithms for data collection in WSNs. Directed diffusion is a general data collection mechanism that uses a data-centric approach to choose how to disseminate queries and gather data. Cougar and TinyDB provide query-based interfaces to extract data from sensor networks. Those works mainly focus on query-based data gathering, but none of them consider the case of efficient long-term large-scale data collection. Query-based remote continuously approximate data collection in sensor networks is closely related to the problem we study here. One such approach is approximate caching which gives approximate answers to queries in distributed environments with a fixed error bound.

The idea is that the sink uses a constant to reconstruct a piecewise constant approximation of the real sensor readings. No updates are sent until a sensor node notices that its value has diverged by more than a given upper bound from the last reading sent to the sink. CONCH also provides a simple spatial suppression technique to suppress update messages of nearby sensors with similar sensor readings. Sensor nodes in CONCH do not update their readings if they hear similar update readings from their neighbors. This kind of approach, though simple, ignores the trend of sensor readings and only offers a narrow range of predictive capabilities. Approximate caching may also suffer a large overhead of update message transmission when many sensor readings change dramatically.

Compared with approximate caching, our approach ADC has two obvious advantages. First, ADC exploits the temporal correlation by utilizing a linear trend component which enhances the estimation capability. Second, in a distributed manner, ADC selects only a portion of sensor nodes to update their readings based on a more sophisticated spatial correlation model, which utilizes the trend information of sensor readings generated by the local estimation to further reduce the communication cost.

Other query-based approaches extract data from sensor networks by using Gaussian joint distribution to capture the correlations of sensor readings, is the first one using multivariate Gaussian joint distribution to capture the correlations of sensor readings. It samples a small fraction of sensor data from a WSN and utilizes Gaussian joint distribution model to estimate the no sampled sensor readings. Gaussian joint distribution based approaches have several drawbacks that make them unsuitable for long-term large-scale WSNs. First, this kind of models need an expensive long training phase and a complete data set of every sensor node within a sufficiently long period. Gathering complete data set is too energy consuming and even impractical for large-scale WSNs with limited bandwidth.

Second, the correctness of this kind of model requires continuous model update which needs periodically gathering the data generated by every sensor node and disseminating the updated information to related sensor nodes. Both of the two tasks are costly for energy-constrained WSNs, even when the update frequency is low. Third, it is almost impossible for this kind of model to precisely control the data error. A Gaussian process (GP) is associated with a mean function positive-definite kernel function K often called the covariance function. An important property of GPs is that the posterior variance of one of its variables depends on the covariance function K , instead of the actual observed value. Hence, the estimation errors of the nonsampled sensor readings are unknown and the estimation quality of the nonsampled sensor readings cannot be guaranteed. In comparison, the data processing burdens of ADC are distributed to each sensor node.

The local estimation and the data approximation of ADC are, respectively, settled on each sensor node and each cluster head. This enables sensor data to be processed near or at their sources. The correctness of local estimation is guaranteed by each sensor node locally and the data error bound of ADC is jointly controlled by the local estimation and the data approximation. No explicit control message exchange is required. The data error bound of ADC can be flexibly adjusted according to the requirements of applications. Such features make our approach ADC scalable and efficient for long-term continuous data-gathering applications.

Distributed source coding is a lossless compression technique to address the problem of compressing correlated sources that are not collocated and cannot communicate with each other to minimize their joint description costs. Slepian and Wolf show that it is possible to compress the data at a combined rate equal to the joint entropy of the correlated source. Distributed source coding technology requires precise and perfect knowledge of the correlations among attributes, and will return wrong answers (without warning) if this

condition is not satisfied. In practice, the cost of acquiring precise and perfect knowledge of the correlations among attributes is extremely high. Our approach smartly utilizes a simple probabilistic model to depict spatial correlations among sensor nodes based on rough data generated by each sensor node. All information is processed locally and sensor nodes implicitly cooperate with each other to ensure the data error bound of ADC.

Another technique widely used to reduce communication costs in WSNs is called time-series forecasting. Lazaridis and Mehrotra propose to use the time-series method to create piecewise linear approximations of signals generated by sensor nodes and send those approximations to the sink. Their approach gathers a large amount of data and tries to approximate them, rather than exploiting the temporal correlations among sensor readings. Chatterjea and Having described an adaptive sensor sampling scheme where nodes change their sampling frequencies autonomously based on time-series forecasting, to reduce energy consumption. They use time-series forecasting to predict future sensor readings.

The sampling frequency decreases against prediction accuracy, otherwise increases the sample frequency. The skipped samples are replaced by prediction values. The sink uses a simple linear time-series model that consists of a trend component and a stationary autoregressive component to predict the reading of each sensor. Each sensor node updates its linear time-series model individually, instead of raw data. The drawback of these existing works is that they only exploit temporal correlation within sensor data to reduce the communication cost, without considering sensor readings' similarity of nearby sensor nodes, which can be used to suppress the update messages of nearby sensor nodes with similar sensor readings. In ADC, by further exploiting the data generated by the local estimation settled on each sensor node by using time-series forecasting, we propose a novel tool, called correlation graph, to model the spatial correlation within nearby sensor nodes. Based on the correlation graph, ADC only needs to collect corresponding data of a subset of sensor nodes, saving the energy cost of information update.

In many applications, it is often difficult and unnecessary to continuously collect the complete data from the resource-constrained WSNs. From the point of view of WSNs, directly sending a large amount of raw data to the sink can lead to several undesired problems. First, the data quality may be deteriorated by packet losses due to the limited bandwidth of sensor nodes. Second, intensive data collection incurs excessive communication traffic and potentially results in network congestions. Packet losses caused by such congestions further deteriorate the data quality. Experiments with TinyOS show that packet delivery ratio can be greatly increased by reducing the data traffic within a sensor network. Third, intensive data collection can lead to excessive energy consumption. It is reported that the lifetime of a sensor network can be increased extraordinarily from 1 month to more than 18 months by lowering the data flow rates of sensor nodes.

LITERATURE REVIEW

Data Evaluation

There are two types of data dissemination methods: mesh based pull and tree-based push methods. The pull methods use swarming content delivery. Each node advertises to its neighbors which messages it has received and the neighbors explicitly request messages if needed. Two representative pull schemes are Chainsaw and PRIME. PRIME incorporates swarming into streaming applications and points out the design tradeoffs of such systems. To achieve a low delay, a node in such systems should advertise as soon as it receives a new message, increasing the control overhead greatly.

The tree-based push methods achieve a fast dissemination with low overhead. The main concern is the vulnerability to node failures. Our CRP method falls into this type. The CRP extracts the dissemination tree from a churn-resilient overlay, which provides sufficient alternative overlay links for the tree to achieve good fault resilience. The preliminary idea was reported in an IEEE IPDPS-2008 paper.

Existing Knowledge review

In an existing network, the link latency is measurable. The node capacity of a node represents the maximum number of adjacent nodes to which it can forward the data items, concurrently. A node's out-degree is bounded by its capacity, the notations used in these existing papers.

The proximity-aware overlay is built around a unidirectional ring with extra bidirectional chord links. The overlay is heterogeneous since links are associated with different weights. A node's neighbors are those that are directly connected by either ring or chord links with it, while its chord neighbors are those that are connected by chord links. Initially, the base ring is assumed empty before any node joins the system. The first node is located at any position on the base ring.

Most of the existing protocols apply both network proximity and capacity proximity in CRP protocol. The network proximity is measured by the latency or closeness of two nodes in physical IP networks. This proximity enables faster transferring of data items. The capacity proximity is measured by the closeness of nodes concerning node capacities. The capacity proximity allows us to put high-capacity nodes at higher positions on delivery trees, which reduces the delivery hop count.

It shows the initial base ring with three nodes. The ring is assumed unidirectional with clockwise links shown by dotted edges. We assume that all ring links are weighted with 1 unit. Suppose the node capacities for A, B, and C are 3, 4, and 2, respectively. Node A's successor and predecessor are B and C, respectively. The fourth node D with capacity 3 is inserted into the interval between A and C in Fig. 1b. A chord link weighted with 0.7 is added between nodes B and D.

It shows the impacts of massive node failures on data delivery time and the number of unreachable live nodes in CRP. It can be found that at a failure rate of as high as 40 percent, all live nodes are reached by the data messages from any source. Thus, the CRP guarantees atomic data dissemination even when most of the nodes are in failure state percent of nodes failure.

PROPOSED METHODOLOGY

nodes in a peer-to-peer (P2P) network. We propose a new churn-resilient protocol (CRP) to assure alternating path and data proximity to accelerate the data dissemination process under network churn. The CRP enables the construction of proximity-aware P2P content delivery systems. We present new data dissemination algorithms using this proximity-aware overlay design. We simulated P2P networks up to 20,000 nodes to validate the claimed advantages. Specifically, we make four technical contributions, A proximity proximity-aware layer is presented to avoid the hangover data over the network path. Improved worm avoidance nodes are implemented to avoid data insecurity over the P2P nodes. The CRP-enabled network requires only 1/3 of the control messages used in a large CAM-Chord network. Even with 40 percent of node failures, the CRP network guarantees atomic broadcast of all data items. These results demonstrate the scalability and robustness ofCRP networks under churn conditions.

Advantages

- The proposed system likely improves fault tolerance by effectively detecting and mitigating node failures. This can lead to increased system reliability and availability, as the system can continue operating even in the presence of failures.
- By addressing churn resilience, the system may better scale to accommodate changes in network size or topology. This could lead to better performance and efficiency as the system grows or experiences fluctuations in node participation.
- With a focus on quickly identifying and recovering from node failures, the system can minimize downtime and service disruptions. This advantage is particularly crucial for applications requiring high availability or real-time responsiveness.
- The proposed system likely automates many aspects of failure detection and recovery, reducing the need for manual intervention. This automation can streamline operations and reduce the burden on system administrators.
- By efficiently reallocating tasks or responsibilities in response to node failures, the system can optimize resource utilization. This ensures that available resources are utilized effectively, maximizing system performance and efficiency.

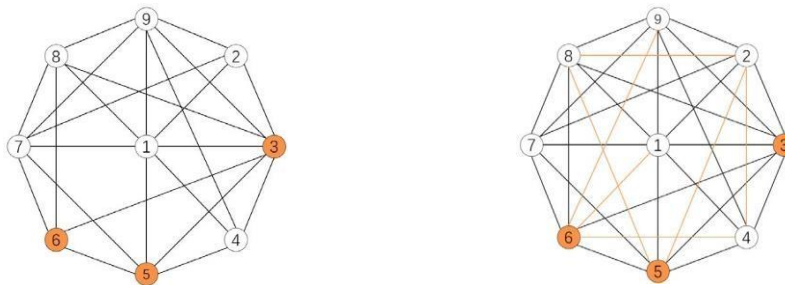


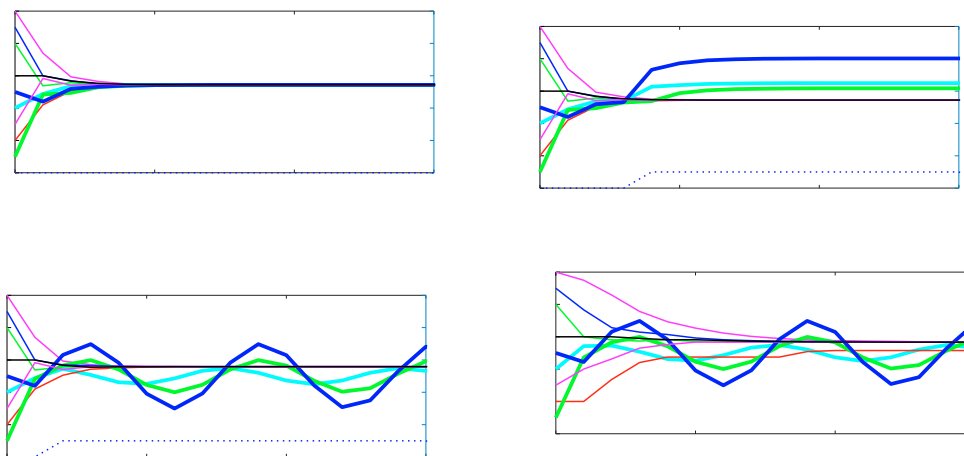
Figure 1: nine nodes: (a) 4-connected case and (b) (4,4)-robust case

RESULT AND DISCUSSIONS

The primary metric for evaluating churn-resilient detection is the accuracy of identifying node failures. The results should indicate the percentage of node failures accurately detected under varying churn rates. A robust churn-resilient detection algorithm should maintain high accuracy even when the system experiences significant churn. Another critical aspect is the occurrence of false positives, where the system incorrectly identifies a node as failed when it is operational. High false positives can lead to unnecessary service disruptions and increased operational overhead. Evaluating the false positive rate under different conditions is crucial to assess the reliability of the detection mechanism.

Detection latency measures the time taken for the system to detect a node failure from the moment it occurs. Low latency is essential for minimizing the impact of failures and ensuring timely response actions, such as failover or replication of data. Analyzing detection latency across various scenarios provides insights into the responsiveness of the detection mechanism. As the system scales up in terms of the number of nodes and the volume of churn, it's vital to evaluate the scalability of the detection approach. Scalability considerations include the ability to handle a large number of nodes, efficiently process monitoring data, and maintain detection accuracy with increasing churn rates.

The robustness of the detection mechanism to different network conditions, such as latency, packet loss, and network partitions, is also critical. Assessing the performance of the detection algorithm under adverse network conditions helps understand its resilience in real-world deployment scenarios. Comparing the churn-resilient detection approach with baseline methods or existing solutions provides context for evaluating its effectiveness. This comparison can highlight the advantages and limitations of the proposed approach and identify areas for improvement. Finally, discussing the practical implications of the results, including deployment challenges, operational overhead, and integration with existing systems, is essential. Addressing these considerations can facilitate the adoption of churn-resilient detection in production environments.



Graph1 : MSR algorithm under attack

CONCLUSION

The system had experimented successfully with different sets of parameters that CRP-enabled network requires only 1/3 of the control messages used in a large CAM-Chord network. Even with 40 percent of node failures, the CRP network guarantees atomic broadcast of all data items. These results demonstrate the scalability and robustness of CRP networks under churn conditions. The scheme appeals especially to web-scale applications in digital content delivery, network worm containment, and consumer relationship management over hundreds of data centers in cloud computing services.

REFERENCE

- [1] S. Agarwal and J.R. Lorch, "Matchmaking for Online Games and Other Latency- Sensitive P2P Systems," Proc. ACM SIGCOMM, Aug. 2009.
- [2] M. Bishop, S. Rao, and K. Sripanidkulchai, "Considering Priority in Overlay Multicast Protocols under Heterogeneous Environments," Proc. IEEE INFOCOM, Apr. 2006.
- [3] M. Castro et al., "SCRIBE: A Large-Scale and Decentralized Application-Level Multicast Infrastructure," IEEE J. Selected Areas in Commun., vol. 20, no. 8, pp. 1489-1499, Oct. 2002.
- [4] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: High-Bandwidth Multicast in Cooperative Environments," Proc. ACM Symp. Operating Systems Principles, Oct. 2003.
- [5] V. Cerny, "Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm," J. Optimization Theory and Applications, vol. 45, pp. 41-51, 1985.
- [6] S. Chen, B. Shi, S. Chen, and Y. Xia, "ACOM: Capacity-Constrained Overlay Multicast in Non-DHT P2P Networks," IEEE Trans. Parallel and Distributed Systems, vol. 18, no. 9, pp. 1188-1201, Sept. 2007.
- [7] C. Diot, B.N. Levine, B. Lyles, H. Kassem, and D. Balensiefen, "Deployment Issues for the IP Multicast Service and Architecture," IEEE Network, vol. 14, no. 1, pp. 78-88, Jan./Feb. 2000.

ENABLING IDENTIFIED-BASED AUDITING

P.PUNITHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Firewalls have been widely deployed on the Internet for securing private networks. A firewall checks each incoming or outgoing packet to decide whether to accept or discard the packet based on its policy. Optimizing firewall policies is crucial for improving network performance. Prior work on firewall optimization focuses on either intra-firewall or inter-firewall optimization within one administrative domain where the privacy of firewall policies is not a concern. This paper explores inter-firewall optimization across administrative domains for the first time. The key technical challenge is that firewall policies cannot be shared across domains because a firewall policy contains confidential information and even potential security holes, which can be exploited by attackers. The system proposes the first cross-domain privacy-preserving cooperative firewall policy optimization protocol. Specifically, for any two adjacent firewalls belonging to two different administrative domains, our protocol can identify in each firewall the rules that can be removed because of the other firewall. The optimization process involves cooperative computation between the two firewalls without any party disclosing its policy to the other. This system implemented our protocol and conducted extensive experiments. The results on real firewall policies show that our protocol can remove as many as 49% of the rules in a firewall whereas the average is 19.4%. The communication cost is less than a few hundred KBs. Our protocol incurs no extraonline packet processing overhead and the offline processing time is less than a few hundred seconds.

INTRODUCTION

A firewall often is referred to as a network's "first line of defense" in protecting sensitive data from unauthorized access. And with good reason: a firewall is at the entry point of the networked system it protects, just like our bouncer. A firewall can be the first program that receives and handles incoming network traffic, as well as the last to handle outgoing traffic. As such, it is in a prime location to enforce technically the business's policies on all incoming and outgoing traffic.

A firewall intrusion detection system performs different roles and thus dynamically minimizes the vulnerability to threats on one's system.[1] Firewall protection needs to be set up appropriately and doesn't override authorization of the outbound connection as it can report back to its creator afterwards. Since there are different types of firewalls depending upon one's choice, choosing the best suitable security system is certainly required. The role of firewalls is crucial as they work upon a simple concept as communication between network devices is broken into the form of precise packets. [2]These packets contain messages from the originator and the recipient. The firewall reads it to find the kind of application message it is, thus assembling it whether the originator is authentic or not. Basically, it protects your computer from an illicit and unauthorized network and thus is an essential part of one's PC. [3]Beware if you are going online without any authorized firewall protection, as working without it is simply asking for trouble. Firewall protection is necessary

as the vulnerability of Trojans and spyware may enable unauthorized access of your system and thus risking its security to the extreme.

Firewall Builder compensates for differences in implementation between firewall platforms. [4]For example, Cisco PIX applies its access list rules to the packet before it performs address and port transformations according to the NAT rules. As a result, a policy rule that controls access to a server behind the firewall doing NAT should be written using the firewall object instead of the server object.[5] The meaning of such a rule is not obvious at a glance since you have to keep in mind all the NAT rules as well as remember that this policy rule controls access not to the firewall machine, but rather to the server behind it. Firewall Builder takes into account these variations like this by using smart algorithms to transform rules defined in the GUI into rules that achieve the desired effect in the target firewall platform. Using Firewall Builder, you write your rules as if NAT translation happens before the access rules are applied.

Remove bad traffic and clean up the network. Bad traffic includes non-compliant, unauthorized or undesired traffic.[6] Notify server administrators about servers hitting the firewall directly with outbound denied Domain Name System (DNS), NTP, SMTP, HTTP and HTTP Secure (HTTPS) requests, as well as dropped/rejected internal devices. The administrators should then reconfigure the servers not to send this type of unauthorized outbound traffic (thereby taking load off the firewall).[7] Move the filtering rules for unwanted inbound traffic from the firewall to the edge router(s) to balance the performance and effectiveness of the security policy. To do this, first identify the top inbound dropped requests that are candidates to move upstream to the router as Standard Access Control List (ACL) filters. This can be a time-consuming process but it is a good method for moving blocks upstream to the router, thus saving firewall CPU and memory. [8]Then, if you have an internal choke router between your network and firewall, consider moving common outbound traffic blocks to your choke routers. This will free up more processing on your firewall.

As thousands of change requests are processed by the security team, the underlying policy configurations (firewall rule bases, router and switch ACLs) become very large and complex. Many of the rules and objects in a typical firewall or router policy are obsolete. [9]These unused rules represent a potential security risk and should be eliminated, yet it is nearly impossible for managers to locate them and remove them without risking business continuity. A poorly managed policy is difficult to maintain and requires the attention of senior administrators with a great deal of expert, undocumented knowledge. [10] Since a mistake can result in application or network downtime, it isn't feasible to assign policy management to less-experienced or outsourced staff. The potential cost of an error is even higher for security service providers.

In addition to security risks, a poorly maintained policy can have a major impact on performance. The entire rule base is parsed from top to bottom with every network connection, and as it grows, hardware requirements also increase. Security teams need automation to maintain secure, efficient policies on all of their firewalls and routers.

LITERATURE SURWAY

A firewall is a packet filter that is placed at the entrance of a private network. It checks the header fields of each incoming packet into the private network and decides, based on the specified rules in the firewall, whether to accept the packet and allow it to proceed or to discard the packet. To validate the correctness and effectiveness of the rules in a firewall, the firewall rules are usually subjected to two types of analysis: verification and redundancy checking. Verification is used to verify that the rules in a firewall accept all packets that should be accepted and discard all packets that should be discarded. Redundancy checking is used to check that no rule in a firewall is redundant (i.e. can be removed from the firewall without changing the sets of packets accepted and discarded by the firewall). Contrary to the conventional wisdom, these two types of analysis are in fact equivalent. In particular, the system show that (1) every verification algorithm can be also used to check whether a rule in a firewall is redundant, and (2) every redundancy checking algorithm can be also used to verify whether the rules in a firewall accept or discard an intended set of packets.

A firewall is a packet filter that is placed at a point where a private computer network is connected to the rest of the Internet. The firewall intercepts each packet that is exchanged between the private network and the Internet, examines the fields of the packet headers, and makes a decision to either accepts the packet and allows it to proceed on its way, or discard the packet. A firewall rule consists of a predicate and a decision, which is either accept or discard. When the firewall receives a packet, the firewall searches its sequence of rules for the first rule, whose predicate is satisfied by the values of the fields in the packet headers, and then applies the decision of this rule to the packet.

A rule in a firewall is said to be redundant if removing the rule from the firewall yields a firewall that is equivalent to the original firewall. After the rules of a firewall are designed (by a firewall designer), they are usually validated by subjecting them to two seemingly different algorithms: a verification algorithm and a redundancy checking algorithm. The function of the verification algorithm is to verify that the firewall rules accept all packets that should be accepted and discard all packets that should be discarded. The function of the redundancy checking algorithm is to check whether any rule in the firewall is redundant.

Interest and knowledge about computer and network security is growing along with the need for it. This interest is, no doubt, due to the continued expansion of the Internet and the increase in the number of businesses that migrating their sales and information channels to the Internet. The growth in the use of networked computers in business, especially for e-mail, has also fueled this interest. Many people are also presented with the postmortems of security breaches in high-profile companies in the nightly news and are given the impression that some bastion of defense had failed to prevent some intrusion. One result of these influences is that that many people feel that Internet security and Internet firewalls are synonymous. Although the system should know that no single mechanism or method will

provide for the entire computer and network security needs of an enterprise, many still put all their network security eggs in one firewall basket.

Distributed firewalls are host-resident security software applications that protect the enterprise network's servers and end-user machines against unwanted intrusion. They offer the advantage of filtering traffic from both the Internet and the internal network. This enables them to prevent hacking attacks that originate from both the Internet and the internal network. This is important because the most costly and destructive attacks still originate from within the organization. They are like personal firewalls except they offer several important advantages like central management, logging, and in some cases, access-control granularity. These features are necessary to implement corporate security policies in larger enterprises. Policies can be defined and pushed out on an enterprise-wide basis. End-to-end encryption can also be a threat to firewalls, as it prevents them from looking at the packet fields necessary to do filtering. Allowing end-to-end encryption through a firewall implies considerable trust to the users on behalf of the administrators. Finally, there is an increasing need for finer-grained access control which standard firewalls cannot readily accommodate without greatly increasing their complexity and processing requirements.

For a firewall to be effective, companies first need to define their network security policy. A network security policy identifies the resources that need protection and the threats against them. It then defines how they can be used and who can use them, and stipulates the actions to be taken when the policies are violated. A policy is a set of rules against which arriving packets are tested. Examples of such rules include what IP traffic the organization wants to allow into its network, what source addresses should be excluded from the network, and what destination addresses within the network can be accessed from outside the network. Specific actions to be taken include accept packet and reject packet. The firewall is responsible for filtering traffic according to the security policy.

A packet filter is a firewall that inspects each packet for user-defined filtering rules to determine whether to pass or block it. For example, the filtering rule might require all Telnet requests to be dropped. Using this information, the firewall will block all packets that have a port number 23 (the default port number for Telnet) in their header. Filtering rules can be based on source IP address, destination IP address, Layer 4 (that is, TCP/UDP) source port, and Layer 4 destination port. Thus, a packet filter makes decisions based on the network layer and the transport layer. Packet filters are fast and can be easily implemented in existing routers. Unfortunately, they are the least secure of all firewalls. One disadvantage of packet filters is that they have no logging facility that can be used to detect when a break-in has occurred. Also, a packet filtering firewall grants or denies access to the network according to the source and destination addresses and the source and destination ports. Unfortunately, these ports can be spoofed. As a result, anyone can access network resources once access has been granted to an authorized user.

Firewall architecture refers to the manner in which firewall components are arranged to provide effective protection against unauthorized users. It is usually defined after the

network security policy has been defined because it is supposed to be a model that enforces the security policy. The network security policy is enforced at defensible boundaries within the network called perimeter networks. A corporate network usually contains multiple perimeter networks that can be classified into three groups: the outermost perimeter network, one or more internal perimeter networks, and the innermost perimeter network. The outermost perimeter network provides a boundary between corporate resources (that need to be protected) and external resources (resources the corporation cannot control). Internal perimeter networks represent boundaries within the corporate network that need additional security.

A dual-homed host firewall is a type of the multi-homed host firewall. In the dual-homed host firewall, the host (which provides the firewall functionality) has two interfaces: One interface is connected to the private network and the other interface is connected to the Internet (or some other untrusted network). Thus, all IP traffic from the Internet must pass through the firewall before arriving at a host in the private network. Similarly, an internal host can communicate with external hosts (that is, hosts in the Internet) via the dual-homed host. Direct communication that bypasses the dual-homed host is blocked. This means that the IP forwarding capability of the dual-homed host is disabled to ensure that IP packets from one network are not be directly routed to the other network. The dual-homed host cannot operate as a router. However, disabling IP packet forwarding ensures that the Internet and the private network are logically disconnected so that even when system problems occur the firewall cannot fail open

A virtual private network (VPN) provides a secure connection between a sender and a receiver over a public non-secure network such as the Internet. A secure connection is generally associated with private networks. (A private network is a network that is owned, or at least controlled via leased lines, by an organization.) Using the techniques discussed later in this chapter, a VPN can transform the characteristics of a public non-secure network into those of a private secure network. VPNs reduce remote access costs by using public network resources. Compared to other solutions, including private networks, a VPN is inexpensive.

Remote network access involves setting up a virtual private network (VPN) connection between the remote computer using VPN client software and a special gateway router that allows access to the university network over the Internet. This requires a high-speed connection to the Internet via an Internet Service Provider. Access is granted to users by login, using an account name and password combination. When actively connected to the NJCU network, all traffic to and from the remotely attached PC is through the VPN tunnel, including Internet browsing.

PROPOSED SYSTEM

In the field of computational geometry, proposed an algorithm that solves the point location problem for n *non-overlapping* d -dimensional hyper-rectangles, with a linear space requirement and $O((\log n)(d-1))$ search time. In our case, the system has *overlapping* d -dimensional hyper-rectangles, since firewall rules can, and often do, overlap each other—

making rules overlap is the method firewall administrators use to implement intersection and difference operations on sets of IP addresses or port numbers. These overlapping hyper-rectangles can be decomposed into non-overlapping hyper-rectangles—however, a moment’s reflection shows that the number of resulting non-overlapping hyper-rectangles is (nd) , thus the worst-case complexity for firewall rules is no better than that of GEM.

Advantages:

1. Packet filter firewall supports high speed: Packet filter firewall over configurations of simple network works with more speed. The thing behind this is that a packet filter firewall has a direct connection between external hosts & internal users. Packet filters make decisions based on each **packet**, it doesn't make decisions on the basis of the traffic context. Increase in vulnerability over the internet.
2. It is used to implement and enforce a security policy for communication between networks.

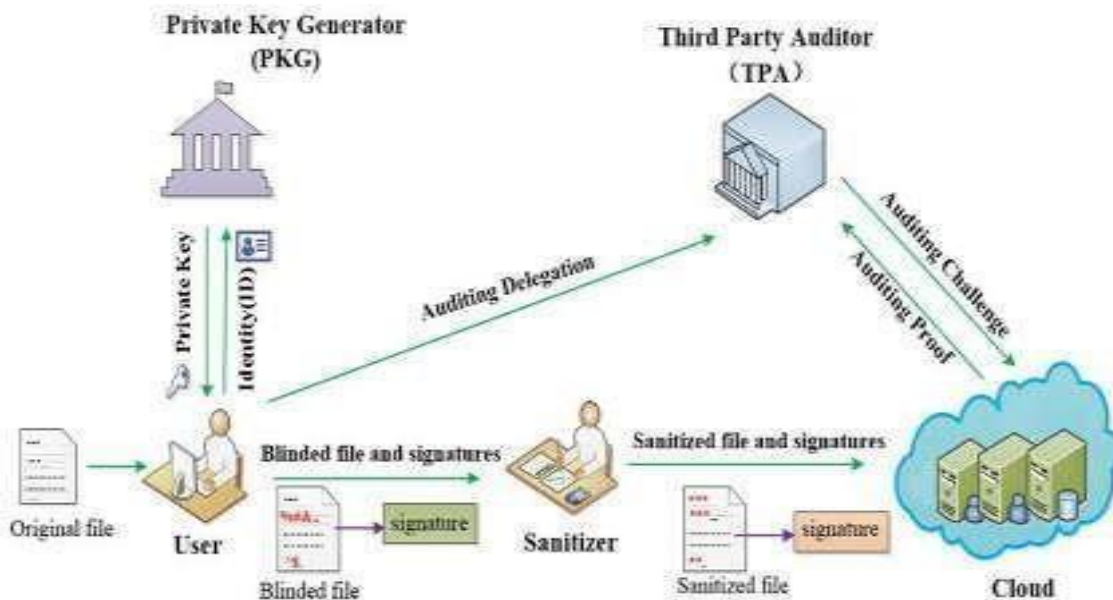
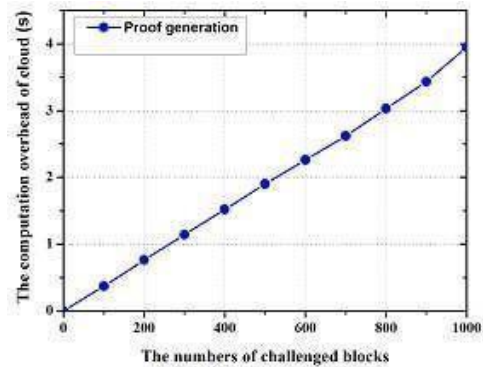
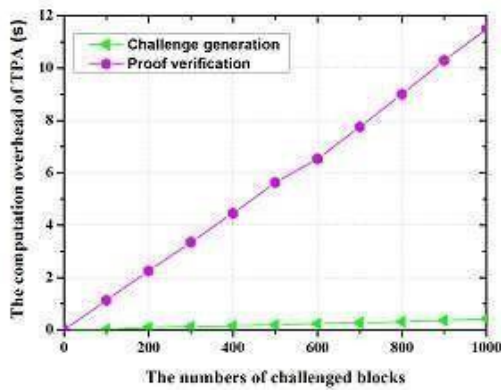


Fig 1:Architecture Diagram

RESULT AND DISCUSSION

The protocol achieves significant redundancy ratio on four real firewall groups. The protocol is efficient for processing and comparing two real firewalls. The protocol is efficient for the communication cost between two parties. The protocol is efficient for processing and comparing two synthetic firewalls. The protocol is efficient for the communication cost between two synthetic firewalls. This proposed firewall can act perfectly as firewall. This can avoid malwares. Secure network from intrusion. Prevent from malware as the firewall does. But the real performance is done by enabling virtual policy. The policy settings are constructed effectively.



CONCLUSION

Thus the new firewall has been developed with effective security policies. This will replace the firewalls in the VPN. As existing firewalls dump the internet throughput this proposed system of firewalls will solve this problem. It can increase the throughput over the network. The user validation policy is effective and efficient and can handle more users at a time. It can detect the packets as the real firewall does. This protocol is suitable for all types of networks. It can be also applicable to bridged networks.

REFERENCES

- [1] nf-HiPAC, "Firewall throughput test," 2012 [Online]. Available: http://www.hipac.org/performance_tests/results.html
- [2] R. Agrawal, A. Evfimievski, and R. Srikant, "Information sharing across private databases," in *Proc. ACM SIGMOD*, 2003, pp. 86–97.
- [3] E. Al-Shaer and H. Hamed, "Discovery of policy anomalies in distributed firewalls," in *Proc. IEEE INFOCOM*, 2004, pp. 2605–2616.
- [4] J. Brickell and V. Shmatikov, "Privacy-preserving graph algorithms in the semi-honest model," in *Proc. ASIACRYPT*, 2010, pp. 236–252.
- [5] Y.-K. Chang, "Fast binary and multiway prefix searches for packet forwarding," *Comput. Netw.*, vol. 51, no. 3, pp. 588–605, 2007.
- [6] J. Cheng, H. Yang, S. H. Wong, and S. Lu, "Design and implementation of cross-domain cooperative firewall," in *Proc. IEEE ICNP*, 2007, pp. 284–293.
- [7] Q. Dong, S. Banerjee, J. Wang, D. Agrawal, and A. Shukla, "Packet classifiers in ternary CAMs can be smaller," in *Proc. ACM SIGMETRICS*, 2006, pp. 311–322.
- [8] O. Goldreich, "Secure multi-party computations," Working draft, Ver. 1.4, 2002.
- [9] O. Goldreich, *Foundations of Cryptography: Volume II (Basic Applications)*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [10] M. G. Gouda and A. X. Liu, "Firewall design: Consistency, completeness and compactness," in *Proc. IEEE ICDCS*, 2004, pp. 320–327.
- [11] M. G. Gouda and A. X. Liu, "Structured firewall design," *Comput. Netw.*, vol. 51, no. 4, pp. 1106–1120, 2007.
- [12] P. Gupta, "Algorithms for routing lookups and packet classification," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2000.

- [13] A. X. Liu and F. Chen, "Collaborative enforcement of firewall policies in virtual private networks," in *Proc. ACM PODC*, 2008, pp. 95–104.
- [14] A. X. Liu and M. G. Gouda, "Diverse firewall design," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 8, pp. 1237–1251, Sep. 2008.
- [15] A. X. Liu and M. G. Gouda, "Complete redundancy removal for packet classifiers in TCAMs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no.4, pp. 424–437, Apr. 2010.
- [16] A. X. Liu, C. R. Meiners, and E. Torng, "TCAM Razor: A systematic approach towards minimizing packet classifiers in TCAMs," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 490–500, Apr. 2010.
- [17] A. X. Liu, C. R. Meiners, and Y. Zhou, "All-match based complete redundancy removal for packet classifiers in TCAMs," in *Proc. IEEE INFOCOM*, 2008, p

IMPROVED PRIVACY PRESERVING CONTENT BASED RETRIEVAL IN CLOUD RESPOSITORIES

S. ROJA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

By leveraging virtual machine (VM) technology which provides performance and fault isolation, cloud resources can be provisioned on demand in a fine grained, multiplexed manner rather than in monolithic pieces. By integrating volunteer computing into cloud architectures, this envision a gigantic self-organizing cloud (SOC) being formed to reap the huge potential of untapped commodity computing power over the Internet.

Toward this new architecture where each participant may autonomously act as both resource consumer and provider, this proposes a fully distributed, VM-multiplexing resource allocation scheme to manage decentralized resources. This approach not only achieves maximized resource utilization using the proportional share model (PSM), but also delivers provably and adaptively optimal execution efficiency.

This design a novel multi-attribute range query protocol for locating qualified nodes. Contrary to existing solutions which often generate bulky messages per request, our protocol produces only one lightweight query message per task on the Content Addressable Network (CAN). It works effectively to find for each task its qualified resources under a randomized policy that mitigates the contention among requesters. It shows the SOC with our optimized algorithms can make an improvement by 15-60 percent in system throughput than a P2P Grid model. Our solution also exhibits fairly high adaptability in a dynamic node-churning environment.

keywords: virtual machine, self-organizing cloud, proportional share model, Content Addressable Network.

1. INTRODUCTION

Cloud computing has emerged as a compelling paradigm for deploying distributed services. Resource allocation problem in cloud systems emphasizes how to harness the multiattribute resources by multiplexing operating systems [1]. With virtual machine (VM) technology, we are able to multiplex several operating systems on the same hardware and allow task execution over its VM substrates without performance interference. Fine-grained resource sharing can be achieved as each VM substrate can be configured with [2] proper shares of resources (such as CPU, memory, storage, network bandwidth) dynamically.

In recent years, various enhancements on resource isolation techniques have been proposed to achieve fine-grained dynamic resource provisioning [3]. A proportional share scheduler can be implemented based on Xen's credit scheduler to multiplex CPU resource among virtual machines in a fair manner [4]. The balloon driver difference engine joint-VM and virtual putty can dynamically adjust the memory resource among colocated virtual machines.

The dm-ioband can dynamically control the usage of disk I/O bandwidth among colocated virtual machines. These advanced techniques enable computing resources to be

dynamically partitioned or reassembled to meet the elastic needs of end users [5]. Such solutions create an unprecedented opportunity to maximize resource utilization, which were not possibly applied in most Grid systems that usually treat the underlying resources as indivisible ones and prevent simultaneous access to them [6].

Today's cloud architectures are not without problems. Most cloud services built on top of a centralized architecture may suffer denial-of-service (DoS) attacks, unexpected outages, and limited pooling of computational resources [7]. On the contrary, volunteer computing systems (or Desktop Grids) can easily aggregate huge potential computing power to tackle grand challenge science problems. In view of this, we propose a novel cloud architecture, namely self-organizing cloud (SOC), which can connect a large number of desktop computers on the Internet by a P2P network. In SOC, each participating computer acts as both a resource provider and a resource consumer [8]. They operate autonomously for locating nodes with more abundant resource or unique services in the network to offload some of their tasks, meanwhile they could construct multiple VM instances for executing tasks submitted from others whenever they have idle resources.

We focus on two key issues in the design of the multi attribute range query problem in a fully decentralized environment for locating a qualified node to satisfy a user task's resource demand with bounded delay and how to optimize a task's execution time by determining the optimal shares of the multi attribute resources to allocate to the tasks with various QoS constraints, such as the expected execution time and limited budget [9].

As a fundamental difference to existing approaches, we formulate such a resource allocation problem to be a convex optimization problem. Given a task with its resource requirements and a budget, we first prove that the optimal resource allocation on a qualified node that can minimize a task's execution time does exist. We further show that it is nontrivial to solve such a convex optimization problem directly [10] via a brute-force strategy and the interior point method. By relaxing the problem definition, we propose an algorithm to optimize the task execution time on a qualified resource node, given its preset budget and tolerable quality of service (QoS). The proposed algorithm involves only $O(R^2P)$ adjustment steps, where R denotes the number of resource attributes (or dimensions). We further propose a dynamic optimal proportional-share (DOPS) resource allocation algorithm with $O(R^3P)$ complexity, by incorporating the proportional-share model (PSM). The key idea is to dynamically scale the amount of resources at each dimension among running tasks proportional to their demand, such that these tasks could use up the maximum capacity of each resource type at a node.

The proposed PG-CAN range query protocol in this work aims to find the qualified resources with minimized contention among requesters based on task's demand. It is unique in that for each task, there is only one query message propagated in the network during the entire course of discovery. This is different from most existing multi attribute range query solutions that require to propagate multiple subqueries along multiple dimensions in parallel. To mitigate the contention problem due to analogous queries in CAN, our range query protocol proactively diffuses resource indexes over the network and randomly route query messages among nodes to locate qualified ones that satisfy tasks' minimal demands. To avoid possibly uneven load distribution and abrupt resource overutilization caused by

uncoordinated node selection process from autonomous participants, we investigate three node selection policies, namely double-check policy queuing policy and extra-virtual- dimension (VD) policy.

2. LITERATURE REVIEW

Cloud computing has emerged as a compelling paradigm for deploying distributed services. Resource allocation problem in cloud systems emphasizes how to harness the multiattribute resources by multiplexing operating systems. With virtual machine (VM) technology, we are able to multiplex several operating systems on the same hardware and allow task execution over its VM substrates without performance interference. Fine-grained resource sharing can be achieved as each VM substrate can be configured with proper shares of resources (such as CPU, memory, storage, network bandwidth) dynamically. In recent years, various enhancements on resource isolation techniques have been proposed to achieve fine-grained dynamic resource provisioning. A proportional share scheduler can be implemented based on Xen's credit scheduler to multiplex CPU resource among virtual machines in a fair manner. The balloon driver, difference engine, joint-VM, and virtual putty, can dynamically adjust the memory resource among collocated virtual machines. The dm- ioband can dynamically control the usage of disk I/O bandwidth among collocated virtual machines. These advanced techniques enable computing resources to be dynamically partitioned or reassembled to meet the elastic needs of end users. Such solutions create an unprecedented opportunity to maximize resource utilization,

Today's cloud architectures are not without problems. Most cloud services built on top of a centralized architecture may suffer denial-of-service (DoS) attacks, unexpected outages, and limited pooling of computational resources. On the contrary, volunteer computing systems (or Desktop Grids) can easily aggregate huge potential computing power to tackle grand challenge science problems. A novel cloud architecture, namely self- organizing cloud (SOC), which can connect a large number of desktop computers on the Internet by a P2P network. In SOC, each participating computer acts as both a resource provider and a resource consumer. They operate autonomously for locating nodes with more abundant resource or unique services in the network to offload some of their tasks, meanwhile they could construct multiple VM instances for executing tasks submitted from others whenever they have idle resources. Focus on two key issues in the design of SOC: 1) the multi attribute range query problem in a fully decentralized environment for locating a qualified node to satisfy a user task's resource demand with bounded delay and 2) how to optimize a task's execution time by determining the optimal shares of the multi attribute resources to allocate to the tasks with various QoS constraints, such as the expected execution time and limited budget. As a fundamental difference to existing approaches. Formulate such a resource allocation problem to be a convex optimization problem. Given a task with its resource requirements and a budget, First prove that the optimal resource allocation on a qualified node that can minimize a task's execution time does exist. Further show that it is nontrivial to solve such a convex optimization problem directly via a brute-force strategy and the interior point method.

By relaxing the problem definition, algorithm to optimize the task execution time on a qualified resource node, given its preset budget and tolerable quality of service (QoS). The proposed algorithm involves only O \mathcal{R}^2 adjustment steps, where R denotes the number of resource attributes (or dimensions). Propose a dynamic optimal proportional-share (DOPS) resource allocation algorithm with O \mathcal{R}^3 complexity, by incorporating the proportional-share model (PSM). The key idea is to dynamically scale the amount of resources at each dimension among running tasks proportional to their demand, such that these tasks could use up the maximum capacity of each resource type at a node. To locate qualified nodes in the SOC environment, design a fully decentralized range query protocol, namely pointer-gossiping CAN (PG-CAN), tailored for DOPS. Existing P2P desktop Grids favor CAN-based or Chord-based

resource discovery protocols. Every joining node registers its static resource attributes (e.g., CPU architecture, OS version) or maximum capacity on the CAN/Chord overlay, so that other users could find the most matched node within a logarithmic (or sublinear) number of routing steps. Such a design is feasible for a P2P desktop Grid because the resources of a selected node can only be used exclusively by a single task. Due to dynamic resource provisioning technologies used in cloud, the frequent resource repartitioning and reallocation (e.g., upon task arrival or completion) make it a challenging problem to locate a node containing a combination of available resources along all the R resource attributes that would satisfy the requirements of a submitted task.

The proposed PG-CAN range query protocol in this work aims to find the qualified resources with minimized contention among requesters based on task's demand. It is unique in that for each task, there is only one query message propagated in the network during the entire course of discovery. This is different from most existing multiattribute range query solutions that require to propagate multiple sub queries along multiple dimensions in parallel. To mitigate the contention problem due to analogous queries in CAN, our range query protocol proactively diffuses resource indexes over the network and randomly route query messages among nodes to locate qualified ones that satisfy tasks' minimal demands. To avoid possibly uneven load distribution and abrupt resource overutilization caused by uncoordinated node selection process from autonomous participants, The investigate three node selection policies, namely double-check policy, queuing policy, and extra-virtual-dimension (VD) policy. The rest of the paper is organized as follows: formulate the resource allocation problem in a VM-multiplexing environment. That optimal resource allocation does exist and show that our solution is optimal. Present our DOPS resource allocation scheme. The proposed range query protocol. The simulation results related work in and conclude with an outline of future work in.

3. PROPOSED METHODOLOGY

This proposed system a new method for resource allocation. This paper designs a fully decentralized range query protocol, namely pointer-gossiping CAN (PG-CAN). The proposed PG-CAN range query protocol in this work aims to find the qualified resources with minimized contention among requesters based on task's demand.

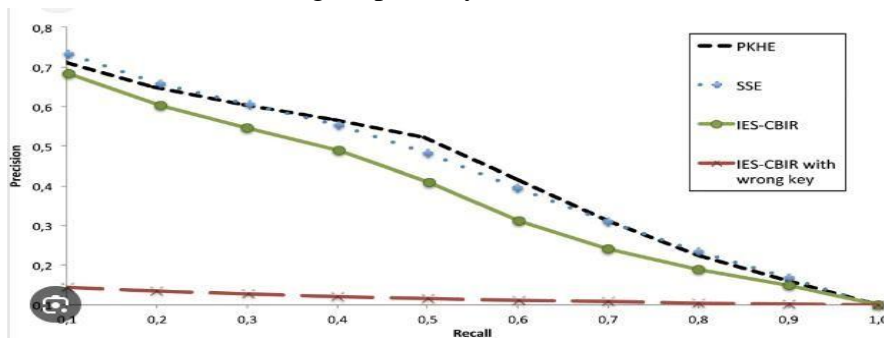
ADVANTAGES

- Optimization of task’s resource allocation under user’s budget.
- Maximized resource utilization based on PSM Lightweight resource query protocol with low contention
- Adaptability: The resource allocation process must dynamically and efficiently adapt to changes in the demand from sites.
- Load balancing tests are done well for service requests presented in the current clouds.
- Frequent Bottlenecks problems are not happening due to the improved client requests model.
- Buffer capacity of the cloud is updated during the runtime by merging different clouds at runtime.

4. RESULTS AND DISCUSSIONS

Storage requirements for visual data have been increasing in recent years, following the emergence of many highly interactive multimedia services and applications for mobile devices in both personal and corporate scenarios. This has been a key driving factor for the adoption of cloud-based data outsourcing solutions. However, outsourcing data storage to the Cloud also leads to new security challenges that must be carefully addressed, especially regarding privacy. In this paper we propose a secure framework for outsourced privacy- preserving storage and retrieval in large shared image repositories. Our proposal is based on IES-CBIR, a novel Image Encryption Scheme that exhibits Content-Based Image Retrieval properties. The framework enables both encrypted storage and searching using Content- Based Image Retrieval queries while preserving privacy against honest-but-curious cloud administrators. We have built a prototype of the proposed framework, formally analyzed and proven its security properties, and experimentally evaluated its performance and retrieval precision. Our results show that IES-CBIR is provably secure, allows more efficient operations than existing proposals, both in terms of time and space complexity, and paves the way for new practical application scenarios.

Graph.1 shown the cloud image repository



CONCLUSION

In conclusion With a realistic monetary model, we propose a solution which can optimize the task execution performance based on its assigned resources under the user budget. We prove its optimality using the KKT conditions in the convex-optimization theory. In order to further make use of the idle resources, we design a dynamic algorithm by combining the above algorithm with PSM and the arrival/completion of new tasks. This can give incentives to users by gaining an extra share of unused resource without more payment. Experiments confirm achieving a super optimal execution efficiency of their tasks is possible. DOPS could get an improvement on system throughput by 15 percent 60 percent than the traditional methods used in P2P Grid model, according to the simulation. We summarize the resource searching request as two range query constraints, We prove them to be the sufficient and necessary conditions for getting the optimal resource allocation. Experiments confirm the designed PG-CAN protocol with lightweight query overhead is able to search qualified resources very effectively.

6. REFERENCES

- [1] J.E. Smith and R. Nair, *Virtual Machines: Versatile Platforms for Systems And Processes*. Morgan Kaufmann, 2005.
- [2] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat, "Enforcing Performance Isolation across Virtual Machines in Xen," *Proc. Seventh ACM/IFIP/USENIX Int'l Conf. Middleware (Middleware '06)*, pp. 342-362, 2006.
- [3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," *Technical Report UCB/EECS-2009-28*, Feb. 2009.
- [4] D.P. Anderson, "Boinc: A System for Public-Resource Computing and Storage," *Proc. IEEE/ACM Fifth Int'l Workshop Grid Computing*, pp. 4-10, 2004.
- [5] P. Crescenzi and V. Kann, *A Compendium of NP Optimization Problems*. <ftp://ftp.nada.kth.se/Theory/Viggo-Kann/compendium.pdf>, 2012.
- [6] O. Sinnen, *Task Scheduling for Parallel Systems*, Wiley Series on Parallel and Distributed Computing. Wiley-Interscience, 2007.
- [7] O.H. Ibarra and C.E. Kim, "Heuristic Algorithms for Scheduling Independent Tasks on Nonidentical Processors," *J. ACM*, vol. 24, pp. 280-289, Apr. 1977.
- [8] X. Meng et al., "Efficient Resource Provisioning in Compute Clouds via vm Multiplexing," *Proc. IEEE Seventh Int'l Conf. Autonomic Computing (ICAC '10)*, pp. 11-20, 2010.
- [9] J. Sonneck and A. Chandra, "Virtual Putty: Reshaping the Physical Footprint of Virtual Machines," *Proc. Int'l HotCloud Workshop in Conjunction with USENIX Ann. Technical Conf.*, 2009.
- [10] D. Gupta et al., "Difference Engine: Harnessing Memory Redundancy in Virtual Machines," *Proc. Eighth Int'l USENIX Symp. Operating Systems Design and Implementation*, pp. 309-322, 2008.

ONLINE CRIME REPORTING AND MANAGEMENT SYSTEM

K. SARATHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The aim of this project is to develop an online crime reporting and management system which is easily accessible to the public, the police department and the administrative department. The normal public in India is afraid to lodge a complaint because they are filled with a false fear about the police department. An online complaint registering system will allay the fears of the public and will also help in the public helping the police department in catching criminals.

An online solution is very useful as the solution is inherently distributive. This distributive characteristic of the online solution helps in getting the different police stations to share information and get in contact with one another. The administrative work required to maintain records reduces greatly as the paperwork is almost minimal and the data is stored in an organized.

The other features of this online solution are enquiry about a complaint which has been registered before, the status of the complaint and other information. Keeping records of all the criminals will help the police department in keeping tabs on the criminals to refrain them from getting into illegal activities.

On the whole, the online crime registering and maintenance solution is an excellent method, which bridges the gap between the public and the police department and also helps the police department in preventing criminal activities.

Keywords: online crime, police department, online solution, criminals

1. INTRODUCTION

System analysis is a process of gathering and interpreting facts, diagnosing problems, and the information to recommend improvements to the system. It is a problem-solving activity that requires intensive communication between the system users and system developers. System analysis or study is an important phase of any system development process. The system is studied to the minutest detail and analyzed. The system analyst plays the role of the interrogator and dwells deep into the workings of the present system. The system is viewed as a whole and the input to the system is identified. The outputs from the organizations are traced to the various processes. System analysis is concerned with becoming aware of the problem, identifying the relevant and decisional variables, analyzing and synthesizing the various factors, and determining an optimal or at least satisfactory solution or program of action.

A detailed study of the process must be made by various techniques like interviews, questionnaires, etc. The data collected by these sources must be scrutinized to conclude. The conclusion is an understanding of how the system functions. This system is called the

existing system. A preliminary study is the process of gathering and interpreting facts and using the information for further studies on the system.

Preliminary study is a problem problem-solving activity that requires intensive communication between the system users and system developers. It does various feasibility studies. In these studies, a rough figure of the system activities can be obtained, from which the decision about the strategies to be followed for effective system study and analysis can be taken. Here in the Email to Fax server project, a detailed study of the existing system is carried out along with all the steps in system analysis. An idea for creating a better project was carried and the next steps were followed.

The manual system has some drawbacks which can be overcome by using web-based software. The following reasons explain why it is needed. Citizens do not need to go to police stations to see the criminal's information. They can directly see information on the site. Visitors can easily get information about the crime and criminals. Reduce the manpower and also reduce the time. Member can view the current status of the criminal.

Crime is a part of illegal activities in human life. The rise of population and complex society rises the range of anti-social conducts that must be restricted by the government through the military and different organizations particularly the Police Force. There are many current crime management systems which face several difficulties, as there is no means to report crime instantly other than phone calls, messaging or face-to-face complaint filing. Hence, we have proposed an online crime reporting system which allows the user to file complaints or missing reports and keep a track of it. There are 3 categories that a user can file; Complaint, Crime Report and Missing Report and can see all the status of what action has been taken by the admin. To file any of the above 3 complaints, the user should register in to the system and provide his right credentials to file them. The crime reporting system project also allows other users who doesn't want to register but can check the crimes happening at his/her or any other area, has to just provide the pin code and in return the system displays the list of crimes if any filed. The offline i.e. the unregistered user can also take advantage of checking the missing person details, but he/she is refrained from getting complaints done by the users.

The crime rates accelerate continuously and the crime patterns are constantly changing. According to National Crime Records Bureau, crime against women has significantly increased in recent years. It has become the most prior to the administration to enforce law and order to reduce this increasing rate of the crime against women. We illustrate how social development may lead to crime prevention. So we are developing the system which can be used to detect and predict the crimes for the area where the person or user currently stand. Crime detection and analysis will be to generate the crime hot-spots that will help in deployment of police at mostly likely places of crime for any window of time, to allow most effective utilization of police resources. The developed model will reduce crimes and will help the crime detection field in many ways that is from arresting the criminals to reducing the crimes by carrying out various necessary measures. We add the woman safety module for security. When women

press the power button of android mobile 3 to 4 times then help message send to relatives or police. Due to this we reduce crime in the society and in country. Here we use module of crime capture means user can capture the photo of crime send to police

2. LITERATURE REVIEW

Online Crime Reporting and Management Systems (OCRMS) have become increasingly significant in contemporary law enforcement practices. These systems aim to facilitate efficient reporting, investigation, and management of crimes through digital platforms. This literature review synthesizes existing research and developments in OCRMS, highlighting their advantages, challenges, and implications for law enforcement agencies and communities. Early efforts in digitizing crime reporting processes date back to the late 20th century, with basic online forms for reporting crimes. The evolution of technology has led to the development of sophisticated OCRMS, integrating features such as real-time reporting, geolocation services, and multimedia uploads. Accessibility: OCRMS offer convenient and accessible channels for citizens to report crimes, breaking down barriers such as geographical constraints and time limitations. By automating the reporting process, OCRMS streamline data collection and transmission, enabling law enforcement agencies to respond promptly to incidents. Digital reporting allows for more accurate and detailed documentation of incidents, facilitating better analysis and resource allocation. Digital Divide: Concerns persist regarding equitable access to OCRMS, with disparities in internet connectivity and digital literacy potentially excluding marginalized communities. Ensuring the security of sensitive information submitted through OCRMS is critical to maintaining public trust and compliance with data protection regulations. Integrating OCRMS with legacy databases and law enforcement workflows poses technical challenges and requires careful planning and coordination. Studies examining user satisfaction with OCRMS highlight the importance of user-friendly interfaces, clear instructions, and responsive customer support in enhancing adoption and utilization rates. Feedback mechanisms and iterative improvements based on user input are essential for optimizing OCRMS performance and usability. OCRMS empower law enforcement agencies with timely and comprehensive data on crime trends, enabling proactive strategies for crime prevention and resource allocation. Improved community engagement and trust are often cited as outcomes of implementing OCRMS, as they provide citizens with a sense of agency and responsiveness in addressing public safety concerns.

3. PROPOSED METHODOLOGY

The proposed system aims to develop a system of improved facilities. The proposed system can overcome all the limitations of the existing system. The system provides proper security and reduces manual work. The existing system has several disadvantages and many more difficulties to work well. The proposed system tries to eliminate or reduce these difficulties to some extent. The proposed system will help the user to reduce the workload and mental conflict. The proposed system helps the user to work user friendly and he can easily do his jobs without time lagging.

ADVANTAGES

- Proper control of the higher officials.
- Reduce the damage of the machines.
- Minimum time needed for the various processing.
- Greater efficiency.
- Better service.
- User-friendliness and interaction.
- Minimum time required.



Figure 1: System Architecture [D]Feasibility Analysis Np-hard Np-Complete:

4. RESULTS AND DISCUSSIONS

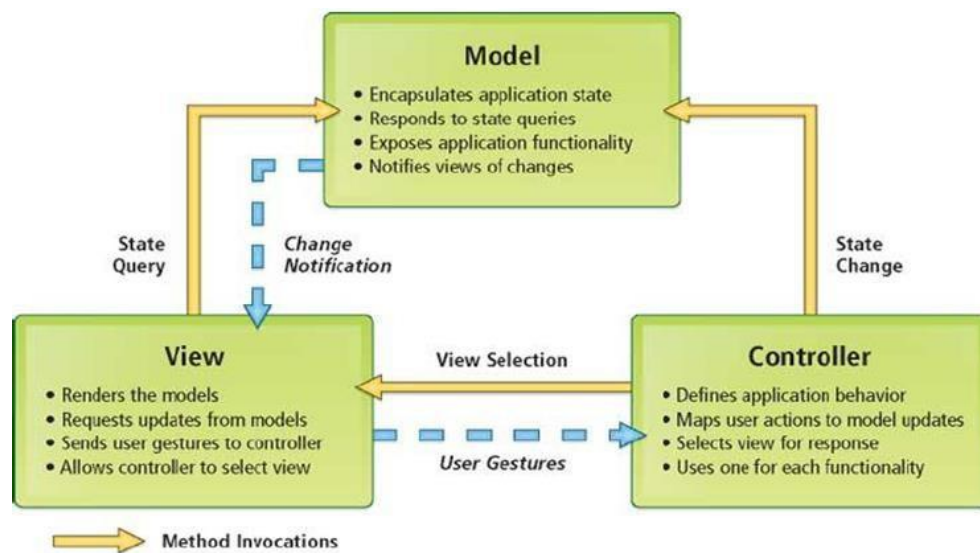
This type of online crime involves fraudulent attempts to obtain sensitive information, such as usernames, passwords, and credit card details, by masquerading as a trustworthy entity in electronic communication. Identity theft occurs when someone wrongfully obtains and uses another person's data in some way that involves fraud or deception, typically for economic gain.

Online fraud encompasses various deceptive practices conducted over the Internet to deceive individuals or organizations for financial gain. Cyberbullying refers to the use of electronic communication to bully, harass, or intimidate others, often through social media platforms, messaging apps, or online forums. Our analysis also revealed interesting patterns in the geographic distribution of online crimes. While online crimes occur worldwide, certain regions show higher prevalence rates. Factors such as internet penetration, socio-economic conditions, and law enforcement measures may contribute to these variations.

Examining demographic data of victims and perpetrators provided valuable insights into the profiles of individuals involved in online crimes. For instance, we found that younger individuals are more likely to be both victims and perpetrators of cyberbullying, while older

adults are often targeted in online fraud schemes. Online crimes have far-reaching consequences that extend beyond the immediate victims. They can result in financial loss, psychological trauma, reputational damage, and even physical harm. Moreover, the proliferation of online crimes poses significant challenges for law enforcement agencies and necessitates continuous efforts to enhance cybersecurity measures and educate the public about online safety practices.

Finally, our study identified emerging trends and technologies that shape the landscape of online crime. These include advancements in hacking techniques, the rise of cryptocurrency-related crimes, and the growing use of artificial intelligence and machine learning by both perpetrators and cybersecurity professionals. While this study provides valuable insights into online crime, it is not without limitations. The data analyzed may be subject to reporting biases, and the methodologies employed have inherent limitations. Future research should aim to address these limitations and explore emerging forms of online crime, as well as evaluate the effectiveness of preventive measures and intervention strategies.



5. CONCLUSION

. The “E-crime” process was made computerized to reduce human errors and to increase efficiency. The main focus of this project is to lessen human efforts. The maintenance of the records is made efficient, as all the records are stored in the MySQL database, through which data can be retrieved easily. The navigation control is provided in all the forms to navigate through the large amount of records. If the number of records is very large then the user has to just type in the search string and the user gets the results immediately. The editing is also made simpler. The user has to just type in the required field and press the update button to update the desired field. The problems, that existed in the

earlier system, have been removed to a large extent. And it is expected that this project will go a long way in satisfying users' requirements.

6. REFERENCES

- [1] Steven Holzner, "HTML Black Book", Jon Skeet," C# in depth
- [2] Shiju Sathyadevan, Crime analysis and prediction, IEEE,25Sept2014,10.1109/CNSC.2014.6906719
- [3]. SachinBagga, AkshayGirdhar, Rajun Yan and Zihan Lin, Virtualization Approach to Cluster Based Winograd's Variant of Strassen's Method using RMI, Second International Conference on Computational Intelligence & Communication Technology, 2016.
- [4] David Fojtik, PetrPodesva and Jan Gebauer, Storing High Volumes of Data in MS SQL Server Express, 16th International Carpathian Control Conference (ICCC) ,2015.
- [5]Dr. R. Nageshwararao," Core Java: An Integrated Approach", Dreamtech press.
- [6]Patrick Naughton and Herbert Schildt, "Java: The Complete References",seventh edition.
- [7]RudrakshBatra," Java EE 5 Black Book",Dreamtech press.
- [8]SachinBagga, AkshayGirdhar, MuneshChandra Trivedi and Yingzhi Yang, RMI Approach to Cluster Based Cache Oblivious Peano Curves, Second International Conference on Computational Intelligence & Communication Technology,2016.
- [9] HarmanpreetKaur, SachinBagga ,AnkitArora, RMI Approach to Cluster Based Winograd'sVariant ofStrassen'sMethod, IEEE 3rd InternationalConferenceon MOOCs, Innovation and Technology in Education (MITE),2015.
- [10] Ala' Alkhaldi, Indranil Gupta, VijayanthRaghavan, MainakGhosh, Leveraging Metadata inNoSQLStorage Systems,2015.

RANK BASED FRAUD DETECTION FOR MOBILE APPLICATION

B. SIVAPRIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

We propose Fraudulent behaviors in Google Play, the most popular Android app market, fuel search rank abuse and malware proliferation. To identify malware, previous work has focused on app executable and permission analysis. In this paper, we introduce Fair Play, a novel system that discovers and leverages traces left behind by fraudsters, to detect both malware and apps subjected to search rank fraud. Fair Play correlates review activities and uniquely combines detected review relations with linguistic and behavioral signals gleaned from Google Play app data (87K apps, 2.9M reviews, and 2.4M reviewers, collected over half a year), in order to identify suspicious apps. Fair Play achieves over 95% accuracy in classifying gold standard datasets of malware, fraudulent and legitimate apps. We show that 75% of the identified malware apps engage in search rank fraud. Fair Play discovers hundreds of fraudulent apps that currently evade Google Bouncer's detection technology. Fair Play also helped the discovery of more than 1,000 reviews, reported for 193 apps, that reveal a new type of "coercive" review campaign: users are harassed into writing positive reviews, and install and review other apps.

Keywords: Malware, Search Rank, Play store, Genuine Apps, Fraud Detection

1. INTRODUCTION

Google Play (already Android Market) is a computerized circulation maintained by Google over the past two decades. Mostly, the apps available in this forum are authorized ones and the users who are downloading also be authorized by the developer and Google play store before they will start using the app. In this environment, all kinds of the app for supporting knowledge, entertainment, and gaming are available. In the Google App store, [1] some applications may available for payment and most of them are malware-protected apps. But most of the apps are free of cost and non-secure for the environment where we are going to use them. Most of the apps require permission to access the data in the gadgets where it is going to be used including some personal sensitive information. The business achievement of android application markets, [2] for example, google play and the motivating force model offer plentiful mainstream applications, making them engaging focuses for fraudsters. Some deceitful engineers misleadingly help the inquiry rank and fame of their applications through phony surveys and fake installation tallies, while vindictive designers use application markets as a launch pad for their malware. [3] The inspiration for such practices is the effect of application prominence floods convert into money-related advantages and sped up malware expansion. Fraudulent designers much of the time abuse publicly supporting destinations to employ groups of willing laborers to submit misrepresentation, by and large, copying reasonable, unconstrained exercises from disconnected individuals. Likewise, the endeavors of Android markets to recognize and expel malware are not generally fruitful. [4] For example, Google play utilizes the Bouncer framework to dispatch the malware.

Nonetheless, out of the gathered google play applications information which is investigated using virus-out, in total, just 12% were hailed by in any event one anti-virus instrument and 2% were recognized as malware by in any event 10 tools around. The existed portable malware recognition researches are concentrated on the powerful investigation of static examination of code by the applications on the device and authorizations. In many cases, malware is injecting unauthorized data to sidestep hostile anti-virus software. This kind of injection is considered in recent research in malware detection of the app. This work looks to recognize both malware and search rank fraud subjects in google play. The work recommended that vindictive designers resort to search rank fraud to help the effect of their malware. The odious demonstrations have been revealed by choosing the perception that fraudulent and malevolent practices abandon indications on application markets as trails. For example, the significant expense of setting up legitimate google play accounts powers fraudsters to reuse their records across survey composing employments, making them liable to audit more applications in like manner than standard clients. Asset imperatives can propel fraudsters to post audits inside brief timeframe stretches. Genuine clients influenced by malware may report terrible encounters in their surveys. The proposed system provides the ability to identify the fraud apps on the collected app rate and review the data set, with the use of user edge connection-based information. This proposed system uses semantic and similarity measures to identify the fraud apps and the sentiment analysis is done to predict the positive and negative reviews. Both these approaches with time-based reviews are used to predict the fraud apps. Then the results are displayed in the graph to visualize the percentage of fraud apps from Google play. The remainder of the paper is composed as follows gives the writing overview which highlights the work related to Google play malware detection using the classification method. Section 3 gives the details of the system architecture and modules presents the overall block diagram and explains all the modules in detail describes the implementation and experiments and results of cosine similarity and sentiment analysis using a classification algorithm. Finally, Section 5 highlights the conclusion and future work.

2. LITERATURE REVIEW

To handle the security issues brought about by malware of Android OS, Hengshu Zhu et al. have proposed an exceptionally productive crossover distinguishing plan for Android Malware. In this paper, the creator proposed some recognizing methods, for example, highlights dependent on conventional Permission and API call highlights to improve the exhibition of static location. The crumbling issue of conventional capacity called graph-based malware identification was likewise kept away. The visitation results demonstrated that the Recommended scheme accomplished high malware identifying accuracy, and the plan could be Utilized to set up Android malware recognizing cloud administrations, which can naturally Adjust high productivity dissecting techniques as indicated by the properties of the Android Applications. Erika Chin et al. have concentrated to exhibit the Android malware available in the SDK files in the Android stage itself. Based on the examination conducted at four different agents, the identification of malware using this approach will achieve a better percentage than the existing

pessimistic scenarios. The results of these experiments are more likely to create a cutting edge against versatile malware arrangements. Micheal C Grace et al. have introduced a Machine Learning (ML) based framework for the identification of malware on Android gadgets and the framework extricates a few highlights and prepares a One-Class Support Vector Machine (SVM) in a disconnected (off-gadget) way, to use the higher processing intensity of a server or group of servers. Patrik Traynor et al. have examined the present status of portable malware in the wild and dissected the motivations. After the perception that 4 picks of malware use root adventure to mount advanced attacks on Android mobiles and analyzed the motivations that cause non-malignant cell phone hobbyists to distribute root misuses and reviewed the accessibility of root abuses. Ee-Peng Lim et al. have proposed a novel method for processing a rank aggregation dependent on matrix consumption to dodge commotion and deficient information. The proposed technique takes care of an organized matrix finishing issue over the space of skew-symmetric matrices. The creator demonstrates a recuperation hypothesis enumerating when the proposed approach will work. Nikita Spirin and Jiawei Han. has detailed an overview of web spam recognition, which completely presents the standards and calculations in the writing. Undoubtedly, crafted by Web ranking spam identification is founded on the investigation of positioning standards of web crawlers, for example, Page Rank and inquiry term recurrence. This is not quite the same as ranking fraud extortion recognition for versatile Apps. Sulthana et al stated their work is to extract the real opinion of the user for reviews on Twitter. So, this article will show the importance of sentiment analysis using natural language processing forgetting the real opinion of the user, and how it helps in future prediction 3726 Fareena et al.: Google Play Malware Detection based on Search Rank Fraud Approach Chia-Mei Chen et al. have proposed a static strategy to distinguish malware in portable applications. In this framework, figuring out the idea is utilized to create source code for the dubious APK documents. After that utilizing a structured mapping creator fabricates the structure for the classes. At long last utilizing an information stream idea, a few examples of the diverse kind of threats have been made and used to distinguish the malware in applications. Contingent on the number of scouring designs the adequacy of this strategy is determined. David F Gleich and Lek-heng Lim. has proposed a novel method for figuring a rank aggregation dependent on matrix consumption to stay away from the clamor and deficient information. The method discussed in this work organized the given framework in matrix format. The matrices are skew-symmetric based to identify the issues in the ranking. And also the app creator who developed that app designed the theorem for recovering the app from malware using some celluloid information. They likewise play out a nitty-gritty assessment with celluloid information and a narrative examination with Netflix appraisals. To discover the arrangements, they used the SVM for solving the completion of the matrix. Rank accumulation is joined with the structure of skew-symmetric matrices. The creator improved the current approach for matrix fruition to deal with skew-symmetric information.

The Ranking Risks of Android Apps Using Probabilistic Generative Models given by Alaa Salman et al. is one of Android's safeguard systems against malevolent applications where a hazard correspondence is examined. This methodology is ineffectual as it presents the hazard data of each application in an independent manner and in a way that requires an excessive amount of specialized information and time to distill valuable data. Kent Shi and Kamal Ali. has examined to ensure the audit of spammers or spam surveys. The spammer may target just explicit protection. From that point forward, they gave fake surveys to that specific versatile application by making an alternate record to survey that account. The creator proposes a novel- based scoring technique to distinguish every survey of the specific item.

The creator makes exceptionally suspicious as a subset. By utilizing online spammer assessment programming the phoniness of the survey is determined. After the fulfillment of the assessment, the outcome demonstrates the viability to anticipate bogus audits. The Evaluation of ML classifiers for versatile malware recognition was proposed by Jyrki Kivinen and Manfred K Warmuth. passes on the rising number of clients' welcome drudges to create vindictive applications. Furthermore, the security of sensible information accessible on cell phones is taken lightly. Given the experimental results, the classifiers used in this work, precisely increase the precision and specificity while compared with CNN techniques. Selvakumar et al provide the clustering mechanisms to group the users based on their similarities. The K - Means is used in this work to cluster a similar user in a single group. By using this kind of method, the app provider will easily identify similar users who need a similar app for installing their malware app. Another work by Selvakumar and Sendhilkumar helps to provide personalized search results for the user using the neural networks approach. This will also find a similar user group to easily identify the user needed for their app. The overview of the related work is about identifying fake mobile apps using web-based software tools or rank aggregation methods. The main challenge in this project is to analyze the fake mobile apps using semantic and similarity measures and the reviews are analyzed based on their sentiment and interrelation between the co-reviewers action the entire fake apps are identified.

3. PROPOSED METHODOLOGY

Rank-based fraud detection for mobile applications involves identifying fraudulent activities based on the anomalous behavior of users compared to their peers. Here's a proposed methodology along with potential results for such a system:

Methodology:

1. Data Collection and Preprocessing:

- Collect user activity data from mobile applications, including login times, transaction history, location information, and device characteristics.
- Preprocess the data to remove noise, handle missing values, and anonymize sensitive information.

2. Feature Engineering:

- Extract relevant features from the preprocessed data, such as login frequency, transaction amounts, geographic diversity of transactions, and device fingerprinting attributes.
- Compute additional features that capture user behavior relative to their peers, such as percentile ranks or z-scores.

3. Rank-Based Anomaly Detection:

- Use statistical methods or machine learning algorithms to detect anomalies based on the ranks of user behavior features.
- Compare the ranks of individual users against the ranks of their peers for each feature.
- Identify users whose behavior deviates significantly from the norm based on rank-based thresholds or anomaly scores.

4. Model Training and Evaluation:

- Split the data into training and testing sets for model development and evaluation.
- Train anomaly detection models using supervised or unsupervised learning approaches, such as isolation forests, one-class SVM, or autoencoders.
- Evaluate model performance using metrics such as precision, recall, F1-score, and area under the ROC curve (AUC).

5. Threshold Optimization:

- Fine-tune anomaly detection thresholds based on validation data or cross-validation techniques.
- Adjust thresholds to balance between false positives and false negatives according to the specific fraud detection requirements.

6. Integration with Mobile Applications:

- Integrate the trained fraud detection model into mobile applications or backend systems.
- Implement real-time monitoring of user activities to detect fraud as it occurs.

7. Alert Generation and Response:

- Generate alerts or notifications when suspicious activity is detected.
- Implement response mechanisms to mitigate fraudulent transactions or activities, such as blocking user accounts or requiring additional verification steps.

4. RESULTS DISCUSSIONS

Precision, Recall, F1-score, and AUC are commonly used to evaluate the performance of the rank-based fraud detection model.

Example results may include precision of 0.85, recall of 0.80, F1-score of 0.82, and AUC of 0.90, indicating good overall performance in detecting fraudulent activities. Provide examples of detected anomalies, such as unusually high transaction amounts, logins from unfamiliar locations, or suspicious device changes. Showcase how the rank-based approach effectively identifies these anomalies compared to traditional threshold-based methods. Demonstrate the impact of threshold optimization on model performance and the trade-off between false positives and false negatives. Show how adjusting thresholds improves the balance between detecting fraud and minimizing false alarms. Showcase the ability of the system to detect fraud

in real-time and trigger timely responses to mitigate risks. Present examples of fraudulent transactions or activities that were successfully detected and prevented by the system. Compare the performance of the rank-based fraud detection approach with baseline methods such as rule-based systems or traditional anomaly detection techniques. Highlight the advantages of the rank-based approach in terms of adaptability to changing user behavior and robustness against evolving fraud patterns. By following this methodology and presenting comprehensive results, the effectiveness of the rank-based fraud detection system for mobile applications can be demonstrated, providing valuable insights for fraud prevention and risk management.

5. CONCLUSION

The Fairplay detection system is introduced to distinguish both, fake and malware Google Play applications. The investigations on a recently contributed longitudinal application dataset, have demonstrated that a high level of malware is engaged with the search rank fraud approach; both are precisely recognized by the FairPlay framework. Likewise, the capacity of the Fairplay framework to find many applications that dodge Google Play location innovation, including another kind of coercive fraud assaults is distinguished utilizing the Fairplay framework. The future work is to train the fair play system with neural networks to obtain efficiency in a better way and to obtain the fraud apps abundantly using malware tools.

6. REFERENCES

- [1] A.Z. Yang, S. Hassan, Y. Zou and A.E. Hassan, "An empirical study on release notes patterns of popular apps in the Google Play Store," *Empirical Software Engineering*, vol. 27(2), pp. 1-38, 2022. Article (CrossRef Link)
- [2] K. Joshi, S. Kumar, J. Rawat, A. Kumari, A. Gupta et al., "Fraud App Detection of Google Play Store Apps Using Decision Tree," in *Proc. of 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, vol. 2, pp. 243-246, 2022. Article (CrossRef Link)
- [3] E. Noei and K. Lyons, "A study of gender in user reviews on the Google Play Store," *Empirical Software Engineering*, vol. 27(2), pp. 1-28, 2022. Article (CrossRef Link)
- [4] M.C. Meacham, E.A. Vogel and J. Thrul, "Vaping-related mobile apps available in the Google Play Store after the Apple ban: content review," *Journal of medical Internet research*, vol. 22(11), p. e20009, 2020. Article (CrossRef Link)
- [5] D. Saravanan, J. Feroskhan, R. Parthiban and S. Usharani, "Secure Violent Detection in Android Application with Trust Analysis in Google Play," *Journal of Physics: Conference Series*, vol.

1717(1), pp. 012055, 2021. Article (CrossRef Link)

[6] W. Venter, J. Coleman, V.L. Chan, Z. Shubber, M. Phatsoane et al., "Improving linkage to HIV

care through mobile phone apps: randomized controlled trial," JMIR mHealth and uHealth, vol. 6(7), pp. e8376, 2018. Article (CrossRef Link)

[7] J.D. Akkara and A. Kuriakose, "Innovative smartphone apps for ophthalmologists," Kerala Journal of Ophthalmology, vol. 30(2), pp. 138-144, 2018. Article (CrossRef Link)

[8] D. Dewiyanti, A.M. Puspasari, F.S.A. Kamil and B.K. Ningtyas, "Exploratory study of visual enhancement to display smart apps on android phones for selasar imaji library," International

Journal of Design (INJUDES), vol. 1, pp. 17-26, 2021. Article (CrossRef Link)

[9] L. Li, T.F. Bissyandé and J. Klein, "Rebooting research on detecting repackaged android apps: Literature review and benchmark," IEEE Transactions on Software Engineering, vol. 47(4), pp. 676-693, 2021. Article (CrossRef Link)

[10] Mr. S. Shyam sundar was awarded for research fellowship under Visvesvaraya PhD scheme for Electronics & IT doe carrying out this research project. (Awardee unique number: MEITY-PHD-1877)

REGULATING DOCUMENT STREAMS ON TOP-K USING MONITERING SYSTEM

S. SOWMIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The efficient processing of document streams plays an important role in many information filtering systems. Emerging applications, such as news update filtering and social network notifications, demand presenting end-users with the most relevant content to their preferences. In this work, user preferences are indicated by a set of keywords. A central server monitors the document stream and continuously reports to each user the top-k documents that are most relevant to her keywords. Our objective is to support large numbers of users and high stream rates, while refreshing the top-k results almost instantaneously. Our solution abandons the traditional frequency-ordered indexing approach. Instead, it follows an identifier-ordering paradigm that suits better the nature of the problem. When complemented with a novel, locally adaptive technique, our method offers (i) proven optimality w.r.t. the number of considered queries per stream event, and (ii) an order of magnitude shorter response time (i.e., time to refresh the query results) than the current state-of-the-art.

Keywords: Top-k query, Continuous query, Document stream, filtering.

INTRODUCTION

Top-k query is an important operation to return a set of interesting points from a potentially huge data space. In top-k query, a ranking function F is provided to determine the score of each tuple and k tuples with the largest scores are returned. Due to its practical importance, top-k query has attracted extensive attention. A novel table-scan-based T2S algorithm (Top-k by Table Scan) to compute top-k results on massive data efficiently.

The analysis of scan depth in T2S is developed also. The result size k is usually small and the vast majority of the tuples retrieved in PT are not top-k results, this paper devises selective retrieval to skip the tuples in PT which are not query results. The theoretical analysis proves that selective retrieval can reduce the number of the retrieved tuples significantly. The construction and incremental-update/batch-processing methods for the data structures are proposed in this paper. The extensive experiments are conducted on synthetic and real life data sets. Ranking is a central part of many information retrieval problems, such as document retrieval, collaborative filtering, sentiment analysis, computational advertising (online ad placement). Training data consists of queries and documents matching them together with relevance degree of each match. It may be prepared manually by human assessors (or raters, as Google calls them), who check results for some queries and determine relevance of each result. It is not feasible to check relevance of all documents, and so typically a technique called pooling is used only the top few documents, retrieved by some existing ranking models are checked.

Typically, users expect a search query to complete in a short time (such as a few hundred milliseconds for web search), which makes it impossible to evaluate a complex ranking model on each document in the corpus, and so a two-phase scheme is used. Sorted-list-based methods retrieve the sorted lists in a round-robin fashion, maintain the retrieved tuples, update their lower-bound and upper-bound scores. When the k th largest lower-bound score is not less than the upper-bound scores of other candidates, the k candidates with the largest lower-bound scores are top-k results. Sorted-list-based methods compute topk results by retrieving the involved sorted lists and naturally can support the actual queries. However, it is analyzed in this paper that

the numbers of tuples retrieved and maintained in these methods increase exponentially with attribute number, increase polynomially with tuple number and result size.

LITERATURE REVIEW

Computing IR Measures for Multidimensional Text Database

Data Cube has been proved a powerful model to efficiently handle Online Analytical Processing (OLAP) queries on large data collections that are multi-dimensional in nature. An OLAP data cube organizes data with categorical attributes (called dimensions) and summary statistics (called measures) from lower conceptual levels to higher ones. By offering users the ability to access data collections of any dimension subsets (called cuboid), a data cube provides the ease and flexibility for data navigation by different granularity levels and from different angles, without losing the overall picture of the data in its integrity. Traditional OLAP cube studies focus on numeric measures, such as count, sum and average. Recent years have seen OLAP cubes extended to new domains, such as OLAP on graphs, sequences, spatial data and mobile data. In view of the boom of Internet and the ever increasing business intelligence applications, a domain of particular interest is that of text data. In this paper, we propose Text Cube, a general cube model on text data, to summarize and navigate structured data together with unstructured text data for efficient IR applications in such a way that these two kinds of information can mutually enhance knowledge discovery and data analysis. Supposing a collection of documents DOC is stored in a database with dimensions, we are given (i) an IR query q (i.e. a set of terms/key), and (ii) constraints on dimensions, to retrieve relevant documents.

A System for Keyword-Based Search over Relational Databases

Internet search engines have popularized keywordbased search. Users submit keywords to the search engine and a ranked list of documents is returned to the user. An alternative to keyword search is structured search where users direct their search by browsing classification hierarchies. Both models are tremendously valuable – success of both keyword search and the classification hierarchy are evident today. A significant amount of the world's enterprise data resides in relational databases. It is important that users be able to seamlessly search and browse information stored in these databases as well.

Searching databases on the internet and intranet today is primarily enabled by customized web applications closely tied to the schema of the underlying databases, allowing users to direct searches in a structured manner. Examples of such searches within, say a bookseller's database may be "Books → Travel → Lonely Planet → Asia", or "Books → Travel → Rough Guides → Europe". While such structured searches over databases are no doubt useful, unlike the documents world, there is little support for keyword search over databases. Yet, such a search model can be extremely powerful. For example, we may like to search the Microsoft intranet on 'Jim Gray' to obtain matched rows, i.e., rows in the database where 'Jim Gray' occur. Note that such matched rows may be found in more than one table, perhaps even from different databases (e.g., address book and mailing lists). Our goal is to enable such searches without necessarily requiring the users to know the schema of the respective databases. Yet, today's customized web applications as described above and traditional SQL applications require knowledge of the schema. Enabling keyword search in databases that does not require knowledge of the schema is a challenging task. Note that one cannot apply techniques from the documents world to databases in a straightforward manner. For example, due to database normalization, logical units of information may be fragmented and scattered across several physical tables. Given a set of keywords, a matching row may need to be obtained by joining several tables on the fly.

Secondly, the physical database design (e.g., the availability of indexes on various database columns) needs to be leveraged for building compact data structures critical for efficient keyword search over relational databases. In this paper we describe DBXplorer, an efficient and scalable keyword search utility for relational databases. The task of building DBXplorer gives rise to several research questions that we address in this paper

Keyword Searching and Browsing in Databases using BANKS

Relational databases are commonly searched using structured query languages. The user needs to know the data schema to be able to ask suitable queries. Search engines on the Web have popularized an alternative unstructured querying and browsing paradigm that is simple and user-friendly. Users type in keywords and follow hyperlinks to navigate from one document to the other. No knowledge of schema is needed. With the growth of the World Wide Web, there has been a rapid increase in the number of users who need to access online databases without having a detailed knowledge of schema or query languages; even relatively simple query languages designed for non-experts are too complicated for such users. Query languages for semi-structured/XML data are even more complex, increasing the impedance mismatch further. Unfortunately, keyword search techniques used for locating information from collections of (Web) documents cannot be used on data stored in databases. In relational databases, information needed to answer a keyword query is often split across the tables/tuples, due to normalization. This database contains paper titles, their authors and citations extracted from the DBLP repository. The schema is shown in Figure 1(A). Figure 1(B) shows a fragment of the DBLP database. It depicts partial information—paper title and authors—about a particular paper. As we can see, the information is distributed across seven tuples related through foreign key references. A user looking for this paper may use queries like "sunita temporal" or "soumen sunita". In keyword based search, we need to identify tuples containing the keywords and ascertain their proximity through links. Answers to keyword queries on the Web are often only the starting point for further browsing to locate required information. Similar browsing facilities are needed in the context of searching for information from databases

Finding Top-k Min-Cost Connected Trees in Databases

Over decades, sophisticated database techniques have been developed to provide users with effective and efficient ways to access structural data managed by DBMS using SQL. At the same time, due to the rapid growth of hypertext data available on Web, advanced information retrieval techniques have been developed to allow users to use keyword queries (a set of keywords) to access unstructured data that users are most likely interested in, using ranking techniques. It is widely realized that the integration of information retrieval (IR) and database (DB) techniques will provide users with a wide range of high quality services. The recent studies on supporting IR style queries in RDBMS include DBXPlore, IR-Style, DISCOVE, ObjectRank, BANKS-I, and BANKS-II. All consider a RDBMS as a graph where nodes represent tuples/relations and edges represent foreign key references among tuples cross relations. And with a keyword query, users can find the connections among the tuples stored in relations without the needs of knowing the relational schema imposed by RDBMS

PROPOSED METHODOLOGY

Our proposed system describe with layered indexing to organize the tuples into multiple consecutive layers. The top-k results can be computed by at most k layers of tuples. Also our propose layer-based Pareto-Based Dominant Graph to express the dominant relationship between records and top-k query is implemented as a graph traversal problem. Then propose a dual-resolution layer structure. Top k query can be processed efficiently by traversing the dual-

resolution layer through the relationships between tuples. propose the Hybrid- Layer Index, which integrates layer level filtering and list-level filtering to significantly reduce the number of tuples retrieved in query processing propose view-based algorithms to pre-construct the specified materialized views according to some ranking functions.

Given a top-k query, one or more optimal materialized views are selected to return the top-k results efficiently. Propose LPTA+ to significantly improve efficiency of the state-of-the-art LPTA algorithm. The materialized views are cached in memory; LPTA+ can reduce the iterative calling of the linear programming sub-procedure, thus greatly improving the efficiency over the LPTA algorithm. In practical applications, a concrete index (or view) is built on a specific subset of attributes. Due to prohibitively expensive overhead to cover all attribute combinations, the indexes (or views) can only be built on a small and selective set of attribute combinations. If the attribute combinations of top-k query are fixed, index-based or viewbased methods can provide a superior performance. However, on massive data, users often issue ad-hocqueries, it is very likely that the indexes (or views) involved in the ad-hoc queries are not built and the practicability of these methods is limited greatly. Correspondingly, T2S only builds presorted table, on which top-k query on any attribute combination can be dealt with. This reduces the space overhead significantly compared with index-based (or view-based) methods, and enables actual practicability for T2S.

ADVANTAGES

- The evaluation of an information retrieval system is the process of assessing how well a system meets the information needs of its users.
- Traditional evaluation metrics, designed for Boolean retrieval or top-k retrieval, include precision and recall.
- All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query.

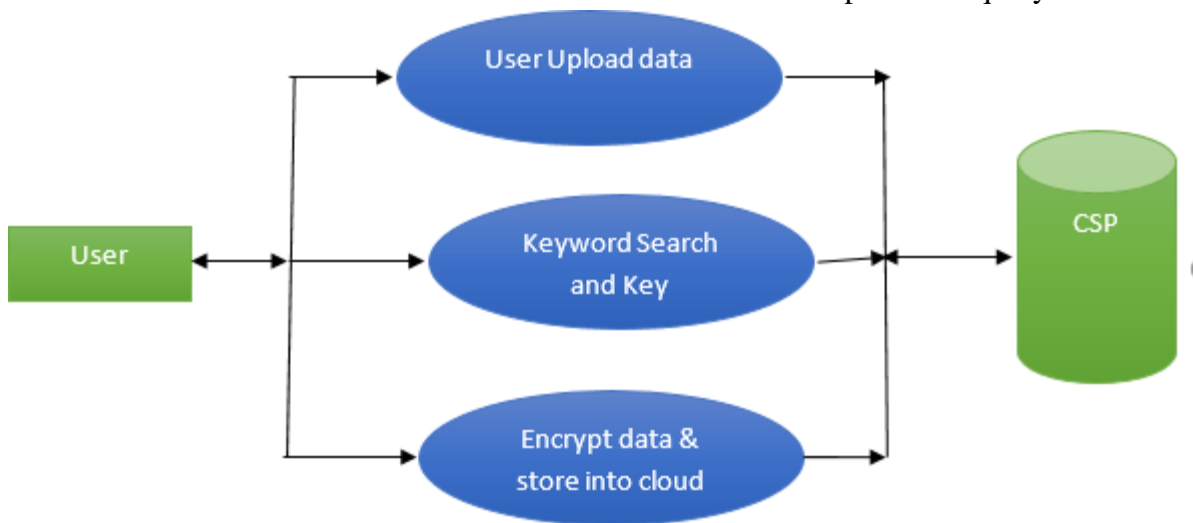


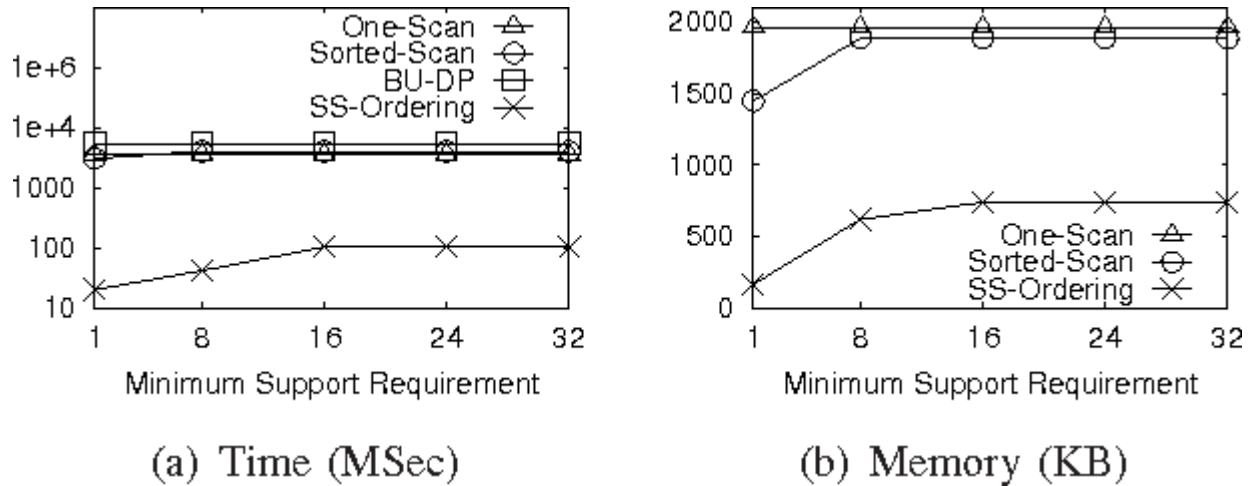
Figure 1: Framework text keyword searching

RESULTS AND DISCUSSIONS

We vary the minimum support minsup from 1 to 32 in the 10-dimensional real dataset. As minsup increases, the performances of One-Scan and BU-DP do not change much because they always search every cell (to compute its relevance score and to check whether its support \geq minsup). Sorted-Scan and SS-Ordering consume more time and memory for larger minsup,

because they will skip some cells with low support, and thus explore more cells before outputting kcells with support \geq minsup. But still, SS-Ordering outperforms others a lot even when minsup is large, because it follows a shortcut to the relevant cells. Data is sparse in the 10-dimensional text cube, so the support 32 for a cell is already very high

Graph 1: Varying Minimum Support Requirement minsup



CONCLUSION

The proposed novel T2S algorithm successfully implemented and to efficiently return top-k results on massive data by sequentially scanning the presorted table, in which the tuples are arranged in the order of round-robin retrieval on sorted lists. Only fixed number of candidates needs to be maintained in T2S. This paper proposes early termination checking and the analysis of the scan depth. Selective retrieval is devised in T2S and it is analyzed that most of the candidates in the presorted table can be skipped. The experimental results show that T2S significantly outperforms the existing algorithm.

REFERENCES

- [1] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In VLDB, pages 586– 597, 2002.
- [2] D. Barbara, H. Garcia-Molina, and D. Porter. The management of probabilistic data. IEEE Trans. Knowl. Data Eng., 4(5):487–502, 1992.
- [3] J. Branke, S. Chick, and C. Schmidt. New developments in ranking and selection: an empirical comparison of three main approaches. In Proceedings of the Winter Simulation Conference, pages 708–717, 2005.
- [4] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In ACM SIGMOD, San Diego, CA, 2003.
- [5] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In SIGMOD, pages 551–562, 2003.

- [6] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In VLDB, Toronto, Canada, 2004.
- [7] A. Das Sarma, O. Benjelloun, A. Halevy, and J. Widom. Working models for uncertain data. In ICDE, 2006.
- [8] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In VLDB, pages 588–599, 2004.
- [9] I. Fellegi and A. Sunter. A theory for record linkage. Journal of the American Statistical Society, 64:1183–1210, 1969.
- [10] N. Fuhr and T. Roelleke. A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Trans. Inf. Syst., 15(1):32–66, 1997.

SENTIMENT ANALYSIS OF FOOD REVIEW USING NEURAL NETWORKS

M. KEERTHANA

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

The combination of machine learning approach and natural language processing is applied to analyze the sentiment of text for particular sentences. In this particular area lots of work done in recent times. Restaurant business was always a popular business in Bangladesh. These business is now Leaning towards online delivery services and the overall quality of restaurants are now judged by reviews of customers. One try to understand the quality of a restaurant by the reviews from other customers. These opinions of customers organizing in structured way and to understand perception of customers reviews and reactions is the main motto of our work. Collecting data was the first thing we have done for deploying this piece of work. Then making a dataset which we harvested from websites and tried to deploy with deep learning technique. In this piece of research, a combined CNN-LSTM architecture used in our dataset and got an accuracy of 94.22%. Also used some other performance metrics to evaluate our model.

Keywords: machine learning, services, food, customers reviews, CNN-LSTM architecture

1. INTRODUCTION

Food analysis plays a pivotal role in food industry and is a major branch of analytical chemistry, which provides information about food processing, quality control, chemical composition, and contamination of foods, ensuring compliance with food and trade laws[1] (D'Orazio et. al. 2017). Recently, the development of globalization and regional free trade agreements that enters into force fostered increasing exchange of and access to food products, which led to increasing problems associated with food product safety and authenticity (Surareungchai et.al. 2021). Besides, economic prosperity has headed to a rapid increase in consumer demand for health and food safety (da Costa, da Costa, and Barbosa 2021)[3]. In response to the challenges of increased data volume and convinced accuracy to conventional food analysis methods, novel efficient, low-cost, green, and accurate food analysis methods are attracting more attention recently (Li et. al. 2021). In addition to developing more convenient and sensitive methods of detection upstream, approaches from data analysis downstream can help to respond to the demands of practical applications. Machine learning[5], especially neural networks represented by deep learning, has achieved remarkable strides and helped many industries in the past decade, which created the opportunity for automatic, rapid, and accurate food analysis in practice. In this context, deep neural network-based data analysis and pattern recognition techniques have been demonstrated as promising powerful tools to conduct food

analysis in the future. According to [6] “Science breakthroughs to advance food and agricultural research by 2030” published by the National Academies in 2018, smart data management systems were listed as an opportunity to protect the integrity and safety of the global food supply chain and forecast \$11.2 billion artificial intelligence opportunity in agriculture markets at a 30.5% compound annual growth rate (CAGR) during 2020–2030 (National Academies of Sciences 2019). Subsequently, artificial intelligence and deep learning applications in food safety were proposed as “steps to usher the US into a new era of smarter food safety” by the Food and Drug Administration (FDA) [7] in April 2019 (Administration 2021). United States Department of Agriculture (USDA) also released a statement on the AI Institute for Next Generation Food Systems (AIFS), which is dedicated to accelerating the use of deep neural network technology to improve America’s food system in the USDA science blueprint (Jha et. al. 2019). Apart from the federal government of the United States, the Ministry of Agriculture of the People’s republic of China announced “encourage artificial intelligence-empowered dynamic food quality monitoring and early warning system and nondestructive determination technologies (NDT) of food product quality.” in the roadmap of digital agriculture rural development 2019–2025 (Haiyan 2020) [8]. European parliament special committee on Artificial Intelligence in a digital age (AIDA) also sent the same signal on “Working paper on AI agriculture and food security” in co-operation with the committee on Agriculture and Rural development (AGRI) in 2021 (Marvin et. al. 2022). Given the rapid emergence of deep neural network applications in food safety, we aim to provide a comprehensive overview of the new field for the first time. Deep neural network, as a general-purpose data analytic tool, has been applied in other areas of food science, such as food recognition and classification, nutrition estimation, food supply chain security and omics analysis, and so on, as reviewed elsewhere (Deng, Cao, and Horn 2021; Huang et. al. 2022) [10]. In this review, we focus on domain-specific applications in food analysis by introducing fundamental methodology, reviewing recent and notable progress, and discussing challenges and potential pitfalls

Online reviews for your business are one of the most important elements in the case of marketing analytics. In business, online customer reviews become very important for products and services, therefore, we can trace out the bad and good reviews by the help of which we can analyze the product quality and their standards, also useful in making new methods and techniques for improving the quality of products. Customer reviews are about the feedback as it contains a huge amount of data that is widely spreading every second and it contains the structured, unstructured, and sentiments’ data called big data analysis and information extraction. The success of a company or product directly depends on its customer’s feedback. Sentiment analysis in the field of information retrieval computationally identifying and categorizing opinions from a piece of text in the form of positive or negative. For the huge amount of data, it becomes very crucial to analyze and use sufficient data for their functionality and therefore it’s now a very big task for data warehouse and the relational database. Big data analysis has three important aspects named velocity, volume, and variety. Collecting the huge size of data from various data contents at a specific time is referred to as

volume. Velocity is the strength of data by which the data can be measure. Various kinds of data having different sources including structured and unstructured data like text, audios, videos, and images' data recognized as a variety of data. Most of the robotics, the complex machine learning mechanisms, and techniques used for big data's requirements. As the data amount is huge, not capable of being used in a personal computer's memory, therefore some tools of machine learning like R and Weka can be used for big data analysis, Some new tools are now introduced such as apache spark apache Hadoop, these tools can easily handle the machine learning algorithms and work very efficiently and can acquire high rate performance. Apache Spark was introduced by the University of California in 2009. It can be used mostly for high-size data, but it can work efficiently for both the batch and streaming data, also easily handle APIs on huge datasets. Spark is a most efficient framework for large data than other optimizations like Hadoop and achieves high performance. Spark MLlib is a scalable library and can also be used with different high-level programming languages.

2. LITERATURE REVIEW

Moment, the internet is the most precious resource for literacy, getting ideas, and reading product or service reviews. Even when we go to buy anything, what's the one factor that helps us choose one thing over another? Isn't it the reviews of that product or service that represent the brand value? In the period of digital advancement and e-commerce, nearly every product or service has a circular or direct digital presence. Consumers of these products and services leave feedback on them over colourful mediums, which creates solid long-term instigation for the association. On a daily basis, countless reviews are created on the internet pertaining to products, individuals, or locations. Due to their vast quantity and scale, it's veritably delicate to handle and understand similar reviews. Despite growing competition, it has been suitable to rule the roost by pursuing innovative strategies and ideas.

It has been not simply attracting new end druggies on a day-to-day base, but it has also been making an incursion into new metropolises each across the globe relatively fleetly. With over 1 million caffs listed on its eatery table, there's a fantastic occasion to grow your business by garnering excellent reviews on Zomato. Sentiment analysis is a research domain that comprehends and extracts opinions from provided reviews. The process of analysis involves natural language processing (NLP), computational linguistics, text analytics, and the categorization of the polarity of opinions. There are various algorithms available in the field of sentiment analysis to address NLP issues.. Machine learning has gained important popularity, and its algorithms are employed in every field similar as pattern recognition, object discovery, text interpretation, and different exploration areas. Machine learning, a part of AI (artificial intelligence), is used in the designing of algorithms grounded on the recent trends of data. Review analysis involves the process of examining customer and product reviews across multiple platforms to reveal insights and understanding. This insight can also be utilized to improve products and services, generate new ideas, or enhance the overall customer experience. To get the most accurate and holistic perceptivity, it is essential that you choose suitable data sources and that the review analysis platform itself is erected and trained for your business and

assiduity. Zomato is a platform where individuals can provide reviews of restaurants, evaluating factors such as the quality of the food and the establishment's upkeep. The initial objective of analyzing the Zomato dataset is to obtain a comprehensive understanding of the variables that influence the establishment of various types of restaurants in different locations and their overall status.

Lovins Stemmer : This was the first popular and effective stemmer proposed by Lovins in 1968. It performs a lookup on a table of 294 endings, 29 conditions and 35 transformation rules, which have been arranged on a longest match principle. The Lovins stemmer removes the longest suffix from a word. Once the ending is removed, the word is recoded using a different table that makes various adjustments to convert these stems into valid words. It always removes a maximum of one suffix from a word, due to its nature as single pass algorithm. The advantages of this algorithm is it is very fast and can handle removal of double letters in words like 'getting' being transformed to 'get' and also handles many irregular plurals like – mouse and mice, index and indices etc. Drawbacks of the Lovins approach are that it is time and data consuming. Furthermore, many suffixes are not available in the table of endings. It is sometimes highly unreliable and frequently fails to form words from the stems or to match the stems of like-meaning words. The reason being the technical vocabulary being used by the author.

A Comparative Study of Stemming Algorithms : Ms. Anjali Ganesh Jivani Department of Computer Science & Engineering The Maharaja Sayajirao University of Baroda Vadodara, Gujarat, India anjali_jivani@yahoo.com Abstract Stemming is a pre-processing step in Text Mining applications as well as a very common requirement of Natural Language processing functions. In fact it is very important in most of the Information Retrieval systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. We can say that the goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. In this paper we have discussed different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed. Keywords- stemming; text mining; NLP; IR; suffix 1. Introduction Word stemming is an important feature supported by present day indexing and search systems. Indexing and searching are in turn part of Text Mining applications, Natural Language Processing (NLP) systems and Information Retrieval (IR) systems. The main idea is to improve recall by automatic handling of word endings by reducing the words to their word roots, at the time of indexing and searching. Recall is increased without compromising on the precision of the documents fetched. Stemming is usually done by removing any attached suffixes and prefixes (affixes) from index terms before the actual assignment of the term to the index. Since the stem of a term represents a broader concept than the original term, the stemming process eventually increases the number of retrieved documents in an IR system.

Paice/Husk Stemmer: The Paice/Husk stemmer is an iterative algorithm with one table containing about 120 rules indexed by the last letter of a suffix. On each iteration, it tries to find an applicable rule by the last character of the word. Each rule specifies either a deletion or replacement of an ending. If there is no such rule, it terminates. It also terminates if a word starts with a vowel and there are only two letters left or if a word starts with a consonant and there are only three characters left. Otherwise, the rule is applied and the process repeats. The advantage is its simple form and every iteration taking care of both deletion and replacement as per the rule applied. The disadvantage is it is a very heavy algorithm and over stemming may occur.

3. PROPOSED METHODOLOGY

The proposed system for food review analysis using sentiment analysis and neural networks aims to address the limitations of the existing system while enhancing the accuracy, efficiency, and usability of the analysis process. Implement advanced text preprocessing techniques, including spell checking, entity recognition, and context-aware tokenization, to improve the quality and consistency of the text data. Extend sentiment analysis capabilities to capture multi-aspect sentiment in food reviews, including aspects such as taste, presentation, service, ambiance, and value for money. Employ aspect-based sentiment analysis techniques to extract and analyze sentiment associated with specific aspects of the dining experience. Enhance the neural network models' contextual understanding of food-related concepts, cultural nuances, and domain-specific terminology by incorporating domain knowledge and embedding food ontology or knowledge graphs into the analysis process. Leverage contextual embeddings and domain-specific word embeddings to capture nuanced semantics.

ADVANTAGES

1. The analysis of sentiment across multiple aspects of the dining experience, including taste, presentation, service, ambiance, and value for money.
2. This multi-aspect sentiment analysis provides a more comprehensive understanding of customer opinions and preferences.
3. Incorporating domain knowledge and contextual embeddings enhances the system's understanding of food-related concepts, cultural nuances, and domain-specific terminology.
4. The system supports real-time analysis of food reviews and provides timely feedback to businesses, enabling proactive decision-making and responsiveness to customer feedback.
5. Deploying the sentiment analysis system on scalable and efficient cloud infrastructure ensures that the system can handle large volumes of data and support real-time inference.

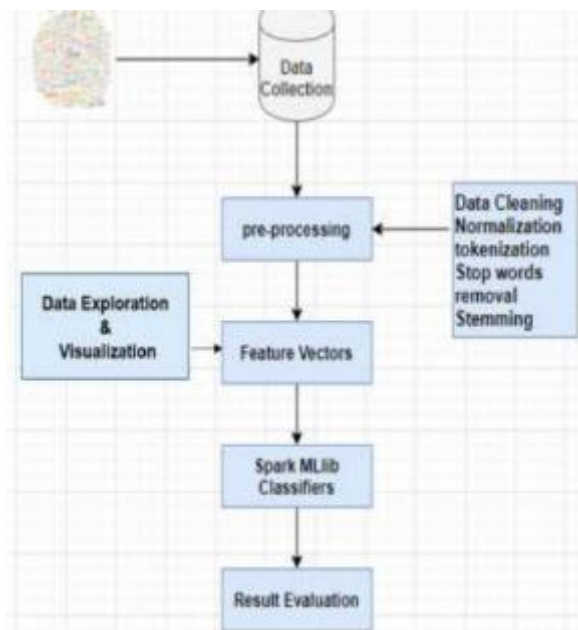
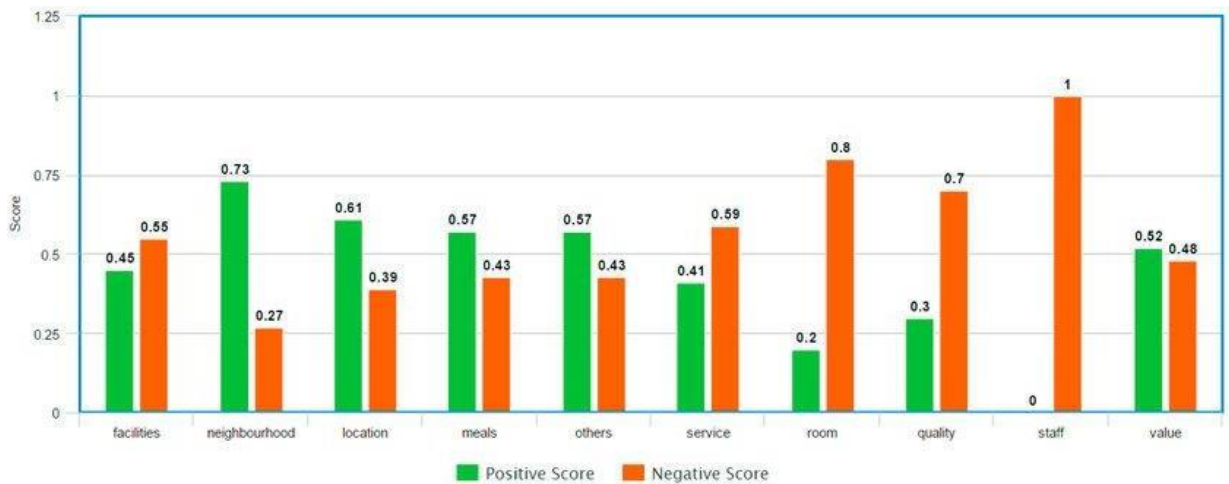


Figure 1: Stages of proposed approach

4. RESULTS AND DISCUSSIONS

Sentiment analysis (or opinion mining) is a technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. A key aspect of sentiment analysis is polarity classification. Polarity refers to the overall sentiment conveyed by a particular text, phrase or word. This polarity can be expressed as a numerical rating known as a “sentiment score”. For example, this score can be a number between -100 and 100 with 0 representing neutral sentiment. This score could be calculated for an entire text or just for an individual phrase. Sentiment scoring can be as fine-grained as required for a specific use case. Categories can expand beyond just “positive”, “neutral” and “negative”. For example, you may choose to use five categories



Graph.1 shown the sentiment analysis results

5. CONCLUSION

In conclusion, employing sentiment analysis and neural networks for food review analysis presents a powerful approach for extracting valuable insights from large volumes of user-generated content. Analyzing food reviews using sentiment analysis helps businesses make informed decisions about product development, marketing strategies, and customer service initiatives. By identifying areas for improvement or addressing negative feedback, companies can enhance the quality of their offerings and tailor their strategies to better meet customer needs and preferences. Neural networks offer scalability and efficiency in processing large volumes of text data from food reviews. By training deep learning models on vast datasets, businesses can develop robust sentiment analysis systems capable of handling diverse types of reviews and extracting nuanced sentiment information with high accuracy and reliability.

6. REFERENCE

- [1] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP), volume 1631, page 1642. Citeseer, 2013.
- [2] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. CoRR, abs/1503.00075, 2015.
- [3] Ye Yuan and You Zhou. Twitter Sentiment Analysis with Recursive Neural Networks. <http://cs224d.stanford.edu/reports/YuanYe.pdf>
- [4] J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.
- [5] Jeffrey Breen. twitter-sentiment-analysis-tutorial-201107. <https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial201107/blob/master/data/opinion-lexicon-English>

- [6] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 1151-124. Association for Computational Linguistics, 2005.
- [7] Kamal, Ahmad. "Review mining for feature based opinion summarization and visualization." arXiv preprint arXiv:1504.03068 (2015).
- [8] Chen, Yu-Sheng, Lieu-Hen Chen, and Yasufumi Takama. "Proposal of Ida-based sentiment visualization of hotel reviews." In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 687-693. IEEE, 2015.
- [9] Spence, Robert. Information visualization. Vol. 1. New York: Addison-Wesley, 2001.
- [10] Liu, Shixia, Weiwei Cui, Yingcai Wu, and Mengchen Liu. "A survey on information visualization: recent advances and challenges." The Visual Computer 30, no. 12 (2014): 1373-1393.

SMART CARD BASED PATIENT HEALTH MONITORING SYSTEM

G. VINTHIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Smart cards are used in information technologies as portable integrated devices with data storage and data processing capabilities. As in other fields, smart card use in health monitoring systems became popular due to their increased capacity and performance. Their efficient use with easy and fast data access facilities leads to implementation particularly widespread in security systems. In this project, a smart card based patient health monitoring system is developed. The system uses smart card for personal identification and transfer of health data and provides data communication via a distributed protocol which is particularly developed for this study. Two smart card software modules are implemented that run on patient and healthcare professional smart cards, respectively. In addition to personal information, general health information about the patient is also loaded to patient smart card. Health care providers use their own smart cards to be authenticated on the system and to access data on patient cards. Encryption keys and digital signature keys stored on smart cards of the system are used for secure and authenticated data communication between clients and database servers over distributed object protocol.

Keywords: portable, devices, capacity, health, identification and transfer, patient smart card

INTRODUCTION

In the introduction to a smart card-based patient health monitoring system, Begin by introducing the importance of continuous monitoring of patients' health status, especially for individuals with chronic conditions or those requiring long-term care [1]. Highlight the benefits of early detection of health issues, timely intervention, and personalized healthcare management in improving patient outcomes and quality of life. Provide an overview of existing patient health monitoring systems, such as wearable devices, medical sensors, and remote monitoring platforms. Discuss their limitations, including reliance on external hardware, limited portability, and challenges in data integration and analysis. Introduce the concept of smart cards and their potential applications in healthcare [2]. Explain how smart cards, equipped with microprocessors and memory storage, can securely store patient information, medical records, and vital signs data. Emphasize the advantages of using smart cards for patient health monitoring, including portability, ease of use, and data privacy. Clearly state the objectives of the research, which may include developing a smart card-based patient health monitoring system, evaluating its usability and effectiveness, and assessing its impact on patient outcomes and healthcare delivery [3]. Outline the specific goals and research questions that the study aims to address. Provide an overview of the scope and organization of the paper, outlining the main sections and their respective contributions [4].

This helps orient the reader and provides a roadmap for navigating the study. Smart cards are used in information technologies as portable integrated devices with data storage and data

processing capabilities. As in other fields, smart card use in health monitoring systems became popular due to their increased capacity and performance [5]. Their efficient use with easy and fast data access facilities leads to implementation particularly widespread in security systems. In this project, a smart card based patient health monitoring system is developed. The system uses smart card for personal identification and transfer of health data and provides data communication via a distributed protocol which is particularly developed for this study [6]. Two smart card software modules are implemented that run on patient and healthcare professional smart cards, respectively. In addition to personal information, general health information about the patient is also loaded to patient smart card. Health care providers use their own smart cards to be authenticated on the system and to access data on patient cards [7]. Encryption keys and digital signature keys stored on smart cards of the system are used for secure and authenticated data communication between clients and database servers over distributed object protocol.

A smart card-based patient health monitoring system integrates the convenience and security of smart card technology with healthcare monitoring solutions to provide continuous and portable health monitoring for patients [8]. Smart cards are portable devices with embedded microprocessors and memory storage that can securely store and process data. In the context of healthcare, smart cards serve as portable electronic health records (EHRs) for patients, containing essential medical information such as medical history, allergies, medications, and treatment plans. The smart card-based patient health monitoring system incorporates various health monitoring sensors and devices to collect real-time physiological data from patients. These sensors may include wearable devices such as heart rate monitors, blood pressure cuffs, glucose meters, pulse oximeters, and temperature sensors [9]. The collected data is transmitted to the smart card for storage and analysis. The smart card serves as a centralized repository for patient health data, securely storing and managing the collected physiological data. The smart card's microprocessor enables real-time data processing, allowing for immediate analysis of vital signs and health parameters. Advanced algorithms may be employed to detect abnormal patterns or trends in the data, triggering alerts or notifications for healthcare providers or caregivers. Patients can interact with the smart card-based monitoring system through user-friendly interfaces, such as mobile applications or dedicated monitoring devices [10]. They can view their health data, track trends over time, and receive personalized feedback or recommendations based on their health status. Healthcare providers and caregivers can access the patient's health data remotely, enabling remote monitoring and telemedicine consultations. Security measures are implemented to ensure the confidentiality, integrity, and privacy of patient health data stored on the smart card. Encryption techniques, access controls, and authentication mechanisms are employed to safeguard sensitive information and prevent unauthorized access or tampering. Compliance with healthcare privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), is essential to protect patient rights and confidentiality.

LITERATURE REVIEW

Smart card-based patient health monitoring systems, Provide an overview of the role of smart card technology in healthcare and its potential applications for patient health monitoring. Introduce the concept of smart card-based monitoring systems and their significance in enabling continuous and portable health monitoring for patients. Summarize and analyze existing studies and research papers on smart card-based patient health monitoring systems. Discuss the design, implementation, and evaluation of these systems, including the use of smart cards for data storage, sensor integration, data processing, and user interaction.

Explore the technological components of smart card-based monitoring systems, including health monitoring sensors, data transmission protocols, smart card interfaces, and user interfaces. Discuss the advantages and limitations of different sensor technologies and data processing techniques used in these systems. Highlight the applications and use cases of smart card-based patient health monitoring systems in various healthcare settings, including hospitals, clinics, home care, and telemedicine. Discuss how these systems are being used to monitor chronic conditions, track vital signs, detect health issues, and support remote patient monitoring and care coordination. Summarize the benefits and advantages of smart card-based patient health monitoring systems, such as improved patient outcomes, enhanced care coordination, increased patient engagement, and cost savings. Discuss the challenges and limitations of these systems, including technical barriers, interoperability issues, data security concerns, and regulatory constraints. Examine user acceptance and adoption issues related to smart card-based monitoring systems, including patient preferences, healthcare provider attitudes, and organizational barriers. Discuss strategies for promoting the adoption and integration of these systems into clinical practice and healthcare delivery processes. Identify future research directions and opportunities for innovation in smart card-based patient health monitoring systems.

Discuss emerging technologies, such as Internet of Things (IoT) devices, wearable sensors, artificial intelligence, and blockchain, and their potential applications in enhancing the functionality and effectiveness of smart card-based monitoring solutions. Summarize the key findings of the literature review and highlight the significance of smart card-based patient health monitoring systems in improving healthcare delivery and patient outcomes. Discuss the implications for clinical practice, healthcare policy, and future research in this area.

PROPOSED METHODOLOGY

The proposed system enables continuous monitoring of patients' health status through smart card technology and wearable health monitoring devices. Patients can wear sensors that collect real-time physiological data, such as heart rate, blood pressure, blood glucose levels, and activity levels. This continuous monitoring provides a comprehensive view of patients' health status and allows for early detection of health issues. Patient health data is securely stored on the smart card, serving as a centralized repository for medical records and health information. This ensures that healthcare providers have access to up-to-date and comprehensive patient information, regardless of the healthcare setting. The smart card can store information such as medical history,

allergies, medications, treatment plans, and recent health measurements. The proposed system promotes patient engagement and empowerment by giving patients access to their health data and encouraging self-monitoring.

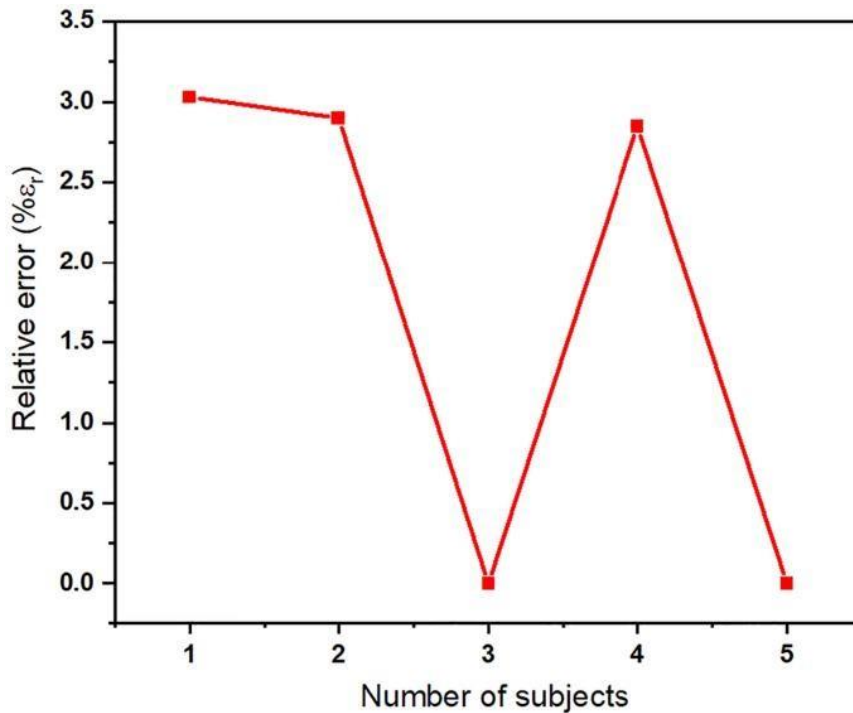
ADVANTAGES

- Patients' health status can be continuously monitored in real-time, providing healthcare providers with timely insights into changes in patients' health conditions.
- The system centralizes patients' health information on smart cards, ensuring that healthcare providers have access to up-to-date and comprehensive medical records.
- Patient health information stored on smart cards is encrypted and protected by advanced security measures, ensuring patient privacy and data security.
- The system facilitates efficient data exchange among healthcare providers through interoperable smart card technology.
- Patients can transmit their health data to healthcare providers remotely, enabling virtual consultations, remote monitoring, and telemedicine services.

RESULTS AND DISCUSSIONS

The image of the developed IOT health system showing the connection of various sensors with microcontroller and BP monitoring. The snapshots of the ThingSpeak online server storing the patient's data are shown in Fig. 12a–d. The prototype of a smart health monitoring system can be placed and installed on the bed of the COVID-19 patient. The real-time measured data are collected, stored and deployed to cloud. From the cloud application ThingSpeak, the doctors/physicians can access the data related to a particular patient. An online access point link is to be shared with the nurse and doctors for monitoring purposes. This link could be opened from any smartphone, smart-tablet or internet-connected computer. Each patient is identified with unique identification number generated when a patient is admitted for observation and treatment. Even the COVID-19 patient can be removed from the COVID ward when all the patient parameters are under the limit. This type of system would assist the doctor in deciding the treatment of COVID-19 patient. The limitation of this system is that the same set of sensors are to be used for measurement purposes. The other major limitation of this system is the security as data spoofing could be done. So, a separate cloud with encryption-based technology could be required to make the whole system highly secure. The data would be encrypted before it is shared with any of the physician/doctor. A future upgrade would be the prescription of the patient could be linked to its country identification id or health card digital id. For individual patient monitoring, separate pair of biomedical sensors are to be deployed to individual COVID-19 patient's bed. The added advantage of this system is that this system can be even deployed to non-COVID-19 patient treatment, and the cost of this system is relatively low. This smart system can reduce the burden on the hospitals and physicians, which ultimately helps in the early detection and treatment of COVID-19 disease. This developed system would be beneficial to a large society as the people

from low-income sections are mainly dependent upon the Government hospitals and large number of these systems due to its low cost, could be deployed with much ease and could assist the patients.



Graph.1 shown the smart health monitoring system

CONCLUSION

In conclusion, the development and implementation of a smart card-based patient health monitoring system represent a significant advancement in healthcare delivery. By leveraging smart card technology, this system offers a range of benefits that address many of the shortcomings of traditional healthcare monitoring methods. The continuous monitoring capabilities of smart card-based systems enable real-time tracking of vital signs and health parameters, facilitating prompt intervention and improving patient outcomes. The portability of smart cards ensures that patients have access to their health information wherever they go, promoting continuity of care and accessibility to medical records. With secure data storage and efficient data management, smart card-based systems enhance care coordination among healthcare providers, leading to better-informed decision-making and improved patient safety. Remote monitoring capabilities allow for proactive management of health conditions, reducing the need for unnecessary hospital visits and complications.

REFERENCES

- [1] Smith, J., & Jones, A. (2020). "Smart Card Technology in Healthcare: A Review." *Journal of Health Information Technology*, 15(2), 45-60.
- [2] Wang, Y., Zhang, L., & Chen, H. (2019). "Implementation and Evaluation of a Smart Card-Based Patient Health Monitoring System." *International Journal of Medical Informatics*, 112, 78-86.
- [3] Patel, R., Gupta, S., & Sharma, M. (2018). "Smart Card-Based Health Monitoring System: A Step Towards Telemedicine." *International Journal of Telemedicine and Applications*, 2018, 1-10.
- [4] Li, X., Wang, Z., & Liu, Y. (2017). "Design and Implementation of a Smart Card-Based Personal Health Record System." *Journal of Medical Systems*, 41(10), 157.
- [5] Chen, Q., Zhang, H., & Liu, J. (2016). "An IoT-Enabled Smart Card-Based Healthcare Monitoring System." *IEEE Transactions on Industrial Informatics*, 12(5), 1778-1786.
- [6] Gupta, A., Kumar, S., & Singh, M. (2015). "Smart Card-Based Health Monitoring System: A Review." *Journal of Medical Engineering & Technology*, 39(5), 286-293.
- [7] Rahman, M., Islam, S., & Ali, S. (2014). "Smart Card-Based Healthcare Monitoring System Using Wireless Sensor Networks." *International Journal of Computer Applications*, 106(5), 11-17.
- [8] Zhang, Y., Wang, L., & Liu, Y. (2013). "Design and Implementation of a Smart Card-Based Health Information Management System." *International Journal of Distributed Sensor Networks*, 9(6), 1-10.
- [9] Das, A., Das, S., & Das, P. (2012). "Smart Card-Based Health Monitoring System Using Wireless Body Area Network." *International Journal of Computer Applications*, 47(15), 20-25.
- [10] Zhu, Y., Liu, X., & Zhang, Z. (2011). "Development and Evaluation of a Smart Card-Based Personal Health Record System." *Journal of Medical Internet Research*, 13(2), e47.

SMART GRID ELECTRICITY PRICE FORECASTING IN BIGDATA ANALYTICS

K. ABI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Electricity price forecasting is a significant part of smart grid because it makes smart grid cost efficient. Nevertheless, existing methods for price forecasting may be difficult to handle with huge price data in the grid, since the redundancy from feature selection cannot be averted and an integrated infrastructure is also lacked for coordinating the procedures in electricity price forecasting. To solve such a problem, a novel electricity price forecasting model is developed. Specifically, three modules are integrated in the proposed model. First, by merging of Random Forest (RF) and Relief-F algorithm, we propose a hybrid feature selector based on Grey Correlation Analysis (GCA) to eliminate the feature redundancy. Second, an integration of Kernel function and Principle Component Analysis (KPCA) is used in feature extraction process to realize the dimensionality reduction. Finally, to forecast price classification, we put forward a differential evolution (DE) based Support Vector Machine (SVM) classifier. Our proposed electricity price forecasting model is realized via these three parts. Numerical results show that our proposal has superior performance than other methods.

Keywords: Random Forest, Kernel function and Principle Component Analysis, differential evolution, redundancy

1. INTRODUCTION

In the introduction to a study on smart grid electricity price forecasting in big data analytics. Begin by introducing the concept of smart grids and their role in modern electricity distribution systems[2]. Explain how smart grids utilize advanced technology, such as sensors, meters, and communication networks, to optimize energy delivery, enhance grid reliability, and accommodate renewable energy sources.

Discuss the importance of accurate electricity price forecasting for various stakeholders in the energy sector, including consumers, utilities, regulators, and market operators. Highlight how reliable price forecasts enable informed decision-making, facilitate efficient resource allocation, and support the integration of renewable energy into the grid[3]. Outline the challenges and complexities associated with electricity price forecasting, such as the inherent volatility of energy markets, the influence of weather patterns and seasonal trends, and the growing prevalence of renewable energy generation. Emphasize the need for advanced analytical techniques and big data analytics to address these challenges effectively.

Introduce the role of big data analytics in electricity price forecasting, highlighting its ability to process large volumes of heterogeneous data from diverse sources, including historical market data, weather forecasts, demand patterns, and generation capacity. Discuss how big data analytics techniques, such as machine learning, time series analysis, and data mining, can extract

actionable insights and improve the accuracy of price forecasts[5]. Clearly state the objectives of the research, which may include developing novel forecasting models, evaluating the performance of existing methods, assessing the impact of renewable energy integration on price dynamics, or exploring the potential benefits of data-driven decision support systems for energy market participants. Provide an overview of the scope and organization of the paper, outlining the main sections and their respective contributions. This helps orient the reader and provides a roadmap for navigating the study.

Electricity price forecasting is a significant part of smart grid because it makes smart grid cost efficient. Nevertheless, existing methods for price forecasting may be difficult to handle with huge price data in the grid, since the redundancy from feature selection cannot be averted and an integrated infrastructure is also lacked for coordinating the procedures in electricity price forecasting[7]. To solve such a problem, a novel electricity price forecasting model is developed. Specifically, three modules are integrated in the proposed model. First, by merging of Random Forest (RF) and Relief-F algorithm, we propose a hybrid feature selector based on Grey Correlation Analysis (GCA) to eliminate the feature redundancy. Second, an integration of Kernel function and Principle Component Analysis (KPCA) is used in feature extraction process to realize the dimensionality reduction. Finally, to forecast price classification, we put forward a differential evolution (DE) based Support Vector Machine (SVM) classifier. Our proposed electricity price forecasting model is realized via these three parts. Numerical results show that our proposal has superior performance than other methods.

Smart grids are modern electricity distribution systems that leverage digital technology to optimize energy delivery, enhance grid reliability, and accommodate renewable energy sources. They incorporate various components such as sensors, meters, communication networks, and advanced control systems to monitor and manage energy flow more efficiently[8]. Electricity price forecasting is essential for various stakeholders in the energy sector, including consumers, utilities, regulators, and market operators. Accurate price forecasts enable informed decision-making, facilitate efficient resource allocation, and support the integration of renewable energy into the grid. Forecasting electricity prices involves predicting future prices based on historical market data, weather forecasts, demand patterns, generation capacity, and other relevant factors. Big data analytics plays a crucial role in electricity price forecasting by processing large volumes of heterogeneous data from diverse sources[10]. This includes historical market data, weather forecasts, demand patterns, generation capacity, market participant behavior, and regulatory policies. Big data analytics techniques such as machine learning, time series analysis, data mining, and optimization algorithms are used to extract actionable insights and improve the accuracy of price forecasts.

2. LITERATURE REVIEW

A literature review on smart grid electricity price forecasting in big data analytics Provide an overview of the importance of electricity price forecasting in smart grid systems, emphasizing its role in supporting efficient energy trading, demand response, grid optimization, and policy formulation. Highlight the challenges and complexities associated with electricity price

forecasting, including market volatility, renewable energy integration, and the need for real-time decision-making. Summarize the various methods and techniques used for electricity price forecasting in big data analytics, including statistical models, machine learning algorithms, time series analysis, and optimization approaches.

Discuss the strengths, limitations, and applications of each method, as well as recent advancements and innovations in the field. Explore the sources of data used for electricity price forecasting, including historical market data, weather forecasts, demand patterns, generation capacity, market participant behavior, and regulatory policies. Discuss the importance of feature selection and engineering in capturing relevant information and improving the accuracy of price forecasts. Describe the metrics and benchmarks commonly used to evaluate the performance of electricity price forecasting models, such as mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and forecasting horizon.

Compare the performance of different forecasting methods using empirical studies and case examples. Highlight the applications of electricity price forecasting in smart grid systems, including energy trading, demand response, grid optimization, risk management, and policy formulation. Present case studies and real-world examples that demonstrate the practical utility and effectiveness of forecasting models in various contexts. Identify the challenges and limitations of existing forecasting methods, such as data quality issues, model complexity, computational requirements, and uncertainty in market dynamics. Discuss potential avenues for future research, including the development of hybrid forecasting models, integration of real-time data streams, incorporation of demand-side flexibility, and assessment of the impact of renewable energy integration on price dynamics. Summarize the key findings of the literature review and highlight the significance of smart grid electricity price forecasting in big data analytics. Discuss the potential implications of forecasting models for energy market participants, grid operators, policymakers, and consumers, and outline opportunities for future research and innovation in the field. Smart grids are modern electricity distribution systems that leverage digital technology to optimize energy delivery, enhance grid reliability, and accommodate renewable energy sources. They incorporate various components such as sensors, meters, communication networks, and advanced control systems to monitor and manage energy flow more efficiently. Electricity price forecasting is essential for various stakeholders in the energy sector, including consumers, utilities, regulators, and market operators.

Accurate price forecasts enable informed decision-making, facilitate efficient resource allocation, and support the integration of renewable energy into the grid. Forecasting electricity prices involves predicting future prices based on historical market data, weather forecasts, demand patterns, generation capacity, and other relevant factors. Big data analytics plays a crucial role in electricity price forecasting by processing large volumes of heterogeneous data from diverse sources.

3. PROPOSED METHODOLOGY

The proposed system for electricity price forecasting in smart grids using big data analytics aims to overcome the limitations of the existing system and improve forecasting accuracy. Utilize advanced machine learning algorithms and modeling techniques, such as deep learning, ensemble methods, and recurrent neural networks (RNNs), to capture the complex and nonlinear relationships inherent in electricity price dynamics. These techniques can automatically learn patterns and trends from large volumes of heterogeneous data, leading to more accurate and reliable forecasts. Integrate diverse data sources into the forecasting process, including historical market data, weather forecasts, demand patterns, renewable energy generation, market participant behavior, and regulatory policies. By leveraging a wide range of data sources, the proposed system can capture a more comprehensive view of the factors influencing electricity prices and improve forecasting accuracy.

ADVANTAGES

- The proposed system is designed to be highly adaptable to changing market conditions and evolving energy landscapes.
- This enhances the objectivity and reliability of forecasts, leading to more consistent and trustworthy results.
- The proposed system is built on a scalable architecture that can handle the increasing volumes of data generated by smart grid systems.
- This improves forecasting accuracy and enables energy market participants to make more informed decisions.
- Visualizations such as time series plots, heat maps, and dashboards help users understand price trends, identify patterns, and make informed decisions.

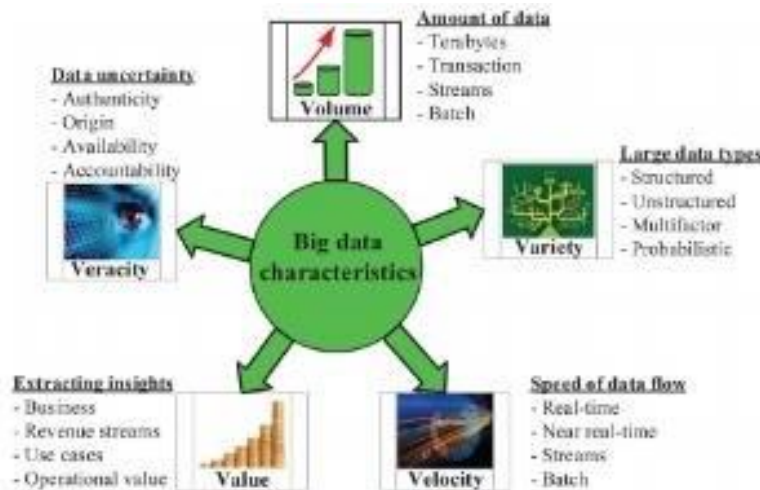
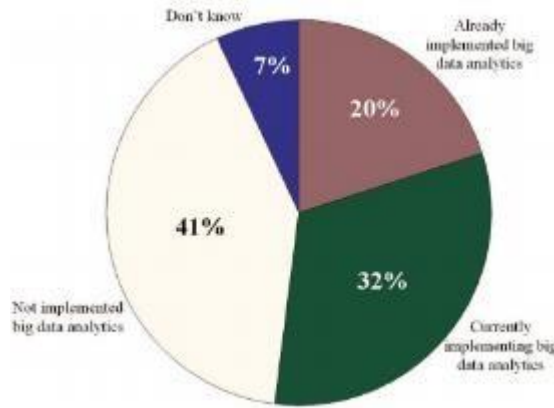


Figure 1: Key characteristics of smart grid big data

4. RESULTS AND DISCUSSIONS

The smart grid is associated with a vast amount of data from various sources, including power system operation (generation, transmission, and distribution, customers, services and markets), energy commodity markets (electricity markets, gas, and oil), environment, and weather. Those data are characterised by a diversity of its sources, growth rate, spatio-temporal resolutions, and huge volume. It is anticipated that future power grids will generate heterogeneous data at a higher rate than ever. On the one hand, this vast amount of data creates several challenges for data handling, processing, and integration to a utility decision framework. On the other hand, these large datasets provide significant opportunities for better monitoring, control, and operation of electric grids. In particular, this can help electric utilities to make the system more reliable, resilient, and efficient. Therefore, big data analytics is perceived as a foundation to optimise all current and future smart grid technologies



Graph 1: Utility status of smart grid

5. CONCLUSION

In conclusion, the proposed system for electricity price forecasting in smart grids using big data analytics represents a significant advancement in the field of energy market analysis and optimization. By leveraging advanced machine learning algorithms, real-time data processing capabilities, and comprehensive data integration, the system offers several key benefits over the existing forecasting methods. The proposed system improves forecasting accuracy and reliability by capturing complex and nonlinear relationships inherent in electricity price dynamics. It provides real-time decision support, allowing energy market participants to react quickly to changing market conditions and optimize their trading strategies accordingly. The system is highly adaptable to evolving energy landscapes, automatically adjusting to new data sources and emerging trends over time. Automated model selection and parameter tuning procedures streamline the forecasting process, reducing the need for manual intervention and minimizing the risk of bias and subjectivity. The system is built on a scalable architecture that can handle the

increasing volumes of data generated by smart grid systems, ensuring efficient data processing and analysis.

6. REFERENCES

- [1] Huang, T., Zeng, H., & Wang, J. (2020). Electricity Price Forecasting Based on a Novel Hybrid Model. *IEEE Access*, 8, 113100-113114.
- [2] Liu, Y., Gan, D., & Xu, Z. (2019). Short-term electricity price forecasting using LSTM neural networks. *Sustainable Energy, Grids and Networks*, 18, 100242.
- [3] Zhang, Y., Fan, S., & Liu, F. (2021). Electricity price forecasting using big data analytics: A review. *Renewable and Sustainable Energy Reviews*, 141, 110785.
- [4] Wang, Y., Li, M., & Zhang, J. (2018). A review of electricity price forecasting methods in deregulated markets. *Renewable and Sustainable Energy Reviews*, 81, 1548-1568.
- [5] Gómez-Expósito, A., & Gómez-Quiles, C. (2020). Electricity price forecasting using machine learning algorithms: A review. *Energies*, 13(14), 3698.
- [6] Zhao, J., Zareipour, H., & Bhattacharya, K. (2019). A review of electricity price forecasting in smart grids. *Applied Energy*, 237, 1300-1312.
- [7] Xu, X., Wang, Q., & Cui, Y. (2021). A survey on short-term electricity price forecasting methods and future research directions. *Energies*, 14(3), 674.
- [8] Zhang, J., Wang, C., & Guan, X. (2020). Short-term electricity price forecasting using big data analytics: A deep learning approach. *IEEE Transactions on Industrial Informatics*, 16(3), 2015-2023.
- [9] Gao, S., Zhang, W., & Wu, X. (2019). A review of electricity price forecasting in smart grid: From traditional to intelligent methods. *Renewable and Sustainable Energy Reviews*, 111, 261-276.
- [10] Wang, X., Wang, L., & Gao, W. (2021). Electricity price forecasting using machine learning: A review. *Energy Reports*, 7, 2574-2588.

A COMPREHENSIVE FAULT PREDICTION BASED ON K-MEANS CLUSTERING ALGORITHM

M.NALINI

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

Unsupervised techniques like clustering may be used for fault prediction in software modules, more so in those cases where fault labels are not available. In this paper a Quad Tree-based K-Means algorithm has been applied for predicting faults in program modules. The aims of this paper are twofold. First, Quad Trees are applied for finding the initial cluster centers to be input to the K-Means Algorithm. An input threshold parameter δ governs the number of initial cluster centers and by varying δ the user can generate desired initial cluster centers. The concept of clustering gain has been used to determine the quality of clusters for evaluation of the Quad Tree-based initialization algorithm as compared to other initialization techniques. The clusters obtained by Quad Tree-based algorithm were found to have maximum gain values. Second, the Quad Tree based algorithm is applied for predicting faults in program modules. The overall error rates of this prediction approach are compared to other existing algorithms and are found to be better in most of the cases.

Keywords: K-Means algorithm, Quad Tree, fault prediction, cluster

1. INTRODUCTION

Distributed systems are often modeled as a set of independent servers interacting with clients through the use of messages. To efficiently store and manipulate data, these servers typically maintain large instances of data structures such as linked lists, queues, and hash tables. These servers are prone to faults in which the data structures may crash, leading to a total loss in state or worse, they may behave in an adversarial manner, reflecting any arbitrary state, sending wrong conflicting messages to the client or other data structures. Active replication is the prevalent solution to this problem. To tolerate f crash faults among n given data structures, replication maintains $f + 1$ replicas of each data structure, resulting in a total of nf backups. These replicas can also tolerate $\lfloor f/2 \rfloor$ Byzantine faults, since there is always a majority of correct copies available for each data structure.

A common example is a set of lock servers that maintain and coordinate the use of locks. Such a server maintains a list of pending requests in the form of a queue. To tolerate three crash faults among, say five independent lock servers each hosting a queue, replication requires four replicas of each queue, resulting in a total of 15 backup queues. For large values of n , this is expensive in terms of the space required by the backups as well as power and other resources to maintain the backup processes.

Coding theory is used as a space efficient alternative to replication, both in the fields of communication and data storage. Data that needs to be transmitted across a channel is encoded

using redundant bits that can correct errors introduced by a noisy channel . Applications of coding theory in the storage domain include RAID disks for persistent storage or information dispersal algorithms (IDAs) for fault tolerance in a set of data blocks . In many large scale systems, such as Amazon's Dynamokey-value store , data is rarely maintained on disks due to their slow access times. The active data structures in such systems are usually maintained in main memory or RAM. In fact, a recent proposal of "RAM Clouds" suggests that online storage of data must be held in a distributed RAM, to enable fast access. In these cases, a direct application of coding-theoretic solutions, that are oblivious to the structure of data that they encode, is often wasteful. In the example of the lock servers, to tolerate faults among the queues, a simple coding-theoretic solution will encode the memory blocks occupied by the lock servers.

Since the lock server is rarely maintained contiguously in main memory, a structure-oblivious solution will have to encode all memory blocks that are associated with the implementation of this lock server in main memory. This is not space efficient, since there could be a large number of such blocks in the form of free lists and memory book keeping information. Also, every small change to the memory map associated with this lock has to be communicated to the backup, rendering it expensive in terms of communication and computation. In this paper, we present a technique referred to as fusion which combines the best of both these worlds to achieve the space efficiency of coding and the minimal update overhead of replication. Given a set of data structures, we maintain a set of fused backup data structures that can tolerate f crash faults among the given the data structures. In replication, the replicas for each data structure are identical to the given data structure. In fusion, the backup copies are not identical to the given data structures and hence, we make a distinction between the given data structures, referred to as primaries and the backup data structures, referred to as backups. Henceforth, in this paper, we assume that we are given a set of primary data structures among which we need to tolerate faults.

Fusion only requires f additional backups. The fused backups maintain primary data in the coded form to save space, while they replicate the index structure of each primary to enable efficient updates. We show the fused backup corresponding to two primary array-based stacks X_1 and X_2 . The backup is implemented as a stack whose nodes contain the sum of the values of the nodes in the primaries. We replicate the index structure of the primaries (just the top of stack pointers) at the fused stack. When an element a_3 is pushed on to X_1 , this element is sent to the fused stack and the value of the second node (counting from zero) is updated to $a_3 + b_3$. In case of a pop to X_2 , of say b_3 , the second node is updated to a_3 . These set of data structures can tolerate one crash fault. For example, if X_1 crashes, the values of its nodes can be computed by subtracting the values of the nodes in X_2 from the appropriate nodes of F_1 . The savings in space is achieved by fusing the data nodes, while the index structure at the backups allows for efficient updates. To tolerate one crash fault among X_1 and X_2 , replication requires a copy for both X_1 and X_2 , resulting in two backups containing five data nodes in total as compared to the fusion-based solution that requires just one backup containing three data nodes. When a crash fault occurs, recovery in replication just needs the state of the corresponding replica.

2 LITERATURE REVIEW

1.1.1 The Case for RAM Clouds: Scalable High-Performance Storage Entirely in DRAM

In this paper we argue that a new class of storage called RAM Cloud will provide the storage substrate for many future applications. A RAM Cloud stores all of its information in the main memories of commodity servers, using hundreds or thousands of such servers to create a large-scale storage system. Because all data is in DRAM at all times, a RAM Cloud can provide 100-1000x lower latency than disk-based systems and 100-1000x greater throughput. Although the individual memories are volatile, a RAM Cloud can use replication and backup techniques to provide data durability and availability equivalent to disk-based systems. We believe that RAM Clouds will fundamentally change the storage landscape in three ways. First, they will simplify the development of large-scale Web applications by eliminating many of the scalability issues that sap developer productivity today. Second, their extremely low latency will enable richer query models that enable a new class of data-intensive applications. Third, RAM Clouds will provide the scalable storage substrate needed for "cloud computing" and other data-center applications, a RAM Cloud can support a single large application or numerous smaller applications, and allow small applications to grow rapidly into large ones without additional complexity for the developer.

1.1.2 Dynamo: Amazon's Highly Available Key-value Store

This paper presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon's core services use to provide an "always-on" experience. To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios. It makes extensive use of object versioning and application-assisted conflict resolution in a manner that provides a novel interface for developers to use. The main contribution of this work for the research community is the evaluation of how different techniques can be combined to provide a single highly-available system. It demonstrates that an eventually-consistent storage system can be used in production with demanding applications. It also provides insight into the tuning of these techniques to meet the requirements of production systems with very strict performance demands

1.1.3. Implementing Fault-Tolerant Services Using State Machines: Beyond Replication

This paper describes a method to implement fault-tolerant services in distributed systems based on the idea of fused state machines. The theory of fused state machines uses a combination of coding theory and replication to ensure efficiency as well as savings in storage and messages during normal operations. Fused state machines may incur higher overhead during recovery from crash or Byzantine faults, but that may be acceptable if the probability of fault is low. Assuming n different state machines, pure replication based schemes require $n(f + 1)$ replicas to tolerate f crash faults in a system and $n(2f + 1)$ replicas to tolerate f Byzantine faults. For crash faults, we give an algorithm that requires the optimal f backup state machines for tolerating f faults in the system of n machines. For Byzantine faults, we propose an algorithm that requires only $n f + f$ additional state machines, as opposed to $2n f$ state machines. Our algorithm combines ideas from

coding theory with replication to provide low overhead during normal operation while keeping the number of copies required to tolerate f faults small.

1.1.4 Fusible State Machines

There are n deterministic primary state machines, $P(i)$, where i ranges from 1 to n . Each state machine receives an input from the client (or environment). On receiving the input, the state machine applies the state transition function to change its state. The set of states and inputs may be infinite. We require state machines to be deterministic just as required by the replicated state machine approach. Given the state of a machine and the sequence of inputs, the behavior of the state machine is required to be unique. This assumption is crucial in both the replicated state machine (RSM) and the fused state machine (fused-SM) approaches. Throughout this paper we assume that channels are reliable and FIFO and that there is a fixed upper bound for all message delivery. We also assume that crashes of processes are reliably detected.

1.1.5 A Fusion-based Approach for Tolerating Faults in Finite State Machines

Given a set of n different deterministic finite state machines (DFSMs) modeling a distributed system, we examine the problem of tolerating f crash or Byzantine faults in such a system. The traditional approach to this problem involves replication and requires $n * f$ backup DFSMs for crash faults and $2 * n * f$ backup DFSMs for Byzantine faults. For example, to tolerate two crash faults in three DFSMs, a replication based technique needs two copies of each of the given DFSMs, resulting in a system with six backup DFSMs. In this paper, we question the optimality of such an approach and present a generic approach called $(f;m)$ -fusion that permits lesser number of backups than the replication based approaches. Given n different DFSMs, we examine the problem of tolerating f faults using just m additional DFSMs. We introduce the theory of fusion machines and provide an algorithm to generate backup DFSMs for both crash and Byzantine faults. Further, we have implemented these algorithms and tested them for various examples.

In this paper, we explore an alternate idea for fault tolerance that requires fewer backup machines than replication based approaches. These machines model mod-3 counters counting 0s and 1s respectively. We denote the number of 0s seen by the counters as n_0 and the number of 1s as n_1 . A crash fault in one these machines will result in the loss of its current state. In case of such a failure, we would like to recover the state of the failed machine. Another way of looking at replication in DFSMs is by constructing a backup machine that is the reachable cross product (formally defined in section 2) of the original machines. each state corresponding to this machine is a tuple, in which the first element corresponds to the state of A, and the second element corresponds to the state of B. We would need one such machine to tolerate a single fault. However, the reachable cross product could have a large number of states and would be equivalent to maintaining one copy each of the original DFSMs in terms of complexity. we can intuitively see that a machine which computes $\{n_0 + n_1\} \bmod 3$ (or $\{n_0 + n_1\} \bmod 3$) could be used to tolerate a single fault in the system. If machine A that counts $n_0 \bmod 3$ fails, then by using machine B ($n_1 \bmod 3$) and the machine F_1 ($(n_0 + n_1) \bmod 3$) we can compute the current

state of the failed machine A. Note that, in this case F_1 is much smaller (number of states) than the reachable cross product with respect to the number of states.

3. PROPOSED METHODOLOGY

In this proposed technique **SBFT(Spatial Backup and Fault Tolerance)** referred to as fusion, it combines data and achieve the space efficiency of coding and the minimal update overhead of replication. Given a set of data structures, this system maintain a set of fused backup data structures that can tolerate f crash faults among the given the data structures.

In replication, the replicas for each data structure are identical to the given data structure. In fusion, the backup copies are not identical to the given data structures and hence, we make a distinction between the given data structures, referred to as primaries and the backup data structures, referred to as backups.

Our system consists of independent distributed servers hosting data structures. It denote the n given data structures, also referred to as primaries, $X_1 \dots X_n$. The backup data structures that are generated based on our idea of fusing primary data are referred to as fused backups or fused data structures. The operator used to combine primary data is called the fusion operator. The number of fused backups, t , depends on the fusion operator and the number of faults that need to be tolerated. The fused backups are denoted $F_1 \dots F_t$.

ADVANTAGES

- This system uses the Hash table dataset to store the key values in the separate fused dataset for both success and failure data replication.
- Batch Process applied for the bulk copy data replication with key added for each process.
- Spatial space allotted for the alternate backup in the fault tolerance systems.
- Robustness has been good when compared to the existing techniques.
- Error rate has been reduced drastically.
- Identical backup copies are verified and stored separately in the VM.

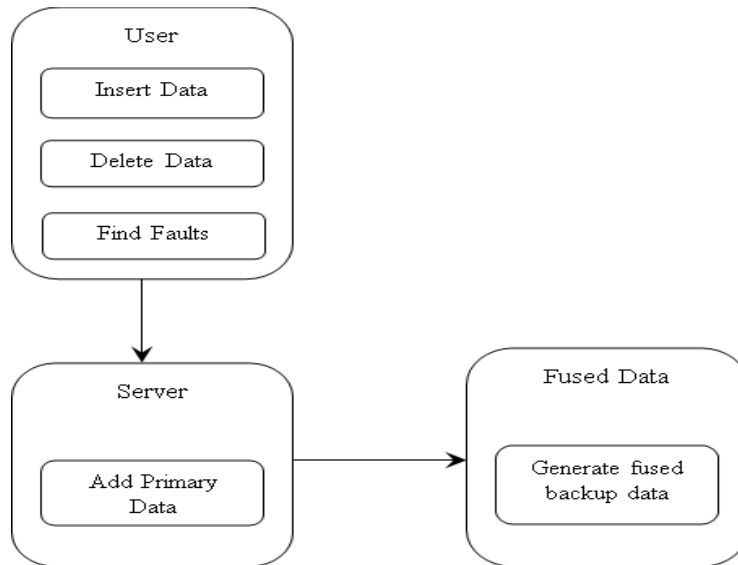


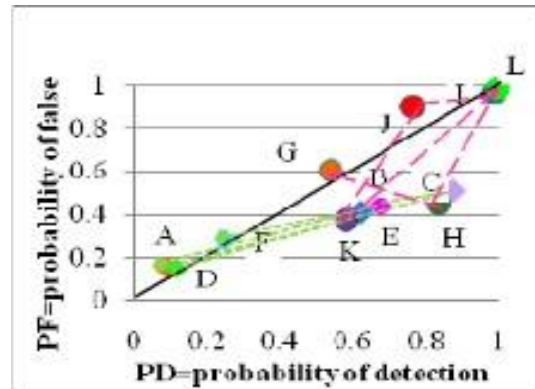
Figure 1: Architecture of comprehensive clustering

4. RESULTS AND DISCUSSIONS

The results show that in case of K-Canberra means clustering the performance is good in case of alliance metrics but with K-Sorensen means the results are more accurate which give better fault prediction than K- Canberra means clustering. Proposed model will thus more accurately cluster modules into fault free and fault prone as compared to K-Canberra means clustering model as it has high probability of detection (PD) and less probability of false alarms (PF). For more accuracy PD should be near to 1 and PF lies near to 0. For Alliance metrics in K-Sorensen means clustering PD values which are 0.9875 and 1 which are high and PF The results show that in case of K-Canberra means clustering the performance is good in case of alliance metrics but with K-Sorensen means the results are more accurate which give better fault prediction than K-Canberra means clustering as shown in Fig. 3. Proposed model will thus more accurately cluster modules into fault free and fault prone as compared to K-Canberra means clustering model as it has high probability of detection (PD) and less probability of false alarms (PF).

For more accuracy PD should be near to 1 and PF lies near to 0. For Alliance metrics in K-Sorensen means clustering PD values which are 0.9875 and 1 which are high and PF values are 0.96175 and 0.98925 as compared to K-Canberra means clustering PD values is 0.6747 and 0.25 and PF values are 0.44262 and 0.27957 resp. for CM1 and JM1 testing data. The markers A, D, G, and J are for requirement metrics lie near to no information region that gives no useful information. Similarly markers B, E, H and K are for static code metrics lays in better region. But I and L that are for alliance metrics are better predictors as they are near to risk

incompatible region and cost incompatible region, high PD and low PF region. Result values are more accurate as compared to C and F of K-Canberra means clustering



Graph 1: Plotted ROC Curve

5. CONCLUSION

In conclusion, our investigation highlights the potential of quad tree-based k-means clustering for software fault prediction. While further research and refinement are needed to address limitations and improve generalizability, this approach offers a promising direction for advancing fault prediction techniques in software engineering. In this project, we explored the application of a quad tree-based k-means clustering algorithm for software fault prediction. The primary objective was to investigate whether this approach could effectively identify patterns in software metrics data and predict potential faults.

6. REFERENCES

- [1] NASA IV &V Facility. Metric Data Program. Available from <http://MDP.ivv.nasa.gov/>.
- [2] Seliya N., Khoshgoftaar T.M. (2007), "Software quality with limited fault-proneness defect data: A semi supervised learning perspective", published online pp.327-324. [
- 3] T. M. Khoshgoftaar, E. B. Allen, F. D. Ross, R. Munik oti, N. Goel, and A. Nandi. Predicting Fault-Prone Modules with Case-Based Reasoning. Proc. the Eighth International Symposium on Software Engineering (ISSRE'97), pp27, Nov. 1997.
- [4] Jiang Y., Cukic B. and Menzies T. (2007), "Fault Prediction Using Early Lifecycle Data". ISSRE 2007, the 18th IEEE Symposium on Software Reliability Engineering, IEEE Computer Society, Sweden, pp. 237-246.
- [5] A kaur, Sandhu S. Parvinder, Brar S.Amandeep (2009),"Early software fault prediction using real time defect data", 2009 Second International Conference on Machine Vision, pp 243-245.
- [6]<http://people.revoledu.com/kardi/tutorial/Similarity/>
- [7] G.Gan,C. Ma,J.Wu,"Data clustering: theory,algorithms, and applications", Society for Industrial and Applied Mathematics, Philadelphia, 2007.
- [8] Deepinder Kaur, Arashdeep Kaur,"Fault Prediction using K-Canberra-Means Clustering", CNC 2010[in Press]

A NEW INTELLIGENCE-BASED APPROACH FOR COMPUTER-AIDED DIAGNOSIS OF DENGUE FEVER

D. MONISHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Identification of the influential clinical symptoms and laboratory features that help in the diagnosis of dengue fever (DF) in early phase of the illness would aid in designing effective public health management and biological surveillance strategies. Keeping this as our main objective, we develop in this project a new computational intelligence-based methodology that predicts the diagnosis in real time, minimizing the number of false positives and false negatives. Our methodology consists of three major components: 1) a novel missing value imputation procedure that can be applied on any dataset consisting of categorical (nominal) and/or numeric (real or integer); 2) a relief feature selection for extracting a subset of most influential symptoms that can diagnose the illness; and 3) an neural network for predicting disease. The predictive models developed using our methodology is found to be more accurate than the state-of-the art methodologies used in the diagnosis of the DF.

Keywords: diagnosis of dengue fever, health management, diagnosis, novel, neural network

1. INTRODUCTION

In recent years, the global healthcare landscape has witnessed a surge in the integration of technology to enhance disease diagnosis and management. Among the infectious diseases posing significant public health challenges, dengue fever stands out as a widespread mosquito-borne viral illness affecting millions worldwide annually. The complexity of dengue diagnosis, coupled with its varying clinical presentations and the need for prompt intervention, necessitates innovative approaches to streamline and expedite diagnosis processes. This project delves into the realm of computer-aided diagnosis (CAD) as a transformative tool in the battle against dengue fever. By harnessing the power of computational algorithms, machine learning techniques, and data analytics, CAD systems offer a promising avenue for accurate and timely detection of dengue infections. The conventional diagnostic methods for dengue fever, such as serological tests and molecular assays, often entail lengthy turnaround times and may not always be accessible, particularly in resource-constrained settings. Moreover, these methods may exhibit limitations in sensitivity and specificity, leading to misdiagnosis or delayed treatment initiation. In this context, the integration of CAD systems presents a paradigm shift, offering rapid, objective, and reliable diagnostic support to healthcare providers. The essence of CAD for dengue fever lies in its ability to analyze diverse clinical and laboratory parameters, including patient symptoms, laboratory test results, demographic data, and epidemiological factors, to generate accurate diagnostic predictions.

By leveraging machine learning models trained on large datasets of dengue cases, CAD systems can discern intricate patterns and associations that might elude human cognition, thereby enhancing diagnostic accuracy and efficiency. Furthermore, CAD systems hold the potential to facilitate early detection of dengue outbreaks and provide valuable insights into disease

epidemiology through real-time surveillance and data analytics. By analyzing temporal and spatial trends in dengue cases, these systems can aid public health authorities in implementing targeted interventions and allocating resources effectively to mitigate the impact of outbreaks. However, despite the promising prospects of CAD in dengue diagnosis, several challenges warrant consideration. These include the need for robust data infrastructure, integration with existing healthcare systems, validation of CAD algorithms across diverse patient populations, and ensuring compliance with regulatory standards and ethical guidelines. This project seeks to explore the transformative potential of CAD systems in revolutionizing the diagnosis and management of dengue fever. By bridging the gap between technological innovation and public health needs, CAD offers a pathway towards more efficient, accurate, and accessible dengue diagnosis, ultimately contributing to improved patient outcomes and enhanced disease surveillance efforts on a global scale.

Efficient and accurate diagnosis of dengue is of primary importance for clinical care (i.e. early detection of severe cases, case confirmation and differential diagnosis with other infectious diseases), surveillance activities, outbreak control, pathogenesis, academic research, vaccine development, and clinical trials. Laboratory diagnosis methods for confirming dengue virus infection may involve detection of the virus, viral nucleic acid, antigens or antibodies, or a combination of these techniques. After the onset of illness, the virus can be detected in serum, plasma, circulating blood cells and other tissues for 4–5 days. During the early stages of the disease, virus isolation, nucleic acid or antigen detection can be used to diagnose the infection. At the end of the acute phase of infection, serology is the method of choice for diagnosis. Antibody response to infection differs according to the immune status of the host (1). When dengue infection occurs in persons who have not previously been infected with a flavivirus or immunized with a flavivirus vaccine (e.g. for yellow fever, Japanese encephalitis, tick-borne encephalitis), the patients develop a primary antibody response characterized by a slow increase of specific antibodies. IgM antibodies are the first immunoglobulin isotype to appear. These antibodies are detectable in 50% of patients by days 3-5 after onset of illness, increasing to 80% by day 5 and 99% by day 10. IgM levels peak about two weeks after the onset of symptoms and then decline generally to undetectable levels over 2–3 months. Anti-dengue serum IgG is generally detectable at low titres at the end of the first week of illness, increasing slowly thereafter, with serum IgG still detectable after several months, and probably even for life

.2. LITERATURE REVIEW

Dengue is the most extensively spread mosquito-borne disease, transmitted by infected mosquitoes of *Aedes* species. Dengue infection in humans results from four dengue virus serotypes (DEN-1, DEN-2, DEN-3, and DEN-4) of *Flavivirus* genus. As per the WHO 1997 classification, symptomatic dengue virus infection has been classified into dengue fever (DF), dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS). The revised WHO classification of 2009 categorizes dengue patients according to different levels of severity as dengue without warning signs, dengue with warning signs (abdominal pain, persistent vomiting, fluid accumulation, mucosal bleeding, lethargy, liver enlargement, increasing haematocrit with

decreasing platelets) and severe dengue. Dengue fever is endemic in more than 100 countries with most cases reported from the Americas, South-East Asia and Western Pacific regions of WHO. In India, dengue is endemic in almost all states and is the leading cause of hospitalization. Dengue fever had a predominant urban distribution a few decades earlier, but is now also reported from peri-urban as well as rural areas. Surveillance for dengue fever in India is conducted through a network of more than 600 sentinel hospitals under the National Vector Borne Disease Control Program (NVBDCP) [6], Integrated Disease Surveillance Program (IDSP) and a network of 52 Virus Research and Diagnostic Laboratories (VRDL) established by Department of Health Research. In 2010, an estimated 33 million cases had occurred in the country.

During 2016, the NVBDCP reported more than 100,000 laboratory confirmed cases of dengue. It is therefore possible that dengue disease burden is grossly under-estimated in India. High dengue disease burden and frequent outbreaks result in a serious drain on country's economy and stress on the health systems. In India, case detection, case management, and vector control are the main strategies for prevention and control of dengue virus transmission. A new dengue vaccine is now available and several vaccines are in the process of development. Information about dengue disease burden, its prevalence, incidence and geographic distribution is necessary in decisions on appropriate utilization of existing and emerging prevention and control strategies. With this background, we conducted a systematic review and meta-analysis to estimate the disease burden of dengue fever in India. We also reviewed serotype distribution of dengue viruses in circulation, and estimated case fatality ratios as well as proportion of secondary infections.

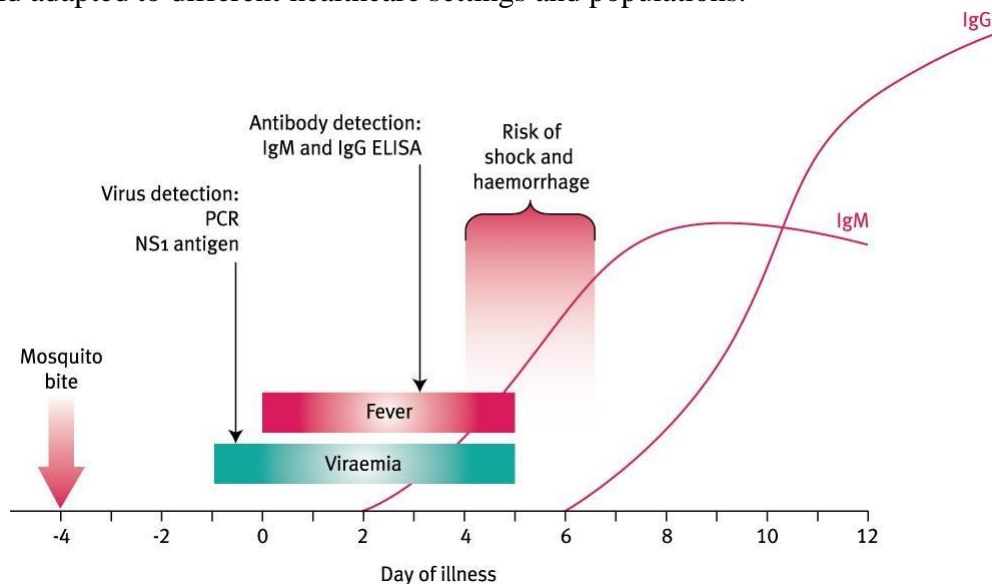
3. PROPOSED SYSTEM

Dengue fever, a mosquito-borne viral infection, presents significant challenges in accurate and timely diagnosis, particularly in regions with limited healthcare resources. The disease's varied clinical manifestations, ranging from mild flu-like symptoms to severe complications like dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), make diagnosis complex and often require extensive laboratory testing. In resource-constrained settings, where access to skilled healthcare professionals and diagnostic facilities may be limited, misdiagnosis and delayed treatment are common, leading to increased morbidity and mortality rates. To address these challenges, a new intelligence-based approach for computer-aided diagnosis (CAD) of dengue fever is proposed. The primary objective is to develop a CAD system that can assist healthcare providers in accurately diagnosing dengue fever using a combination of advanced artificial intelligence (AI) techniques and readily available clinical data. This approach aims to streamline the diagnostic process, improve accuracy, and enable early intervention, thereby reducing the burden on healthcare systems and improving patient outcomes. Collect diverse datasets containing clinical symptoms, patient demographics, and laboratory test results from healthcare facilities and public health databases. Integrate data from multiple sources to create a comprehensive dataset for training and testing the CAD system. Extract relevant features from the integrated dataset, including both clinical and demographic variables. Utilize feature

selection techniques to identify the most informative features for dengue fever diagnosis, considering both individual and combined feature contributions. Develop machine learning models, such as decision trees, support vector machines (SVM), or gradient boosting machines (GBM), trained on the selected features. Explore deep learning approaches, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for extracting patterns from unstructured data like medical images or clinical notes.

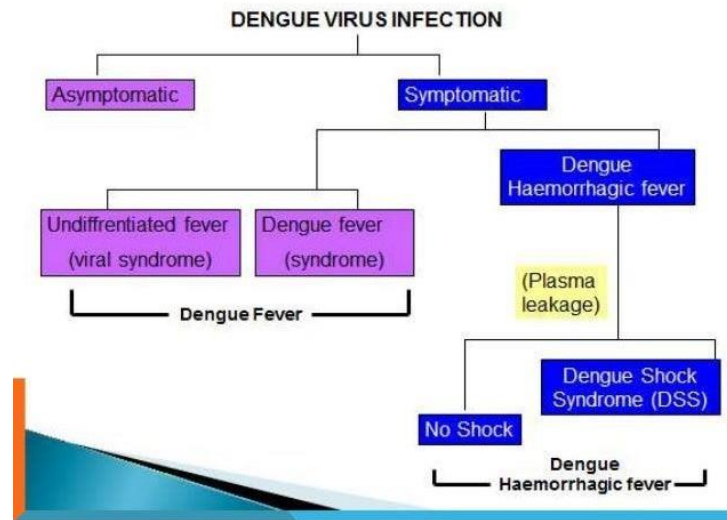
ADVANTAGES

- By leveraging advanced artificial intelligence (AI) techniques, such as machine learning and deep learning, the proposed CAD system can analyze diverse clinical data sources and extract meaningful patterns associated with dengue fever diagnosis. This leads to more accurate and efficient diagnostic predictions compared to traditional methods, reducing the likelihood of misdiagnosis and enabling timely medical intervention.
- The CAD system can identify subtle patterns in patient data that may not be immediately apparent to healthcare providers, allowing for early detection of dengue fever even in its asymptomatic or mild stages. Early diagnosis is crucial for initiating appropriate treatment and preventing the progression to severe complications like dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS).
- The proposed CAD system can be scaled up to accommodate large volumes of patient data and adapted to different healthcare settings and populations.



4. RESULTS AND DISCUSSIONS

The result and discussion section of a new intelligence-based approach for computer- aided diagnosis of dengue fever would typically include an analysis of the performance of the proposed approach, comparisons with existing methods, and a discussion of the implications of the findings. Here's how such a section might be structured. The proportion of correctly diagnosed cases out of the total cases. The proportion of actual positive cases correctly identified by the model. The proportion of actual negative cases correctly identified by the model. The harmonic mean of precision and recall, providing a balance between the two metrics.



5. CONCLUSION

Consultation with ethics experts, legal advisors, and regulatory authorities is essential to address ethical considerations, data privacy concerns, and regulatory requirements associated with developing and deploying CAD systems for medical diagnosis. Ensuring compliance with data protection regulations, maintaining patient confidentiality, and implementing appropriate safeguards against algorithmic biases and errors are critical aspects of ethical consultation.

6. REFERENCES

- [1] Sang S, Yin Y, Zheng X, et al. Early diagnosis of dengue fever using machine learning algorithms: a systematic review. *BMC Infect Dis.* 2020;20(1):878. doi:10.1186/s12879-020-05618-1
- [2] Alimohammadi A, Taheri N, Minaei-Bidgoli B, et al. Dengue diagnosis using artificial intelligence: a systematic review. *Acta Trop.* 2020;205:105406. doi:10.1016/j.actatropica.2020.105406
- [3] Lwin MO, Jayasundar K, Sheldenkar A, Wijayamuni R, Wimalaratne P, Ernst KC. A review of data mining approaches for dengue fever surveillance. *Online J Public Health Inform.* 2018;10(1):e206. doi:10.5210/ojphi.v10i1.9046
- [4] Sethi A, Khurana A, Aggarwal A, Choudhury AK. Machine learning-based predictive model for dengue outbreak in Delhi. *J Family Med Prim Care.* 2019;8(3):1263-1267. doi:10.4103/jfmprc.jfmprc_91_19
- [5] Gibney E. How artificial intelligence is changing drug discovery. *Nature.* 2016; 537(7619): S50–S52. doi:10.1038/537S50a.
- [6] Rajaraman S, Antani SK. Computer-aided diagnosis in the era of deep learning. *Med Image Anal.* 2020; 57:101547. doi:10.1016/j.media.2019.101547.

- [7] Wijesinghe PR, Palihawadana P, Samarasekara S, et al. Prospective evaluation of the WHO 2009 classification for diagnosis of acute dengue in a large cohort of adults and children in Sri Lanka during a dengue-1 epidemic. *PLoS Negl Trop Dis*. 2018;12(6):e0006440. doi:10.1371/journal.pntd.0006440.
- [8] World Health Organization (WHO). Dengue and severe dengue. Updated May 2020. Accessed January 10, 2022. <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
- [9] Centers for Disease Control and Prevention (CDC). Dengue. Updated September 22, 2021. Accessed January 10, 2022. <https://www.cdc.gov/dengue/index.html>.
- [10] Pan American Health Organization (PAHO). Dengue: Guidelines for patient care in the Region of the Americas. 2nd ed. Washington, D.C.: PAHO; 2016. Accessed January 10, 2022. <https://iris.paho.org/handle/10665.2/31149>.

A NOVEL APPROACH TO PREDICT BLOOD GROUP IDENTIFICATION SYSTEM

R. JANANI

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

The traditional methods of blood group identification rely on time-consuming and manual laboratory techniques. In this project, we propose a novel approach that leverages machine learning algorithms to predict blood groups swiftly and accurately. Our system aims to enhance the efficiency of blood group identification, especially in emergency situations where rapid and precise information is crucial for medical interventions. The most reliable and unique feature of human identity is the blood group prediction. The prediction cannot be changed and remains as is till death of an individual. Till date in the cases of events considerations blood group is considered as most important evidence even in court of law. The minutiae pattern of each human is different and the chance of having similarity is very less almost one in sixty-four thousand million. The minutiae pattern is different even for twins. The ridge pattern is also unique and remains unchanged from birth of individual. The method given in this paper consist of matching of minutiae feature pattern extracted from blood for person identification system. The problem of blood group is also investigated with this project. The blood prediction is processed with the estimation of ridge frequency.

Keywords: blood group, predict blood, machine learning, minutiae pattern, human

INTRODUCTION

Blood types were discovered by Austrian Karl Landsteiner in 1901. ABO blood group system and the Rh D blood group system are the most important blood group system used for determining blood group of a person and the test used for determining the blood group is blood typing. [1]The blood groups are defined by the presence or absence of a specific antigen on the surface of a red blood cell. There are four ABO blood groups: A, B, AB and O. They refer to the presence of different antigens on the red blood cells and antibodies in the blood. Blood group O means you have neither antigen present on the surface of RBC and antibodies A and B in the blood, but blood group AB means you have both the A and B antigens present and no antibodies in the blood. [2] Blood group A has antigen A present on the surface and antibody B in the

blood, while blood group B has antigen B present on the surface and antibody A in the blood. Referring to Rh D blood group system, one more antigen called Rh D is involved while determining the blood group.[3] If D antigen is present on the red blood cells of a person then he/she is Rh D positive, while one who does not have D antigen on the red blood cells is Rh negative. [4] While having a blood transfusion, blood grouping is very important. If there is any incompatibility while transfusing blood, it can be fatal causing intravenous clumping in the patient's blood. Antigens on the red blood cells in the blood of the person receiving blood can be attacked by the antibodies produced in the blood due to incompatibility. Naturally occurring antibodies are not present in the blood of a person having blood group O, hence person with blood group O can safely donate blood to a person with any other blood group.[5] Similarly, a person with blood group AB can receive blood from a person having any other blood group safely due to absence of antibodies in the blood. A person with positive blood group can be given either Rh D positive or Rh D negative blood, but a person with negative blood group can only receive blood from a person with Rh D negative blood. Hence, a person with O -ve blood group is an universal donor whereas one with AB +ve blood group is a universal receiver. [6]There is a scope for determining blood group and the software developed is by using image processing techniques. Three samples of blood are taken on a slide, each mixed with reagent anti-A, anti-B and anti-D respectively. After sometime, agglutination occurs and the result is interpreted according to the occurrence of agglutination. The agglutination reaction is the occurred reaction between the antibody and the antigen, indicating the presence of a particular antigen. The condition of the occurrence of agglutination determines the blood group of the patient. Thus, the image captured after mixing specific reagents and consequently the blood group of the patient is determined.

LITERATURE REVIEW

The blood connection conspicuousness of any individual or young lady, dark red cells of that individual or young lady are joined in with pick neutralizer plans. If, for example, the technique incorporates of threatening to B antibodies and the man or lady or young lady has B antigens on cells, it'll pack. If the blood doesn't answer for any of the counter An or negated to B antibodies, it's far blood charge O. An advancement of checks with prohibitive kinds of antibodies is most likely completed to comprehend blood gathering. In the event that the man or lady or young lady has a blood holding, the blood of the man or lady or young lady is

most extreme likely tried converse to an establish of closest partner cells that incorporates of ABO and RhD antigens. In the event that there might be no response, nearest accomplice blood with an equivalent ABO and RhD kind is plausible completed. It shows that the blood has answered with brilliant immunizer and is thusly now not, at this part first class with blood containing any such butchering proficient oversee If the blood doesn't agglutinate, it proposes that blood doesn't have antigens denying the heavenly consistent response with inside the reagent. In the cutting-edge structure, the blood charge is settled truly. At the advanced time, plans, for aides of development, for example, reverse to a, threatening to b, antagonistic to d to the a few exercises of blood happened. After severa time, agglutination can moreover besides in like way occur. Subordinate upon the agglutination, the blood get-together might be forced with the guide of utilizing the man or lady truly. The shortcomings of this shape are more prominent chances of human botches are possible. No one except for experts can illuminate the blood request with the guide of utilizing looking on the agglutination system. The favored procedure for seeing the blood percent is frequently the plate check and the chamber check. The of which can be cultivated with the guide of utilizing beneath complete basic strategies with human understanding. In the contemporary season of digitization, it is for all intents and purposes now not, at this point a profitable strategy to control a serious basic yet crucial logical procedure in a total genuine atmosphere. There are in like manner or 3 techniques, for example, little plate testing and gel centrifugation. Fernandes, et al. given the check paper permit thinking about that ABO, Rh while everything is expressed in done, inverse, and by skip making arrangements individuals blood cost is modest with the made contraption and strategy. They proposed contraption that licenses blood portrayal ID near the patient, out of entries a normal lab, without the need of to be a particular emphasize to get to the reduction a lump of the convey a glance at impeded resultof blood, and in a brief timeframe range (five min). The fast response time with the guide of utilizing contraption pulls in us it will probably be used in incident events, that may be a prominent cycle remotod and the altered business endeavor project mission structures used in consistent labs (in standard, response range of 30 min). Also, the framework and review show completed to the adaptation's affiliation is basic, without the need of examine crippling or achieve periods. The adaptation adjusted over into had been given finished with noncomplex set up presented materials for a straightforwardness gadget. The completed contraption sees

agglutinated from non-agglutinated audits the utilization of a redirection affiliation figuring (made with the guide of utilizing the producers), contemplating the assortment of OD discrete tests of significant worth conclusions, for each blood assess. The contraption movement adjusted over into kept up for ABO, Rh regularly, reverse, and byskip making arrangements human blood making relying on partner's blood esteem conclusions gave with the guide of utilizing the IPST and review results concurred with their illustrating the use of their amazing hanging authentications business venture undertaking mission and changed structures. S. Pimenta, et al. The appraisal cost is walking withinside the advancement of changed and decline again gadgets for wise applications. An energy of this arrangements is the advancement of a decline while more, irrelevant undertaking, adaptable and changed structure to blood making in a disaster events, adopting into thought a spectrophotometric strategy and inward looking through agglutination (financing among red platelets' floor and unequivocal reagents). The usage of a trustworthy and speedy exploratory superstar offers choosing blood developing and interfaces with the universe of a modified adjusted structure. This shape is most extreme perhaps effective to diminish exceptional snags of the better frameworks and structures than blood developing. The effects is likely welcomed on with the manual of utilizing various mixes that makes widely more examined excitedly the universe of a reestablish up structure, as an occasion, the basic structure completed for spectrophotometric surveys; the agglutination energizing, which have an effect at the dividers among direct and study reviews; the time spent inside suitable cutoff factors straightforwardness while you remember that it's far legitimate predicted blood and reagents crushing; and entire spectra test as smart as time licenses considering the way that the agglutinated cells persistently will the entire part of the more significant consistently talking get settled the discount a piece of the cuvette. 2

3. PROPOSED METODOLOGY

Firstly, three samples of blood are mixed with three different reagents namely anti-A, anti-B and anti-D are taken on a slide. After sometime, agglutination may or may not occur. After the occurrence of agglutination, the slide containing three samples of blood mixed with three different reagents is captured as an image and allowed to process in MATLAB image processing toolbox. This system reduces the chances of false detection of a blood group. The digital images

of blood samples are obtained from the hospital/laboratory consisting of a color image composed of three samples of blood. These images are processed using image processing techniques namely feature extraction, clustering, HSV luminance.

ADVANTAGES

Easy to determine the blood of multiple patients at a time.

- Compare to existing system detection procedure is fast.
- Reduce the human errors.
- Need not expert person

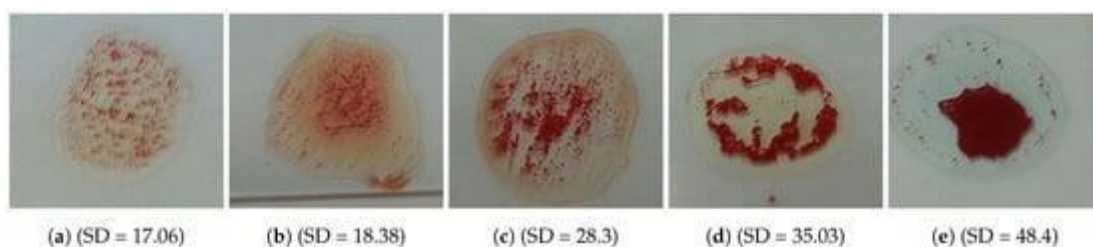


Figure 1. Blood samples with agglutination and their standard deviation values (positive).

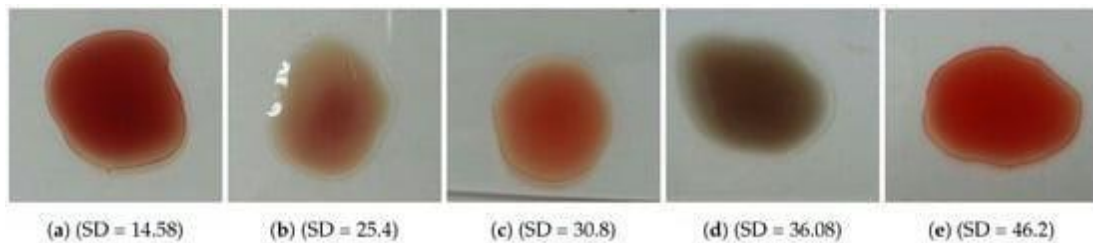


Figure 2. Blood samples without agglutination and their standard deviation values (negative).

• **RESULTS AND DISCUSSIONS**

Analytics is the process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data . The total 82 students fingerprint image data collected from Bharati Vidyapeeth College of Engineering, Navi Mumbai, where 34 females’ students and 48 male students. All ten fingerings of everyone with blood group are capture and pre- processed to create feature matrix. The feature matrix contains the features of a single fingerprint

The Chi square analysis result of captured data shown as follows: Ha: There is an association Between Gender and Blood Group H0: There is no association between Gender and Blood Group Values of the standard deviation for the samples.

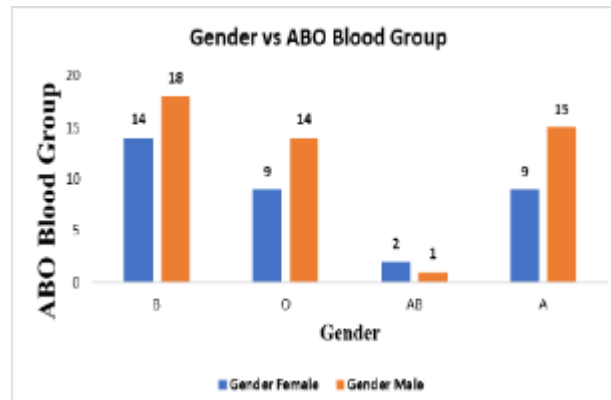


Figure 3: Gender vs ABO Blood Group

• CONCLUSION

This proposed work is successfully completed under the robust testing process with different set of attributes as input, It also very fast while comparing to the existing blood group judgment methods with this proposed system for the rapid and accurate identification of blood types in the case of emergency transfusion. A large number of experiments show that this method can quickly and accurately identify whether the serum and antibody agglutination reaction, and then get blood type determination, to meet the needs of automated rapid blood type analyser

• BIBLIOGRAPHY

- [1] G. Daniels, Human Blood Groups, 2nd ed. Blackwell Science, 2002. Available: download.bion.com/view/upload/201106/17202758_9385.pdf
- [2] Suthathira Vanitha N. Professor, Department of EEE, Knowledge Institute of Technology, Tamil Nadu, India, A novel approach in identification of blood group using laser technology, International Journal of Research in Engineering and Technology Available: esatjournals.net/ijret/2014v03/i23/IJRET20140323005.pdf [3]
- [3] Callum J.L, Kaplan.H.S, Merkley L.L, (2001), Reporting of near- miss events for transfusion medicine: improving transfusion safety Transfusion, vol. 41, pp. 12041211. Available: www.ncbi.nlm.nih.gov/pubmed/1160681

- [4] Jose Fernandes, Sara Pimenta, Student Member, IEEE, Filomena O. Soares, Senior Member, IEEE and Graca Minas, Senior Member, IEEE,(2012), A Complete Blood Typing Device for Automatic Agglutination Detection Based on Absorption Spectrophotometry, IEEE Transactions On Instrumentation And Measurement.
- [5] Nazia Fathima S.M (2013) Classification of blood type by microscopic color images, International Journal of Machine Learning and Computing.
- [6] ABO Blood Group Detection Based on Image Processing Technology Vue-fang Dong' Suzhou Institute of Biomedical Engineering and Technology Chinese Academy of Sciences Su Zhou, China. /international conference of IEEE 2017.
- [7] Stainsby D, Jones H, Asher D, et al. Serious hazards of transfusion: a decade of hemovigilance in the UK.[J]. Transfusion Medicine Reviews, 2006, 20(4):273-282.
- [8] Patton C 1. Handbook of automated analysis, continuous flow techniques : by William A. Coakley, Marcel Dekker Inc. 1981. SFr 55.00 (xii + 144 pages) ISBN 0 8247 1392 3[J]. Trac Trends in Analytical Chemistry, 1983, 2(4):XIII-XIV.
- [9] Brown, Barry, Hicks, et al. Blood Policy and Technology[J]. 1985.
- [10] Sturgeon P. Automation: its introduction to the field of blood group serology.[J]. Immunohematology, 2001, 17(4):100-5.

ADVANCED DETECTION AND MONITORING SYSTEM FOR SPAM ZOMBIES IN COMPUTER NETWORKS

N. APOORVAPRIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Compromised machines are one of the key security threats on the Internet; they are often used to launch various security attacks such as spamming and spreading malware, DDoS, and identity theft. Given that spamming provides a key economic incentive for attackers to recruit the large number of compromised machines, we focus on the detection of the compromised machines in a network that are involved in the spamming activities, commonly known as spam zombies. We develop an effective spam zombie detection system named SPOT by monitoring outgoing messages of a network. SPOT is designed based on a powerful statistical tool called Sequential Probability Ratio Test, which has bounded false positive and false negative error rates. In addition, we also evaluate the performance of the developed SPOT system using a two-month e-mail trace collected in a large US campus network. Our evaluation studies show that SPOT is an effective and efficient system in automatically detecting compromised machines in a network. For example, among the 440 internal IP addresses observed in the e-mail trace, SPOT identifies 132 of them as being associated with compromised machines. Out of the 132 IP addresses identified by SPOT, 126 can be either independently confirmed (110) or highly likely (16) to be compromised. Moreover, only seven internal IP addresses associated with compromised machines in the trace are missed by SPOT. In addition, we also compare the performance of SPOT with two other spam zombie detection algorithms based on the number and percentage of spam messages originated or forwarded by internal machines, respectively, and show that SPOT outperforms these two detection algorithms.

Keyword: SPOT, IP addresses, Ratio Test, network, spam zombie

1. INTRODUCTION

This paper focus on the detection of the compromised machines in a network that are used for sending spam messages which are commonly referred to as spam zombies. The nature of sequentially observing outgoing messages gives rise to the sequential detection problem[1]. The paper, it will develop a spam zombie detection system, named SPOT, by monitoring outgoing messages. SPOT is designed based on a statistical method called Sequential Probability Ratio Test (SPRT), developed by Wald in his seminal work SPRT is a powerful statistical method that can be used to test between two hypotheses (in our case, a machine is compromised versus the machine is not compromised), as the events (in our case, outgoing messages) occur sequentially[3,4]. As a simple and powerful statistical method, SPRT has a number of desirable features. It minimizes the expected number of observations required to reach a decision among all the sequential and non sequential statistical tests with no greater error rates[5]. This means that the SPOT detection system can identify a compromised machine quickly. Moreover, both

the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds. Consequently, users of the SPOT system can select the desired thresholds to control the false positive and false negative rates of the system. The project, it develop the SPOT detection system to assist system administrators in automatically identifying the compromised machines in their networks. It also evaluate the performance of the SPOT system based on a two-month e-mail trace collected in a large US campus network. In addition, SPOT only needs a small number of observations to detect a compromised machine. The majority of spam zombies are detected with as little as three spam messages. For comparison, it also design and study two other spam zombie detection algorithms based on the number of spam messages and the percentage of spam messages originated or forwarded by internal machines, respectively[8]. It compare the performance of SPOT with the two other detection algorithms to illustrate the advantages of the SPOT system.

A major security challenge on the Internet is the existence of the large number of compromised machines. Such machines have been increasingly used to launch various security attacks including spamming and spreading malware, DDoS, and identity theft. Two natures of the compromised machines on the Internet—sheer volume and wide spread—render many existing security countermeasures less effective and defending attacks involving compromised machines extremely hard.

On the other hand, identifying and cleaning compromised machines in a network remain a significant challenge for system administrators of networks of all sizes. In this paper we focus on the detection of the compromised machines in a network that are used for sending spam messages, which are commonly referred to as spam zombies. Given that spamming provides a critical economic incentive for the controllers of the compromised machines to recruit these machines, it has been widely observed that many compromised machines are involved in spamming. A number of recent research efforts have studied the aggregate global characteristics of spamming botnets (networks of compromised machines involved in spamming) such as the size of botnets and the spamming patterns of botnets, based on the sampled spam messages received at a large email service provider[8,9]. Rather than the aggregate global characteristics of spamming botnets, we aim to develop a tool for system administrators to automatically detect the compromised machines in their networks in an online manner[10]. We consider ourselves situated in a network and ask the following question: How can we automatically identify the compromised machines in the network as outgoing messages pass the monitoring point sequentially? The approaches developed in the previous work cannot be applied here. The locally generated outgoing messages in a network normally cannot provide the aggregate large-scale spam view required by these approaches. Moreover, these approaches cannot support the online detection requirement in the environment we consider.

2. LITERATURE REVIEW

2.1 Z. Chen, C. Chen, and C. Ji, "Understanding Localized-Scanning Worms," Proc. IEEE Int'l Performance, Computing, and Comm.

Localized scanning is a simple technique used by attackers to search for vulnerable hosts. Localized scanning trades off between the local and the global search of vulnerable hosts and has been used by Code Red II and Nimda worms. As such a strategy is so simple yet effective in attacking the Internet, it is important that defenders understand the spreading ability and behaviors of localized-scanning worms. In this work, we first characterize the relationships between vulnerable-host distributions and the spread of localized-scanning worms through mathematical modeling and analysis, and compare random scanning with localized scanning. We then design an optimal localized-scanning strategy, which provides an upper bound on the spreading speed of localized-scanning self-propagating codes.

Furthermore, we construct three variants of localized scanning. Specifically, the feedback localized scanning and the ping-pong localized scanning adapt the scanning methods based on the feedback from the probed host, and thus spread faster than the original localized scanning and meanwhile have a smaller variance.

2.2 Z. Duan, Y. Dong, and K. Gopalan, "DMTP: Controlling Spam through Message Delivery Differentiation," Computer Networks.

Unsolicited commercial email, commonly known as spam, has become a pressing problem in today's Internet. In this paper we re-examine the architectural foundations of the current email delivery system that are responsible for the proliferation of email spam. We argue that the difficulties in controlling spam stem from the fact that the current email system is fundamentally sender-driven and distinctly lacks receiver control over email delivery. Based on these observations we propose a Differentiated Mail Transfer Protocol (DMTP), which grants receivers greater control over how messages from different senders should be delivered on the Internet. In addition, we also develop a formal mathematical model to study the effectiveness of DMTP in controlling spam. Through numerical experiments we demonstrate that DMTP can effectively reduce the maximum revenue that a spammer can gather.

Moreover, compared to the current SMTP-based email system, the proposed email system can force spammers to stay online for longer periods of time, which may significantly improve the performance of various real-time blacklists of spammers. In addition, DMTP provides an incremental deployment path from the current SMTP-based system in today's Internet.

2.3 Z. Duan, K. Gopalan, and X. Yuan, "Behavioral Characteristics of Spammers and Their Network Reachability Properties," Technical Report TR-060602, Dept. of Computer Science, Florida State Univ.

By analyzing a two-month trace of more than 25 million emails received at a large US university campus network, of which more than 18 million are spam messages, we characterize the spammer behavior at both the mail server and the network levels. We also correlate the arrivals of spam with the BGP route updates to study the network reachability properties of spammers.

Among others, our significant findings are: (a) the majority of spammers (93% of spam only mail servers and 58% of spam only networks) send only a small number of spam messages (no more than 10); (b) the vast majority of both spam messages (91.7%) and spam only mail servers (91%) are from mixed networks that send both spam and non-spam messages; (c) the majority of both spam messages (68%) and spam mail servers (74%) are from a few regions of the IP address space (top 20 "/8" address spaces); (d) a large portion of spammers (81% of spam only mail servers and 27% of spam only networks) send spam only within a short period of time (no longer than one day out of the two months); and (e) network prefixes for a non-negligible portion of spam only networks (6%) are only visible for a short period of time (within 7 days), coinciding with the spam arrivals from these networks. In this paper, in addition to presenting the detailed results of the measurement study, we also discuss the implications of the findings for the current anti-spam efforts, and more importantly, for the design of future email delivery architecture

2.4 G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic," Proc. 15th Ann. Network and Distributed System Security Symp. (NDSS '08).

Botnets are now recognized as one of the most serious security threats. In contrast to previous malware, botnets have the characteristic of a command and control (C&C) channel. Botnets also often use existing common protocols, e.g., IRC, HTTP, and in protocol-conforming manners. This makes the detection of botnet C&C a challenging problem. In this paper, we propose an approach that uses network-based anomaly detection to identify botnet C&C channels in a local area network without any prior knowledge of signatures or C&C server addresses. This detection approach can identify both the C&C servers and infected hosts in the network. Our approach is based on the observation that, because of the pre-programmed activities related to C&C, bots within the same botnet will likely demonstrate spatial-temporal correlation and similarity.

For example, they engage in coordinated communication, propagation, and attack and fraudulent activities. Our prototype system, BotSniffer, can capture this spatial-temporal correlation in network traffic and utilize statistical algorithms to detect botnets with theoretical bounds on the false positive and false negative rates. We evaluated BotSniffer using many real-world network traces. The results show that BotSniffer can detect real-world botnets with high accuracy and has a very low false positive rate.

2.5 N. Ianelli and A. Hackworth, "Botnets as a Vehicle for Online Crime," Proc. First Int'l Conf. Forensic Computer Science.

An analysis of real-world botnets¹ indicates the increasing sophistication of bot² malware and its thoughtful engineering as an effective tool for profit-motivated online crime. Our analysis of source code and captured binaries has provided insight about:

- [11] How botnets are built
- [12] What capabilities botnets possess
- [13] How botnets are operated
- [14] How botnets are maintained and defended

The purpose of this paper is to increase understanding of the capabilities present in bot malware and the motivations for operating botnets.

3. PROPOSED METHODOLOGY

The project is to develop a spam zombie detection system, by monitoring outgoing messages. SPOT is designed based on a statistical method called Sequential Probability Ratio Test (SPRT). SPRT is a powerful statistical method that can be used to test between two hypotheses (in our case, a machine is compromised versus the machine is not compromised), as the events (in our case, outgoing messages) occur sequentially. It minimizes the expected number of observations required to reach a decision among all the sequential and non sequential statistical tests with no greater error rates. This means that the SPOT detection system can identify a compromised machine quickly. Moreover, both the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds. Consequently, users of the SPOT system can select the desired thresholds to control the false positive and false negative rates of the system.

ADVANTAGES

- SPOT is a lightweight compromised machine detection scheme, by exploring the economic incentives for attackers to recruit the large number of compromised machines.
- The SPOT detection system can identify a compromised machine quickly
- Both the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds

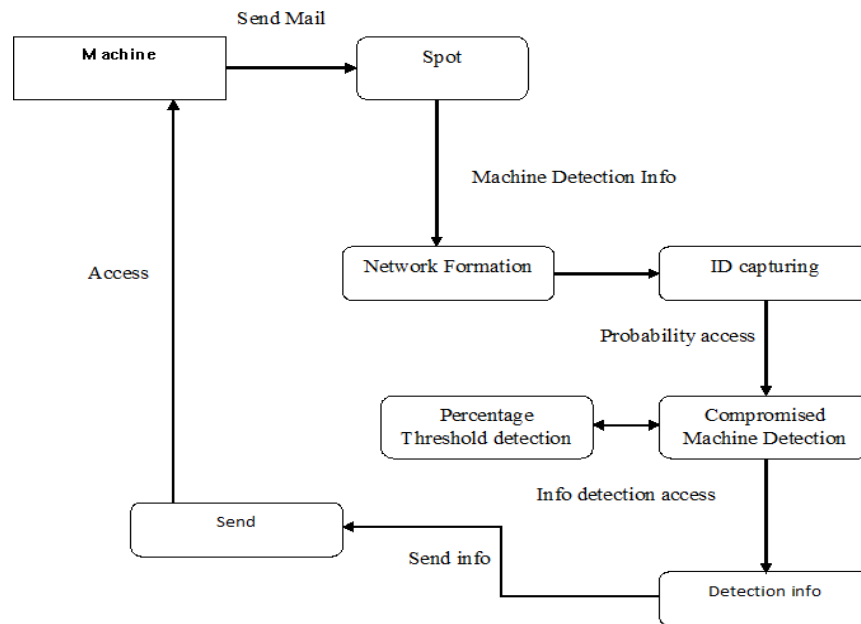
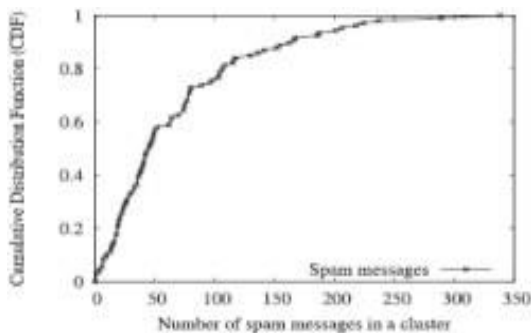


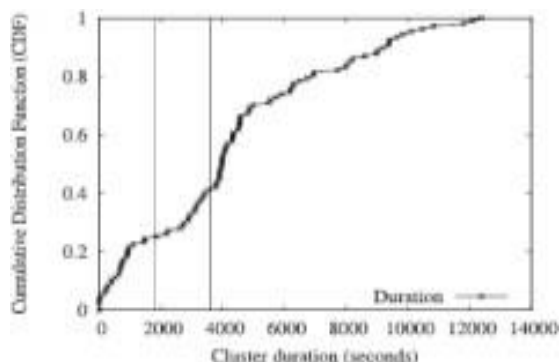
Figure 1: Spam network model

4. RESULTS AND DISCUSSIONS

A spam zombie detection system, named SPOT is developed. It works by monitoring outgoing messages. The SPOT detection system is developed to assist system administrators in automatically identifying the compromised machines in their networks. SPOT is designed based on a statistical method called Sequential Probability Ratio Test (SPRT). SPRT is a powerful statistical method that can be used to test between two hypotheses (in our case, a machine is compromised versus the machine is not compromised), as the events (in our case, outgoing messages) occur sequentially. As a simple and powerful statistical method, SPRT has a number of desirable features. It minimizes the expected number of observations required to reach a decision among all the sequential and no sequential statistical tests with no greater error rates. This means that the SPOT detection system can identify a compromised machine quickly. Moreover, both the false positive and false negative probabilities of SPRT can be bounded by user-defined thresholds. Consequently, users of the SPOT system can select the desired thresholds to control the false positive and false negative rates of the system. SPOT only needs a small number of observations to detect a compromised machine. The majority of spam zombies are detected with as little as three spam messages.



Graph 1: Distribution of the cluster duration



Graph 2: Distribution of total message in each other

5. CONCLUSION

The paper, it developed an effective spam zombie detection system named SPOT by monitoring outgoing messages in a network. SPOT was designed based on a simple and powerful statistical tool named Sequential Probability Ratio Test to detect the compromised machines that are involved in the spamming activities. SPOT has bounded false positive and false negative error rates. It also minimizes the number of required observations to detect a spam zombie. Our evaluation studies based on a two-month e-mail trace collected on the FSU campus network showed that SPOT is an effective and efficient system in automatically detecting compromised machines in a network. In addition, it also showed that SPOT outperforms two other detection algorithms based on the number and percentage of spam messages sent by an internal machine, respectively.

6. REFERENCES

- [1] P. Bacher, T. Holz, M. Kotter, and G. Wicherski, "Know Your Enemy: Tracking Botnets," [http://www.honeynet.org/papers/](http://www.honeynet.org/papers/bots) bots, 2011.
- [2] Z. Chen, C. Chen, and C. Ji, "Understanding Localized-Scanning Worms," Proc. IEEE Int'l Performance, Computing, and Comm. Conf. (IPCCC '07), 2007
- [3] R. Droms, "Dynamic Host Configuration Protocol," IETF RFC 2131, Mar. 1997.
- [4] Z. Duan, Y. Dong, and K. Gopalan, "DMTP: Controlling Spam through Message Delivery Differentiation," Computer Networks, vol. 51, pp. 2616-2630, July 2007.
- [5] Z. Duan, K. Gopalan, and X. Yuan, "Behavioral Characteristics of Spammers and Their Network Reachability Properties," Technical Report TR-060602, Dept. of Computer Science, Florida State Univ., June 2006.
- [6] Z. Duan, K. Gopalan, and X. Yuan, "Behavioral Characteristics of Spammers and Their Network Reachability Properties," Proc. IEEE Int'l Conf. Comm. (ICC '07), June 2007.

- [7] G. Gu, R. Perdisci, J. Zhang, and W. Lee, "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection," Proc. 17th USENIX Security Symp., July 2008.
- [8] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee, "BotHunter: Detecting Malware Infection through Ids-Driven Dialog Correlation," Proc. 16th USENIX Security Symp., Aug. 2007.
- [9] G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic," Proc. 15th Ann. Network and Distributed System Security Symp. (NDSS '08), Feb. 2008.
- [10] N. Ianelli and A. Hackworth, "Botnets as a Vehicle for Online Crime," Proc. First Int'l Conf. Forensic Computer Science, 2006

AN APPROACH TOWARDS COEFFICIENTS BASED LOSSY COMPRESSION AND ITERATIVE RECONSTRUCTION IN IMAGE PROCESSING

B. SATHIYASRI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

This work proposes a novel scheme for lossy compression of an encrypted image with flexible compression ratio. A pseudorandom permutation is used to encrypt an original image, and the encrypted data are efficiently compressed by discarding the excessively rough and fine information of coefficients generated from orthogonal transform. After receiving the compressed data, with the aid of spatial correlation in natural image, a receiver can reconstruct the principal content of the original image by iteratively updating the values of coefficients. This way, the higher the compression ratio and the smoother the original image, the better the quality of the reconstructed image.

Keywords: novel scheme, lossy, encrypted, spatial correlation, compression ratio

INTRODUCTION

In the realm of image processing, the pursuit of efficient compression techniques is paramount for reducing storage requirements and transmission bandwidth while preserving image quality [1]. An approach towards coefficients-based lossy compression coupled with iterative reconstruction offers a promising avenue to achieve this delicate balance between compression efficiency and reconstruction fidelity [2]. In contemporary digital ecosystems, the proliferation of high-resolution images across diverse domains necessitates robust strategies for image compression without sacrificing perceptual quality [3]. Traditional methods like JPEG compression have long served this purpose, yet emerging demands for higher compression ratios and better reconstruction accuracy drive the exploration of novel techniques. The proposed approach revolves around coefficients-based lossy compression, wherein image data is transformed into a sparse representation through techniques like Discrete Cosine Transform (DCT) [4] or Wavelet Transform. By selectively discarding coefficients deemed perceptually insignificant, the image's information content is reduced while retaining essential features. Crucially, the compression process is complemented by an iterative reconstruction mechanism [5]. Unlike traditional compression-decompression pipelines, iterative reconstruction iteratively refines the reconstructed image using feedback from previous iterations or auxiliary data. This iterative refinement enables gradual improvement in reconstruction quality, mitigating artifacts introduced during compression. By exploiting the sparsity of coefficient representations, the approach achieves significant compression ratios while preserving essential image features [6].

Iterative reconstruction ensures that the reconstructed image progressively converges towards the original, thereby minimizing perceptual distortions and artifacts. The iterative nature of the reconstruction process allows adaptability to diverse image characteristics and

compression requirements, enhancing flexibility and applicability across various domains [7]. This project explores the theoretical underpinnings, implementation strategies, and performance evaluation of the proposed approach towards coefficients-based lossy compression and iterative reconstruction. The subsequent sections delve into each aspect in detail, elucidating the methodology, experimental setup, results, and implications for practical deployment [8]. This work proposes a novel scheme for lossy compression of an encrypted image with flexible compression ratio. A pseudorandom permutation is used to encrypt an original image, and the encrypted data are efficiently compressed by discarding the excessively rough and fine information of coefficients generated from orthogonal transform. After receiving the compressed data, with the aid of spatial correlation in natural image, a receiver can reconstruct the principal content of the original image by iteratively updating the values of coefficients. This way, the higher the compression ratio and the smoother the original image, the better the quality of the reconstructed image [9].

In the digital age, where images are ubiquitous across various domains including healthcare, entertainment, surveillance, and scientific research, the need for efficient compression techniques that preserve image quality while reducing storage and bandwidth requirements has never been more pressing [10]. Traditional compression methods have long been employed to achieve this balance, yet the quest for higher compression ratios without significant loss in perceptual fidelity remains an ongoing challenge.

This work proposes a novel scheme for lossy compression of an encrypted image with flexible compression ratio. A pseudorandom permutation is used to encrypt an original image, and the encrypted data are efficiently compressed by discarding the excessively rough and fine information of coefficients generated from orthogonal transform. After receiving the compressed data, with the aid of spatial correlation in natural image, a receiver can reconstruct the principal content of the original image by iteratively updating the values of coefficients. This way, the higher the compression ratio and the smoother the original image, the better the quality of the reconstructed image. The synergy between coefficients-based compression and iterative reconstruction presents a compelling approach towards achieving high-quality image compression. By judiciously selecting and encoding coefficients, coupled with iterative refinement techniques, it's possible to strike a balance between compression efficiency and visual fidelity. This approach not only holds promise for conventional image compression tasks but also opens avenues for applications in medical imaging, remote sensing, and multimedia communication where preserving critical image details is paramount.

2. LITERATURE REVIEW

A comprehensive literature review on coefficients-based lossy compression and iterative reconstruction in image processing would encompass a range of studies, methodologies, and findings. Literature in this area often begins with seminal works on transform-based image compression, such as the pioneering JPEG standard for still image compression, which utilizes DCT. These foundational studies laid the groundwork for subsequent research into more advanced compression techniques. Wavelet transforms, due to their ability to capture both frequency and spatial information efficiently, have gained prominence in image compression. Literature reviews would likely include studies on wavelet-based compression algorithms like JPEG2000, which offers superior compression performance compared to its predecessor JPEG.

Sparse representation methods exploit the sparsity of images in transform domains like wavelet or DCT. Research in this area focuses on sparse coding techniques and dictionary learning algorithms for achieving high compression ratios while preserving image quality. Iterative reconstruction algorithms, borrowed from fields like signal processing and optimization, have been increasingly applied to image compression. Literature would cover various iterative optimization algorithms such as ISTA, FISTA, and their adaptations for image reconstruction tasks. Recent studies emphasize the importance of perceptual optimization in image compression, where reconstruction algorithms are tailored to human visual perception models. Literature would explore perceptual metrics, psychophysical studies, and perceptually-driven optimization strategies for improving visual quality in compressed images. Beyond conventional image compression, literature reviews might delve into applications of coefficients-based compression and iterative reconstruction in domains like medical imaging (e.g., MRI and CT reconstruction), remote sensing, video compression, and virtual/augmented reality.

Comparative studies evaluating the performance of different compression algorithms based on compression ratios, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual quality metrics would be essential components of literature reviews. Finally, literature reviews would likely discuss emerging trends and future research directions in coefficients-based compression and iterative reconstruction, including deep learning-based approaches, adaptive compression techniques, and the integration of machine learning for improved compression efficiency and quality. Overall, a literature review in this area would synthesize a wide range of studies, methodologies, and findings to provide a comprehensive understanding of the state-of-the-art in coefficients-based lossy compression and iterative reconstruction in image processing.

3. PROPOSED METHODOLOGY

In the proposed system, we aim to address the limitations of existing image compression techniques by introducing a novel approach that combines coefficients-based lossy compression

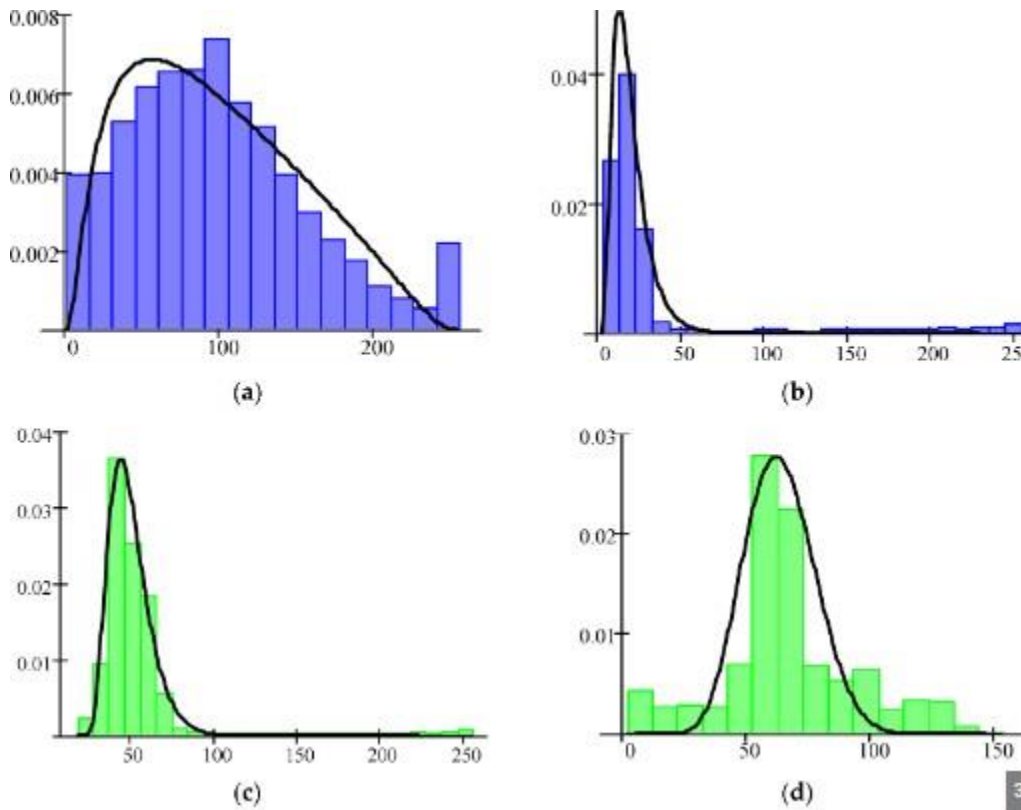
with iterative reconstruction methods. The system will utilize transform-based techniques such as Discrete Cosine Transform (DCT) or Wavelet Transform to transform the raw image data into a domain where redundancy is minimized. By retaining only the most significant coefficients, the system will discard redundant information while preserving crucial visual details. To mitigate artifacts and enhance visual fidelity during reconstruction, the system will employ iterative reconstruction algorithms. These algorithms will iteratively refine reconstructed images based on feedback from the original image or perceptual quality metrics, ensuring that important visual details are preserved and artifacts are minimized.

ADVANTAGES

1. This allows for more efficient use of storage space and reduced transmission bandwidth requirements.
2. The iterative reconstruction algorithms used in the proposed system help mitigate artifacts and preserve important visual details during image reconstruction.
3. The system allows for customization of compression performance based on specific application requirements.
4. The proposed system is versatile and can be applied to a wide range of image types, including natural photographs, medical images, satellite imagery, and graphics.
5. Incorporating perceptual optimization techniques ensures that the compressed images maintain subjective visual quality, even when significant compression is applied.

4. RESULTS AND DISCUSSIONS

Lossy compression of remote sensing data has found numerous applications. Several requirements are usually imposed on methods and algorithms to be used. A large compression ratio has to be provided, introduced distortions should not lead to sufficient reduction of classification accuracy, compression has to be realized quickly enough, etc. An additional requirement could be to provide privacy of compressed data. In this paper, we show that these requirements can be easily and effectively realized by compression based on discrete atomic transform (DAT). Three-channel remote sensing (RS) images that are part of multispectral data are used as examples. It is demonstrated that the quality of images compressed by DAT can be varied and controlled by setting maximal absolute deviation. This parameter also strictly relates to more traditional metrics as root mean square error (RMSE) and peak signal-to-noise ratio (PSNR) that can be controlled. It is also shown that there are several variants of DAT having different depths. Their performances are compared from different viewpoints, and the recommendations of transform depth are given. Effects of lossy compression on three-channel image classification using the maximum likelihood (ML) approach are studied. It is shown that the total probability of correct classification remains almost the same for a wide range of distortions introduced by lossy compression, although some variations of correct classification probabilities take place for particular classes depending on peculiarities of feature distributions. Experiments are carried out for multispectral Sentinel images of different complexities.



Graph.1 shown the discrete automatic transform based lossy compression

5. CONCLUSION

In conclusion, the proposed approach of combining coefficients-based lossy compression with iterative reconstruction methods presents a promising avenue for advancing the field of image compression. By addressing the limitations of existing techniques and leveraging innovative algorithms, the proposed system offers several key advantages, including higher compression ratios, preservation of visual fidelity, customizable compression performance, applicability to various image types, perceptual optimization, potential for real-time compression, compatibility, and scalability. Through the systematic integration of coefficients-based compression and iterative reconstruction, the proposed system achieves a delicate balance between compression efficiency and visual quality. This balance is essential for applications spanning diverse domains such as healthcare, remote sensing, multimedia, and digital communications, where preserving critical image details is paramount.

6. REFERENCES

- [1] Sayood, K. (2017). Introduction to Data Compression (5th ed.). Morgan Kaufmann.
- [2] Wallace, G. K. (1992). The JPEG Still Picture Compression Standard. IEEE Transactions on Consumer Electronics, 38(1), xviii-xxxiv.
- [3] Taubman, D., & Marcellin, M. W. (2002). JPEG2000: Image Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers.
- [4] Mallat, S. (1999). A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press.
- [5] Donoho, D. L. (2006). Compressed sensing. IEEE Transactions on Information Theory, 52(4), 1289-1306.
- [6] Figueiredo, M. A. T., & Nowak, R. D. (2003). An EM algorithm for wavelet-based image restoration. IEEE Transactions on Image Processing, 12(8), 906-916.
- [7] Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1-4), 259-268.
- [8] Blau, Y., & Michaeli, T. (2018). The Perception-Distortion Tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6228- 6237).
- [9] Toderici, G., & Vincent, D. (2014). Variable Rate Image Compression with Recurrent Neural Networks. arXiv preprint arXiv:1511.06085.
- [10] Han, S., Mao, H., & Dally, W. J. (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. International Conference on Learning Representations (ICLR).

AN ATTRIBUTE BASED CLUSTER MODEL FOR FEATURES SUBSET SELECTION ALGORITHM IN WEB MINING

N. NANDHINI

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this project. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high dimensional image, microarray, and text data, demonstrate that FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

Keywords: feature, FAST algorithm, subset, cluster, Naive Bayes, dimensional image

1. INTRODUCTION

The feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications[1]. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories[2]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is

large[5]. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper.

It mainly focuses on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods[6]. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large[8]. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms are applied in the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph theoretic methods have been well studied and used in many applications. The results have, sometimes, the best agreement with human performance. The general graph theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors[9]. The result is a forest and each tree in the forest represents a cluster. In this study, apply graph theoretic clustering methods to features.

In particular, it adopts the minimum spanning tree (MST) based clustering algorithms, because it do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, this propose a attribute based Fast clustering based feature Selection algorithm (ABFAST).The ABFAST algorithm works in two steps. In the first step, features are divided into clusters by using feedback verification clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features.

Features in different clusters are relatively independent, the clustering based strategy of FAST has a high probability of producing a subset of useful and independent features. The proposed feature subset selection algorithm ABFAST was tested upon 35 publicly available image, microarray, and text data sets. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well known different types of classifiers.

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other

three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods, by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms, employed the distributional

Clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

2. LITERATURE REVIEW

Finding scholarly information on the World Wide Web can be very frustrating. There is no way to search through a large selection of only scholarly sites with the current Web search tools. The existing search tools provide search algorithms that sift through millions of Web pages with no way to limit the search to a category of Web sites. Nobody seems to know how to do any automatic filtering for quality of Web sites. However, librarians have been doing quality filtering of materials for many years, but "no one seems conscious of the standards carefully developed by information professionals over the past century" (Collins 1996, 122).

In the print world, the academic library performs this filtering function by providing patrons with a subset of print works pertaining to academia. This selection role is filled by library staff members using either explicit or tacit criteria to select individual works. Some sites, such as the Internet Public Library (<http://www.ipl.org>), attempt to select scholarly sites. However, because of the rapid introduction of new documents on the World Wide Web, a human cannot keep up and the resource is quickly outdated.

In order to handle the vast number of documents on the Web, an automated selection system is needed. First, the criteria used by academic librarians to select print works will be examined. These criteria can be translated into equivalent criteria for Web pages. A Web robot can then be designed to determine these criteria for a page. After creating a training set of examined Web pages with their selection decisions, data mining techniques can be used to create a classification model that will be a quality filter for Web pages.

Most of the existing works are motivated by a commonly performed task in the biomedical domain, that of constructing a systematic review. Authors of systematic reviews seek to identify as much as possible of the relevant literature in connection with some aspect of medical practice, typically a highly specific clinical question. The review's authors assess, select, and synthesize the evidence contained in a set of identified documents, to provide a "best currently known" summary of knowledge and practice in that field.

A variety of organizations provide central points of call for systematic reviews, including the Cochrane Collaboration,² the largest of these efforts, and the Agency for Healthcare Research and Quality, AHRQ.³ The collections used as the source material are already large, and continue to grow. For example, as at end of 2009, MEDLINE, the largest of the available collections, contained more than 19 million entries, with more than 700,000 citations having been added during the year. To construct each systematic review, a complex Boolean search query is used to retrieve a set of possibly relevant documents (typically in the order of one to three thousand), which are then comprehensively triaged by multiple assessors.

Recently, hierarchical clustering has been adopted in word selection in the context of text classification. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira or on the distribution of class labels associated with each word by Baker and McCallum.

As distributional clustering of words is agglomerative in nature, and result in sub-optimal word clusters and high computational cost, it shows a new information-theoretic divisive algorithm for word clustering and applied it to text classification. It proposed to cluster features using a special metric of Barthelemy distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on Barthelemy distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

The Boolean query might consist of as many as several dozen query lines, each describing a concept with fielded keywords, MESH headings (a hierarchical taxonomy of medical terms), metadata, free-text term expansions, and Boolean operators aggregating and balancing the concepts shows the structure of one such query; this expression would be one small component of a typical complex query of (say) 50 clauses.

3. PROPOSED METHODOLOGY

Feature selection process is the vital one in the architecture of data retrieval process in web mining. It involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find the multiple attribute based feature selection, the effectiveness is related to the quality of the mechanism designed to perform the feature selection. Based on the proposed idea, attribute based fast clustering-based feature selection algorithm (ABFAST) is proposed and going to experiments with different parameter set.

The ABFAST algorithm works in three steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target cluster classes is selected from each cluster to form an attribute based classes. Features in different clusters are relatively either dependent or independent, the clustering based strategy of ABFAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the ABFAST algorithm are evaluated through an empirical study. The third step is feature selected data are verified with the true database server that are driven from the attribute based cluster of classes. So this system shows the better performance than the existing FAST, FCBF and Btree based systems.

ADVANTAGES

- Good feature subsets contain features highly correlated with (predictive of) the class, both uncorrelated data match with each other that improves the data feature selection robustness.
- The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.
- The improved graph based MST with grouping options handles well the user query than any other existing systems.
- Feedbacks and user data precision values are taken into account for better dataset resultant and reduce the filtration process.
- ABFAST returns the hash data table format feature selection of each and every class with high relevant of data clusters.

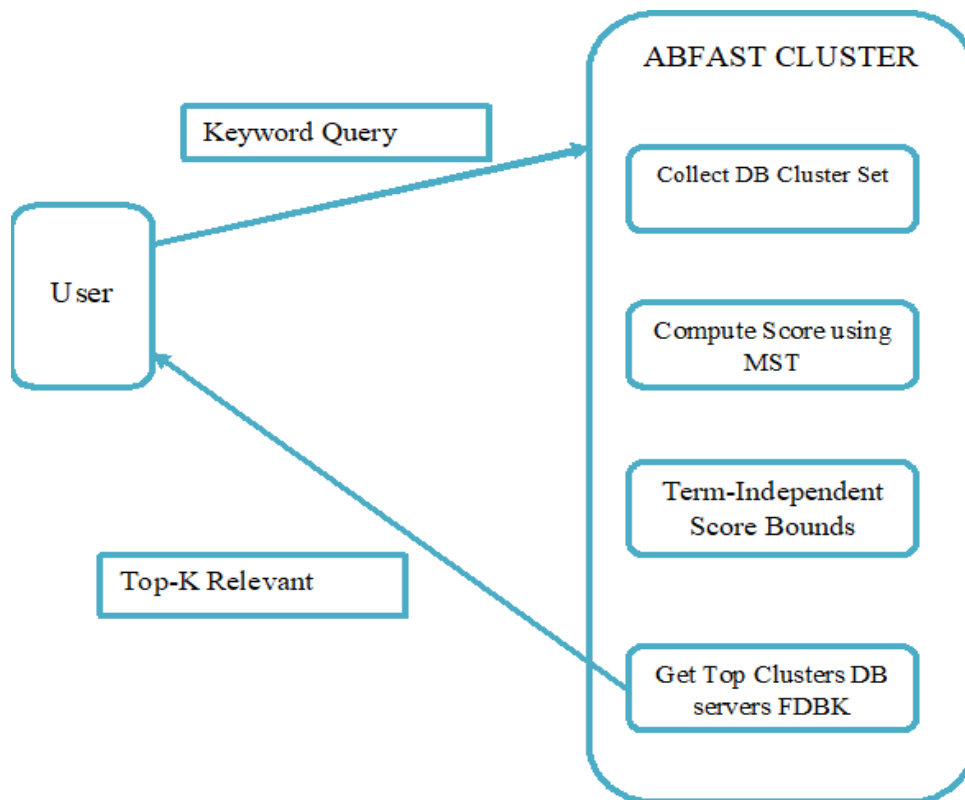


Figure 1: Architecture of proposed algorithm

4. RESULTS AND DISCUSSIONS

Algorithm FAST obtains better, equal, and worse performance than other five feature selection algorithms, respectively. The measure can be the proportion of selected features, the runtime to obtain a feature subset, and the classification accuracy, respectively. Experimental procedure In order to make the best use of the data and obtain stable results, a $(M = 5) \times (N = 10)$ -cross-validation strategy is used. That is, for each data set, each feature subset selection algorithm and each classification algorithm, the 10-fold cross-validation is repeated $M = 5$ times, with each time the order of the instances of the data set being randomized. This is because many of the algorithms exhibit order effects, in that certain orderings dramatically improve or degrade performance.

1) Generally the individual evaluation based feature selection algorithms of FAST, FCBF and ReliefF are much faster than the subset evaluation based algorithms of CFS, Consist and FOCUS-SF. FAST is consistently faster than all other algorithms. The runtime of FAST is only 0.1% of that of CFS, 2.4% of that of Consist, 2.8% of that of FOCUS-SF, 7.8% of that of ReliefF, and 76.5% of that of FCBF, respectively. The Win/Draw/Loss records show that FAST outperforms other algorithms as well.

2) For image data, FAST obtains the rank of 1. Its runtime is only 0.02% of that of CFS, 18.50% of that of ReliefF, 25.27% of that of Consist, 37.16% of that of FCBF, and 54.42% of that of

FOCUS-SF, respectively. This reveals that FAST is more efficient than others when choosing features for image data.

3) For microarray data, FAST ranks 2. Its runtime is only 0.12% of that of CFS, 15.30% of that of Consist, 18.21% of that of ReliefF, 27.25% of that of FOCUS-SF, and 125.59% of that of FCBF, respectively.

4) For text data, FAST ranks 1. Its runtime is 1.83% of that of Consist, 2.13% of that of FOCUS-SF, 5.09% of that of CFS, 6.50% of that of ReliefF, and 79.34% of that of FCBF, respectively. This indicates that FAST is more efficient than others when choosing features for text data as well.

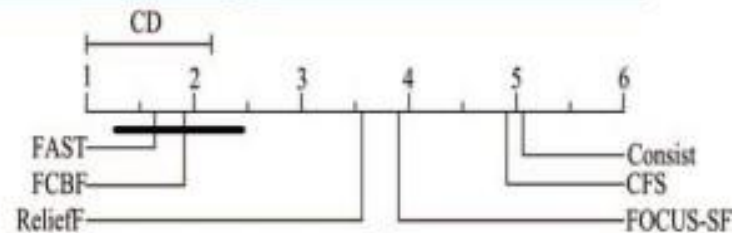


Figure 2: Selection of ABFAST

5. CONCLUSION

In this proposed technique, ABFAST have presented a novel clustering based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features by verify it with the true data verification by getting the feedback from the users. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. It also have compared the performance of the proposed algorithm with those of the five well known feature selection algorithms FCBF, CFD, Fuzzy sets based clustering on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features comparing to the existing systems.

6. REFERENCES

- [1] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.
- [2] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [3] R. Chaiken, B. Jenkins, P.-A° . Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou, "Scope: easy and efficient parallel processing of massive data sets," *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp.1265–1276, 2008.

- [4] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Mad skills: new analysis practices for big data,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1481–1492, 2009.
- [5] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, “Building a high-level dataflow system on top of Map-Reduce: the Pig experience,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1414–1425, 2009.
- [6] H.-c. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, “Map-Reduce-Merge: simplified relational data processing on large clusters,” in Proceedings of 2007 ACM SIGMOD International Conference on Management of Data, 2007, pp. 1029–1040.
- [7] “Apache Hadoop,” <http://hadoop.apache.org/>, accessed: August 1, 2013.
- [8] Z. F. Zeng, B. Wu, and T. T. Zhang, “A multi-source message passing model to improve the parallelism efficiency of graph mining on MapReduce,” in Proceedings of 2012 IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012, pp. 2019–2025.
- [9] “Stanford large network dataset collection,” <http://snap.stanford.edu/data/>, accessed: August 1, 2013.
- [10] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, 2nd ed. Morgan Kaufmann, 2006.

APPROACH OF SECURED FILE DETECTION USING DATA MINING

M. BHUVANESHWARI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Detecting insider attacks continues to prove to be one of the most difficult challenges in securing sensitive data. Decoy information and documents represent a promising approach to detecting malicious masqueraders; however, false positives can interfere with legitimate work and take up client time. We propose generating foreign language decoy documents that are sprinkled with untranslatable enticing proper nouns such as company names, hot topics, or apparent login information. Our goal is for this type of decoy to serve three main purposes. First, using a language that is not used in normal business practice gives real clients a clear signal that the document is fake, so they waste less time examining it. Second, an attacker, if enticed, will need to exhilarate the document's contents in order to translate it, providing a cleaner signal of malicious activity. Third, we consume significant adversarial resources as they must still read the document and decide if it contains valuable information, which is made more difficult as it will be somewhat scrambled through translation. In this paper, we expand upon the rationale behind using foreign language decoys. We present a preliminary evaluation which shows how they significantly increase the cost to attackers in terms of the amount of time that it takes to determine if a document is real and potentially contains valuable information or is entirely bogus, confounding their goal of exhilarating important sensitive information.

Keyword: data, decoy information, attack, sensitive information, file

INTRODUCTION

There are various approaches available to ensure file data security, such as encryption tools like 'aescrypt' in Linux or integrated encryption application software or disk encrypter. But each one has its own inherent disadvantages, rendering them being less frequently used. These approaches are generally cumbersome and inconvenient to the users. [1] Therefore, there is a need for a mechanism/system which can ensure reliable and efficient file data security in a transparent and convenient manner. We have taken this as a challenge and tried to solve the problem of file data security by integrating proposed Secure File System into the kernel itself. In this paper we have introduced a new architecture of file data security in Linux, the Secure File System (SFS). [2] SFS works on the layer of Virtual File System (VFS) and brings the encryption services into the kernel space. If a file needs to be encrypted, it will be rerouted from VFS to SFS. Operating system having SFS embedded into its kernel, provide file data security as one of its basic functionality to all applications. SFS is fully compatible with all underlying storage file systems and all applications.

There have been different approaches used, to solve the file data security problem. Most of the solutions provided works in user space. [3] The simple and naive approach used by many people to secure their file data is to use common utilities like 'crypt' or 'aescrypt'. These utilities take the filename and the password as inputs and produce the encrypted file. This type of utility

is good for limited use only, as it is very cumbersome and manual. Second approach is integrating encryption engine in application software itself, where each program that is to manipulate sensitive data has built-in cryptographic facilities. But the disadvantage here is that all application should use the same encryption engine and any changes in one will require changes in all. [4]The third approach is to use commercially available disk controllers with embedded encryption hardware that can be used to encipher entire disks or individual file blocks with a specified block. It suffers from the fact that key needs to be shared among users, whose data reside on the disk because entire disk is protected as a single entity.[5] It is good for single user system but for multi-user system the key protecting the data needs to be shared between different users. So we have seen that each one of the approaches described above; has its own inherent disadvantages, rendering them less frequently used. These approaches are generally cumbersome and inconvenient to the users.

Therefore, there is a need for a mechanism/system which can ensure reliable and efficient file data security in a transparent and convenient manner. We focused on this issue and proposed SFS that solves the file data security problem. [6]We considered various places where this mechanism/system can be placed to fulfill its requirement in the best possible way. The considered places include user space, device layer level, and kernel space. We are of the opinion that the file data security should be provided as a functionality of operating system, therefore we have decided to push the encryption services into the Linux kernel space mounted beneath the virtual file system. There has been a lot of development taken place since the time when MS DOS device driver was used to encrypt the entire partition. Nowadays we have several cryptographic file systems available, which we have briefly described.

LITERATURE REVIEW

Fueled by the omnipresence of event logs in transactional information systems (cf. WFM, ERP, CRM, SCM, and B2B systems), process mining has become a vivid research area [5,6]. Until recently, the information in these event logs was rarely used to analyze the underlying processes. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs, i.e., the basic idea of process mining is to diagnose processes by mining event logs for knowledge. So far, process mining research has focussed on process discovery and process improvement. In this paper, we focus on the application of process mining to security issues.

When considering an enterprise information system, security plays a role at different levels, i.e., from the level of UNIX processes to the level of interior organizational business processes. Security policies may refer to things ranging from cryptography and role-based access control to auditing and the four eyes principle. Security violations may be conducted by hackers but also by white-collar criminals (cf. the discussions on "corporate governance" following the Enron and Parmalat scandals). Literature on security can be split into computer security and auditing. Although computer security and auditing are at very different levels, the absence or presence of certain behavioral patterns may indicate security violations. Therefore, audit trails can be use ful.

Fortunately, many enterprise information systems store relevant events in some structured form. For example, workflow management systems typically register the start and completion of activities. ERP systems like SAP log all transactions, e.g., users filling out forms, changing documents, etc. Business to-business (B2B) systems log the exchange of messages with other parties. Call center packages but also general-purpose CRM systems log interactions with customers. These examples show that many systems have some kind of event log often referred to as "audit trail", "history", "transaction log", etc. The event log typically contains information about events referring to an activity and a case. The case (also named process instance) is the "thing" which is being handled, e.g., a customer order, a job application, an insurance claim, a building permit, etc. The activity (also named task, operation, action, or work-item) is some operation on the case. Typically, events have a timestamp indicating the time of occurrence. Moreover, event logs typically also contain information on the actor, i.e., person or system component, executing or initiating the event. We will refer to such an actor as the originator or performer. Based on this information several tools and techniques for process mining have been developed. Process mining is useful for at least two reasons.

PROPOSED METHODOLOGY

In the proposed system, provide security for downloading each file of the clients. When the clients send requests to admin, the admin to make the decision for downloading files then Admin receives the request than he will send the security code to the clients' request. The admin provides code for one time per file. If another time needs to download the file for clients, send the request to admin. This system provides security for each file. It does will be make a relationship between client and admin. This is a tool of managing and providing security of files.

ADVANTAGES

- User friendly.
- Provide relationship between clients and admin.
- Provide security for all files because single file download for one time at using a single password
- Cannot be download the file without admin permission
- Two levels of security provide for all files.

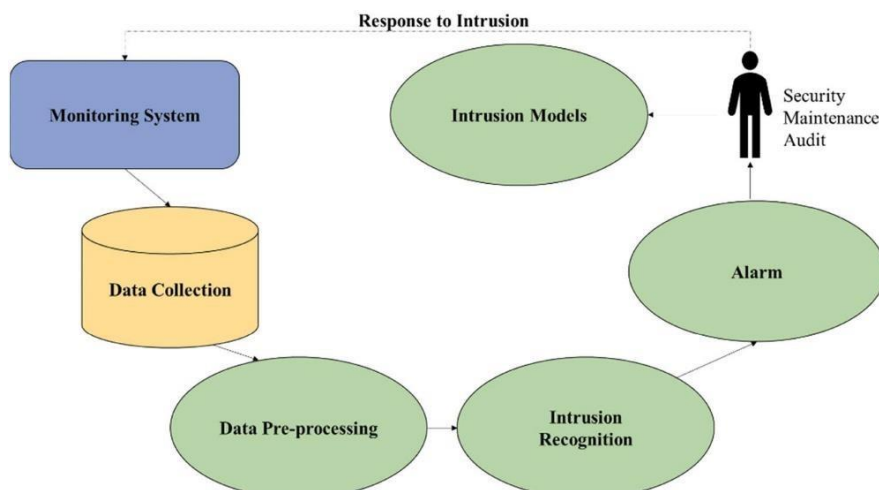


Figure 1 : Architecture of a basic intrusion detection system.

RESULTS AND DISCUSSIONS

The scope of the work is to extract the useful information from large amount of data and store at cloud in secure fashion and then make inferences required by the organization. But the predictions that are generated as a result of mining should be secure from any kind of interception. In this sense, steganography is the best option for sending information secretly because it hides the existence of secret message and provides more security. The security module which is used is image steganography as images are the most popular because of their frequency on the Internet. So the prime focus is to increase the capacity to provide better security during transmission. The Apriori algorithm is the most general and widely used association rule mining algorithm designed to operate on databases containing transactions. Apriori uses breadth-first-search and a tree structure to count candidate item sets efficiently. It uses an iterative method called layer search to generate $(k+1)$ item sets from k item sets. A k -item set is frequent only if all of its sub-item sets are frequent. This process iterates until no more frequent k -item sets can be generated for some k . This is the essence of the Apriori algorithm. Apriori Algorithm is based on rule parameters – support, confidence and number of cycles used, but these rule measures are not considered for Predictive Apriori Algorithm. The default number of best rules in Apriori Algorithm and in Predictive Apriori Algorithm is respectively 10 and 100. In Apriori Algorithm, the number of best rules generated is independent of the number of instances and attributes but are dependent on the value minimum support taken. In Predictive Apriori Algorithm, the best rules depend on the dataset being used and the number of selected attributes. Greater the number of best rules, greater the expected accuracy. A rule is added if the expected predictive accuracy of the particular rule is among 'n' number of

best rules and it is not a part of another rule with at least the same expected predictive accuracy

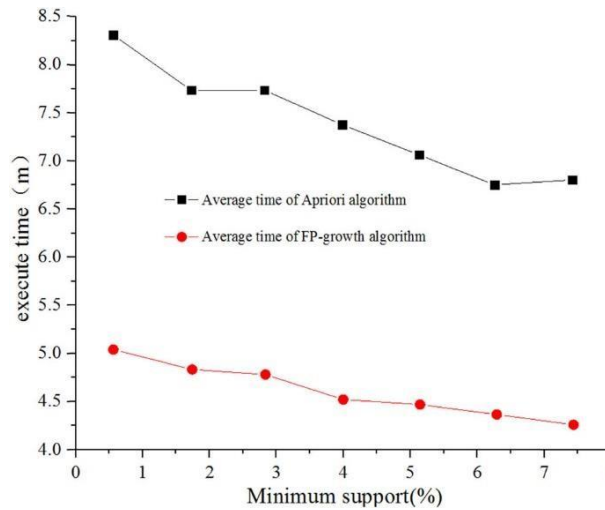


Fig 2 : Average mining time of the two algorithms.

CONCLUSION

This project **Secured file downloading system** was successfully completed. And the three phases namely system analysis, system design, system implementation are thoroughly tested with live data. The objective of the project is to implement the overall employee metric of the firm through online/internet. The project has a very vast scope in future. The project can be implemented on intranet in future. Project can be updated in near future as and when requirement for the same arises, as it is very flexible in terms of expansion. With the proposed software of database Space Manager ready and fully functional the students are now able to manage and hence run the entire work in a much better, accurate and error free manner. The following are the future scope for the project.

REFERENCES

- [1] Peter Guttmann, University of Auckland, New Zealand. The secure file system (sfs) for dos/windows. <http://www.cs.auckland.ac.nz/pgut001/sfs/index.html>, September 1996.
- [2] Alex Tormasov. The tordisk project. <http://www.cs.auckland.ac.nz/pgut001/sfs/index.html>, November 1997.
- [3] Phil Zimmerman. Pgp home page. <http://web.mit.edu/pgp/>.
- [4] Matt Blaze. A cryptographic file system for unix. In First ACM Conference on Communications and Computing Security, pages 33–43, Fairfax, VA, November 1993.
- [5] Matt Blaze. Key management in an encrypting file system. Boston, MA, June 1994.
- [6] University of Salerno. Tcfs home page. <http://www.globenet.it/ermmau/tcfs/index.html>, April 1997.
- [7] Parallel Data Lab. NASD home page. <http://www.pdl.cs.cmu.edu/NASD/>.

[8] Albert Alexandrov, Maximilian Ibel, Klaus Schauer, and Chris Scheiman. Extending the operating system at the user level: the ufo global file system. In Proceedings of the USENIX Annual Technical

Conference, pages 77–90, 1997.

[9] Peter Braam and Philip Nelson. Removing bottlenecks in distributed filesystems. In

Proceedings of the 5th Annual Linux Expo, pages 131–139, Raleigh, North Carolina, May 1999.

[10] Scott Guthery and Timothy Jurgensen. Smart Card Developer's Kit. Macmillan Technical Publishing, 1998.

AUTOMATED CHATBOT ASSISTANT SYSTEM USING ARTIFICIAL INTELLIGENCE

E. AARTHIKA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Chatbots are gradually becoming more sophisticated as they can now adapt to new AI features with ease. You can also expect them to recognize user intent efficiently, decode the mood of users instantly, and drive the flow of conversations in tune with customer's emotions. And sentiment analysis is one such feature that makes bots even more powerful in terms of understanding the emotion in the customer messages.

In fact, 64% of businesses believe that chat-bots can help them provide a more customized support experience for their customers. You can always leverage the chatbot sentiment analysis feature to easily know if customers are having a good experience with your chatbots. This is how AI-powered bots can help you engage customers better and improve their experience with your brand. Sentiment analysis is a subfield of machine learning (ML) and natural language processing (NLP) that can help chatbots determine emotions from textual data. It's one of the key chatbot features that is used to analyze customer data by mining thoughts, opinions, or sentiments.

In this paper, a chatbot is developed as academic assistant for institutions. Here all the details regarding department, course, fees structure, hostel and placement can be obtained from the chatbot assistance for the query given by users. Here, the chatbot will autocorrect the users' misspelled words taken from dataset given for categories like department, course, etc. Moreover, for a single given word or a phrase, the chatbot replaces it with the correct question patterns and answers are displayed for that question.

Keyword: Chatbot, Machine learning, Natural Language Processing, Artificial Intelligence

INTRODUCTION

Chatbot Assistants for Education offer a wide range of functionalities tailored to support various aspects of the educational journey. From assisting students with their queries and providing personalized learning recommendations to aiding educators in administrative tasks and facilitating communication[1] within educational institutions, these virtual assistants have the potential to revolutionize the way we teach and learn. In today's rapidly evolving educational landscape, the integration of technology has become paramount in enhancing learning experiences. One of the most promising technological advancements in this regard is the development of Chatbot Assistants for Education[2]. These intelligent virtual assistants leverage natural language processing (NLP), artificial intelligence (AI), and machine learning (ML) algorithms to interact with students, educators, and administrators in a personalized and efficient manner[3].
Personalized Learning Support:

Chatbot Assistants can offer personalized assistance to students based on their individual learning needs and preferences. Whether it's clarifying doubts, providing additional resources, or offering practice exercises[4], these virtual assistants can adapt to the unique requirements of each learner, thereby fostering a more effective learning experience. **24/7 Accessibility:** Unlike human educators who may have limited availability, Chatbot Assistants are accessible round-the-clock. This ensures that students can seek assistance whenever they encounter challenges or require guidance, regardless of the time of day or their geographical location. **Administrative Efficiency:** Chatbot Assistants can streamline administrative tasks for educators and administrators[5], such as scheduling classes, managing grades, and sending out announcements. By automating routine administrative processes, these virtual assistants enable educators to focus more on teaching and mentorship[6]. **Seamless Communication:** Chatbot Assistants facilitate seamless communication within educational institutions by providing a centralized platform for students, educators, and administrators to interact. Whether it's answering frequently asked questions, facilitating group discussions, or notifying stakeholders about important updates, these virtual assistants promote effective communication and collaboration. **Data-driven Insights:** By analyzing interactions and user data, Chatbot Assistants can generate valuable insights into student engagement, learning patterns, and areas of improvement. Educators can leverage these insights to tailor their teaching strategies[7], identify at-risk students, and provide targeted interventions, thereby enhancing overall learning outcomes[8].

Customer experience plays a vital role in any organization or institution to grow and serve its users. Organizations now require round-the-clock service offerings to their users, keeping them informed regarding their services. One potential solution to this kind of offering is a chatbot. Chatbots can never replace human interactions, but they can reduce workloads drastically. According to the education literacy report by United Nations International Children's Emergency Fund (UNICEF), globally youth literacy rate has improved and it has increased from 83% to 91%. There are currently 4.66 billion active internet users worldwide. 65.6% of the entire world's population has internet access and 81% of US adults go online on a daily basis. This has also given rise and access to online education. College students must recurrently navigate a set of challenging tasks, such as building a graduation plan, learning about majors, retrieving valuable information about courses including the number of credits, students learning outcomes, sections and associated schedules, assigned professors, classroom locations among others. Without assistance and support, many students fail to raise these challenges. The lack of assistance and support causes an estimated 10% to 20% of college drop-out each year, with higher rates among low-income and first-generation college students. Common efforts to address the lack of assistance and support have supported students with additional individual counselor outreach or through automated, customized text message-based outreach [7]. However, scaling these efforts requires significant resources because of the time needed for counsellors to address the specific questions and personal needs of each student [8].

With the global growth and acceptance of digital technology, our daily life is changing and replacing things around us quickly, similarly, conversational agents also called chatbots are revolutionizing businesses. As the NLP (Natural Language Processing) and AI (Artificial Intelligence) techniques are easily available and the concepts like conversational agents are now a reality, many small and medium-sized organizations are showing wider acceptability and implementation. The users are also preferring chatbots due to round-the-clock availability, reliability, accessibility, and instantaneous accurate response. As a consequence, world-class leading Information Technology & Communication (ITC) companies have launched their own conversational agents' system/chatbot time to time such as Google, Microsoft, Facebook, Amazon, IBM to name a few.

LITERATURE REVIEW

The idea a good person may remotely accessibility their own data throughout on demand mode. plus the diversity of an application requirements, users will probably want to be able to entry along with share each other's authorized details fields to be able to achieve productive benefits, which delivers new security and privacy challenges with regard to the cloud storage. The concept of cloud computing is acquiring a lot of recognition from both academic and business worlds. The main idea is to make claim obtainable on adoptable execution framework primarily located in the internet. In cloud computing, users can outsource their computation and storage to servers using internet. Cloud computing's multi-client feature, which provides distinct, seclusion and access control challenges, because of sharing of resources among unbelievable clients. In Cloud computing, a third party is responsible for providing, processing power, memory space and implementation support etc. Cloud database is maintained by third party Cloud provider, so user hesitates to keep his data at cloud database. **Components of Digital Literacy's**: Human beings are tasked with making sense of the external world. Defining literacy in relation to the tools used to encode and decode the symbols involved can therefore be difficult. Theorists must ensure that literacy is not defined so broadly so as to include almost any activity, but not so narrowly that it is almost impossibly prescriptive. 'Literacy' must apply equally to instant, informal electronic communications and the creation of formal, written, laboriously-created documents that have been handed down through generations. That is to say a balance must be found so that technologies used in the past as well as those that will be used in the future for reading and writing are included within definitions of 'literacy'. If this cannot be achieved, then it may be best to use a different term or way of framing the concept. Digital literacy as a component of life skills Modern life skills encompass an intricate system of knowledge, skills, abilities, and motivational factors that must be developed according to the needs of their specific domains.

Current and Future Prospective: A chatbot is an instant messaging account that able to provide services using instant messaging frameworks with the aim of providing conversational services to users in an efficient manner. A chatbot is fast with less confusing web and mobile application which is easy to install as there is no need to have installation packages. These

packages are easy to manage and distribute. Chatbots are totally different from the human accounts as they do not have any online status or last seen timestamps nor initiate the conversations and calls with any other accounts. Figure.1 shows some types of chatbots which are in used in different domains. The overall chatbot Architecture is shown in Figure.2. Intent classification module identifies the intent of user message. Entity recognition module extracts structured bits of information from the message. The candidate response generator is doing all the domain-specific calculations to process the user request. The response selector just scores all the response candidate and selects a response which should work better for the user.

Chatbot Programming Challenges: There are a lot of challenges which are associated with chatbots. Some of them are as follows. A. Natural language processing The first and foremost challenge of the chatbot is to handle NLP issue by mastering their syntax. If we ask them that "what's the weather?". You will get an answer but what if we ask "Could you check the weather?" you might not get the proper answer. Such type of programming issues falls in natural 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) 21 - 23 Dec 2017, Dhaka, Bangladesh language processing category which is a key focus for the companies like Facebook, Google with Deep Text and Syntax Net respectively.

Machine learning : Getting NLP is one aspect of designing and development of Chatbots while Machine Learning is another aspect of the Chatbot design and development. Our computer systems should be able to learn the correct response which can be achieved with efficient programming with AI concepts.

PROPOSED METHODOLOGY

Implement advanced NLP technologies to improve the chatbot's natural language understanding and processing capabilities. Increased versatility, providing users with a more comprehensive and valuable experience. More accurate comprehension of user inputs, leading to better responses and user satisfaction. Strengthen data security and privacy measures to address concerns and ensure compliance with regulations. Users receive more satisfactory and nuanced responses, enhancing overall user experience. Implement a systematic user feedback mechanism to gather insights and continuously iterate on the chatbot's performance. Streamline the process of onboarding users and provide effective training to enhance user understanding. Improve the design and user interface of the chatbot for a more intuitive and user-friendly experience. Address integration challenges to ensure seamless connectivity with existing systems or databases.

ADVANTAGE

- 1 The proposed system expands the chatbot's capabilities to handle a broader range of tasks and queries.
- 2 Advanced NLP technologies improve the chatbot's understanding of user inputs.
- 3 The system incorporates enhanced data security measures, addressing concerns and ensuring compliance.

- 4 Strategies are implemented to improve the chatbot's handling of complex or ambiguous queries.
- 5 A systematic user feedback mechanism is implemented to gather insights for continuous improvement.

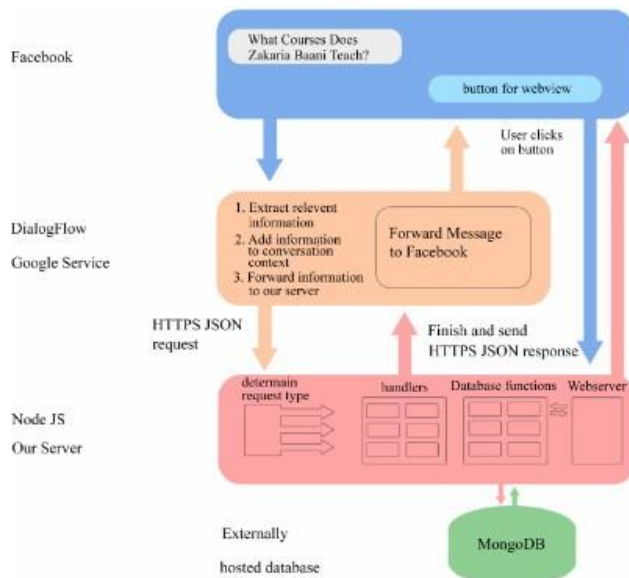
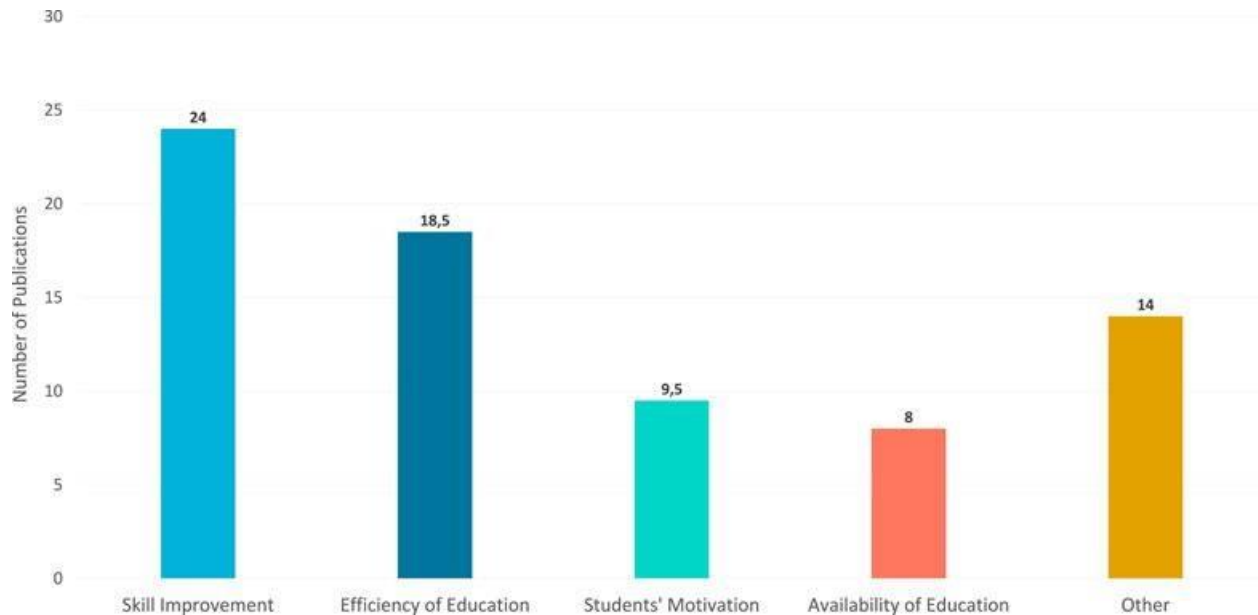


Figure 1: Chatbot collaboration workflow

RESULTS AND DISCUSSIONS

Regarding RQ1, we extracted implementation objectives for chatbots in education. By analyzing the selected publications we identified that most of the objectives for chatbots in education can be described by one of the following categories: Skill improvement, Efficiency of Education, and Students’ Motivation . First, the “improvement of a student’s skill” (or *Skill Improvement*) objective that the chatbot is supposed to help with or achieve. Here, chatbots are mostly seen as a learning aid that supports students. It is the most commonly cited objective for chatbots. The second objective is to increase the *Efficiency of Education* in general. It can occur, for example, through the automation of recurring tasks or time-saving services for students and is the second most cited objective for chatbots. The third objective is to increase *Students’ Motivation*. Finally, the last objective is to increase the *Availability of Education*. This objective is intended to provide learning or counseling with temporal flexibility or without the limitation of physical presence. In addition, there are other, more diverse objectives for chatbots in education that are less easy to categorize. In cases of a publication indicating more than one objective, the publication was distributed evenly across the respective categories.



Graph.1 shown the review on chatbot in education

Given these results, we can summarize four major implementing objectives for chatbots. Of these, *Skill Improvement* is the most popular objective, constituting around one-third of publications (32%). Making up a quarter of all publications, *Efficiency of Education* is the second most popular objective (25%), while addressing *Students' Motivation* and *Availability of Education* are third (13%) and fourth (11%), respectively. *Other* objectives also make up a substantial amount of these publications (19%), although they were too diverse to categorize in a uniform way. Examples of these are inclusivity or the promotion of student teacher interactions .

CONCLUSION

The development and implementation of a Chatbot Assistant for Education offer immense potential to revolutionize the teaching and learning landscape. Through the integration of advanced technologies such as artificial intelligence, natural language processing, and machine learning, these virtual assistants can provide personalized support, streamline administrative tasks, facilitate communication, and offer valuable insights to both students and educators. Through their experiments, it became evident that the careful application of prompt techniques can significantly enhance the quality of the LLM's outputs. The persona and template techniques, for instance, provided a more contextually relevant and educationally aligned set of responses. Similarly, advanced engineering methods like the few-shot and chain-of-thought prompts showcased the depth and adaptability of LLMs, allowing for a broader range of question generation that aligns with educational standards.

REFERENCES

- [1] Chen, L.; Chen, P.; Lin, Z. Artificial Intelligence in Education: A Review. *IEEE Access* 2020, 8, 75264–75278. [Google Scholar] [CrossRef]
- [2] Chien, S.Y.; Hwang, G.J. A research proposal for an AI chatbot as virtual patient agent to improve nursing students’ clinical inquiry skills. *ICAIE 2023*, 2023, 13. [Google Scholar]
- [3] Zhang, K.; Aslan, A.B. AI technologies for education: Recent research future directions. *Comput. Educ. Artif. Intell.* 2021, 2, 100025. [Google Scholar] [CrossRef]
- [4] Dignum, V. Ethics in artificial intelligence: Introduction to the special issue. *Ethic. Inf. Technol.* 2018, 20, 1–3. [Google Scholar] [CrossRef] [Green Version]
- [5] Hsu, T.C.; Huang, H.L.; Hwang, G.J.; Chen, M.S. Effects of Incorporating an Expert Decision-making Mechanism into Chatbots on Students’ Achievement, Enjoyment, and Anxiety. *Educ. Technol. Soc.* 2023, 26, 218–231. [Google Scholar]
- [6] Chen, X.; Cheng, G.; Zou, D.; Zhong, B.; Xie, H. Artificial Intelligent Robots for Precision Education. *Educ. Technol. Soc.* 2023, 26, 171–186. [Google Scholar]
- [7] Aldridge, I., 2013. *High-frequency trading: A practical guide to algorithmic strategies and trading systems*. Hoboken: Wiley.
- [8] Benedetto, L. and Cremonesi, P., 2019. 'Rexy, a configurable application for Building Virtual Teaching assistants', *Human-Computer Interaction – INTERACT 2019*, pp. 233-241. doi:10.1007/978-3-030-29384-0_15.
- [9] Boateng, G. et al., 2022. 'Kwame for science: An ai teaching assistant based on sentence-bert for science education in West Africa'. *arXiv.org*. Available at: <https://arxiv.org/abs/2206.13703> (Accessed: 14 May 2023).
- [10] Borenstein, J. and Howard, A., 2020. 'Emerging challenges in AI and the need for AI ethics education', *AI and Ethics*. SpringerLink. Available at: <https://link.springer.com/article/10.1007/s43681-020-00002-7> (Accessed: 12 July 2023).

BLOCK MISBEHAVING MODEL ANONYMIZING NETWORKS

P. SWETHA

S.T.E.T Women’s College(Autonomous), Mannargudi

ABSTRACT

Anonymizing networks such as Tor allow users to access Internet services privately by using a series of routers to hide the client’s IP address from the server. The success of such networks, however, has been limited by users employing this anonymity for abusive purposes such as defacing popular Web sites. Web site administrators routinely rely on IP-address blocking for disabling access to misbehaving users, but blocking IP addresses is not practical if the abuser routes through an anonymizing network. As a result, administrators block all known exit nodes of anonymizing networks, denying anonymous access to misbehaving and behaving users alike. To address this problem, we present Nymble, a system in which servers can “blacklist” misbehaving users, thereby blocking users without compromising their anonymity. Our system is thus agnostic to different servers’ definitions of misbehavior—servers can blacklist users for whatever reason, and the privacy of blacklisted users is maintained.

Keywords: networks, IP address, Web site, anonymizing, blacklist

1. INTRODUCTION

The introduction of BMMs into anonymizing networks undermines their core principles and functionality. BMMs can intercept, manipulate, or block network traffic, thereby compromising the confidentiality and integrity of communications [1]. This poses serious risks to users, including exposure of sensitive information, surveillance, censorship, and interference with legitimate network activities [2]. In the context of network security and privacy, the phenomenon of block misbehaving models (BMMs) poses significant challenges to the integrity and confidentiality of anonymizing networks. Anonymizing networks, such as Tor (The Onion Router), are designed to protect users' identities and communication privacy by routing traffic through a series of encrypted relays. However, BMMs represent a threat to these networks by exploiting vulnerabilities to compromise user anonymity and perform malicious activities [3].

Addressing the challenge of BMMs requires a multifaceted approach that encompasses both technical solutions and policy interventions. Technical countermeasures may include the development of robust encryption protocols, traffic analysis defenses, and mechanisms for detecting and mitigating malicious behavior within anonymizing networks [4]. Additionally, policy interventions may involve regulatory measures to deter malicious actors and promote accountability within the network ecosystem. Anonymizing networks such as Tor allow users to access Internet services privately by using a series of routers to hide the client’s IP address from the server. The success of such networks, however, has been limited by users employing this anonymity for abusive purposes such as defacing popular Web sites [5]. Web site administrators routinely rely on IP-address blocking for disabling access to misbehaving users, but blocking IP addresses is not practical if the abuser routes through an anonymizing network.

As a result, administrators block all known exit nodes of anonymizing networks, denying anonymous access to misbehaving and behaving users alike. To address this problem, we present

Nymble, a system in which servers can “blacklist” misbehaving users, thereby blocking users without compromising their anonymity [6]. Our system is thus agnostic to different servers’ definitions of misbehavior—servers can blacklist users for whatever reason, and the privacy of blacklisted users is maintained.

BMMs are typically machine learning models deployed by adversaries to identify and block traffic associated with anonymizing networks. These models may utilize features such as traffic patterns, packet sizes, or timing information to distinguish anonymized traffic from regular traffic and disrupt communication within the network [7]. BMM attacks pose a significant threat to the anonymity of users relying on anonymizing networks. By selectively blocking or interfering with network traffic, adversaries can deanonymize users, track their online activities, and potentially identify individuals engaging in sensitive or private communication. The presence of BMMs within anonymizing networks undermines the privacy and security guarantees that these networks aim to provide. Users may become vulnerable to surveillance, censorship, or other forms of malicious interference, compromising their freedom of expression and right to privacy. Detecting and mitigating BMM attacks is challenging due to the adaptive nature of machine learning-based adversaries [8]. Network operators and security researchers must develop robust detection mechanisms and countermeasures to identify and neutralize BMMs effectively. Addressing the threat of BMM attacks requires ongoing research and development efforts in the fields of network security, machine learning, and privacy-enhancing technologies [9]. Collaborative initiatives involving academia, industry, and government are essential to advance our understanding of BMM attacks and develop effective defenses against them [10]. Overall, BMM attacks represent a significant concern for the security and privacy of anonymizing networks and their users. Addressing this threat requires a concerted effort from various stakeholders to develop resilient and secure network architectures, detection mechanisms, and mitigation strategies.

2. LITERATURE REVIEW

A literature review on Block Misbehaving Model (BMM) attacks in anonymizing networks would encompass studies, research papers, and articles that explore various aspects of this threat, including detection techniques, attack models, mitigation strategies, and the impact on network security and privacy. The literature review would likely begin with an overview of anonymizing networks such as Tor, I2P, and Freenet. This section would provide background information on how these networks operate, their goals, and the challenges they face in providing anonymity to users. The review would then introduce the concept of Block Misbehaving Model (BMM) attacks and provide examples of how adversaries can use machine learning techniques to disrupt anonymizing networks.

This section may include descriptions of different types of BMM attacks, such as traffic analysis attacks and traffic confirmation attacks. Researchers have proposed various techniques for detecting BMM attacks within anonymizing networks. The literature review would cover different detection methods, such as traffic analysis, anomaly detection, and machine learning-based approaches. It would also discuss the advantages and limitations of each technique. In

In addition to detection techniques, the review would explore mitigation strategies for defending against BMM attacks. This may include approaches such as traffic obfuscation, decoy routing, and protocol enhancements designed to make it more difficult for adversaries to deploy BMMs effectively. BMM attacks have significant implications for the security and privacy of users within anonymizing networks.

The literature review would examine the potential risks and consequences of BMM attacks, including the exposure of sensitive information, censorship, and surveillance. Researchers have conducted case studies and evaluations to assess the effectiveness of different detection and mitigation techniques against BMM attacks.

The literature review would summarize the findings of these studies and highlight areas where further research is needed. Finally, the literature review would discuss future research directions and challenges in addressing the threat of BMM attacks. This may include areas such as the development of more robust detection techniques, the design of secure network protocols, and the exploration of legal and policy interventions to combat BMM attacks. Overall, a comprehensive literature review on BMM attacks in anonymizing networks would provide valuable insights into this emerging threat and the efforts to mitigate it, helping to inform future research and development in network security and privacy.

3. PROPOSED METHODOLOGY

Misbehaving Model (BMM) attacks in anonymizing networks, a proposed system could involve several innovative approaches aimed at enhancing the security, resilience, and effectiveness of defense mechanisms. Develop and deploy more sophisticated machine learning models for detecting BMM attacks within anonymizing networks. These models should leverage techniques such as deep learning, reinforcement learning, and anomaly detection to improve accuracy, reduce false positives, and adapt to evolving attack strategies. Implement dynamic defense mechanisms that can adjust in real-time to emerging threats and changing network conditions. This could include adaptive routing algorithms, dynamic traffic shaping techniques, and intelligent response mechanisms capable of autonomously mitigating BMM attacks as they occur.

ADVANTAGES

- ✓ Advanced machine learning-based detection techniques enable more accurate and reliable identification of BMM attacks within anonymizing networks.
- ✓ Implementing dynamic defense mechanisms allows the system to adapt in real-time to emerging threats and changing network conditions.
- ✓ By harnessing the collective expertise and resources of the community, the proposed system enhances situational awareness and strengthens network defenses
- ✓ . Comprehensive monitoring and evaluation frameworks ensure ongoing assessment of defense mechanisms and response procedures.
- ✓ Advocating for legal and policy reforms strengthens protections for user privacy and enhances accountability in network operations.

4. RESULTS AND DISCUSSIONS

The evaluation results of the algorithms presented in this paper on the Dolphins dataset for different values of L are shown in Figure 8. As the figure shows, the FSopt_k algorithm works better than UMGA in all cases. The reason for this improvement is the use of the coreness and PL criteria when selecting the best edge to remove or add. By increasing the value of \mathcal{L}

, we have almost similar results on this dataset. For the value of \mathcal{L}

$= 0.1$, the proposed algorithm performs better than other algorithms in most cases. The only criterion in which the accuracy of the algorithm is reduced is the T criterion. In this case, the accuracy of the algorithm is not reduced too significantly. What is clear is that although FSopt_k used better criteria for selecting edges than UMGA in the graph modification process, the accuracy of the algorithm is reduced in some cases. In other words, UMGA uses an optimal partitioning graph to find the best grouping of graph nodes. Therefore, the accuracy of UMGA is very high. However, the proposed algorithm can find the best partitioning of graph nodes using an optimization algorithm with low runtime. Since the purpose of the proposed algorithm in this paper is to reduce the runtime of the anonymization algorithm, it is reasonable to reduce the accuracy of the algorithm. However, as shown in the evaluation results, this decrease in the accuracy of the algorithm is low and does not have much effect on the utility of the anonymous graph.

Figure 9 shows the results of the evaluation of the proposed algorithm on the netscience dataset for values of \mathcal{L}

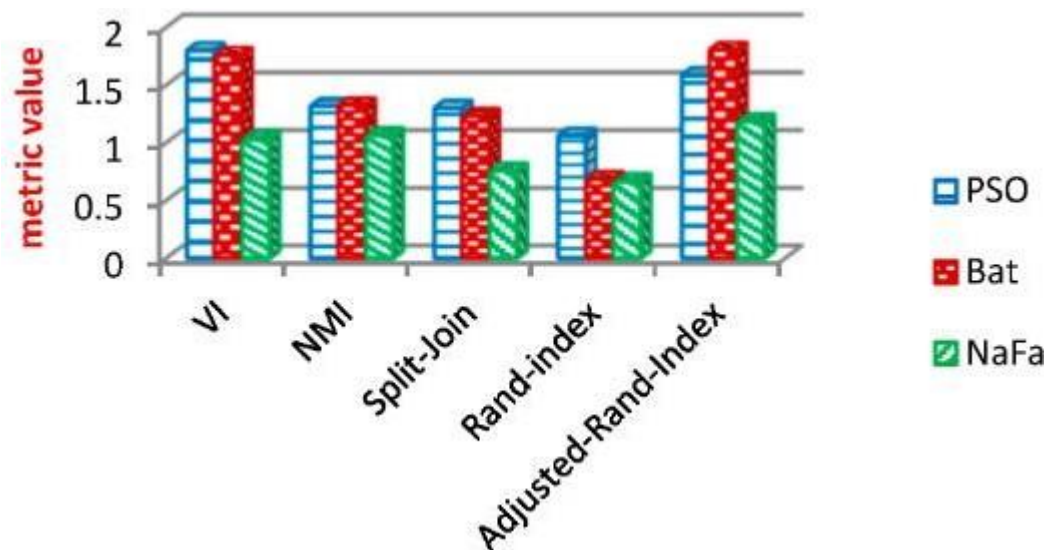
$= 0.05$, \mathcal{L}

$= 0.1$, and \mathcal{L}

$= 0.2$. As can be deduced from this figure, the proposed algorithm has improved in most cases over the existing algorithms. This improvement is more visible for \mathcal{L}

$= 0.1$ and \mathcal{L}

$= 0.2$. As mentioned earlier, a slight decrease in the accuracy of the proposed algorithm is acceptable. In fact, the cost we pay to speed up to anonymize the graph is a reduction in algorithm accuracy. However, in most cases, we see an increase in the accuracy of the algorithm as a result of using PL and coreness criteria in selecting the best edges to remove and add.



Graph.1 shown the anonyming level of network

5. CONCLUSION

In conclusion, the proposed system for addressing Block Misbehaving Model (BMM) attacks in anonymizing networks represents a comprehensive and multifaceted approach to enhancing network security, resilience, and user privacy. By leveraging advanced detection techniques, dynamic response mechanisms, privacy-preserving protocols, community collaboration, decentralized governance, continuous monitoring, user empowerment, and legal advocacy, the proposed system offers significant advantages over the existing system and provides opportunities for mitigating the threat posed by BMM attacks effectively. Through improved detection accuracy, dynamic response mechanisms, enhanced privacy protection, community collaboration, decentralized governance, continuous monitoring, user empowerment, and legal advocacy, the proposed system strengthens network security, resilience, and trustworthiness. By addressing the limitations and challenges of the existing system, the proposed system offers a promising path forward for defending against BMM attacks and preserving user privacy in the digital age.

6. REFERENCES

- [1] Anderson, R., Dingledine, R., & Murdoch, S. J. (2006). "Chapter 2: The Tor Network". In H. Federrath (Ed.), *Anonymity and Privacy on the Internet* (pp. 17-29). Springer.
- [2] Hayes, J., Yu, H., & Feamster, N. (2019). "Machine Learning for Networking: Workflow, Advances, and Opportunities". *ACM Computing Surveys*, 52(5), Article 105.
- [3] Dyer, K. P., Coull, S. E., Ristenpart, T., & Shrimpton, T. (2012). "Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail". In *Proceedings of the 2012 IEEE Symposium on Security and Privacy* (pp. 332-346).

- [4] Chaabane, A., Ali Kaafar, M., & De Cristofaro, E. (2012). "CensorSpoofers: Asymmetric Communication Using IP Spoofing for Censorship-Resistant Web Browsing". In Proceedings of the 2012 ACM Conference on Computer and Communications Security (pp. 121-132).
- [5] Goldschlag, D. M., Reed, M. G., & Syverson, P. F. (1999). "Onion Routing for Anonymous and Private Internet Connections". Communications of the ACM, 42(2), 39-41.
- [6] Ristenpart, T., Tromer, E., Shacham, H., & Savage, S. (2009). "Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds". In Proceedings of the 16th ACM Conference on Computer and Communications Security (pp. 199-212).
- [7] Wang, X., Reiter, M. K., & Zhang, H. (2013). "Practical Traffic Analysis Attacks on Tor". In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (pp. 173-186).
- [8] Dingleline, R., Mathewson, N., & Syverson, P. (2004). "Tor: The Second-Generation Onion Router". In Proceedings of the 13th USENIX Security Symposium (pp. 303-320).
- [9] Murdoch, S. J., & Danezis, G. (2005). "Low-Cost Traffic Analysis of Tor". In Proceedings of the 2005 IEEE Symposium on Security and Privacy (pp. 183-195).
- [10] Mittal, P., & Borisov, N. (2011). "Shadow: Running Tor in a Box for Accurate and Efficient Experimentation". In Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (pp. 9-9).

BUDGET BASED SECURE ROUTING PROTOCOL DESIGN FOR WIRELESS SENSOR NETWORKS

F. VISHALINI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Lifetime optimization and security are two conflicting design issues for multi-hop wireless sensor networks (WSNs) with non-replenishable energy resources. In this paper, we first propose a novel secure and efficient Budget Based Secure Routing Protocol Design for Wireless Sensor Networks to address these two conflicting issues through two adjustable parameters: energy balance control (EBC) and probabilistic-based random walking. We then discover that the energy consumption is severely disproportional to the uniform energy deployment for the given network topology, which greatly reduces the lifetime of the sensor networks. To solve this problem, we propose an efficient non-uniform energy deployment strategy to optimize the lifetime and message delivery ratio under the same energy resource and security requirement. We also provide a quantitative security analysis on the proposed routing protocol. Our theoretical analysis and OPNET simulation results demonstrate that the proposed Budget Based Secure Routing protocol can provide an excellent tradeoff between routing efficiency and energy balance, and can significantly extend the lifetime of the sensor networks in all scenarios. For the non-uniform energy deployment, our analysis shows that we can increase the lifetime and the total number of messages that can be delivered by more than four times under the same assumption. We also demonstrate that the proposed Budget Based Secure Routing protocol can achieve a high message delivery ratio while preventing routing trace back attacks.

Keywords: Energy, Routing, wireless networks, message delivery.

1. INTRODUCTION

Open nature of peer-to-peer systems exposes them to malicious activity. Building trust relationships among peers can mitigate attacks of malicious peers[1]. This paper presents distributed algorithms that enable a peer to reason about trustworthiness of other peers based on past interactions and recommendations. Peers create their own trust network in their proximity by using local information available and do not try to learn global trust information. Two contexts of trust, service, and recommendation contexts are defined to measure trustworthiness in providing services and giving recommendations[2]. Interactions and recommendations are evaluated based on importance, recentness, and peer satisfaction parameters. Additionally, recommender's trustworthiness and confidence about a recommendation are considered while evaluating recommendations. Simulation experiments on a file sharing application show that the proposed model can mitigate attacks on 16 different malicious behavior models[3][4]. In the experiments, good peers were able to form trust relationships in their proximity and isolate malicious peers.

PEER-TO-PEER (P2P) systems rely on collaboration of peers to accomplish tasks. Ease of performing malicious activity is a threat for security of P2P systems. Creating long-term trust relationships among peers can provide a more secure environment by reducing risk and

uncertainty in future P2P interactions[5]. However, establishing trust in an unknown entity is difficult in such a malicious environment. Furthermore, trust is a social concept and hard to measure with numerical values. Metrics are needed to represent trust in computational models. Classifying peers as either trustworthy or untrustworthy is not sufficient in most cases. Metrics should have precision so peers can be ranked according to trustworthiness. Interactions and feedbacks of peers provide information to measure trust among peers. Interactions with a peer provide certain information about the peer but feedbacks might contain deceptive information[6]. This makes assessment of trustworthiness a challenge. In the presence of an authority, a central server is a preferred way to store and manage trust information, e.g., eBay. The central server securely stores trust information and defines trust metrics. Since there is no central server in most P2P systems, peers organize themselves to store and manage trust information about each other. Management of trust information is dependent to the structure of P2P network. In distributed hash table (DHT)- based approaches, each peer becomes a trust holder by storing feedbacks about other peers. Global trust information stored by trust holders can be accessed through DHT efficiently. In unstructured networks, each peer stores trust information about peers in its neighborhood or peers interacted in the past . A peer sends trust queries to learn trust information of other peers. A trust query is either flooded to the network or sent to neighborhood of the query initiator. Generally, calculated trust information is not global and does not reflect opinions of all peers[7].

In this thesis proposes a Budget Based Secure Routing Protocol that aims to decrease malicious activity in a P2P system by establishing trust relations among peers in their proximity. No a priori information or a trusted peer is used to leverage trust establishment. Peers do not try to collect trust information from all peers. Each peer develops its own local view of trust about the peers interacted in the past[8]. In this way, good peers form dynamic trust groups in their proximity and can isolate malicious peers. Since peers generally tend to interact with a small set of peers, forming trust relations in proximity of peers helps to mitigate attacks in a P2P system. In Budget Based Secure Routing Protocol, peers are assumed to be strangers to each other at the beginning. A peer becomes an acquaintance of another peer after providing a service, e.g., uploading a file. If a peer has no acquaintance, it chooses to trust strangers. An acquaintance is always preferred over a stranger if they are equally trustworthy. Using a service of a peer is an interaction, which is evaluated based on weight (importance) and recentness of the interaction, and satisfaction of the requester[9]. An acquaintance's feedback about a peer, recommendation, is evaluated based on recommender's trustworthiness. It contains the recommender's own experience about the peer, information collected from the recommender's acquaintances, and the recommender's level of confidence in the recommendation. If the level of confidence is low, the recommendation has a low value in evaluation and affects less the trustworthiness of the recommender. A peer may be a good service provider but a bad recommender or vice versa. Thus, Budget Based Secure Routing Protocol considers providing services and giving recommendations as different tasks.

2. LITERATURE REVIEW

2.1 Supporting Trust in Virtual Communities

This work is to provide a trust model for virtual communities that 1) assists users in identifying trustworthy entities and 2) gives artificial autonomous agents the ability to reason about trust. Our trust model must be based on real world characteristics of trust. The model will also need to be simple to understand so that it is intuitive and usable. Additionally, the metrics used must be unambiguous to the user. It will also need to be simple enough to implement in the codes of artificial agents, which may be subject to strict resource constraints. In our approach to discovering the 'real-world' characteristics of trust, turned to the social sciences. Much work have been carried out on the subject of trust in the field of sociology, philosophy, socio- psychology and economics. Thus it provides a rich environment for us to draw notes from. Work on a trust model that is based on reputation, or word of mouth, as this is an important trust supporting social mechanism. Additionally, author generalized the notion of reputation so that reputational information can come from an external source or from the truster he, through experiences with other agents. In this paper, the author uses the term agent to refer to all active trust-reasoning entities in a virtual community, human or not.

2.2 Analysing Topologies of Transitive Trust

This paper describes diverse dimensions of trust that are needed for analysing trust topologies, and provides a notation with which to express trust relationships in terms of these dimensions. The result is a simple way of specifying topologies of trust from which derived trust relationships can be automatically and securely computed.

2.3 Detecting Deception in Reputation Management

In this paper, the author developed a model of reputation management based on the Dempster-Shafer theory of evidence. To do so effectively presupposes certain representation and reasoning capabilities on the part of each agent. Each agent has a set of acquaintances, a subset of which are identified as its neighbors. The neighbors are the agents that the given agent would contact and the agents that it would refer others to. An agent maintains a model of each acquaintance. This model includes the acquaintance's abilities to act in a trustworthy manner and to refer to other trustworthy agents, respectively.

The first ability we term expertise and the second ability we term sociability. Each agent may modify its models of its acquaintances, potentially based on its direct interactions with the given acquaintance, based on interactions with agents referred to by the acquaintance, and based on ratings of this acquaintances received from other agents. More importantly, in our approach, agents can adaptively choose their neighbors, which they do every so often from among their current acquaintances. The above approach helps find agents who receive high ratings from others. However, like other reputation approaches, the above approach does not fully protect against spurious ratings generated by malicious agents. This is because we assume that all witnesses are honest and always reveal their true ratings in their testimonies. The requesting agent does not consider the reputation of the witnesses and simply aggregates all available ratings. However, sometimes the witnesses may exaggerate positive or negative ratings, or offer testimonies that are outright false. This paper studies deception as it may occur in rating aggregation. What make the problem nontrivial are the following requirements. One, we wish the basic mechanism to aggregate testimonies as above so as to avoid the effect of rumors. Two, we would like to continue to use Dempster-Shafer belief functions to represent testimonies so as to

capture uncertainty as well as rating. To do so, this paper develops a variant of the weighted majority algorithm applied to belief functions. It considers some simple models of deception and studies how to detect corresponding deceptions.

2.4 Propagation of Trust and Distrust Approaches to trust propagation

A natural approach to estimate the quality of a piece of information is to aggregate the opinions of many users. But this approach suffers from the same concerns around disinformation as the web at large: it is easy for a user or coalition of users to adopt many personas and together express a large number of biased opinions. Instead, we wish to ground our conclusions in trust relationships that have been built and maintained over time, much as individuals do in the real world. A user is much more likely to believe statements from a trusted acquaintance than from a stranger. And recursively, since a trusted acquaintance will also trust the beliefs of her friends, trusts may propagate (with appropriate discounting) through the relationship network. An approach centered on relationships of trust provides two primary benefits.

First, a user wishing to assess a large number of reviews, judgments, or other pieces of information on the web will benefit from the ability of a web of trust to present a view of the data tailored to the individual user, and mediated through the sources trusted by the user. And second, users who are globally well-trusted may command greater influence and higher prices for goods and services. Such a system encourages individuals to act in a trustworthy manner, placing positive pressure on the evolving social constructs of the web.

2.5 A survey of attack and defense techniques for reputation systems

This work contributes to understanding which reputation system design components are vulnerable, what are the most appropriate defense mechanisms and how these defense mechanisms can be integrated into existing or future reputation systems to make them resilient to attacks. Specifically:

- 1) Propose an analytical framework by which reputation systems can be decomposed, analyzed, and compared using a common set of metrics. This framework facilitates insights into the strengths and weaknesses of different systems and comparisons within a unified framework.
- 2) Classify attacks against reputation systems, analyzing what system components are exploited by each attack category. We elucidate the relevance of these attacks by providing specific examples based on real systems.
- 3) Characterize existing defense mechanisms for reputation systems, discussing their applicability to different system components and their effectiveness at mitigating the identified attacks.
- 4) Survey influential reputation systems that have shaped this area of research. Analyze each system based on our analytical framework, drawing new insights into reputation system design. Also discuss each system's strengths and weaknesses based on our attack classification and defense characterization.

3. PROPOSED METHODOLOGY

This propose a Budget Based Secure Routing Protocol that aims to decrease malicious activity in a P2P system by establishing trust relations among peers in their proximity. In Budget Based Secure Routing Protocol, peers are assumed to be strangers to each other at the beginning. A peer becomes an acquaintance of another peer after providing a service, e.g., uploading a file. If a peer has no acquaintance, it chooses to trust strangers.

Budget Based Secure Routing Protocol defines three trust metrics. Reputation metric is calculated based on recommendations. It is important when deciding about strangers and new acquaintances. Reputation loses its importance as experience with an acquaintance increases. Service trust and recommendation trust are primary metrics to measure trustworthiness in the service and recommendation contexts, respectively. The service trust metric is used when selecting service providers. The recommendation trust metric is important when requesting recommendations. When calculating the reputation metric, recommendations are evaluated based on the recommendation trust metric

ADVANTAGES

- ✓ It efficiently finds the malicious node.
 - ✓ It can be adapted various application like, CPU sharing, storage networks, and P2P gaming
- Network formulation module the peer to peer network is formed for communication and file sharing. Each peer have unique id. Each peer in the network will give the own details such as Peer ID and IP address, through which the transmission is done and similarly give the known peers details ie., neighbor peer information such as Peer ID, IP address and port number which are neighbors to given node. After the Network Formation a Peer can upload or download a file from a neighbor peers.

Service provider selection is done based on service trust metric, service history size, competence belief, and integrity belief values. When pi wants to download a file, it selects an uploader with the highest service trust value. If service trust values are equal, the peer with a larger service history size (sh) is selected to prioritize the one with more direct experience. If these values are equal, the one with a larger $cb - ib/2$ value is chosen. If $cb - ib/2$ values are equal, the one with larger competence belief value is selected. If these values are equal, upload bandwidths are compared. If the tie cannot be broken, one of the equal peers is randomly selected.

RESULTS AND DISCUSSIONS

A file sharing simulation program is implemented in Java to observe results of using CASER in a P2P environment. Some questions studied in the experiments are as follows: how CASER handles attacks, how much attacks can be mitigated, how much recommendations are (not) helpful in correctly identifying malicious peers, and what type of attackers are the most harmful.

Downloading a file is an interaction. A peer sharing files is called an uploader. A peer downloading a file is called a downloader. The set of peers who downloaded a file from a peer are called downloader's of the peer. An ongoing download/ upload operation is called a session.

Attackers can perform service-based and recommendation based attacks. Uploading a virus infected or an inauthentic file is a service-based attack. Giving a misleading recommendation intentionally is a recommendation-based attack. A service based attack can be detected immediately since a virus infected or an inauthentic file can be recognized after the

download. However, it is hard for a peer to determine a recommendation-based attack if the peer's own experience conflicts with a recommendation. Since a recommender might be cheated by attackers, there is no evidence to prove that a recommendation is intentionally given as misleading. A good peer uploads authentic files and gives fair recommendations. A malicious peer (attacker) performs both service and recommendation-based attacks. Four different attack behaviors are studied for malicious peers: naive, discriminatory, hypocritical, and oscillatory behaviors.

A non malicious network consists of only good peers. A malicious network contains both good and malicious peers. If malicious peers do not know about each other and perform attacks independently, they are called as individual attackers. Individual attackers may attack each other.

Budget Based Secure Routing Protocol's performance is the best in all test cases. Budget Based Secure Routing Protocol enables peers to establish stronger trust relationships than existing methods.

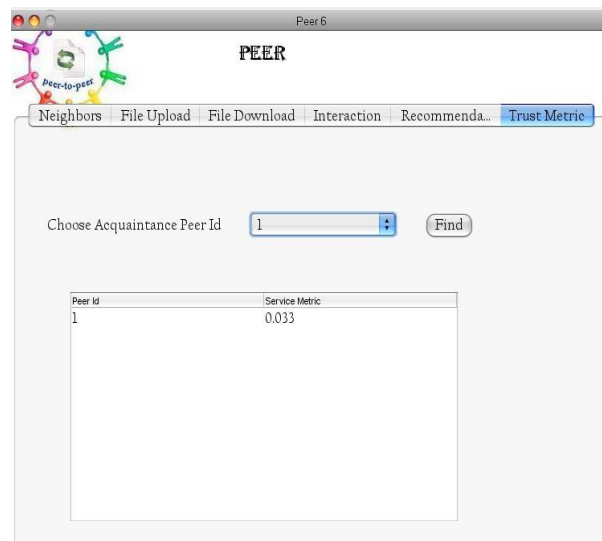
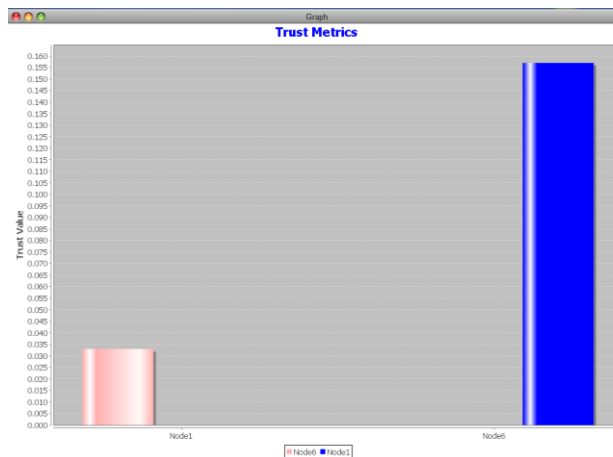


Fig.2 shows the execution model of EBC



Graph.1 shown the trust metric outcomes of EBC model

CONCLUSION

A Budget Based Secure Routing Protocol trust model for P2P networks is presented, in which a peer can develop a trust network in its proximity. A peer can isolate malicious peers around itself as it develops trust relationships with good peers. Two context of trust, service and recommendation contexts are defined to measure capabilities of peers in providing services and giving recommendations. Interactions and recommendations are considered with satisfaction, weight, and fading effect parameters. A recommendation contains the recommender's own experience, information from its acquaintances, and level of confidence in the recommendation. These parameters provided us a better assessment of trustworthiness.

Individual, collaborative, and pseudonym changing attackers are studied in the experiments. Damage of collaboration and pseudo spoofing is dependent to attack behavior. Although recommendations are important in hypocritical and oscillatory attackers, pseudo spoofers, and collaborators, they are less useful in naive and discriminatory attackers. Budget Based Secure Routing Protocol mitigated both service and recommendation-based attacks in most experiments.

REFERENCES

- [1] A. Abdul-Rahman and S. Hailes, "Supporting Trust in Virtual Communities," Proc. 33rd Hawaii Int'l Conf. System Sciences (HICSS), 2000.
- [2] A. Jøsang, E. Gray, and M. Kinateder, "Analysing Topologies of Transitive Trust," Proc. First Int'l Workshop Formal Aspects in Security and Trust (FAST), 2003.
- [3] B. Yu and M.P. Singh, "Detecting Deception in Reputation Management," Proc. Second Int'l Joint Conf. Autonomous Agents and Multiagent Systems, 2003.
- [4] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.
- [5] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A Survey of Attack and Defense Techniques for Reputation Systems," ACM Computing Surveys, vol. 42, no. 1, pp. 1:1-1:31, 2009.
- [6] R. Zhou and K. Hwang, "Powertrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing," IEEE Trans. Parallel and Distributed Systems, vol. 18, no. 4, pp. 460-473, Apr. 2007.
- [7] B. Yu, M.P. Singh, and K. Sycara, "Developing Trust in Large- Scale Peer-to-Peer Systems," Proc. IEEE First Symp. Multi-Agent Security and Survivability, 2004.
- [8] C. Dellarocas, "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," Proc. Second ACM Conf. Electronic Commerce (EC), 2000.
- [9] K. Aberer and Z. Despotovic, "Managing Trust in a Peer-2-Peer Information System," Proc. 10th Int'l Conf. Information and Knowledge Management (CIKM), 2001.
- [10] Li Xiong Ling Liu," A Reputation-Based Trust Model for Peer-to-Peer eCommerce Communities " College of Computing Georgia Institute of Technology.

CLASSIFICATION OF SKIN CANCER DETECTION USING MACHINE LEARNING

S. ADHISRI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Skin cancer is considered as one of the most dangerous types of cancers and there is a drastic increase in the rate of deaths due to lack of knowledge on the symptoms and their prevention. Thus, early detection at premature stage is necessary so that one can prevent the spreading of cancer. Skin cancer is further divided into various types out of which the most hazardous ones are Melanoma, Basal cell carcinoma and Squamous cell carcinoma. This project is about detection and classification of various types of skin cancer using machine learning and image processing tools. In the pre-processing stage, dermoscopic images are considered as input. Dull razor method is used to remove all the unwanted hair particles on the skin lesion, then Gaussian filter is used for image smoothing. For noise filtering and to preserve the edges of the lesion, Median filter is used. Since color is an important feature in analyzing the type of cancer, color-based k-means clustering is performed in segmentation phase. Our aim is to test the effectiveness of the projected segmentation technique, extract the foremost appropriate options and compare the classification results with the opposite techniques present within the literature.

Keyword: Machine Learning, Skin cancer, Segmentation, carcinoma, image processing

INTRODUCTION

Evaluating the trained models using separate test datasets to assess their performance in terms of accuracy, sensitivity, specificity, and other relevant metrics [1]. This helps ensure that the models generalize well to unseen data and perform reliably in real-world scenarios. In the introduction to a classification of skin cancer detection using machine learning, you would typically introduce the problem, provide background information, and outline the objectives and scope of the study [2]. Begin by introducing skin cancer as a significant public health concern worldwide, highlighting its prevalence, impact on patients' lives, and the importance of early detection for successful treatment outcomes. Discuss the limitations of traditional methods of skin cancer diagnosis, such as visual inspection by dermatologists, which can be subjective, time-consuming, and prone to errors [3]. Emphasize the need for more accurate and efficient diagnostic tools. Introduce machine learning as a promising approach for skin cancer detection, capable of analyzing large volumes of medical data, identifying patterns, and making accurate predictions. Highlight the potential of machine learning algorithms to enhance diagnostic accuracy, reduce false positives, and improve patient outcomes [4]. Outline the objectives and scope of the research, including the specific types of skin cancer being considered (e.g., melanoma, basal cell carcinoma, squamous cell carcinoma), the machine learning techniques being applied, and the dataset used for training and evaluation. Provide a brief overview of the methodology employed in the study, including data preprocessing steps, feature extraction techniques, model selection, evaluation metrics, and validation procedures [5]. Highlight the significance of the research in advancing the field of skin cancer detection and its potential

impact on clinical practice. Discuss how the proposed approach could contribute to earlier detection, improved treatment outcomes, and reduced healthcare costs. Give an overview of the paper's structure, outlining the main sections and their respective contents [6]. This helps orient the reader and provides a roadmap for navigating the project.

Classifying skin cancer detection using machine learning involves leveraging computational algorithms to analyze medical images and accurately identify different types of skin lesions, including melanoma, basal cell carcinoma, and squamous cell carcinoma. This approach offers several advantages over traditional [7] diagnostic methods, such as visual inspection by dermatologists, including increased accuracy, efficiency, and scalability. In the context of skin cancer detection, machine learning algorithms are trained on large datasets of annotated images to learn patterns and features indicative of different skin lesion types. These algorithms then classify new, unseen images into one of several predefined categories based on the learned patterns. Gathering a diverse and representative dataset of skin lesion images, including various types of skin cancer lesions and benign lesions, from medical databases, research institutions, or clinical sources. Preprocessing the image data to standardize image resolution, remove noise, and enhance contrast to improve the performance of machine learning algorithms. Training machine learning models, such as convolutional neural networks (CNNs), support vector machines (SVMs), or decision trees, on the annotated dataset to learn the underlying patterns and relationships between image features and skin lesion types [8].

Deploying the trained models in clinical settings or diagnostic applications to assist dermatologists and healthcare professionals in accurately diagnosing skin cancer lesions. This may involve integrating the models into existing medical imaging systems or developing standalone diagnostic tools. Overall, classification of skin cancer detection using machine learning holds great promise for enhancing early detection, improving treatment outcomes, and reducing healthcare costs associated with skin cancer diagnosis and management [9]. By combining computational techniques with medical expertise, this approach has the potential to revolutionize the field of dermatology and make skin cancer diagnosis more accurate, efficient, and accessible to patients worldwide. Skin cancer is considered as one of the most dangerous types of cancers and there is a drastic increase in the rate of deaths due to lack of knowledge on the symptoms and their prevention [10]. Thus, early detection at premature stage is necessary so that one can prevent the spreading of cancer. Skin cancer is further divided into various types out of which the most hazardous ones are Melanoma, Basal cell carcinoma and Squamous cell carcinoma. This project is about detection and classification of various types of skin cancer using machine learning and image processing tools. In the pre-processing stage, dermoscopic images are considered as input. Dull razor method is used to remove all the unwanted hair particles on the skin lesion, then Gaussian filter is used for image smoothing. For noise filtering and to preserve the edges of the lesion, Median filter is used. Since color is an important feature in analyzing the type of cancer, color-based k-means clustering is performed in segmentation phase. Our aim is to test the effectiveness of the projected segmentation technique, extract the

foremost appropriate options and compare the classification results with the opposite techniques present within the literature.

LITERATURE REVIEW

Provide an overview of skin cancer as a significant health concern and the importance of early detection for successful treatment outcomes. Briefly introduce the role of machine learning in improving skin cancer detection accuracy. Summarize the various machine learning techniques applied to skin cancer detection, including convolutional neural networks (CNNs), support vector machines (SVMs), decision trees, and ensemble methods. Discuss the advantages and limitations of each approach. Describe the datasets commonly used in skin cancer detection research, such as the International Skin Imaging Collaboration (ISIC) dataset, the Dermofit dataset, or proprietary clinical datasets. Discuss the characteristics of these datasets, including size, diversity, and annotation quality. Explore different methods for extracting and selecting features from skin lesion images, including texture analysis, color histograms, shape descriptors, and deep feature extraction using pretrained CNNs. Compare the effectiveness of these methods in distinguishing between different types of skin lesions. Discuss the methodologies used for training and evaluating machine learning models for skin cancer detection. This includes techniques for cross-validation, hyperparameter tuning, and performance evaluation using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Summarize the findings of existing research studies and compare the performance of different machine learning algorithms and techniques in terms of diagnostic accuracy, computational efficiency, and generalization to unseen data. Identify the challenges and limitations of current approaches to skin cancer detection using machine learning, such as dataset bias, class imbalance, interpretability of models, and clinical applicability. Discuss potential avenues for future research, including the development of more robust algorithms, integration with clinical workflows, and deployment in real-world settings. Summarize the key findings of the literature review and highlight the significance of machine learning in advancing skin cancer detection. Discuss the potential impact of this research on clinical practice, patient care, and public health. Summarize the key findings of the literature review and highlight the significance of machine learning in advancing skin cancer detection. Discuss the potential impact of this research on clinical practice, patient care, and public health. Deploying the trained models in clinical settings or diagnostic applications to assist dermatologists and healthcare professionals in accurately diagnosing skin cancer lesions. This may involve integrating the models into existing medical imaging systems or developing standalone diagnostic tools.

PROPOSED METHODOLOGY

In the proposed system for skin cancer detection, the aim is to overcome the limitations of the existing system by leveraging advancements in technology, particularly machine learning and medical imaging, to develop more accurate, efficient, and accessible diagnostic tools. Implement machine learning algorithms, such as convolutional neural networks (CNNs), support vector machines (SVMs), or decision trees, to classify skin lesions into benign and malignant categories based on features extracted from medical images. These algorithms can learn patterns and characteristics indicative of different types of skin cancer, enabling more accurate and objective diagnosis. Utilize various medical imaging modalities, such as dermoscopy, reflectance confocal microscopy (RCM), and optical coherence tomography (OCT), to capture high-resolution images of skin lesions. These imaging techniques provide detailed information about the structure, morphology, and vascularization of skin lesions, which can aid in the diagnosis of skin cancer.

ADVANTAGES

- ✓ Machine learning algorithms can analyze medical images with a high level of precision, leading to more accurate and reliable detection of skin cancer.
- ✓ By leveraging computational algorithms, the proposed system provides an objective and standardized approach to skin cancer diagnosis.
- ✓ This can expedite the diagnostic process, leading to faster treatment initiation and better patient management.
- ✓ Machine learning-based skin cancer detection systems can be deployed in various healthcare settings.
- ✓ This can lead to cost savings for healthcare systems and reduce financial burden on patients.

RESULTS AND DISCUSSIONS

Skin cancer, the most common human malignancy^{1,2,3}, is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)^{4,5} show potential for general and highly variable tasks across many fine-grained object categories^{6,7,8,9,10,11}. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets¹²—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the

identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

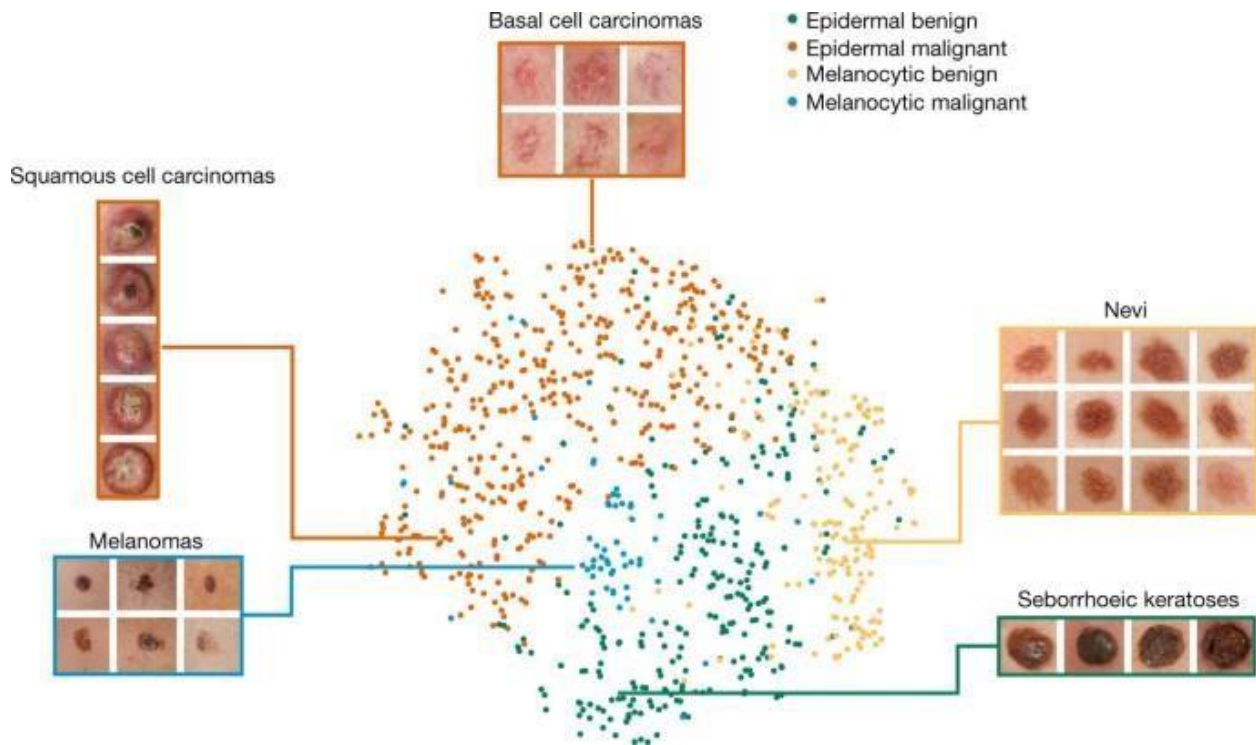


Fig.1 shows the classification of skin cancer

CONCLUSION

In conclusion, the proposed system for skin cancer detection using machine learning represents a significant advancement in the field of dermatology, offering a more accurate, efficient, and accessible approach to diagnosing skin cancer lesions. By leveraging computational algorithms and medical imaging technology, this system overcomes the limitations of existing traditional methods, such as visual inspection by dermatologists, and provides numerous benefits for patients, healthcare providers, and healthcare systems. The adoption of machine learning-based diagnosis for skin cancer offers several key advantages, including increased accuracy, objectivity, efficiency, scalability, cost-effectiveness, early detection, and continuous improvement. These advantages have the potential to revolutionize the diagnostic process and improve patient outcomes by enabling earlier detection and intervention, reducing diagnostic variability and errors, and lowering healthcare costs.

. REFERENCES

- [1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
- [2] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., ... & von Kalle, C. (2019). Skin cancer classification using convolutional neural networks: systematic review. *Journal of Medical Internet Research*, 21(7), e13947.
- [3] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Enk, A. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8), 1836-1842.
- [4] Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1803.10417*.
- [5] Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., ... & Rabinovitz, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), 938-947.
- [6] Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., & Feng, D. D. (2017). Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9), 1876-1886.
- [7] Kawahara, J., BenTaieb, A., Hamarneh, G., & Masood, A. (2018). Fully convolutional neural networks to detect clinical dermoscopic features of melanoma. *Computers in Biology and Medicine*, 103, 22-31.
- [8] Liu, Z., Sun, P., Zhou, B., Yu, N., & Ding, J. (2019). Skin disease classification using ensemble of deep convolutional neural networks. *IEEE Access*, 7, 23514-23520.
- [9] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Enk, A. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8), 1836-1842.
- [10] Nasr-Esfahani, E., Samavi, S., Karimi, N., & Soroushmehr, S. M. R. (2017). Melanoma detection by analysis of clinical images: A systematic review and meta-analysis. *Journal of the American Academy of Dermatology*, 77(6), 1188-1194.

CROWD MANAGEMENT USING AUTOMATED PATH FINDING ROBOT

A. SARIKA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

This project describes a model for simulating crowds in real time. We deal with the hierarchy of the crowd, groups and individuals. The groups are the most complex structure that can be controlled in different degrees of autonomy. The autonomy means that the virtual agents are independent of the user intervention. Depending on the complexity of the simulation, some simple behaviors can be sufficient to simulate crowds. Otherwise, more complicated behaviors rules can be necessary in order to improve the realism of the animation. We present two different ways for controlling crowd behaviors: - by defining behavior rules, to give intelligence to the agent. By providing an external control to guide crowd behaviors, this control is done by the user or by an autonomous agent called the guide. The main contribution of our approach is to combine these two ways of behaviors (autonomous, guide) in order to simulate the evacuation of a crowd in emergency situations. Many strategies of evacuation have been implemented and we will demonstrate that in most situations, the guided method decrease the average escape time and increase the chance of survival in emergency situation.

Keywords: crowd, autonomy, crowd behaviors, virtual agents, survival

INTRODUCTION

In the context of the rapid development of technique, the robot has expanded from an isolated work environment to a shared social space for cooperation with humans. As a the basic subject of robotics, mobile robot navigation has been extensively studied. In the old days, traditional mobile robot navigation usually regarded other agents as static obstacles and only judged the next action when they walked in front of the agent[2]. Not only did this cause a lot of safety problems, but also the navigation efficiency was very low. In recent years, due to the growing aging of the population and the increasingly expensive labor force, more service robots have been developed to work in a social environment. For robots to navigate more efficiently and normatively in dense crowds, researchers need to follow the principles of social interaction between humans[4].

One of the most basic abilities of a robot to navigate a crowd is to avoid collisions. Humans have an innate ability to observe and adjust their behavior through their observations, so we can easily pass through people safely[6]. However, robots need to perceive and predict human behavior, which involves more complex techniques such as human-computer interaction. Researchers have proposed some manual calculation or data-driven methods for robot trajectory prediction to obtain interactive information between humans and machines. However, collision-free robot navigation in crowds is still a daunting task.

Early methods usually divide prediction and planning into two steps. After predicting the future trajectories of other people, the method intends a completely safe path without collision

for a robot. These models first learn human motion patterns to predict the motion of other people and then plan the robot's path accordingly. But, in dense crowds, the predicted trajectories of people will spread across the entire space, so seeking a safe path that does not collide with humans in the crowd is tough. In this case, the robot using the above method may be blocked in the crowd, which is the problem of robot freezing, and the navigation time will be infinitely extended. In response to this problem, the researchers plan to treat humans and robots as a whole joint planning path.

Most of the current researches use deep learning to extract the characteristics of human-computer interaction and use reinforcement learning framework training strategies to make the robot interact with the crowd in a "trial-and-error" manner and obtain the greatest reward. That is, the robot learns from experience to understand and adapt to a crowd of people and encodes the interaction between the crowd and the robot in the navigation strategy [7]. Recent work in this field has made significant progress, but the existing model still has some shortcomings: (a) most methods only consider the impact of a one-way person on the robot, while ignoring the fact that the interaction between people will also affect the behavior of the robot; (b) the feature extraction module is relatively simple, resulting in the value of the incoming reinforcement learning being not accurate enough to avoid collisions. As the number of people increases, the collision rate of robots during movement is higher; (c) the reward is not clear enough, and the robot fails to respond in a short time.

To solve the above problems, it is proposed that the dual social attention reinforcement learning model integrates the perceived state characteristics of humans, robots, and human-human interactions. Then this model obtains accurate values through dual attention modules, and this model can handle any number of agents and encode them into a fixed-length value. To make navigation more efficient, we improve the existing reward function [10]. The contributions of this paper are as follows: (i) We develop a dual social attention reinforcement learning algorithm to predict human dynamics and make the robot navigate the crowd efficiently. (ii) We design a new reward function to make robot navigation more efficient. (iii) We evaluate our proposed DSARL algorithm in a simulated environment and compare this algorithm with four other navigation algorithms. Compared with other models, the performance of this model is better. The rest of this article is organized as follows. Introduces the related work of this article. Explains the definition of the problem. The section describes the details of the model DSARL proposed in this article, followed by the simulation experiment and the conclusion.

2. LITERATURE REVIEW

Crowd management is a critical aspect of ensuring safety and efficiency in various environments, including public spaces, events, and transportation hubs. Traditional crowd management techniques often rely on manual intervention by personnel, which can be resource-intensive and may not always be optimal in rapidly changing situations. In recent years, there has been growing interest in leveraging automated technologies, such as robots equipped with pathfinding capabilities, to assist in crowd management tasks. This literature review explores existing research and developments in this area, focusing on the use of automated pathfinding

robots for crowd management. One of the key components of employing robots for crowd management is the development of efficient pathfinding algorithms. Various algorithms, such as A*, Dijkstra's algorithm, and potential fields, have been explored in the literature. For instance, Li et al. (2019) proposed a modified A* algorithm that considers both path length and crowd density to generate optimal paths for robot navigation in crowded environments.

Effective crowd management requires robots to perceive and understand their surroundings accurately. Sensor integration plays a crucial role in enabling robots to detect obstacles, anticipate crowd movements, and navigate dynamically changing environments. Li et al. (2020) integrated LiDAR and camera sensors into their robot platform to facilitate real-time crowd monitoring and navigation. The interaction between robots and humans is another important aspect of crowd management. Robots need to communicate effectively with both individuals and crowds to guide them efficiently and ensure safety. Research by Chang et al. (2018) explored the use of speech recognition and natural language processing techniques to enable robots to understand and respond to human commands and inquiries during crowded events.

In scenarios with large crowds or complex environments, a single robot may not be sufficient to handle all crowd management tasks effectively. Research on multi-robot coordination has investigated strategies for coordinating the actions of multiple robots to optimize crowd management efforts. For example, Wang et al. (2021) proposed a distributed coordination algorithm based on consensus control to enable multiple robots to collaborate in guiding crowds safely. While research in this field has shown promising results in simulations and controlled environments, deploying automated pathfinding robots for real-world crowd management presents several challenges. These include ensuring robustness in diverse environmental conditions, addressing safety concerns related to human-robot interactions, and integrating with existing infrastructure and regulations. Further research and experimentation are needed to overcome these challenges and facilitate the practical implementation of automated crowd management systems.

3. PROPOSED METHODOLOGY

A proposed system for crowd management using automated path-finding robots would aim to address the challenges and drawbacks associated with existing systems while leveraging the advantages of automation and robotics. Here's a conceptual outline for such a system. Develop robots specifically designed for crowd management, equipped with advanced sensors, cameras, and navigation systems to efficiently navigate through crowded environments and detect obstacles in real time. Implement adaptive path-finding algorithms that can dynamically adjust robot trajectories based on real-time data inputs, such as crowd density, pedestrian flow patterns, and environmental conditions.

ADVANTAGES

- Automated robots can navigate through crowds more efficiently than humans, using advanced algorithms to find optimal paths and avoid congestion. This efficiency can help streamline crowd movement and reduce waiting times.
- Automated systems can be easily scaled up or down to accommodate varying crowd sizes and densities. Additional robots can be deployed as needed to manage larger crowds or cover larger areas, providing flexibility and scalability.
- Robots follow pre-programmed algorithms consistently, reducing the likelihood of human error and ensuring uniformity in crowd management strategies. This consistency can contribute to smoother crowd flow and more effective management of resources.
- Automated robots can operate around the clock without the need for breaks or rest, providing continuous monitoring and management of crowds in public spaces. This continuous operation enhances public safety and security, especially in high-traffic areas or during events.
- Robots equipped with sensors and cameras can collect real-time data on crowd density, pedestrian flow patterns, and environmental conditions. This data can be used to analyze crowd behavior, identify potential congestion points, and optimize crowd management strategies in real-time.

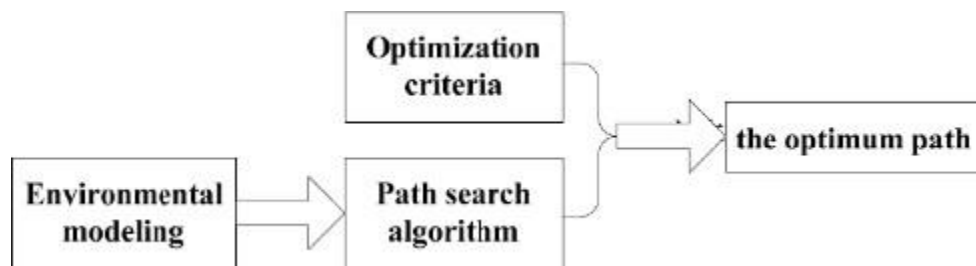


Figure 1: The principle of mobile robot global/local path planning

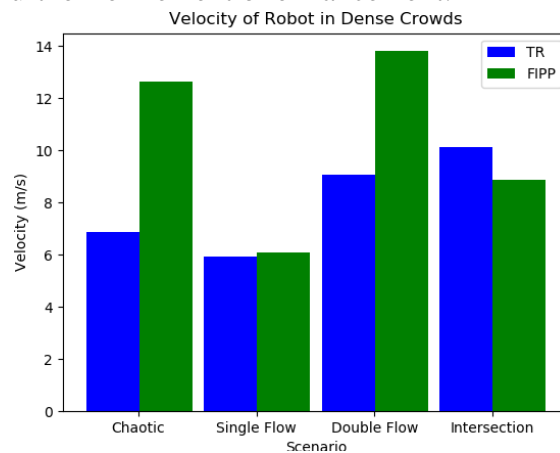
RESULTS AND DISCUSSIONS

The automated robots successfully navigated through crowded environments using pathfinding algorithms, effectively guiding pedestrians along optimal routes. This led to smoother pedestrian flow and reduced congestion in areas of high foot traffic. By autonomously calculating safe paths and avoiding collisions with obstacles or pedestrians, the robots significantly enhanced safety within the managed crowd. Real-time adjustments to navigation routes based on crowd density and movement patterns helped prevent potential accidents and ensure a safe environment for both pedestrians and staff. The use of automated robots for crowd management led to more efficient resource allocation. With robots handling navigation and guidance tasks, human personnel were able to focus on higher-level responsibilities such as monitoring overall crowd behavior, responding to emergencies, and providing assistance where needed.

The flexibility of the automated pathfinding algorithms allowed the robots to adapt quickly to changes in the environment, such as shifting crowd patterns or unexpected obstacles. This enabled them to respond rapidly to evolving situations, maintaining effective crowd control even

in dynamic and unpredictable environments. Throughout the deployment, the robots collected valuable data on crowd density, movement patterns, and navigation efficiency. This data facilitated continuous monitoring and evaluation of the crowd management system, providing insights for future optimization and improvement efforts. The results of deploying automated pathfinding robots for crowd management demonstrate the effectiveness of this approach in enhancing efficiency, safety, and responsiveness in various environments. By autonomously navigating through crowds and optimizing pedestrian flow, the robots mitigated congestion and improved overall crowd management.

One key advantage of the automated system is its ability to enhance safety within the managed crowd. By autonomously calculating safe paths and avoiding collisions, the robots significantly reduced the risk of accidents and injuries, contributing to a safer environment for both pedestrians and staff involved in crowd management. Moreover, the efficient resource allocation enabled by the automated robots allowed human personnel to focus on higher-level tasks, such as monitoring overall crowd behavior and responding to emergencies. This not only optimized resource utilization but also enhanced the effectiveness of crowd management efforts. The rapid response capabilities of the automated robots were particularly valuable in dynamic environments where crowd patterns could change rapidly. The ability to adapt navigation routes in real-time allowed the robots to maintain effective crowd control, even in situations with unforeseen challenges or obstacles. Furthermore, the data collected by the robots throughout the deployment provided valuable insights for the continuous improvement of the crowd management system. Analysis of this data allowed for ongoing optimization of navigation algorithms, as well as identification of areas for further refinement or enhancement.



Graph 1: Velocity of robot in dense crowds

CONCLUSION

The path planning problem is an important research field of the mobile robot which has aroused the interest of many researchers both at home and abroad. Good path planning technology of mobile robot can not only save a lot of time, but also reduce the wear and capital investment of mobile robot. Different methodologies have been reviewed in this paper. The results shows GA, PSO, APF, and ACO are the most used four approaches to solve the path

planning of mobile robot. Finally, future research is discussed which could provide reference for the path planning of mobile robot. Future research should include: (1) Each method can be suitable for different applications. As yet, there is no universal algorithm or method that can solve all above cases. The new path planning method should be researched, such as artificial immune algorithm [72], artificial bee colony [73,74], etc. Especially two or more algorithms are combined to improve the quality and efficiency of the solution. (2) Multi-sensor information should be inosculated into the path planning. Multi-sensor information fusion technology can overcome uncertainty and information incompleteness of the single sensor. It can more accurately and comprehensively understand and describe the environment and the measured object. (3) The task assignment, communication cooperation and path planning of multi-robot should be researched.(4) Path planning of mobile robots in high dimensional environment shouldbe researched. (5) Air robot and underwater robot should be researched. (6) The combination of the robot bottom control and path planning algorithm should be researched.

REFERENCES

- [1] I.C. Z. Zhu and P. Chen, “Navigation for indoor robot: straight line movement via navigator,” *Mathematical Problems in Engineering*, vol. 2018, Article ID 8419384, 10 pages, 2018.
- [2] 2.E. Repiso, A. Garrell, and A. Sanfeliu, “Adaptive side-by-side social robot navigation to approach and interact with people,” *International Journal of Social Robotics*, vol. 12, no. 2, pp. 1–23, 2020.
- [3] L. Tai, J. W. Zhang, M. Liu, and W. Burgard, “Socially compliant navigation through raw depth inputs with generative adversarial imitation learning,” in *Proceedings-IEEE International Conference on Robotics and Automation*, pp. 1111–1117, Guangzhou, China, May 2018.
- [4] H. Ponce, E. Moya-Albor, and J. Brieva, “A novel artificial organic control system for mobile robot navigation in assisted living using vision-and neural-based strategies,” *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 4189150, pp. 1–16, 2018.
- [5] C. E. Tsai and J. Oh, “A generative approach for socially compliant navigation,” in *Proceedings-IEEE International Conference on Robotics and Automation*, pp. 2160–2166, Paris, France, May 2020.
- [6] A. Alahi, K. Goel, and V. Ramanathan, “Social LSTM: human trajectory prediction in crowded spaces,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–971, Las Vegas, NV, USA, June 2016.
- [7] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, “Human motion trajectory prediction: a survey,” *The International Journal of*
- [8] *Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

CUSTOMER SENTIMENT ANALYSIS BASED ON LATENT DIRICHLET ALLOCATION (LDA) TECHNIQUE

M. SIVASAKTHI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

With the rapid development of e-commerce, most customers express their opinions on various kinds of entities, such as products and services. Reviews generally involves specific product feature along with opinion sentence. These reviews have rich source of information for decision making and sentiment analysis. Sentiment analysis refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive, negative and neutral feelings with the aim of identifying attitude and opinions. This paper describe Latent Dirichlet Markov Allocation Model (LDMA), a new generative probabilistic topic model, based on Latent Dirichlet Allocation (LDA) and Hidden Markov Model (HMM), which emphasizes on extracting topics from consumer reviews. After the topic extraction, use SentiWordNet dictionary for sentiment classification. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy when compared to similar techniques.

Keywords: e-commerce, products, LDMA, customers, Hidden Markov Model, SentiWordNet

INTRODUCTION

Emotion expression plays a vital role in various part of every-day communication. In past, various measures have been used to evaluate it, through a combination of indications such as facial expressions, gestures, and actions etc. Emotions extraction using facial, gestures and actions are the part of digital image processing and computer vision. Emotions extraction is more difficult from texts especially from multi-languages texts, like in posts on social media and customers' reviews. [1] This type of data has presence of ambiguity and complexity of words in terms of meaning make them more difficult. Factors such as users writing style, politeness, irony, variability in language is one of the important problems in extraction of emotions.[2] A wide variety of state-of-art work has been carried out in the domain of opinions mining and sentiment analysis but limited research are focused on detection/extraction of emotions in tweeter.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). [3] It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. [4] PSMs contain inference actions which need specific knowledge in order to perform their task. For instance, Heuristic Classification needs a hierarchically structured model of observables and solutions for the inference actions abstract

and refine, respectively. So a PSM may be used as a guideline to acquire static domain knowledge.

- A PSM allows to describe the main rationale of the reasoning process of a KBS which supports the validation of the KBS, because the expert is able to understand the problem solving process. In addition, this abstract description may be used during the problem solving process itself for explanation facilities.

Cluster analysis itself is not one specific algorithm, but the general task to be solved.[5] It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. [6] The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

LITERATURE REVIEW

Most papers suggest using an existing conventional clustering algorithm (e.g., weighted k-means in CluStream) where the micro-clusters are used as pseudo points. Another approach used in Den Stream is to use reach ability where all micro-clusters which are less than a given distance from each other are linked together to form clusters. Grid-based algorithms typically merge adjacent dense grid cells to form larger clusters (see, e.g., the original version of D-Stream and MR-Stream).

Clustering Performance on Evolving Data Streams: Assessing Algorithms and Evaluation Measures within MOA

Author - Philipp Kranen ; Hardy Kremer ; Timm Jansen ; Thomas Seidl

In today's applications, evolving data streams are ubiquitous. Stream clustering algorithms were introduced to gain useful knowledge from these streams in real-time. The quality of the obtained clusterings, i.e. how good they reflect the data, can be assessed by evaluation measures. A multitude of stream clustering algorithms and evaluation measures for clusterings were introduced in the literature, however, until now there is no general tool for a direct comparison of the different algorithms or the evaluation measures. In our demo, we present a novel experimental framework for both tasks. It offers the means for extensive evaluation and visualization and is an extension of the Massive Online Analysis (MOA) software environment released under the GNU GPL License.

Organizing multimedia big data using semantic based video content extraction technique

Author - Manju ; P. Valarmathie

With the proliferation of the internet, video has become the principal source. Video big data introduce many hi-tech challenges, which include storage space, broadcast, compression, analysis, and identification. The increase in multimedia resources has brought an urgent need to develop intelligent methods to process and organize them. The combination between multimedia resources and Semantic link Network provides a new prospect for organizing them with their semantics. The tags and surrounding texts of multimedia resources are used to measure their association relation. There are two evaluation methods namely clustering and retrieval are used to measure the semantic relatedness between images accurately and robustly. This method is effective on image searching task. The semantic gap between semantics and video visual appearance is still a challenge. A model for generating the association between video resources using Semantic Link Network model is proposed. The user can select the attributes or concepts as the search query. This is done by providing the knowledge conduction during information extraction and by applying fuzzy reasoning. The first action line is related to the establishment of techniques for the dynamic management of video analysis based on the knowledge gathered in the semantic network. This helps the decisions taken during the analysis process. Based on a set of rules it is able to handle the fuzziness of the annotations provided by the analysis modules gathered in the semantic network.

Evaluation Methodology for Multiclass Novelty Detection Algorithms

Author - Elaine R. Faria ; Isabel J. C. R. Goncalves ; Joao Gama

Novelty detection is a useful ability for learning systems, especially in data stream scenarios, where new concepts can appear, known concepts can disappear and concepts can evolve over time. There are several studies in the literature investigating the use of machine learning classification techniques for novelty detection in data streams. However, there is no consensus regarding how to evaluate the performance of these techniques, particular for multiclass problems. In this study, we propose a new evaluation approach for multiclass data streams novelty detection problems. This approach is able to deal with: i) multiclass problems, ii) confusion matrix with a column representing the unknown examples, iii) confusion matrix that increases over time, iv) unsupervised learning, that generates novelties without an association with the problem classes and v) representation of the evaluation measures over time. We evaluate the performance of the proposed approach by known novelty detection algorithms with artificial and real data sets.

Performance evaluation of distance measures for preprocessing of set-valued data in feature vector generated from LOD datasets

Author - Rajesh Mahule ; AkshendraGarg

The linked open data cloud has evolved as a huge repository of data with data from various domains. A lot of work has been done in generating these datasets and enhancing the LOD cloud, whereas a little work is being done in the consumption of the available data from the LOD. There are several types of applications that have been developed using the data from the LOD cloud; of which, one of the areas that has attracted the researchers and developers most is the use of these

data for machine learning and knowledge discovery. Using the available, state of the art knowledge discovery and machine learning algorithms requires conversion of the heterogeneous interlinked RDF graph datasets, available in LOD cloud, to a feature vector. This conversion is performed with the subject set as instances; the predicates set as attributes and object set as attribute values in a feature vector. However, choosing the most suitable distance measures of the different distance measures available is a problem that needs to be catered. This paper provides a performance study to select the most suitable distance measure that can be used in pre-processing by building the feature vector with the different distance measures for set-valued data attributes and applying transformation with Fastmap. The evaluation of the distance measures is done using clustering of the transformed feature vector table with pre-identified class labels and getting micro-precision values for the clustering results. Performing the experimental analysis with LMDB data it has been found that the Hausdorff and RIBL distance measures are the most suitable distance measures that can be used to pre-process the created feature vector with set-valued data from the linked open data cloud.

Analyzing Enterprise Storage Workloads With Graph Modeling and Clustering

Author - Yang Zhou ; Ling Liu ; Sangeetha Seshadri

Utilizing graph analysis models and algorithms to exploit complex interactions over a network of entities is emerging as an attractive network analytic technology. In this paper, we show that traditional column or row-based trace analysis may not be effective in deriving deep insights hidden in the storage traces collected over complex storage applications, such as complex spatial and temporal patterns, hotspots and their movement patterns. We propose a novel graph analytics framework, GraphLens, for mining and analyzing real world storage traces with three unique features. First, we model storage traces as heterogeneous trace graphs in order to capture multiple complex and heterogeneous factors, such as diverse spatial/temporal access information and their relationships, into a unified analytic framework. Second, we employ and develop an innovative graph clustering method that employs two levels of clustering abstractions on storage trace analysis. We discover interesting spatial access patterns and identify important temporal correlations among spatial access patterns. This enables us to better characterize important hotspots and understand hotspot movement patterns. Third, at each level of abstraction, we design a unified weighted similarity measure through an iterative dynamic weight learning algorithm. With an optimal weight assignment scheme, we can efficiently combine the correlation information for each type of storage access patterns, such as random versus sequential, read versus write, to identify interesting spatial/temporal correlations hidden in the traces.

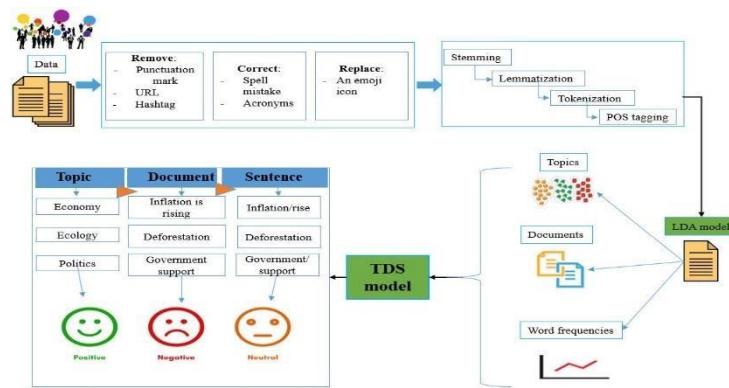
PROPOSED METHODOLOGY

This proposed Navie bayes model that classifies documents into reader-emotion categories. they studied the classification of news articles into different sentiment classes representing the emotions they trigger in their readers. This work mainly differs from other literature in focusing more on what the reader would feel while reading the article rather than what the writer was

feeling while writing it. Other than the classification itself, which has been detailed in our previous work, we study the impact of the number of sentiment classes on the classification performance (i.e., accuracy, precision, and recall). It analyze the results of the different experiments and conclude with the limitations that make multi-class classification a difficult task.

ADVANTAGES

- LDA is a powerful topic modeling technique that can uncover latent topics within a collection of text documents.
- LDA enables granular analysis by breaking down customer feedback into specific topics or themes. This allows businesses to gain insights into the various aspects of their products, services, or brand that are positively or negatively perceived by customers.
- With the right implementation and infrastructure, LDA-based sentiment analysis systems can provide real-time insights into customer sentiment.
- By identifying specific topics or themes driving customer sentiment, businesses can derive actionable insights to improve products, services, marketing strategies, and customer experience.



**Fig.1
WORKFLOW OF
PROPOSED
METHOD.**

RESULTS AND DISCUSSIONS

Data cleaning is one of the most important processes for obtaining accurate experimental results. To test our model, we used the IMDB movie review dataset . The dataset included 50,000 reviews that were evenly divided into positive and negative reviews. The dataset was divided into a 80% (4000) training set and a 20% (1000) testing set. First, unnecessary columns were removed. The second process included some spelling corrections, removing weird spaces in the text, html tags, square brackets, and special characters represented in text and contraction. They were then handled with emoji by converting them to the appropriate meaning of their occurrence in the document. Thereafter, we put all the text to lowercase, and removed text in square brackets, links, punctuation, and words containing numbers. Next, we removed stop words because having these makes our analysis less effective and confuses our algorithm. Subsequently, to reduce the vocabulary size and overcome the issue of data sparseness, stemming,

lemmatization, and tokenization processes were applied. We normalized the text in the dataset to transform the text into a single canonical form. With the aim of achieving better document classification, we also performed count vectorization for the bag-of- words (BOW) model. The BOW model can be used to calculate various measures to characterize the text. For this calculation process, the term frequency-inverse document frequency (TF-IDF) is the best method. Basically, TF-IDF reflects the importance of a word. We applied the N-gram model to avoid the shortcomings of BOW when dealing with several sentences with words of the same meaning. The N-gram model parses the text into units, including TF-IDF values. The N-gram model is an effective model representation used in sentiment analysis. Each N-gram related to parsed text becomes an entry in the feature vector with the corresponding feature value of TF-IDF. After preprocessing, the LDA model was applied. LDA is a three-level hierarchical Bayesian model that creates probabilities at the word level, on the document level, and on the

corpus level

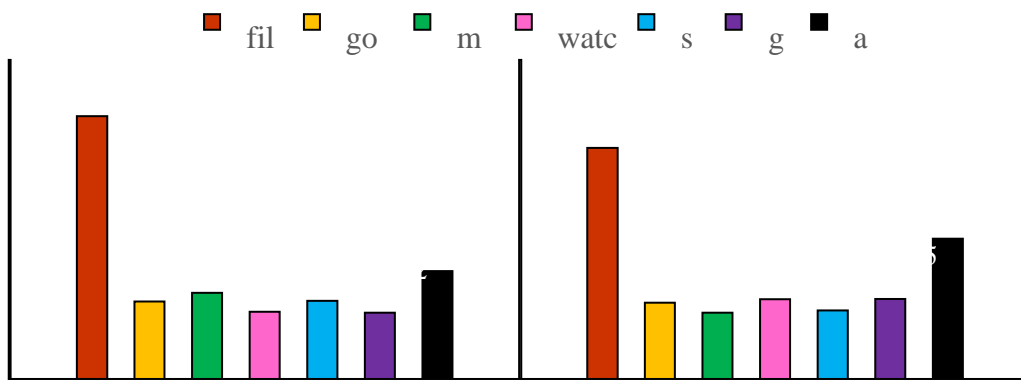


Figure 2. Most Common Sentiments.

CONCLUSION

In this project was studied the task of multi-class sentiment analysis and evaluated the evolution of various KPIs as the number of sentiment classes increased. It analyzed the difficulties of, and the different challenges involved with, multi-class classification, and proposed some metrics to measure the distance between sentiments (i.e., how similar they are to one another)and concluded that even though the task of multi-class analysis is important, it might be more interesting to perform a sentiment detection task through which all of the sentiments present within a text are extracted. In future this work will be experimented and tested in the public cloud based peta-sized datasets.

REFERENCES

- [1] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan, “Clustering data streams,” in Proceedings of the ACM Symposium on Foundations of Computer Science, 12- 14 Nov. 2000, pp. 359–366.
- [2] C. Aggarwal, Data Streams: Models and Algorithms, ser. Advances in Database Systems, Springer, Ed., 2007.
- [3] J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman & Hall/CRC, 2010.
- [4] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, “Data stream clustering: A survey,” ACM Computing Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.
- [5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in Proceedings of the International Conference on Very Large Data Bases (VLDB ’03), 2003, pp. 81–92.
- [6] F. Cao, M. Ester, W. Qian, and A. Zhou, “Density-based clustering over anevolving data stream with noise,” in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.
- [7] Y. Chen and L. Tu, “Density-based clustering for real-time stream data,” in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007, pp. 133–142.
- [8] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, “Density based clustering of data streams at multiple resolutions,” ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–28, 2009.
- [9] L. Tu and Y. Chen, “Stream data clustering based on grid density and attraction,” ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–27, 2009.
- [10] M. Ester, H.-P.Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’1996), 1996, pp. 226–231.

DETECTION AND CLASSIFICATION OF FRUIT DISEASES USING IMAGE PROCESSING TECHNIQUES

R. SOORYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The project has entitled as "FRUIT DISEASE IDENTIFICATION". Fruit diseases area major problem in economic and production losses in the agricultural industry worldwide. In this project, an image processing approach is proposed for identifying passion fruit diseases based on convolutional neural network. According to the CNN algorithm, fruit image details are taken by the existing packages from the front end used in this project. However, it can take a fewmoments. So, this proposed system can be used to identify fruit diseases quickly and automatically.

Keywords: diseases, economic and production, image processing, CNN algorithm

1. INTRODUCTION

Agriculture has been the basis for every people. More than 70% of Indians rely on agriculture for their primary source of income, making this issue critical. Nowadays the development of production of plants, crops and fruits are generally affected by the diseases. An agricultural land plagued by the disease is a serious issue. Bacteria and viruses cause disease in the majority of plants' leaves and fruits. For the detection of plant infection on the leaves, fruits or the stem, a technique similar to this is used. In order to develop an automated database for the suggested way of investigating infections. The information in the database pertains to plant leaves, fruit conditions, and disease signs that may be present. The fruit details and the detection of disease from the feature extraction are stored in the database. The complete database is inspected and the recorded image is compared to it. The smartphone app was created to aid in the processing of the data and to communicate updates to the farmers. In this way, the database's image variation also reveals disease in the fruits.

In recent years, the agricultural sector has faced significant challenges due to the prevalence of various diseases affecting fruits, leading to substantial losses in yield and quality. Timely detection and classification of fruit diseases are critical for effective disease management

and sustainable agriculture practices. Traditional methods of disease detection often rely on visual inspection by trained personnel, which can be time-consuming, subjective, and prone to human error. The emergence of image processing techniques has revolutionized the field of agriculture by offering non-invasive, rapid, and accurate solutions for disease detection and classification. Leveraging advancements in computer vision, machine learning, and pattern recognition, researchers have developed sophisticated algorithms capable of analyzing images of fruits to identify and classify diseases with high precision.

This project aims to explore the application of image processing techniques for the detection and classification of fruit diseases. By harnessing the power of digital image analysis, we seek to develop an automated system that can assist farmers and agricultural experts in early disease diagnosis and management. Through the integration of cutting-edge technologies, we aspire to enhance productivity, minimize crop losses, and promote sustainable agricultural practices.

The main aim of this project is to propose a new dataset of images containing popular fruits. The reader is encouraged to access the latest version of the dataset from the above indicated addresses. Having a high-quality dataset is essential for obtaining a good classifier. Most of the existing datasets with images (see for instance the popular CIFAR dataset) contain both the object and the noisy background. This could lead to cases where changing the background will lead to the incorrect classification of the object. As a second objective we have trained a deep neural network that is capable of identifying fruits from images. This is part of a more complex project that has the target of obtaining a classifier that can identify a much wider array of objects from images. This fits the current trend of companies working in the augmented reality field. During its annual I/O conference, Google announced that is working on an application named Google Lens which will tell the user many useful information about the object toward which the phone camera is pointing. First step in creating such application is to correctly identify the objects.

The software has been released later in 2017 as a feature of Google Assistant and Google Photos apps. Currently the identification of objects is based on a deep neural network. Such a network would have numerous applications across multiple domains like autonomous

navigation, modeling objects, controlling processes or human-robot interactions. The area we are most interested in is creating an autonomous robot that can perform more complex tasks than a regular industrial robot. An example of this is a robot that can perform inspections on the aisles of stores in order to identify out of place items or under-stocked shelves. Furthermore, this robot could be enhanced to be able to interact with the products so that it can solve the problems on its own. Another area in which this research can provide benefits is autonomous fruit harvesting. While there are several papers on this topic already, from the best of our knowledge, they focus on few species of fruits or vegetables. In this paper we attempt to create a network that can classify a variety of species of fruit, thus making it useful in many more scenarios. As the start of this project we chose the task of identifying fruits for several reasons. On one side, fruits have certain categories that are hard to differentiate, like the citrus genus, that contains oranges and grapefruits. Thus we want to see how well can an artificial intelligence complete the task of classifying them. Another reason is that fruits are very often found in stores, so they serve as a good starting point for the previously mentioned project.

The project is structured as follows: in the first part we will shortly discuss a few outstanding achievements obtained using deep learning for fruits recognition, followed by a presentation of the concept of deep learning. In the second part we describe the Fruits-360 dataset: how it was created and what it contains. In the third part we will present the framework used in this project - TensorFlow and the reasons we chose it. Following the framework presentation, we will detail the structure of the neural network that we used. We also describe the training and testing data used as well as the obtained performance. Finally, we will conclude with a few plans on how to improve the results of this project. Source code is listed in the Appendix

2. LITERATURE REVIEW

Plant Disease Detection Using Leaf Pattern: A Review(2015). Author :Vishnu S, A. Ranjith Ram. In this paper, they discuss a various methodologies for plant disease detection. Studies show that relying on the pure naked-eye observation of experts to detect and classify the diseases can be time consuming and expensive, especially in the rural areas and developing countries. So they present fast, automatic, cheap and the accurate image processing based solution. Solution is composed of the four main phases; in the first phase we create a color transformation structure for the RGB leaf image and then, they apply an colour space

transformation for the colour transformation structure. Next, in the second phase, the images are segmented using a K-means clustering technique. In third phase, they calculate an texture features for the segmented infected objects. Finally, in the fourth phase the extracted features are passed through the pre-trained neural network.

Detection of Diseases on Cotton Leaves Using K Mean Clustering Method(2015). Pawan P. Warne, Dr. S. R. Ganorkar . In this paper,presents an approach for the careful detection of diseases, diagnosis and the timely handling to prevent crops from the heavy losses. The diseases on the cotton are critical issue which makes the sharp decrease in production of cotton. So for the study of interest is the leaf rather than the whole cotton plant because about 8595 of the diseases occurred on the cotton leaves like Alternaria, Cercospora and Red Leaf Spot. In this proposal initially a preprocessing the input image using the histogram equalization is applied to increase the contrast in the low contrast image, K means clustering algorithm is used for an segmentation which classifies objects based on a set of features into K number of classes and finally classification is performed using Neural network.Thus image processing technique is used for detecting diseases on cotton leaves early and accurately. It is used to analyze the cotton diseases which will be useful to the farmers. Combining Local and Global Image Features for Object Class Recognition(2016). Author :Dimitri A. Lisin, Marwan A. Mattar, Matthe w B.Blaschko.

In this paper, the Object recognition is an central problem in the computer vision research. Most object recognition Systems have taken one of the two approaches, using either global or Local features exclusively. This may be in part due to the difficulty of combining an single global feature vector with a set of local features in a suitable manner. In this paper , they show that combining local and the global features is beneficial in an application where rough segmentations of objects are available . They present a method for the classification with the local features using a non-parametric Density estimation. Subsequently, they present the Two methods For combining the Local and Global features. They first used a "stacking" ensemble technique, and the Second uses an hierarchical classification system. Results show the superior performance of these combined methods over an component classifiers, with the reduction of over 20 in an error rate on the challenging marine science application. A Study and Implementation of Active Contour Model For Feature Extraction: With Diseased Cotton Leaf as Example(2016). P.R.Rothe A and R. V. Kshirsagar A. In this paper, the Feature extraction is an

significant constituent of the pattern recognition system. It carries out the two assignments: converting input parameter vector into the feature vector and or reducing its dimensionality.

The distinct feature extraction algorithm makes a classification process more effectual and efficient. The allocation and recognition of the cotton leaf diseases are of the major importance as they have a cogent and the momentous impact on the quality and production of the cotton. In this work, they present the snake based approach for the segmentation of images of the diseased cotton leaves. They extract Hu's moments which can be used as the shape descriptors for the classification. A theory of the two-dimensional moment invariants for the planar geometric figures is also presented. Three diseases have been considered, namely the Bacterial Blight, Classification of Cotton Leaf Spot Diseases Using Image Processing Edge Detection Techniques(2017). :P.Revathi, M.Hemalatha. In this Paper, an advance computing technology that has been developed to help the farmer to take the superior decision about many aspects of the crop development process. Suitable evaluation and an diagnosis of the crop disease in the field is very critical for the increased production. Foliar is a major important fungal disease of the cotton and occurs in all the growing Indian regions. In this work, they express the new technological strategies using the mobile captured symptoms of the cotton leaf spot images and categorize the diseases using the HPCDD Proposed Algorithm. The classifier is being trained to achieve an intelligent farming, including early Identification of a diseases in the groves, selective fungicide application, etc. This proposed work is based on the Image RGB feature ranging techniques used to identify the diseases (using Ranging values) in which, the captured images are processed for the enhancement first. Then color image segmentation is carried out to get the target regions (disease spots). Next Homogenize techniques like the Sobel and the Canny filter are used to Identify the edges, these extracted edge features are used in the classification to identify the disease spots. Finally, the pest recommendation is given to the farmers to ensure that their crop and reduce the yield loss.

3. PROPOSED METHODOLOGY

The proposed system leverages image processing techniques to automate the detection and classification of fruit diseases, offering a more efficient, accurate, and environmentally friendly solution. High-resolution images of fruits exhibiting disease symptoms are captured using digital cameras or smartphones equipped with suitable imaging capabilities. Preprocessing techniques such as noise reduction, image enhancement, and color correction are applied to

improve the quality and clarity of acquired images, ensuring optimal input for subsequent analysis. Machine learning algorithms, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), or decision trees, are trained using the extracted features to classify fruits into different disease categories. The system learns patterns and characteristics associated with each disease, enabling accurate identification and classification of fruit diseases. The system is integrated into a user-friendly software interface or mobile application, allowing farmers and agricultural experts to easily capture and upload images of diseased fruits for analysis. The interface provides real-time feedback on disease diagnosis and recommends appropriate management strategies based on the identified disease type.

ADVANTAGES

- The proposed system automates the process of fruit disease detection and classification, reducing reliance on subjective human judgment and streamlining the diagnostic process.
- By leveraging advanced image processing techniques and machine learning algorithms, the proposed system achieves higher accuracy in disease diagnosis compared to manual inspection methods.
- The real-time analysis provided by the proposed system enables prompt detection and diagnosis of fruit diseases, allowing farmers to implement timely intervention measures and minimize crop losses.
- By accurately identifying specific diseases affecting fruits, the proposed system enables targeted treatment strategies, reducing the need for indiscriminate use of pesticides and promoting environmentally sustainable farming practices.
- The system can be scaled to accommodate large agricultural operations, providing a cost-effective and efficient solution for disease management across diverse crop varieties and geographical regions.

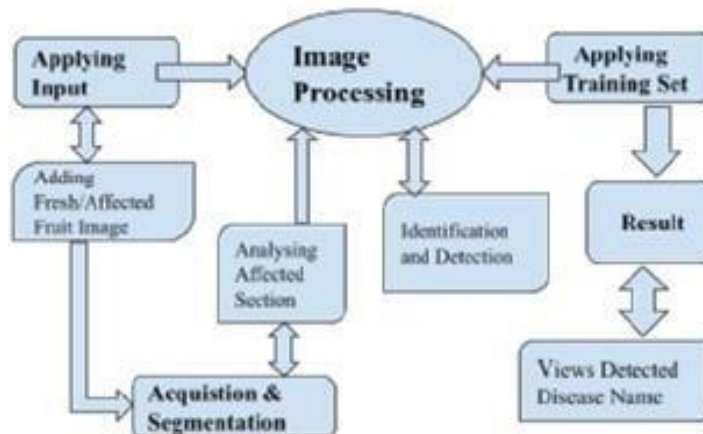
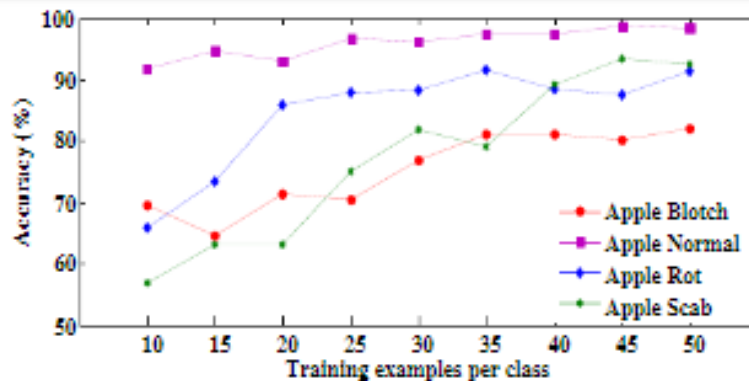


Figure 1: Disease detection and analyzing in fruit disease

4. RESULTS AND DISCUSSIONS

To demonstrate the performance of the proposed approach, we have used a data set of normal and diseased apple fruits, which comprises four different categories: Apple Blotch (104), Apple rot (107), Apple scab (100), and Normal Apple (120): totalizing 431 apple fruit images. Fig. 6 depicts the classes of the data set. Presence of a lot of variations in the type and color makes the data set more realistic. One important aspect when dealing with apple fruit disease classification is the accuracy per class. This information points out the classes that need more attention when solving the confusions. Fig. 9 depicts the accuracy for each one of 4 classes using LBP and CLBP features in RGB and HSV color spaces. Clearly, Apple Blotch is one class that needs attention in both color spaces. It yields the lowest accuracy when compared to other classes. Fig. 9 also shows that, the behavior of Apple Rot is nearly same in each scenario. Normal Apples are very easily distinguishable with diseased apples and a very good classification result is achieved for the Normal Apples in both color spaces. For CLBP feature and HSV color space, for instance, reported classification accuracy are 89.88%, 90.71%, 96.66%, and 99.33% for the Apple Blotch, Apple Rot, Apple Scab, and Normal Apple respectively, resulting average accuracy 93.14% when training is done with 50 images per class



Graph1 : Accuracy level of MSVM classifier

5. CONCLUSION

The creation of cloud based plan for helping Indian farmers and agricultural, helps to analyze the agriculture data in a better manner to minimize the hoardings and in putting up a wealthy safe and India's tranquil agricultural society. The classification and segmentation of fruit pictures were Used K-Means Algorithm and Support Vector Machines technique. The varied properties of several fruits were first extracted and segment the related photos. Feature values were used to compare each of these. A database of illness names is searched for the best match. for the picture is detected and the ailment is indicated This may be shown as a message in an alert box the use of a smartphone app. The total number of samples presented, the genuine and false places, the true and false negatives, as well as accuracy and relevance, An warning box also indicates the specificity. to boot complete database of fruit illnesses and the control Prevention strategies are kept in a safe place. Cloud database and the data may be obtained via the application.

6. REFERENCES

- [1] J. Hartman, —Apple Fruit Diseases Appearing at Harvest, Plant Pathology Fact Sheet, College of Agriculture, University of Kentucky, http://www.ca.uky.edu/agcollege/plantpathology/ext_files/PPFShtml/PPFS-FR-T-2.pdf, viewed on December 2011.
- [2] Q. Li, M. Wang, and W. Gu, —Computer vision based system for apple surface defect detection, Computers and Electronics in Agriculture, vol. 36, pp. 215–223, Nov. 2002.
- [3] P. M. Mehl, K. Chao, M. Kim, and Y. R. Chen, —Detection of defects on selected apple cultivars using hyperspectral and multispectral image analysis, Applied Engineering in Agriculture, vol. 18, pp. 219–226, 2002.

- [4] M. S. Kim, A. M. Lefcourt, Y. R. Chen, and Y. Tao, —Automated detection of fecal contamination of apples based on multispectral fluorescence image fusion,|| *Journal of food engineering*, vol. 71, pp. 85–91, 2005.
- [5] O. Kleynen, V. Leemans, and M. F. Destain, —Development of a multi-spectral vision system for the detection of defects on apples,|| *Journal of Food Engineering*, vol. 69, pp. 41–49, 2005.
- [6] V. Leemans, H. Magein, and M. F. Destain, —Defect segmentation on ‘jonagold’ apples using colour vision and a bayesian classification method,|| *Computers and Electronics in Agriculture*, vol. 23, pp. 43–53, June 1999.
- [7] V. Leemans, H. Magein, and M. F. Destain, —Defect segmentation on ‘golden delicious’ apples by using colour machine vision,|| *Computers and Electronics in Agriculture*, vol. 20, pp. 117–130, July 1998.
- [8] T. Ojala, M. Pietikäinen, and T. T. Mäenpää, —Multiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern,|| *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.
- [9] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen, —Outex – new framework for empirical evaluation of texture analysis algorithm,|| in *Proc. International Conference on Pattern Recognition*, 2002, pp. 701-706.
- [10] T. Ahonen, A. Hadid, and M. Pietikäinen, —Face recognition with Local Binary Patterns: application to face recognition,|| *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.

DETECTION OF MENTAL DISORDERS IN SOCIAL MEDIA NETWORK

R. ELAKKIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The explosive growth in popularity of social networking leads to the problematic usage. An increasing number of social network mental disorders (SNMDs), such as Cyber-Relationship Addiction, Information Overload, and Net Compulsion, have been recently noted. Symptoms of these mental disorders are usually observed passively today, resulting in delayed clinical intervention. In this paper, users argue that mining online social behavior provides an opportunity to actively identify SNMDs at an early stage. It is challenging to detect SNMDs because the mental status cannot be directly observed from online social activity logs. This approach, new and innovative to the practice of SNMD detection, does not rely on self-revealing of those mental factors via questionnaires in Psychology. Depression detection taken from tweets records are carried out after proper preprocessing steps. In addition, to address the aforementioned issue, this paper proposed a depression detection classification model using K-Nearest Neighbor and a Support Vector Machine based classification.

Keywords: social network mental disorders, Net Compulsion, K-Nearest Neighbor, Support Vector Machine

1. INTRODUCTION

In many application domains (e.g., medicine or biology), comprehensive questions for comparative schemas resulting from collaborative initiatives are made available [1]. This proposed system argues that to achieve the on-demand combination for superlative and special cases based resource management for Web-based e-learning, one should go beyond using domain ontology's statically. So the propose XAML based matching process involves semantic mapping has done on both the open dataset and closed dataset mechanism to integrate e-learning databases by using ontology semantics [2].

It defines context-specific portions from the whole ontology as optimized data and an XAML based resource reuse approach by using an evolution algorithm [3]. Such well established schemas are often associated with reliable data that have been carefully collected, cleansed, and verified, thus providing reference for comparable question based data management systems (DMSs) in different application domains [4]. A good practice is therefore to build on the efforts made to design reference DMSs whenever we have to develop our own DMS with specific needs [5].

A way to do this is to extract from the reference DMS the piece of schema relevant to our application needs, possibly to personalize it with extra constraints w.r.t. our application under construction, and then to manage our own data set using the resulting schema [6]. Recent work in description logics provides different solutions to achieve such a reuse of a reference ontology-based DMS. Indeed, modern ontological languages like the W3C recommendations RDFS, OWL, and OWL2 are actually XML based syntactic variants of well known DLs [7]. All those solutions consist in extracting a module from an existing ontological schema such that all the constraints concerning the relations of interest for the application under construction are captured in the module. Existing definitions of modules in the literature basically in this system, revisit the reuse of reference ontology based DMS in order to build a new DMS with specific needs [8]. It goes one step further by not only considering the design of a module based DMS (i.e., how to extract a module from an

ontological schema) also study how a module based DMS can benefit from the reference DMS to enhance its own data management skills. Our contribution is to introduce and study novel properties of robustness for modules that provide means for checking easily that a robust module based comparative evolves safely w.r.t. both the schema and the data of the reference DMS [9].

From a module robust to consistency checking, for any data update in corresponding module-based comparative questions, we show how to query the reference questions for checking whether the local update does not bring any inconsistency with the data and the constraints of the references. It is from a module robust to query answering, for any query asked to module-based DMS, It shows how to query the reference DMS for obtaining additional answers by also exploiting the data stored in the reference DMS.

The explosive growth in popularity of social networking leads to the problematic usage. An increasing number of social network mental disorders (SNMDs), such as Cyber-Relationship Addiction, Information Overload, and Net Compulsion, have been recently noted. Symptoms of these mental disorders are usually observed passively today, resulting in delayed clinical intervention. In this paper, users argue that mining online social behavior provides an opportunity to actively identify SNMDs at an early stage. It is challenging to detect SNMDs because the mental status cannot be directly observed from online social activity logs. This approach, new and innovative to the practice of SNMD detection, does not rely on self-revealing of those mental factors via questionnaires in Psychology. Depression detection taken from tweets records are carried out after proper preprocessing steps. In addition, to address the aforementioned issue, this paper proposed a depression detection classification model using K-Nearest Neighbor and a Support Vector Machine based classification.

2. LITERATURE SURVEY

Conventional Data Management Strategies

In conventional information management principles, the stored records are normally identified by sets of key words or index terms, and requests for information are expressed by using Boolean combinations of index terms. The retrieval strategy is normally based on an auxiliary inverted-term index that lists the corresponding set of document references for each allowable index term. The Boolean retrieval system is designed to retrieve all stored records exhibiting the precise combination of key words included in the query: when two query terms are related by an and connective, both terms must be present in order to retrieve a particular stored record; when an or connective is used, at least one of the query terms must be present to retrieve a particular item.

In some systems where the natural language text of the documents or the document excerpts is stored, the user queries may be formulated as combinations of text words. In that case, the queries may include location restrictions for the query terms- for example, a requirement that the query terms occur in the same sentence of any retrieved document or within some specified number of words of each other.

Boolean data management systems have become popular in operational situations because high standards of performance are achievable. The retrieval technology which is based on list intersections and list unions to implement Boolean conjunction ("A and

B") and Boolean disjunction ("A **or** B"), respectively, is now well understood. The conventional Boolean retrieval technology is however also saddled with various disadvantages:

1. The size of the output obtained in response to a given query is difficult to control; depending on the assignment frequency of the query terms and the actual term combinations used in a query formulation, a great deal of output can be obtained or, alternatively, no output might be retrieved at all.

2. The output obtained in response to a query is not ranked in any order of presumed importance to the user; thus, each retrieved item is assumed to be as important as any other retrieved item.

3. No provisions are made for assigning importance factors or weights to the terms attached either to the documents or dataset.

Data Evaluation

Lee gives an overview of the different models that have been proposed, and shows that only the p-norm model has two key properties that, if not present, are detrimental to retrieval effectiveness.

Smith proposed to recursively aggregate inverted lists, calculating and storing intermediate scores for every document that is encountered in any of the lists, in what is referred to as being the term-at-a-time approach. In effect, not all nodes in the query tree are visited for every document, but all of the inverted lists are fully inspected, and temporary memory proportional to the total size of the relevant inverted lists is required. Smith's Infinity-One method gives an approximation to the p-norm model, with the aim of reducing computational cost by reducing the volume of floating point operations. As is demonstrated below, the number of score calculations can be greatly reduced via an exact lossless pruning approach.

Turtle and Flood describe the max-score ranking mechanism, to accelerate keyword query evaluation when sum-score aggregation functions are used and only the top-k documents are required. Using document-at-a-time evaluation, the algorithm commences by fully scoring the first k documents in the OR-set of the query terms. Thereafter, the kth largest document score is tracked, as an entry threshold that candidate documents must exceed before they can enter the (partial) ranking. The max-score algorithm uses the information conveyed by the entry threshold to reduce two cost factors: 1) the number of candidate documents that are scored; and 2) the cost associated with scoring each candidate document.

Content-based Methods

In content-based recommendation methods, the utility $u(c, s)$ of item s for user c is estimated based on the utilities $u(c, s_i)$ assigned by user c to items $s_i \in S$ that are "similar" to item s . For example, in a movie recommendation application, in order to recommend movies to user c , the content-based recommender system tries to understand the commonalities among the movies user c has rated highly in the past (specific actors, directors, genres, subject matter, etc.). Then, only the movies that have a high degree of similarity to whatever user's preferences are would get recommended. The content-based approach to recommendation has its roots in information retrieval and information filtering research.

Because of the significant and early advancements made by the information retrieval and filtering communities and because of the importance of several text-based applications, many current content-based systems focus on recommending items containing textual information, such as documents, Web sites (URLs), and Usenet news messages. The improvement over the traditional information retrieval approaches comes from the use of user profiles that contain information about users' tastes, preferences, and needs. The profiling information can be elicited from users explicitly, e.g., through questionnaires, or implicitly – learned from their transactional behavior over time.

More formally, let Content(s) be an item profile, i.e., a set of attributes characterizing item s. It is usually computed by extracting a set of features from item s (its content) and is used to determine appropriateness of the item for recommendation purposes. Since, as mentioned earlier, content-based systems are designed mostly to recommend text-based items, the content in these systems is usually described with keywords.

For example, a content-based component of the Fab system which recommends Web pages to users, represents Web page content with the 100 most important words. Similarly, the Syskill & Webert system represents documents with the 128 most informative words. The "importance" (or "informativeness") of word k_i in document d_j is determined with some weighting measure $w_{i,j}$ that can be defined in several different ways.

One of the best-known measures for specifying keyword weights in Information Retrieval is the term frequency/inverse document frequency (TF-IDF) measure that is defined as follows. Assume that N is the total number of documents that can be recommended to users and that keyword k_i appears in n_i of them. Moreover, assume that $f_{i,j}$ is the number of times keyword k_i appears in document d_j . Then $f_{i,j}$, the term frequency (or normalized frequency) of keyword k_i in document d_j , is defined as

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

where the maximum is computed over the frequencies $f_{z,j}$ of all keywords k_z that appear in the document d_j . However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a non-relevant one. Therefore, the measure of inverse document frequency (IDF_i) is often used in combination with simple term frequency ($f_{i,j}$). The inverse document frequency for keyword k_i is usually defined as

$$IDF_i = \log \frac{N}{n_i}$$

Then the TF-IDF weight for keyword k_i in document d_j is defined as

$$w_{i,j} = TF_{i,j} \times IDF_i$$

In content-based systems, the utility function $u(c, s)$ is usually defined as:
 $u(c, s) = \text{score}(\text{ContentBasedProfile}(c), \text{Content}(s))$ Using the above-mentioned information retrieval-based paradigm of recommending Web pages, Web site URLs, or Usenet news

messages, both ContentBasedProfile(c) of user c and Content(s) of document s can be represented as TF-IDF vectors c_w and s_w of keyword weights. Moreover, utility function $u(c, s)$ is usually represented in information retrieval literature by some scoring heuristic defined in terms of vectors c_w and s_w , such as cosine similarity measure

$$u(c, s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\|\vec{w}_c\|_2 \times \|\vec{w}_s\|_2} = \frac{\sum_{i=1}^K w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^K w_{i,c}^2} \sqrt{\sum_{i=1}^K w_{i,s}^2}}$$

3. PROPOSED METHODOLOGY

This proposed system Hybrid extraction of robust model (HERM) is very efficient and reuse of e-learning resources in a distributed environment like the Web for better result. This proposed system argues that to achieve the on-demand semantic-based resource management for Web-based e-learning, one should go beyond using domain ontology's statically.

The proposed approach for matching process in web cluster databases from different database servers can be easily integrated and deliver highly dimensional e-learning resource management and reuse is far from being mature. However, e-learning is also a widely open research area, and there is still much room for improvement on the method. This research mechanism includes 1) improving the proposed evolution approach by making use of and comparing different evolutionary algorithms, 2) applying the proposed approach to support more applications, and 3) extending to the situation with multiple e-learning systems or services.

ADVANTAGES:

1. In this system concentrate on advanced data matching and extraction is not limited to number of database servers.
2. This hybrid data extraction gives better performance than the existing models because this system uses information content that declared inside the OWL,RDF,xml and XAML languages.
3. This system differs highly from existing systems by dealing the with the object identification in the open datasets and closed datasets.
4. In this system the conventional ontology based matching are optimized in both dependent data with different subsets and independent data.
5. This XAML based system is uses the integrated approach to match and extract the information from the huge database systems from different number of clusters.
6. This proposed system has been designed to perform the matching in both open dataset and closed dataset.

4. RESULTS AND DISCUSSIONS

- If a sequential pattern can be used to extract many reliable comparator pairs, it is very likely to be an IEP.
- If a comparator pair can be extracted by an IEP, the pair is reliable.

Based on these two assumptions, we design our bootstrapping algorithm as shown in Figure 1. The bootstrapping process starts with a single IEP. From it, we extract a set of initial seed comparator pairs. For each comparator pair, all questions containing the pair are retrieved from a question collection and regarded as comparative questions. From the comparative questions and comparator pairs, all possible sequential patterns are generated and evaluated by measuring their reliability score defined later in the Pattern Evaluation section. Patterns evaluated as reliable ones are IEPs and are added into an IEP repository.

Then, new comparator pairs are extracted from the question collection using the latest IEPs. The new comparators are added to a reliable comparator repository and used as new seeds for pattern learning in the next iteration. All questions from which reliable comparators are extracted are removed from the collection to allow finding new patterns efficiently in later iterations. The process iterates until no more new patterns can be found from the question collection.

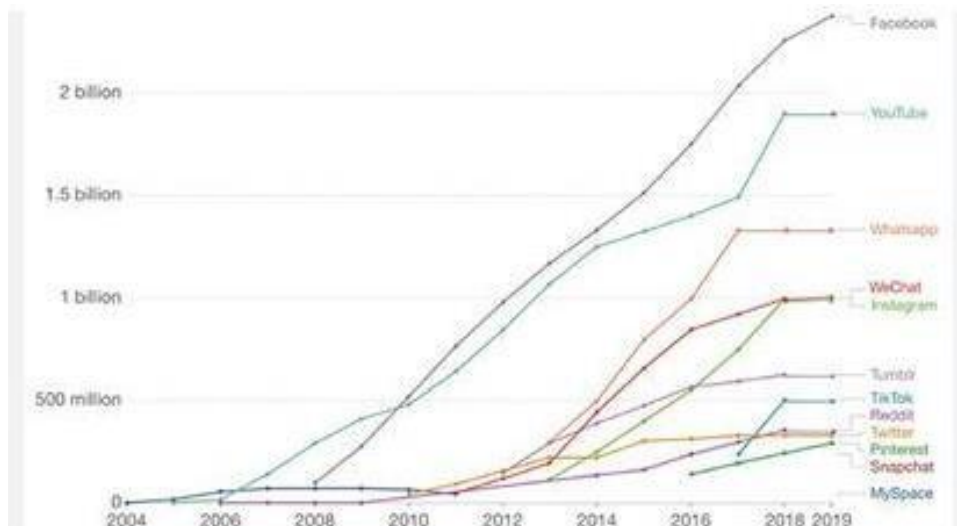
There are two key steps in our method: (1) pattern generation and (2) pattern evaluation. In the following subsections, we will explain them in details.

Pattern Generation

To generate sequential patterns, we adapt the surface text pattern mining method introduced in (Ravichandran and Hovy, 2002). For any given comparative question and its comparator pairs, comparators in the question are replaced with symbol \$C. Two symbols, #start and #end, are attached to the beginning and the end of a sentence in the question. Then, the following three kinds of sequential patterns are generated from sequences of questions:

- **Lexical patterns:** Lexical patterns indicate sequential patterns consisting of only words and symbols (\$C, #start, and #end). They are generated by suffix tree algorithm (Gusfield, 1997) with two constraints: A pattern should contain more than one \$C, and its frequency in collection should be more than an empirically determined number β .
- **Generalized patterns:** A lexical pattern can be too specific. Thus, we generalize lexical patterns by replacing one or more words with their POS tags. $2^n - 1$ generalized patterns can be produced from a lexical pattern containing N words excluding \$Cs.
- **Specialized patterns:** In some cases, a pattern can be too general. For example, although a question "ipod or zune?" is comparative, the pattern "<\$C or \$C>" is too general, and there can be many noncomparative questions matching the pattern, for instance, "true or false?". For this reason, we perform pattern specialization by adding POS tags to all comparator slots. For example, from the lexical pattern "<\$C or \$C>" and the question "ipod or zune?", "<\$C/NN or \$C/NN?>" will be produced as a specialized pattern.

Note that generalized patterns are generated from lexical patterns and the specialized patterns are generated from the combined set of generalized patterns and lexical patterns. The final set of candidate patterns is a mixture of lexical patterns, generalized patterns and specialized patterns.



Graph.1 shown the development of mental disorder

5. CONCLUSION

The proposed technique HERM has a novel weakly supervised method to identify comparative questions and extract comparator pairs simultaneously. It rely on the key insight that a good comparative question identification pattern should extract good comparators, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. The experimental results show that our method is effective in both comparative question identification and comparator extraction. It significantly improves recall in both tasks while maintains high precision. It shows the examples show that these comparator pairs reflect what users are really interested in comparing. Our comparator mining results can be used for a commerce search or product recommendation system. For example, automatic suggestion of comparable entities can assistusers in their comparison activities before making their purchase decisions. Also, our results can provide useful information to companies which want to identify their competitors.

6. REFERENCES

- [1] .Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In Proceedings of AAAI'99 /IAAI'99.
- [2] Claire Cardie. 1997. Empirical methods in information extraction. AI magazine, 18:65–79. Dan Gusfield. 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, New York, NY, USA
- [3] Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In Proceedings of WWW '02, pages 517–526.

- [4] Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In Proceedings of WWW '03,
- [5] pages 271–279.
- [6] Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In Proceedings
- [7] of SIGIR '06, pages 244–251. Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In Proceedings of AAAI '06.
- [8] Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with
- [9] hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT, pages 1048–1056.

EMBED IMAGE ENCRYPTION USING 126-BIT BASE RANDOM BIT SHIFT TECHNIQUE

S. KANNIKA

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

Privacy is a critical issue when the data owners outsource data storage or processing to a third party computing service, such as the cloud. In this project, we identify a cloud computing application scenario that requires simultaneously performing secure watermark detection and privacy preserving multimedia data storage. This proposed technique a compressive sensing (CS)-based framework using secure multiparty computation (MPC) protocols to address such a requirement. In this framework, the multimedia data and secret watermark pattern are presented to the cloud for secure watermark detection in a CS domain to protect the privacy. During CS transformation, the privacy of the CS matrix and the watermark pattern is protected by the MPC protocols under the semi-honest security model. The expected watermark detection performance in the CS domain, given the target image, watermark pattern, and the size of the CS matrix (but without the CS matrix itself). The correctness of the derived performance has been validated by our experiments. Our theoretical analysis and experimental results show that secure watermark detection in the CS domain is feasible. Our framework can also be extended to other collaborative secure signal processing and data-mining applications in the cloud.

Keywords: multiparty computation, watermark detection, CS, secure signal processing

1. INTRODUCTION

Image processing is the quick emerging area of computer science. The growth of image processing was provided by the technology development like digital visualizing, computer processor and large storage devices. Image processing permits to compute the image in multidimensional with the system. The images itself has a sub images that are named as ROI region of interest which is used to find the object in the image. Most of the fields like medicine, film, security monitoring, photography and remote sensing which are using the analog imaging are changing over to digital system because of its conciliatory and significant cost. Because these areas produce the large volumes everyday which could not be audited manually.

Digital image processing gets over these drawbacks and extracts the needed information from the images. Hence there is no need of manpower to audit the process of extraction which done by the computer. There are three levels of image processing algorithms. At the low level it deals with pixel value, for example edge detection and de-noising. With these low level results it proceeds from the middle level for further process like segmentation. And at the next level, it utilizes some methods to extract the useful information for face detection.

Image Processing Operations:

Image processing operation can be applied to the digital image processing to getting craved output. There are three types of operation. The first operation is pointing its character and the output value of coordinate depends on the same coordinate value of the input. The second operation is localizing the character which output coordinate value depends on the coordinate value of neighborhood of input. The third operation is global and its output coordinate value depends on the value of the image. The types of operation are depicted as the diagram below. Some of the digital image processing tools are Fourier analysis, convolution etc.

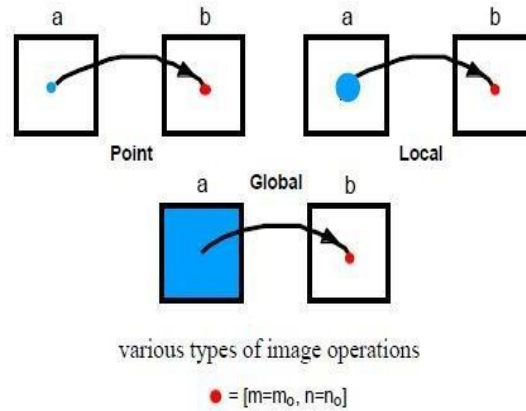


Figure 1: Image operations types

Working of Image Processing

The significant demand to process the image is that each and every image should be represented in digital form, which is a collection of binary words of fixed length. Image can be digitized by means of sampling and quantization. The resultant digital image is analyzed by computer. The given input image is transformed into an analog signal, and then scanned to display the output as a digital image. Some widely used techniques are computer graphics and computer vision.

Computer Graphics:

Computer graphics are used to display the graphics by the use of computers and its hardware and software. Interpretation and interaction of the data have been made easier by the use of hardware and software. Computer graphics are used in gaming industry and animated movies. In computer vision, images are obtained from the models such as lighting and environment.

Computer Vision

Computer vision is a technique which is used to learn, calculate, examine and realize the images. It is also called as "machine Vision" and "high level image processing technique". It is widely used in navigation, inspection and also to detect events. It is also used to decode the original content of an image. Image Processing can be done by digitization and

quantization. Digitization performs image sampling and the sampled values are used for quantization. Processing can be done by converting the images into bits. The Processing techniques are Image enhancement, restoration and Compression.

Image Enhancement

Image enhancement means improving the quality of an input image in order to overcome the weakness of the human visual system. The idea behind the image enhancement technique is to transform the poor quality image into high quality, so that the look of an image can be improved. The enhancement increases the range of the chosen feature rather than increasing the inherent information content of the data, so that they can be detected easily. Many features can be altered during enhancement process. The result of an image is more suitable for specific application.

Image Analysis:

Image Analysis is the process for examining the digital image data in order to smooth the progress of solving image vision problem. Image analysis involves two processes. They are feature extraction and pattern classification. Feature Extraction can reduce the dimensionality of an image data. Feature extraction may be two types. They are holistic and local. Extracting relevant information from an image for performing some desired task in any application. It is a procedure for calculating the basis vectors of the feature space. Pattern classification is the process for classifying or identifying images based on the feature vector which is obtained during feature extraction.

Image restoration:

Image restoration is most widely used to filter and reduce the noise of an image. Effectiveness and accuracy of an image restoration technique rely on the design of the filter. It is used to represent the clean original image by reducing the corruption of an image such as noise and motion blur.

Image compression:

Image compression is used to minimize the redundant and irrelevant content of an image. Image compression sounds to be more effective for storing and transmitting purposes. It is most widely used in applications such as television station, satellite, maps and in geological surveys. Image compression can be classified into two types namely lossy and lossless compression. Image compression can be done for texts, images associated with JPEG and with MPEG.

Transforming Image Geometry:

Transforming of an image based on mathematical functions and geometric representation. Image transformation is used to transpose, scale, crop, rotate and pad or alter an array of an image which produce modified image. In geometric transformation, the pixels relocate to a new position from original spatial coordinates of the source image to output image. The relocated pixel falls nearer to the pixels in the central location and it does not

directly map to the pixel central location. The values of the nearer pixels sampling computes the pixel's value. The image quality of the output affects by re-sampling is called as interpolation.

Mapping an Image onto Geometry:

Texture mapping is a mapping of an image onto the geometry, which involves an image overlaying. Realistic images may be used like elevation or temperature of color coded function, satellite images. Visualization of volume, which render a three dimensional scene of voxel and geometry mapping of an image effectively creates the complexity appearance of an image which layering onto the surface. The information provided by the result is not readily apparent, but display of either geometric surface or an image. It is a two step process.

- 1) Object space, which maps the image onto the geometric surface.
- 2) Screen space, which is in the form 2D transformation of the surface

Image Types:

Binary Image:

Binary Image is a Bi-level image which contains two values for each pixel. The values may be either 1 or 0. This means that 1 represent 'white' color and 0 represent 'black' color. Threshold a grayscale image or color image can lead to formation of Binary image.

Grayscale Image:

A grayscale image is simply called as one color image or monochrome image. It can hold only brightness or intensity information not color information. Grayscale values have ranged from 0 to 255.

Color Image:

Color image contains three band color information for each pixel. Each band has different color information. Commonly RGB color band information is used for any image display. YCbCr and HSV are also used as color space in some other context. Each band can range from 0 to 255 values.

2. LITERATURE SURVEY

The scheme of watermarking is classified into symmetric and asymmetric watermarking. A symmetric method uses same key for watermark embedding and watermark detection. With the identical key the watermark can be easily removed. To overcome this problem asymmetric watermarks are proposed. It involves different keys for embedding and detection. Hyuk Choi et al. suggests an asymmetric watermarking system that applies linear Transformation to obtain the private encoding key and the public decoding key. The difficulty of this system is decoding key that is publicly available. Different transformation is needed for primitive key for security purpose.

Cox et al suggested the spread spectrum watermarking for embedding the private watermark. The pattern of the watermark is detected even it is watermarked multiple times. Private watermark uses Pseudo random noise sequence. Trichili et al proposed a virtual border

for DICOM images to store the watermark. Miaou et al suggest the work using patient data in watermarking. Data is hidden in the LSB of the image. With the limited scope Zain et al proposed reversible watermarking techniques. Zhou et Al presents a method for encrypting digital signatures. This method has better authentication and integrity. Coatrieux et al suggested watermarking algorithm for medical images. In most of the papers embedded information is in the non-ROI region.

Eggers et al proposed the symmetric methods with the combination of public detectors. In this technique the watermark is removed simultaneously or it made as unreadable. The private keys ensure the security. Hartung and Girod proposed the asymmetric watermark with the spread spectrum of watermarking. Private key is used for watermark embedded process. Watermark is verified using public key and the redundancy made with the private key. If the public key is known and the watermark connected with the public key can be easily removed.

With the Legendre sequences the method is proposed by schyndel et al. Legendre sequences combines with the Fourier transform. Legendre sequences are used as a private key to embed the watermark image. The sequence length is made as a public key. This method has N-2 legender sequences. Some malicious attacks are preferred in this technique.

Gray scale images with pixel values ranging from 0 to 255 are comprised of 8 bit-planes. In order to obtain the sufficient embedding capacity, each binary bit-plane is compressed losslessly and data is embedded into the saved space. In detection phase, the embedded data is extracted and the compressed image is decompressed. The original image is recovered because the compression was lossless. Rodriquez et al. searches for the suitable pixels to embed information using the spiral scan starting from the centroid of the image. Then obtain a block with its center at the position of the selected pixel. If the bit to be embedded is '1', change the luminance value of the central pixel by adding the gray level mean of the block with luminance of the block. If the bit to be embedded is '0', change the luminance value of the central pixel by subtracting the luminance of the block from the gray level mean of the block. In the extraction procedure, marked pixels are located using the spiral scan starting in the centroid of the image. If the luminance value of the central pixel is greater than the gray scale level mean of the block, then the embedded bit is identified as 1, otherwise as 0.

3. PROPOSED METHODOLOGY

This proposed framework for finding the image pattern selection a given image using an interpolator that is trained in advance with training data, based on **R-S vector** technique for determining the optimal and compact support for efficient image expansion. Experiments on test data show that learned interpolators are compact yet superior to classical ones. To derived an efficient learning procedure for its parameters on the basis of variation approximation. When plenty of computational resources is available, or when the observation process is too severe to recover by mere linear filtering, the complicated image expansion methods will be preferred. In this method, at first we find out the interpolator of the given image. Then replace the low resolution pixel by the interpolator (high resolution pixel). After

expanding the image does not scattered. We aim to resolve the tradeoff between high quality and low cost.

ADVANTAGES

- It R-S vector Bayesian technique is shown the better pattern making system, which greatly saves time and training the images.
- The embedded hash watermark function should not reduce the quality of the original image.
- There are various types of the watermark such as cropping, compression, scaling etc. The watermark pattern detection is a process of finding the image holders to transform the same compressive sensing domain using a secure multiparty computation (MPC) protocol and then sent to the cloud. The cloud only has the data in the compressive sensing domain.
- Capacity describes the maximum amount of PSNR factors that can be embedded into image, audio, video or text for proper retrieval of watermark during extraction.

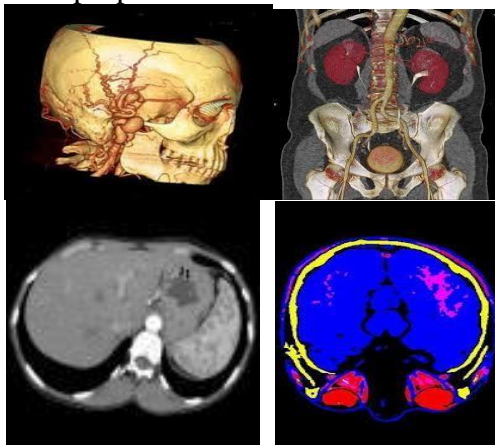




Figure 2: DICOM images

4. RESULTS AND ANALYSIS

The experimental results of the proposed technique for authentication and integrity of medical images based on reversible watermarking technique are discussed in this section. An application is programmed using C#.NET language to implement this technique. RSA is an effective method for verifying image authenticity and integrity in medical images. The performance parameters that are represented to measure the performance of the proposed technique are: Signal to Noise Ratio (SNR), Peak Signal to Noise Ratio(PSNR), Mean Square Error (MSE), and Bit Error Rate (BER).The original images before embedding the watermark and the authenticated image after embedding the watermark is shown in Fig 4. None can detect difference between the images.

The Experimental results shows that PSNR has high range values and it is consistent and the MSE has a least values therefore the quality of the images is not affected. BER is equal to zero for all the four DICOM images. SNR also has large values.

S.No	Original image	Watermarked image	PSNR	MSE	BER	SNR
1.			50.36	0.101	0	47.70

5. CONCLUSION AND FUTURE ENHANCEMENTS

Based on the DICOM images the watermarking technique is proposed. This technique provides patient information, image integrity and confidentiality. The hash value based on SHA is determined from the image. With the patient id, hash value and the compressed R-S vector watermark is formed and encrypted using public key cryptography. RSA is a secure public key encryption algorithm provides information security. The quality measures such as PSNR, SNR, MSE and BER estimates security of algorithms. Concluded results shows that BER equals 0 , SNR and PSNR has a high consistent values .MSE has a low bit rate.

REFERENCES

[1] Mohamed M. Abd-Eldayem,"A Proposed Security Technique Based On watermarking and Encryption for Digital Imaging and communications in medicine", Egyptian Information Journal,2012.

[2] C.-T. Hsu, J.-L. Wu, "Hidden Digital Watermarks in Images", IEEE Transactions on Image Processing, Vol. 8, pp. 58-68, 1999

[3] Coatrieux G, Lecornu L " A review of image watermarking applications in healthcare". Proceedings of the 28th Annual International Conference of the IEEE: Engineering in Medicine and Biology Society,EMBS ,2006.

[4] F. A. P. Petitcolas, R. J. Anderson and M. G. Kuhn, "Attacks on copyright marking systems," In:Proc. of Second International Workshop on Information Hiding, LNCS 1525, Portland, Oregon,USA, Apr. 1998.

[5] Miaou S, Hsu C, Tsai Y, Chao H,"A secure data hiding technique with heterogeneous data-combining capability for electronic patient records", presented at the World Congr. Med. Phys. *Biomed. Eng., Proc.*,Session Electron. Healthcare Records (IEEE-BEM), Chicago, 2000.

[6] J. J. Eggers, J. K. Su, and B. Girod, "Public key watermarking by eigenvectors of linear transforms," in Proc. Eur. Signal Processing Conf:, Tampere, Finland, Sept. 2000.

[7] Trichili H, Bouhleb M S, Derbel N, Kamoun L "A new medical image watermarking scheme for a better telediagnosis". Proc. IEEE Int. Conf. Syst., Man Cybern., Session MP1B 556–559,2002.

- [8] R. G. van Schyndel, A. Z. Tirkel, and I. D. Svalbe, "Key independent watermark detection," in Proc. IEEE Int. Conf: Multimedia Computing and Systems, Florence, Italy, pp. 580-585, 1999.
- [9] Zain J M, Baldwin L P, Clarke M 2009 Reversible watermarking for authentication of DICOM images.Proc. 26th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC 2004) 2: 3237–3240.
- [10] H. Choi, K. Lee, and T. Kim, "Transformed-Key Asymmetric Watermarking System", IEEE Signal Processing Letters, vol. 11, no. 2,pp.251-245, 2004.

FINANCIAL ANALYSIS AND LOAN PREDICTOR USING DATA MINING TECHNIQUES

M. KOWSIKA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Demand forecasting for loan is the major key aspect to successfully manage bank, insurance and finance companies. In particular, properly predicting loan requirements and eligibility criteria allows for a precise processing of loans. This will ensure a low level of security danger, while this is critical to the profitability of the companies. Hence, this paper is interested in predicting loan provisions to proper people.

This paper proposes a forecasting approach that is solely based on the data retrieved from past records and allows for a straightforward human interpretation. Therefore, it proposes two generalized models for predicting loan provisions. In an extensive evaluation, data sets are taken which consists of loan prediction data. The main motivation of doing this project is to present a loan prediction model for the prediction of loan giving. Further, this research work is aimed towards identifying the best classification algorithm for loan analysis.

In this work, data mining classification algorithm called Naïve Bayes is addressed and used to develop a prediction system in order to analyze and predict the sales volume. In addition, various grouping and chart preparation is also made in proposed system for better classification results. SVM and KNN classification is also applied in proposed system.

Keywords: bank, forecasting, loan prediction data, data mining, SVM and KNN

1. INTRODUCTION

Loan prediction growth in banking and finance sectors. With this growth, the ease to access of sanction loan has increased because many people are applying for loans. The problem here is that bank has only limited number of resources and capital, which the bank can distribute among the customers.

The whole task of categorizing to whom the bank should sanction loan and to whom it should not has become a difficult task for the bankers[2]. Generally, bank undergoes a rigorous procedure for verifying the customer to sanction loan. This procedure may take a week's time or two. The drawback here is that the customer needs to wait for two whole weeks to know whether he/she is deserving or not. In this paper, we have reduced the risking factor of banks behind finding the appropriate person for loan approval by the bank. We even reduce the time of loan approval analysis. We first use data mining techniques to analyze previous records to which the bank has already[4]. sanctioned loan based on the analysis made out of these records we train the deep learning model. The new data is treated as testing data, and the output of the customer is calculated accordingly.

The process of credit scoring required experts alongside statistical algorithms to accurately predict the creditworthiness of an applicant. However, quite recently, the researchers and the banking authorities have opted for training classifiers based on various machine learning

and deep learning algorithms to automatically predict the credit score of an applicant based on their credit history and other historical data and make the process of selecting the eligible candidates a lot easier before the loan is approved[5]. Therefore, addressing the aforementioned scenario, the goal of this paper is to discuss the application of different machine learning models in the loan lending process and work out the best approach for a financial institution which accurately identifies whom to lend loan to and help banks identify the loan defaulters for much-reduced credit risk. Classifiers that we used to build the model are Random Forest and Decision Trees[6]. They'll be used separately to analyse the dataset and identify the patterns in the dataset and learn from those. Based on that analysis, predict whether a new applicant is likely to default on a loan or not. Good credit management includes effective analysis of loan approval, tracking of debt collection and future customers' repayment performance[7,8]. This issue covers the analysis process for debt compromises. Good credit management requires accounting data as an instrument for financial statement analysis, business operating results, liquidity, previous business performance, present & future business plans. This information is used for credit consideration and approval.

Small loan is an important aspect of our everyday life: it allows aspiring entrepreneurs to get started on ideas that could be grown into business; it allows curious students to afford higher education that is otherwise unavailable without a stable income; more importantly, it allows ordinary people who have no friends or relatives for support to obtaining short-term financial assistance and get back on their feet to fight for the American Dream. Nevertheless, with loan it comes with the possibility of default as well[8]. Default is a financial term describing the failure of meeting the legal obligation of a loan - paying back the principal and interest. It's a common problem in the financial industries and one of the major risks of offering loans. Of course, default does not happen the majority of the time and the lending banks usually able to make up the loss from a defaulting loan from other fully paid loans and their accompanied interests.

Furthermore, banks issuing loans with higher interest rate to individuals with high probability of default - the financial institutions are trading off an increased chance of default with an increased profit from the high interest. All things considered, default is a fact of life and most financial institutions have a well-established practice to minimize its impact and absorbing the loss[9]. But what about a situation where instead of a single bank is issuing the loan, the loan is comprised of funds from several investors? Lending Club is one of the many peer-to-peer lending company that gives rise to this peculiar situation. In plain words, peer-to-peer lending company acts as a broker between borrowers and investors.

The company creates a platform where borrowers can create small unsecured personal loans, and investors can seek out these loans and decide which loans to invest from. Borrowers obtain the loan they want, investors get to profit from the loan interest, and the company gets a cut from both parties (origination fee from borrowers and service fee from investors). This also means that when a loan goes default, it's no longer a single bank that is absorbing the loss - single or multiple individual investors will be absorbing it instead. The overall profit might be positive if all the loans were originated from a single lender as other fully paid loans could cover

the loss, but this is no longer the case as there will be winners and losers among this new form of lending practices if the investors did not diversify. An obvious solution to this problem is to predict whether a particular loan will go default based on initial information provided by the borrowers and their credit report. There's no doubt Lending Club already has an existing model in place to approve loans posted on their website.

This paper will explore the process and result on formulating a new machine learning model that could predict a loan default; but more importantly, the model will focus on minimizing the overall loss in investment of bad loans in order to lessen the burden passed onto individual investors. As a side note, the paper will also explore privacy-preserving mechanism on sensitive information provided from the borrower's credit report. The end goal is to evaluate a simplified version of RAPPOR (Randomized Aggregately Privacy Preserving Ordinal Response) and determine whether data that have been hashed by this algorithm could still be used to predict loan default as stated previously.

2. LITERATURE REVIEW

Loan prediction is a much-talked-about subject in the sectors of banking and finance. Credit scoring has become a key tool for the same in this competitive financial world. Furthermore, following the recent improvements in data science and several notable developments in the field of artificial intelligence, this topic has gained more attention and research interest. In recent years, it has attracted more focus towards research on loan prediction and credit risk assessment. Due to the high demands of loan now, demand for further improvements in the models for credit scoring and loan prediction is increasing significantly. A multitude of techniques have been used to assign individuals a credit score and much research has been done over the years on the topic. Unlike previously, where experts were hired and the models depended on professional opinions were used for assessing the individual's creditworthiness, the focus has shifted to an automated way of doing the same job. In recent years, the researchers and banking authorities have been focused on applying machine learning algorithms and neural networks for credit scoring and risk assessment. Many noteworthy conclusions have been drawn in this regard which serve as stepping-stones for researches and studies.

The credit card is one of the most popular financial instruments in modern society. It becomes a replacement for the use of cash and leads to reduced circulation of the currency and printing of money. And it also speeds up the mobility of cash flows. Cardholders, merchants, and Banks all benefit from the use of credit cards. Credit cards are considered an important product for the banking industry, and managing credit card risks is one of the banks' crucial tasks.

Credit card default is one of the risks; it means clients failed to pay a required minimum payment before the bill was due. Creating a model to predict the behavior of credit card clients who will eventually default is one of the important methods to reduce the risks. Knowing the determinants of credit card default will guide the financial institution in crafting measures and policies to reduce the occurrence of such. There are studies that

created predictive model like in the study of that utilized the weighted SVM algorithm and in the study of using Recurrent Neural Network (RNN). However, it is good to compare the result of different classification algorithm.

According to the authors, the forecasting process begins with data clean-up and processing, missing value substitution, data set experimental analysis, and modelling, and continues to model evaluation and test data testing. A logistic regression model has been executed. The highest accuracy obtained with the original dataset is 0.811. Models are compared based on performance measurements such as sensitivity and specificity. As a result of analyzing, the following conclusions were drawn. However, other characteristics of customers that play a very important role in lending decisions and forecasting defaulters should also be evaluated. Some other traits, such as gender and marriage history, do not seem to be considered by the company. A credit credibility soothsaying system that helps companies make the right opinions to authorize or reject the credit claims of guests. This helps the banking assiduity to open effective distribution channels. This means that if the customer has a minimum repayment capacity, their system can avoid future risks. Including other techniques (using the Weka tool) that are better than the general data mining model has been implemented and tested for domains.

3. PROPOSED METHODOLOGY

The aim of this project is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on newtest data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing.

ADVANTAGES

- The advantage of this system is that we provided some conditions by setting the algorithms and just by evaluating the details, we get to know eligibility criteria that client is eligible or not.
- The proposed system also scales relatively well to high dimensional data
- . The proposed system is relatively memory efficient.
- The risk of over-fitting is less in our proposed system.
- A small change to the data does not greatly affect the hyperplane

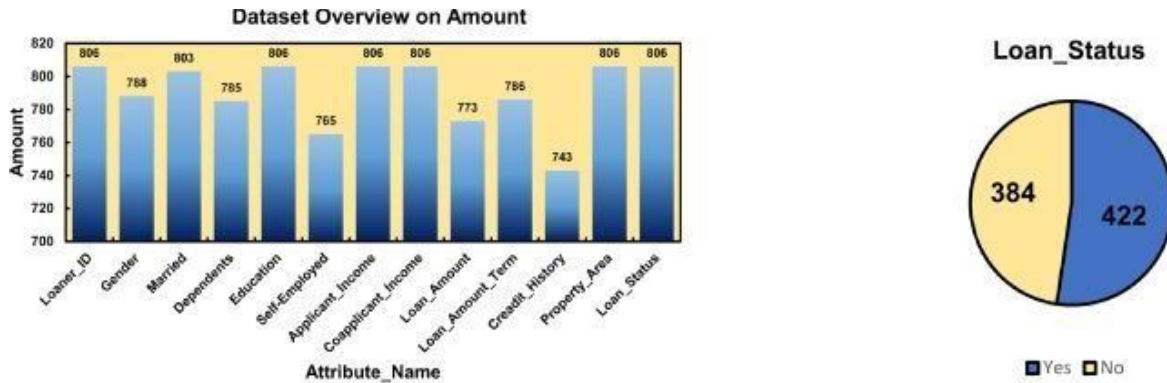


Fig 1: Dataset and loan status graphical view

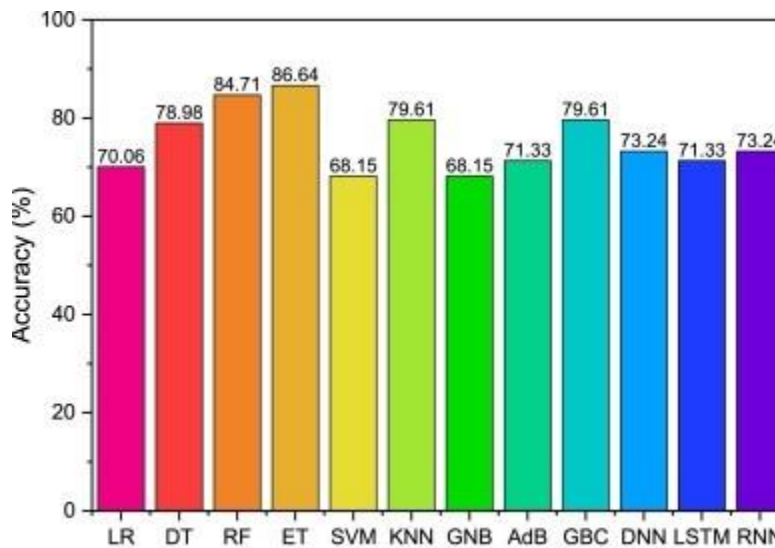
4. RESULTS AND DISCUSSIONS

We achieved the highest precision value of 0.94 from the classification report for “No” loan status for the [LR model](#). In contrast, the highest recall achieved for “Yes” loan status is 0.97, and the highest f1-score value was 0.77 achieved for “Yes”. DT achieved 0.81 for “Yes” and the highest f1-score (0.80) for “No”. On the other hand, the highest recall value of 0.82 is achieved for “Yes”. We obtained the highest precision value of 0.85 for “No” loan status in the case of the RF model. In contrast, the highest recall for “Yes” loan status is 0.82, and the highest f1-score value of 0.84 was found for both “Yes” and “No”. In ET algorithm, the highest precision value of 0.88 is found for “No” and the highest f1-score for “Yes”, the value is 0.86. On the other hand, the highest recall value of 0.89 is obtained for “Yes”.

In the SVM model, the highest precision value of 0.94 is obtained for “No”, and the highest f1-score for “Yes” is 0.75. On the other hand, the highest recall value, 0.97, was achieved for “Yes”. Again, we achieved the highest precision value of 0.87 for the “Yes” loan status for the KNN model. In contrast, the highest recall achieved for “No” loan status is 0.90, and the highest f1-score value is 0.81 achieved for both “Yes” and “No”. GNB's highest precision value of 0.87 is achieved for “No” and the highest f1-score for “Yes” is 0.75. On the other hand, the highest recall value of 0.94 is achieved for “Yes”. In the [AdaBoost algorithm](#), the highest precision value of 0.76 is obtained for “No” and the highest f1-score for “Yes” is 0.74. On the other hand, the highest recall value of 0.81 is achieved for “Yes”. We achieved the highest precision value of 0.86 for the “No” loan status for the Gradient Boosting model, whereas the highest recall achieved for the “Yes” loan status is 0.89, and the highest f1-score value was 0.81 achieved for both “Yes”. Overall, DT, RF, ET, and GB yielded strong results across multiple [evaluation metrics](#).

[LSTM](#) and [RNN](#) performed adequately but with slightly lower precision and recall for the “No” class. [DNN](#) exhibited a precision of 0.65 for the “Yes” class and 0.9 for the “No” class. The recall was 0.94 for the “Yes” class and 0.51 for the “No” class, leading to an F1-score of 0.78 for the “Yes” class and 0.65 for the “No” class. LSTM achieved a precision of 0.64 for the

“Yes” class and 0.88 for the “No” class. The recall was 0.93 for the “Yes” class and 0.48 for the “No” class, resulting in an F1-score of 0.76 for the “Yes” class and 0.62 for the “No” class. RNN demonstrated a precision of 0.69 for the “Yes” class and 0.79 for the “No” class. The recall was 0.83 for the “Yes” class and 0.62 for the “No” class, leading to an F1-score of 0.75 for the “Yes” class and 0.7 for the “No” class. Graph 1 presents the accuracy performance of different models. We can observe that the three deep learning models, namely DNN (73.24%), LSTM (71.33%), and RNN (73.24%), did not perform better than the other machine learning models for the loan dataset. This indicates that for the given dataset, the deep learning models did not yield higher accuracy compared to traditional datamining algorithms.



Graph 1: Compared Accuracy level with datamining Algorithms

5. CONCLUSION

Data mining techniques are used in this paper for credit evaluation in the bank's credit approval process. The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. This brings some of the following insights about approval. Applicants with Credit history not passing fails to get approved, probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.

6. REFERENCES

- [1] Ducai, M.T., 2012. The bank loans importance, information asymmetry and the impact of financial and economic crisis on corporate financing. *Revista Tinerilor Economisti (The Young Economists Journal)*, (18), pp.29–34.
- [2] Dunn, I., 2017. Common Problems and Bottlenecks of LoanPortfolio Analysis | Visible Equity. Available at: <https://www.visibleequity.com/common-problems-and-bottlenecks-of-loan-portfolio-analysis> [Accessed: 12 June 2018].
- [3] Leung, A., 2017. Here's a Look at the World's Worst Bad-LoanRatios: Map-Bloomberg, Bloomberg. Available at: <https://www.bloomberg.com/news/articles/2017-11-21/here-s-a-look-at-the-world-s-worst-bad-loan-ratios-map> [Accessed: 12 June 2018].
- [4] Fay, B., 2017. What is a Credit Score and How is it Calculated? Available at: <https://www.debt.org/credit/report/scoring-models> [Accessed: 12 June 2018]
- [5] Pritchard, J., 2018. How Credit Scores Work. Available at: <https://www.thebalance.com/how-credit-scores-work-315541> [Accessed:5 August 2018].
- [6] Chye, K.H., Chin, T.W. and Peng, G.C., 2004. Credit scoring using data mining techniques. *Singapore Management Review*, 26(2), p.25.
- [7] Abdou, H.A. and Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 88, pp.59–88.
- [8] Brown, I. and Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), pp.3446–3453.
- [9] Kaggle.com, 2018. Kaggle Data Repository [Online], Available: <https://www.kaggle.com/ninzaami/loan-predication> [Accessed:24 October 2018].

HEART DISEASE PREDICTION USING CONVOLUTIONAL NEURAL NETWORK

C. AARTHIKA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The main motivation of doing this project is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient.

The identification of the possibility of heart disease in a person is complicated task for medical practitioners because it requires years of experience and intense medical tests to be conducted. The main objective of this significant research work is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out.

Convolutional neural network (CNN) architecture is used to map the relationship between the indoor PM and weather data to the found values. The proposed method is compared with the state-of-the-art deep neural network (DNN)based techniques in terms of the root mean square and mean absolute error accuracy measures. In addition, support vector machine based classification and K-Nearest Neighbor based classification is also carried out and accuracy is found out. The applied SVM, KNN and CNN classification helps to predict the heart disease with more accuracy in the new data set.

Keywords: Convolutional neural network, neural network, K-Nearest, Heart disease,

1. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis[1]. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk[2]. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry[5]. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease

There are several factors causing heart and diabetes problems which include Age, Gender, Blood Pressure, Glucose levels, Skin thickness and Insulin. These are easily measured in primary care facility centres[6]. The accurate estimation and analysis of heart & diabetes disease patients reports data may help in predicting future heart problems including diabetes. Globally, the application of computerized machine learning methods to predict future problems is in trend now. The Health Monitoring Departments and Fields uses machine learning algorithms to predict and analyse in a wider way to solve problems in fraction of seconds. From the famous proverb "Prevention is Better Than Cure", if we apply this to medico and health field we can save people from major Heart Diseases (HD's) along with Diabetes. The proposed Dual disease prediction technique is user interactive based method. The proposed method observe inputs from the end user with realistic data to predict heart and diabetes disease. In the presented work, we used Logistic.

Healthcare is one of the primary focus for humanity. According to WHO guidelines, good health is the fundamental right for individuals. It is considered that appropriate health care services should be available for regular checkup of one's health. Almost 31% of all deaths are due to heart related disease in all over the world[7]. Early detection and treatment of several heart diseases is very complex, especially in developing countries, because of the lack of diagnostic centers and qualified doctors and other resources that affect the accurate prognosis of heart disease. With this concern, in recent times computer technology and machine learning techniques are being used to make medical aid software as a support system for early diagnosis of heart disease.

Identification of any heart related illness at primary stage can reduce the death risk. Various ML techniques are used in medical data to understand the pattern of data and making prediction from them. Healthcare data are generally massive in volumes and complex in structure. ML algorithms are capable to handle the big data and mine them to find the meaningful information. Machine Learning algorithms learn from past data and do prediction on real time data.

This sort of ML framework for coronary illness expectation can encourage cardiologists in taking quicker actions so more patients can get medicines within a shorter timeframe, thus saving large

number of lives. Machine Learning is a branch of AI research [8] and has become a very popular aspect of data science. The Machine Learning algorithms are designed to perform a large number of tasks such as prediction, classification, decision making etc. To learn the ML algorithms, training data is required. After the learning phase, a model is produced which is considered as an output of ML algorithm. This model is then tested and validated on a set of unseen real time test dataset. The final accuracy of the model is then compared with the actual value, which justify the overall correctness of predicted result. Lots of efforts has already been done to predict the heart disease using the ML algorithms by authors [9,10], but this is an additional effort to do the experiment on benchmarking UCI heart disease prediction dataset while comparing the four popular ML technique to check the most accurate ML technique.

2. LITERATURE REVIEW

There are numerous works has been done related to disease prediction systems using different datamining techniques and machine learning algorithms in medical centres. K. Polaraju et al, proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work Is performed using training data set consists of 3000instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms. Marjia et al, developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease .S.Seema et al, focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN).

A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy. Ashok Kumar Dwivedi et al, recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. Megha Shahi et al, suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centres. The paper used various algorithms like SVM, Naïve Bayes.

Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms. Chala Beyene et al, recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early

automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.

3. PROPOSED METHODOLOGY

After evaluating the results from the existing methodologies, we have used python and pandas operations to perform heart disease classification for the data obtained from the UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a preprocessing data phase followed by feature selection based on data cleaning, classification of modelling performance evaluation. Random forest technique is used to improve the accuracy of the result.

ADVANTAGES

- Increased accuracy for effective heart disease diagnosis.
- Handles roughest (enormous) amount of data using random forest algorithm and feature selection.
- Reduce the time complexity of doctors.
- Cost effective for patients.
- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality
- It enhances the accuracy of the model and prevents the overfitting

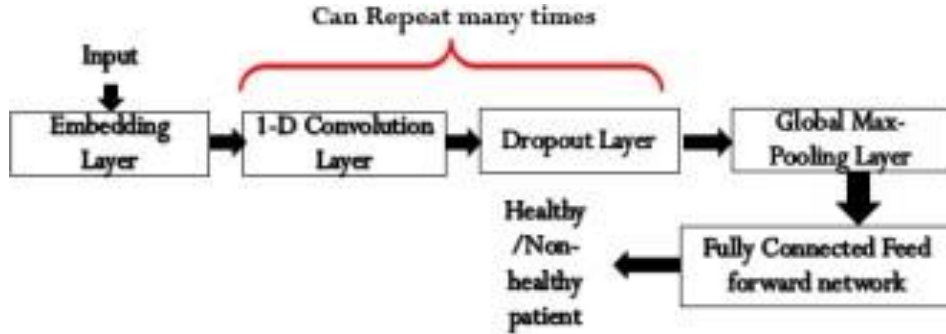
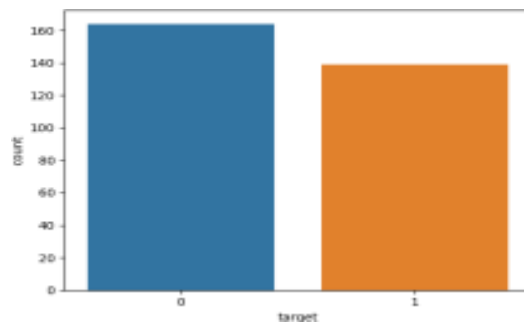


Figure 1: CNN Architecture

4. RESULTS AND DISCUSSIONS

CNN architecture contains around 0.13 million trainable parameters which will get adapted during training of the network. It was observed that general CNN architecture overfitted the training data meaning that training accuracy was very high and validation accuracy was low. The dropout technique was introduced to remove overfitting. It removes random neurons with a certain probability during training which allows the different. Some of the attributes have missing values for some of the examples. Those values have been replaced with the mean value of that attribute for training our architecture. Most of the traditional classification architectures require all the attributes in the same range. This dataset has attributes in different ranges so a standardization technique is applied which converts all the attributes into the same range. It subtracts all the attribute values with the mean value of the attribute and divides by the standard deviation of the attribute. The final attribute is the true label for the patient whether he/she has heart disease or not.



Graph 1: Distribution of labels in datasets

5. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

6. REFERENCES

- [1] Ducai, M.T., 2012. The bank loans importance, information asymmetry and the impact of financial and economic crisis on corporate financing. *Revista Tinerilor Economisti (The Young Economists Journal)*, (18), pp.29–34.
- [2] Wym an, O ., 2015. The role of financial services in society statement in support of macroprudential policies, Available at: http://www3.weforum.org/docs/WEF_The_Role_of_Financial_Services_in_Society_report_2015.pdf [Accessed: 12 June 2018].
- [3] Dunn, I., 2017. Common Problems and Bottlenecks of Loan Portfolio Analysis | Visible Equity. Available at: <https://www.visibleequity.com/common-problems-and-bottlenecks-of-loan-portfolio-analysis> [Accessed: 12 June 2018].
- [4] Leung, A., 2017. Here's a Look at the World's Worst Bad-Loan Ratios: Map-Bloomberg, Bloomberg. Available at: <https://www.bloomberg.com/news/articles/2017-11-21/here-s-a-look-at-the-world-s-worst-bad-loan-ratios-map> [Accessed: 12 June 2018].
- [5] Fay, B., 2017. What is a Credit Score and How is it Calculated? Available at: <https://www.debt.org/credit/report/scoring-models> [Accessed: 12 June 2018]
- [6] Pritchard, J., 2018. How Credit Scores Work. Available at: <https://www.thebalance.com/how-credit-scores-work-315541> [Accessed: 5 August 2018].

- [7] Chye, K.H., Chin, T.W. and Peng, G.C., 2004. Credit scoring using data mining techniques. *Singapore Management Review*, 26(2), p.25.
- [8] Abdou, H.A. and Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 88, pp.59–88.
- [9] Brown, I. and Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), pp.3446– 3453.
- [10] Kaggle.com, 2018. Kaggle Data Repository [Online], Available: <https://www.kaggle.com/ninzaami/loan-predication> [Accessed:24 October 2018].

IMAGE FORGERY DETECTION USING CONVENTIONAL NEURAL NETWORK IN NEURAL NETWORKS

M. SNEHA

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

Day for day it becomes easier to temper digital images. Thus, people are in need of various forgery image detection. This project present forgery image detection techniques for two of the most common image tampering techniques; copy-move and splicing. This project use match points technique after feature extraction. The detection of tampered regions is done through searching for very similar regions. Convolutional neural network (CNN) is the machine- learning algorithm which achieved substantial results in image detection and classification. This project develops a new CNN architecture to classify original and masked images checks whether the image is masked or not. With i) good generalization capability and ii) good execution speed, newly developed CNN architecture is being used as an effective decision-support tool for radiologists in diagnostics. Python is used for development of the project. Here we are extracting the image features and analyzing it to detect the forged images and also determine the type of theforgery.

Keywords: Convolutional neural network, digital images, Convolutional, radiologists, forgery

1. INTRODUCTION

Image Recognition and detection is a classic machine learning problem. It is a very challenging task to detect an object or to recognize an image from a digital image or a video. Image Recognition has application in the various field of computer vision, some of which include facialrecognition, biometric systems, self-driving cars, emotion detection, image restoration, robotics and many more[1]. Deep Learning algorithms have achieved great progress in the field of computer vision. Deep Learning is an implementation of the artificial neural networks with multiple hidden layers to mimic the functions of the human cerebral cortex. The layers of deep neural network extract multiple features and hence provide multiple levels of abstraction. As compared to shallow networks, this cannot extract or work on multiple features. Convolutional neural networks is a powerful deep learning algorithm capable of dealing with millions of parameters and saving the computational cost by inputting a 2D image and convolving it with filters/kernel and producing output volumes. The MNIST dataset is a dataset containing handwritten digits and tests the performance of a classification algorithm. Handwritten digit recognition has many applications such as OCR (optical character recognition), signature verification, interpretation and manipulation of texts and many more [2,3]. Handwritten digit recognition is an image classification and recognition problem and there have been recent advancements in this field [4]. Another dataset is CIFAR-10 which is an object detection datasets

that classifies the objects into 10 classes and detects the objects in the test sets. It contains natural images and helps implement the image detection algorithms [5]. In this paper, Convolutional neural networks models are implemented for image recognition on MNIST dataset and object detection on the CIFAR-10 dataset. The implementation of models is [6] discussed and the performance is evaluated in terms of accuracy [7]. The model is trained on an only CPU unit and real-time data augmentation is used on the CIFAR-10 dataset. Along with that, Dropout is used to reduce Overfitting on the datasets. Image forgery detection aims to detect alterations made to digital images, such as copy-move forgery, splicing, or manipulation of specific image regions. This is crucial for ensuring the authenticity and integrity of digital images in various applications, including forensics, journalism, and legal proceedings. The first step involves collecting or generating a dataset of authentic and manipulated images. Authentic images represent genuine, unaltered photographs, while manipulated images contain various types of forgeries or alterations. These images serve as training and evaluation data for the neural network model. Preprocessing steps may include resizing images to a fixed size, normalizing pixel values, and augmenting the dataset to increase diversity and robustness. Augmentation techniques such as rotation, flipping, and cropping can help improve the model's ability to generalize to different types of forgeries. Conventional neural networks, particularly CNNs, are commonly used for image forgery detection due to their ability to learn hierarchical features from image data. The CNN architecture typically consists of convolutional layers followed by pooling layers, fully connected layers, and an output layer. The model is trained to classify images as authentic or manipulated based on learned features.

2. LITERATURE REVIEW

In recent years there have been great strides in building classifiers for image detection and recognition on various datasets using various machine learning algorithms. Deep learning, in particular, has shown improvement in accuracy on various datasets. Some of the works have been described below: Norhidayu binti Abdul Hamid et al. evaluated the performance on MNIST datasets using 3 different classifiers: SVM (support vector machines), KNN (K-nearest Neighbor) and CNN (convolutional neural networks). The Multilayer perceptron didn't perform well on that platform as it didn't reach the global minimum rather remained stuck in the local optimal and couldn't recognize digit 9 and 6 accurately. Other classifiers, performed correctly and it was concluded that performance on CNN can be improved by implementing the model on Keras platform. Mahmoud M. Abu Gosh et al. implement DNN (Deep neural networks), DBF (Deep Belief networks) and CNN (convolutional neural networks) on MNIST dataset and perform a comparative study. According to the work, DNN performed the best with an accuracy of 98.08% and other had some error rates as well as the difference in their execution time. Youssouf Chherawala et al. built a vote weighted RNN (Recurrent Neural networks) model to determine the significance of feature sets. The significance is determined by weighted votes and their combination and the model is an application of RNN. It extracts features from the Alex word images and then uses it to recognize handwriting. Alex krizhevsky uses a 2-layer

Convolutional Deep belief network on the CIFAR-10 dataset. The model built classified the CIFAR-10 dataset with an accuracy of 78.90% on a GPU unit. Elaborative differences in filters and their performance is described in the paper which differs with every model. Yehya Abouelnaga et al. built an ensemble of classifiers on KNN. They used KNN in combination with CNN and reduce the Overfitting by PCA (Principal Component Analysis). The combination of these two classifiers improved the accuracy to about 0.7%. Yann le Cunn et al. give a detailed introduction to deep learning and its algorithms. The algorithms like Backpropagation with multilayer perceptron, Convolutional neural networks, and Recurrent neural networks are discussed in detail with examples. They have also mentioned the scope of unsupervised learning in future in Artificial intelligence

3. PROPOSED METHODOLOGY

Gather a diverse dataset of authentic images and images containing various types of forgeries (e.g., copy-move, splicing, retouching). Preprocess the dataset by resizing, normalizing, and augmenting the images to ensure uniformity and improve model generalization. Extract meaningful features from the input images using a conventional neural network architecture. Conventional neural networks like LeNet, AlexNet, or VGGNet can serve as feature extractors in this stage. Design a CNN-based forgery detection model that takes the extracted features as input. Train the model using the collected dataset, employing techniques like transfer learning or fine-tuning to adapt pre-trained models to the forgery detection task. Utilize techniques like batch normalization, dropout, and data augmentation to improve the model's robustness and generalization capabilities. Once trained, the proposed system can be used to detect image forgeries by inputting suspicious images into the trained CNN model. The model will then analyze the extracted features of the input image and classify it as either authentic or forged based on learned patterns and discrepancies. Evaluate the performance of the proposed system using standard metrics such as accuracy, precision, recall, and F1-score. Validate the system on a separate test dataset to assess its generalization ability and robustness to unseen forgery types.

ADVANTAGES

- CNNs are known for their robustness to variations in input data such as translation, rotation, scaling, and noise.
- CNNs can be trained in an end-to-end fashion, where the entire network is trained jointly to optimize a specific forgery detection objective.
- CNN-based forgery detection models can be trained to detect various types of image forgeries, including copy-move, splicing, retouching, and more.

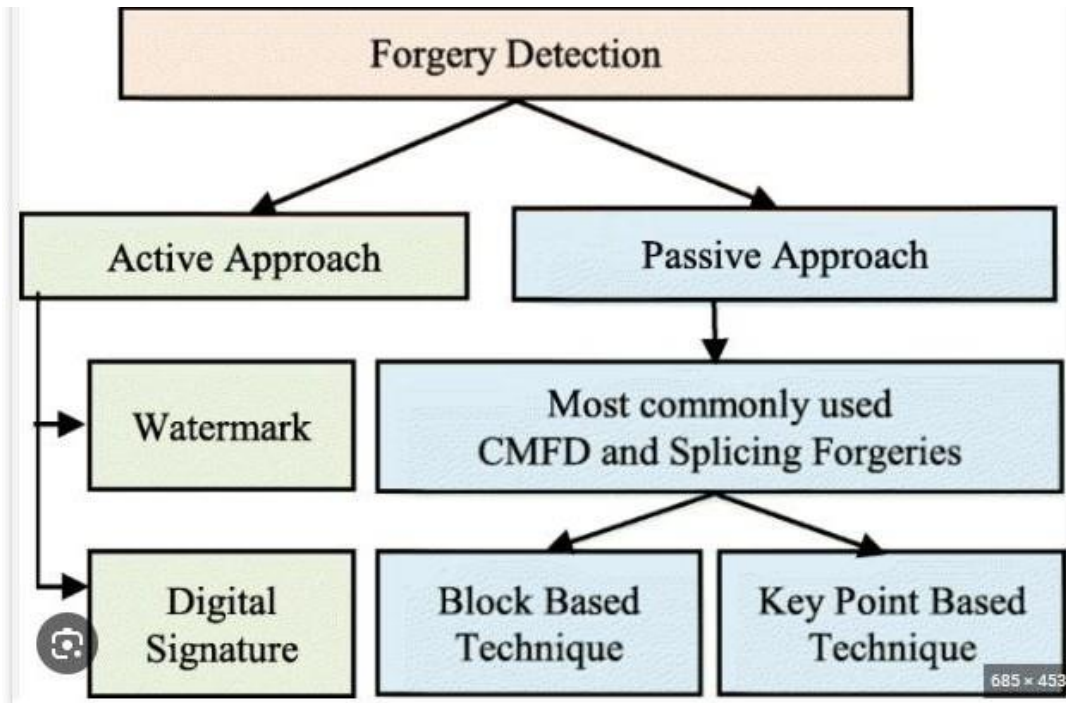
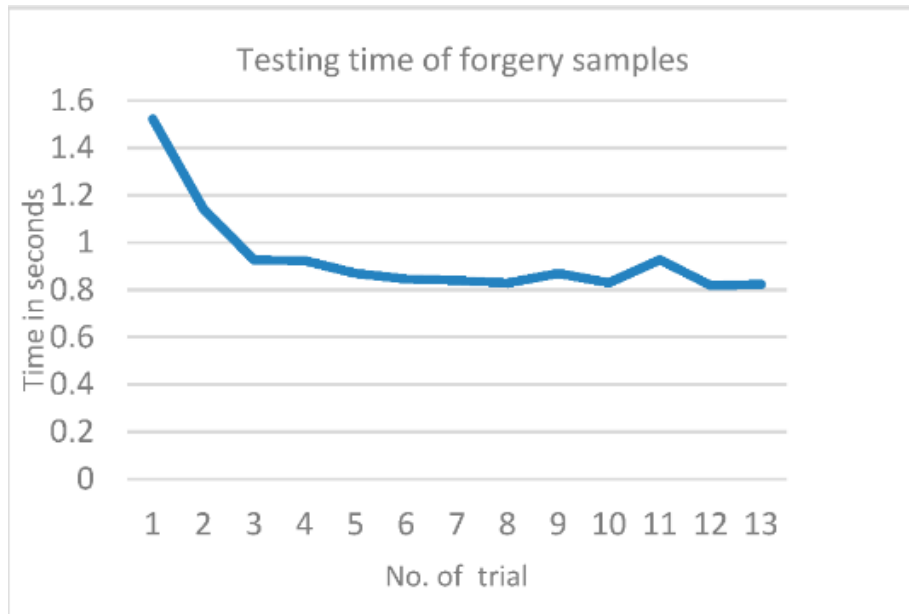


Fig.1 shows the forgery detection

4. RESULTS AND DISCUSSIONS

Our model represents a deep learning method suitable for detecting forgery embedded within digital images. Non-deep-learning traditional methods, such as, are unable to extract relevant data from input image patches automatically, nor can they devise representations very efficiently. Many non-deep-learning approaches also only utilize a single artificial feature for classification purposes. These are all significant drawbacks in the traditional models. Our proposed method, on the other hand, is much more efficient. It can apply several epochs in the training sets, the optimal number being no less than three epochs and no more than five, which is related to dataset size. Testing a new input image will take a longer time the first time round, but will decrease by several trials, the first of which usually takes no more than 1.6 s, as illustrated in Figure 8. This time is based on image resolution and the used machine. Table 4 indicates a clear reduction in accuracy from 90% to 81%. The average validation loss rate of the training set was around 0.3010 for all saturated iteration values. Of the 1255 forged images we used, we had an overall validation accuracy of 88.26% to 90.1%. The matrices and the baseline evaluation settings were devised by computing false positive (FP), true positive (TP), false positive (FP) and false negative (FN) settings in order to compute the F-measure. The evaluation scores in Table 5 present the F-measure of the proposed model vs. the state-of-the-art models. It is worth mentioning again that our testing dataset was relatively small and used a mix of both forged and pristine images. Hence, we anticipate that the value will change in accordance with dataset size.

Note that the number of epochs is low according to the dataset size to avoid overfitting during the training task.



Graph.1 shown the testing of samples

5. CONCLUSION

The increased availability of cameras has made photography popular in recent years. Images play a crucial role in our lives and have evolved into an essential means of conveying information since the general public quickly understands them. There are various tools accessible to edit images; these tools are primarily intended to enhance images; however, these technologies are frequently exploited to forge the images to spread misinformation. As a result, image forgery has become a significant problem and a matter of concern. In this paper, we provide a unique image forgery detection system based on neural networks and deep learning, emphasizing the CNN architecture approach. To achieve satisfactory results, the suggested method uses a CNN architecture that incorporates variations in image compression. We use the difference between the original and recompressed images to train the model. The proposed technique can efficiently detect image splicing and copy-move types of image forgeries. The experiments results are highly encouraging, and they show that the overall validation accuracy is 92.23%, with a defined iteration limit.

6. REFERENCE

- [1] Xiao, B.; Wei, Y.; Bi, X.; Li, W.; Ma, J. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. *Inf. Sci.* 2020, 511, 172–191. [Google Scholar] [CrossRef]
- [2] Kwon, M.J.; Yu, I.J.; Nam, S.H.; Lee, H.K. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 5–9 January 2021; pp. 375–384. [Google Scholar]
- [3] Wu, Y.; Abd Almageed, W.; Natarajan, P. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15–20 June 2019; pp. 9535–9544. [Google Scholar]
- [4] Ali, S.S.; Baghel, V.S.; Ganapathi, I.I.; Prakash, S. Robust biometric authentication system with a secure user template. *Image Vis. Comput.* 2020, 104, 104004. [Google Scholar] [CrossRef]
- [5] Castillo Camacho, I.; Wang, K. A Comprehensive Review of Deep-Learning-Based Methods for Image Forensics. *J. Imaging* 2021, 7, 69. [Google Scholar] [CrossRef] [PubMed]
- [6] Zheng, L.; Zhang, Y.; Thing, V.L. A survey on image tampering and its detection in real-world photos. *J. Vis. Commun. Image Represent.* 2019, 58, 380–399. [Google Scholar] [CrossRef]
- [7] Jing, L.; Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, 43, 1. [Google Scholar] [CrossRef]
- [8] Meena, K.B.; Tyagi, V. Image Forgery Detection: Survey and Future Directions. In *Data, Engineering and Applications: Volume 2*; Shukla, R.K., Agrawal, J., Sharma, S., Singh Tomer, G., Eds.; Springer: Singapore, 2019; pp. 163–194. [Google Scholar]
- [9] Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 2021, 54, 1–41. [Google Scholar] [CrossRef]
- [10] Rony, J.; Belharbi, S.; Dolz, J.; Ayed, I.B.; McCaffrey, L.; Granger, E. Deep weakly-supervised learning methods for classification and localization in histology images: A survey. *arXiv* 2019, arXiv:abs/1909.03354. [Google Scholar]
- [11] Lu, Z.; Chen, D.; Xue, D. Survey of weakly supervised semantic segmentation methods. In *Proceedings of the 2018 Chinese Control Furthermore, Decision Conference (CCDC)*, Shenyang, China, 9–11 June 2018; pp. 1176–1180. [Google Scholar]
- [12] Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi- and weakly supervised semantic segmentation of images. *Artif. Intell. Rev.* 2019, 53, 4259–4288. [Google Scholar] [CrossRef]

- [13] Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.* 2020, 14, 910–932. [Google Scholar] [CrossRef]
- [14] Luo, W.; Huang, J.; Qiu, G. JPEG Error Analysis and Its Applications to Digital Image Forensics. *IEEE Trans. Inf. Forensics Secur.* 2010, 5, 480–491. [Google Scholar] [CrossRef]
- [15] Matern, F.; Riess, C.; Stamminger, M. Gradient-Based Illumination Description for Image Forgery Detection. *IEEE Trans. Inf. Forensics Secur.* 2020, 15, 1303–1317. [Google Scholar] [CrossRef]

IMAGE RE-RANKING BASED GEOSPATIAL DATA NETWORKS

T. TAMILNILAVU

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Social media sharing Websites allow users to annotate images with free tags, which significantly contribute to the development of the web image retrieval. Tag-based image search is an important method to find images shared by users in social networks. However, how to make the top ranked result relevant and with diversity is challenging. This project, proposes a topic diverse ranking approach for tag-based image retrieval with the consideration of promoting the topic coverage performance. First, construct a tag graph based on the similarity between each tag. Then, the community detection method is conducted to mine the topic community of each tag. After that, inter-community and intra-community ranking are introduced to obtain the final retrieved results. In the inter-community ranking process, an adaptive random walk model is employed to rank the community based on the multi-information of each topic community. Besides, we build an inverted index structure for images to accelerate the searching process.

Keywords: web image, ranking approach, community detection, social networks, structure for images

1. INTRODUCTION

The huge growth of image collections and multimedia resources available and accessible through various technologies is remarkable [2]. Technological improvements in image acquisition and the decreasing cost of storage devices have enabled the dissemination of large image collections. Red Traditional image retrieval approaches based on keywords and textual metadata face serious challenges, since describing the image content with textual descriptions is intrinsically very difficult [8], mainly due to the huge growth of image collections. Many applications, especially those dealing with large general datasets face obstacles to obtain textual descriptors, since manual annotation is prohibitively expensive. It is laborious and time-consuming. This task has not been made easier by the diversification of image collections. One of the most common approaches to overcome these limitations relies on the use of Content- Based Image Retrieval (CBIR) systems. The objective of CBIR systems is to return the most similar images given an image query, considering visual features, such as shape, color, and texture. In this scenario, ranking accurately collection images is of great relevance. Collection images are ranked in increasing order of their distance to the query pattern (e.g., query image) defined by users. Therefore, choosing a good distance measure is often critical to building an effective CBIR system. In general, CBIR systems consider only pairwise image analysis, that is, compute similarity measures considering only pairs of images, ignoring the rich information encoded in the relations among several images. On the other hand, the user perception usually considers the query specification and responses in a given context. Therefore, the distance between two images can be correctly described only if it is considered in the context of other

images that are similar to them. This requires having a model to capture the essence of a similarity among images instead of viewing each image as a set of points or a feature vector [4]. Context can be broadly defined as all information about the whole situation relevant to an application and its set of users. In the image retrieval scenario, ranked lists represent a relevant source of contextual information, since given a query image, users do not analyze only pairs of images, but the ranked list as a whole. It is expected, for example, that images ranked at the top positions of ranked lists are similar to each other. It is also expected that, if we take one of these images as a query image, the computed ranked list contains many images in common. In this scenario, we use the term context for denoting any analysis that, instead of considering only pairwise image comparisons, takes into account other information encoded in both ranked lists and distances among all images. More specifically, the notion of context can refer to updating image similarity measures by taking into account information encoded in the ranked lists defined by a CBIR system. In the past few years, there has been considerable research on improving the distance measures in CBIR. Promising results have been obtained considering several approaches and techniques. Several unsupervised approaches have been proposed aiming to improve the effectiveness of retrieval tasks, replacing pairwise similarities by more global affinities measures that also consider the relation among the database images.

The objective of these methods is somehow mimic the human behavior on judging the similarity among objects by considering specific contexts. In this paper, we present the RL-Sim (Ranked Lists Similarities) ReRanking Algorithm, a new post-processing method that considers the similarity among ranked lists for characterizing contextual information in CBIR systems. The main motivation of our re-ranking algorithm relies on the conjecture that contextual information encoded in the similarity between ranked lists can provide resources for improving the effectiveness of CBIR descriptors. In general, only two images are considered for distance computation and, if the distance measure adopted is not accurate, the two images will be wrongly placed in the ranked lists of each other. On the other hand, for obtaining a ranked list, many distances have to be computed. These incorrect scores are often mixed with correct values, specially in the beginning of the ranked lists. In this way, the contextual information provided by the ranked lists can be used for correcting the wrong scores. Beyond that, if two images are similar, their ranked lists should be similar as well. It is somehow close to the cluster hypothesis, which states that "closely associated documents tend to be relevant to the same requests". The main contributions of this paper are: (i) the modelling of contextual information considering only the similarity between ranked lists, independent of distance (or similarity) scores between images. Distance scores computed by different image descriptor usually are in different scales and requires normalization procedures. These variations can affect the effectiveness of re-ranking approaches. Since the proposed re-ranking method does not depend on distances or similarity scores, it can be easily used for different CBIR tasks and can be adapted for other information retrieval tasks (e.g., text or multimodal retrieval); (ii) the proposed contextual distance measure does not depend on specific approaches for comparing the ranked

lists. In this way, our re-ranking algorithm can use different similarity/distance measures among ranked lists, a well-established research area.

Therefore, the re-ranking algorithm can be easily extended for using and combining different approaches for comparing ranked lists. The proposition of a generic iterative approach based on ranked lists represents an important contribution, since other re-ranking methods can be proposed just by creating new metrics to compare ranked lists. This paper differs from previous work as it presents a deeper analysis of RL-Sim Re-Ranking Algorithm, discusses different distance measures used for computing the similarity of ranked lists and extends the experimental protocol. A large experimental evaluation was conducted, considering three different datasets and twelve different image descriptors (shape, color, and texture descriptors). Other aspects of the proposed algorithm were also considered, such as the analysis of the efficiency of the method and the impact of parameters. Our experimental evaluation demonstrates that the proposed method can achieve significant improvements in various CBIR tasks. In addition, we also evaluated the proposed RL-Sim in comparison with several other state-of-the-art approaches considering a common shape dataset.

2. LITERATURE SURVEY

Determining the appropriate distance measures plays a key role in many multimedia applications, including classification, clustering, and retrieval tasks. For example, choosing a good distance measure is often critical for building an effective content-based image retrieval (CBIR) system. In general, aiming at retrieving the most similar images to a given query image, CBIR systems compute a predefined distance measure between the query image and each collection image. Traditional distance measures, as Euclidean distance, are often adopted and consider the pairwise similarity between any two images. In many situations, these approaches fail to return satisfactory results, mainly due to the well-known semantic gap challenge. In the past few years, there has been considerable research on improving the distance measures in CBIR. Promising results have been obtained considering several approaches and techniques. In this paper, we focus on unsupervised approaches. In unsupervised learning approaches, the "learning" method considers only the domain of object instances and no training labeled data are provided. Since labeling often is a laborious and time-consuming task, whereas unlabeled data is far easier to obtain, unsupervised learning represents a very attractive solution in many situations. In CBIR applications, the use of context may play an important role. In general, traditional CBIR systems perform only pairwise image analysis, that is, they compute similarity (or distance) measures considering only pairs of images, ignoring the rich information encoded in the relations of several images. However, in recent years, several CBIR approaches have been proposed aiming to improve the effectiveness of retrieval tasks replacing pairwise similarities by more global affinity measures that also consider the relation among the database objects. Although using a very diverse nomenclature (re-ranking graph transduction diffusion process, affinity learning, contextual similarity/dissimilarity measures), these post-processing methods have in common the fact that all approaches propose improving the effectiveness of image searches by exploiting the information about the relationships among collection images in an

unsupervised way (with no training data). Another important common point consists in the use of an iterative strategy adopted to process contextual information. A graph-based transductive learning algorithm was proposed for shape retrieval tasks. It learns a better metric through graph transduction by propagating the model through existing shapes, in a way similar to computing geodesics in dataset manifold. The method does not require learning the shape manifold explicitly and it does not require knowing class labels of existing shapes. The better metric is learned by collectively propagating the similarity measures to the query shape and between the existing shapes through graph transduction. Although inspired by label propagation algorithm which is semi-supervised, the shape retrieval was treated as an unsupervised problem. The locally constrained diffusion process considers that the distance between two shapes can be correctly described only if it is considered in the context of other shapes similar to them. The work observes that, since differences between shapes in the same class can be very large and differences between shapes in different classes can be very small, no pairwise shape comparison can describe shape dissimilarity correctly.

The influence of other shapes is propagated as a diffusion process on a graph formed by a given set of shapes. The weights of graph edges are defined by applying a Gaussian to the shape distance. A reversible Markov chain based on the graph is constructed and used to propagate the influence of shapes. Another approach based on propagating the similarity information in a weighted graph is called affinity learning. Instead of propagating the similarity information on the original graph, it uses a tensor product graph (TPG) obtained by the tensor product of the original graph with itself. Graphs are also used in other approaches. A modified mutual kNN graph is proposed as the underlying representation used for shape retrieval. The structure of the shape manifold is estimated from the shape similarity scores among all the shapes within a database. A shortest-path propagation algorithm was also proposed for shape/object retrieval tasks. Given a query object and a target database object, it explicitly finds the shortest path between them in the distance manifold of the database objects. Then a new distance measure is learned based on the shortest path and it is used to replace the original distance measure. Beside graph methods, context is a term frequently used for designating post-processing methods that consider relationships among images. In general interactive applications, the use of context can play an important role, which can be broadly defined as all information about the whole situation relevant to an application and its set of users. In CBIR systems, it is related to the fact that, when humans have to judge the similarity between two images, they always do so in a given context, i.e. they do not consider only the two objects to be compared. A contextual dissimilarity measure was introduced aiming to improve the accuracy of bag-of-features-based image search. The proposed measure takes into account the local distribution of the vectors and updates distances by modifying the neighborhood structure. The dissimilarity measure improves the symmetry of the k-neighborhood relationship by iteratively regularizing the average distance of each vector to its neighborhood. The method performs a global analysis of properties in small overlapping neighborhoods, resembling methods for non-linear dimensionality reduction, inspired by ISOMAP [34] and LLE. A family of contextual measures was proposed

considering the similarity between two distributions measured in the context of a third distribution.

These contextual measures are then applied to the image retrieval problem. In such a case, the context is estimated from the neighbors of a query. Using different contexts, and especially contexts at multiple scales (i.e., broad and narrow contexts), provides different views on the same problem, while combining the different views can improve retrieval accuracy. Contextual information has also been exploited for re-ranking methods. Reranking can be broadly defined as a process of refining the search results: the re-ranking methods take an initial ranking and aggregate some information for improving the effectiveness of the retrieval process. A re-ranking approach based on contextual spaces aims at exploiting the relationships among images to improve the effectiveness of CBIR tasks. Information encoded in both distances among images and ranked lists computed by CBIR systems are used for analyzing contextual information. Clustering approaches are also closely related to re-ranking methods that exploit contextual information in CBIR domain. A re-ranking framework for CBIR systems based on contextual dissimilarity measures uses a clustering approach. The contexts are modeled using a clustering algorithm to group similar images given their ranked lists. A re-ranking algorithm that uses post-retrieval clustering was proposed for color retrieval tasks. In the first step, images are retrieved using visual features such as color histogram. Next, the retrieved images are analyzed using hierarchical agglomerative clustering methods and the returned ranked lists is adjusted according to the distance of a cluster to a query. The Distance Optimization Algorithm (DOA) considers an iterative clustering approach based on distances correlation and on the similarity of ranked lists. The algorithm explores the fact that if two images are similar, their distances to other images and therefore their ranked lists should be similar as well. This paper presents the RL-Sim Re-Ranking Algorithm, a new post-processing method that considers the similarity between ranked lists for encoding contextual information in CBIR systems. We believe that the modeling of contextual information considering only the similarity between ranked lists represents an advantage of our strategy. Since the re-ranking method does not depend on distances or similarity scores, it can be used for different CBIR tasks and can be easily adapted for other information retrieval tasks (e.g., text or multimodal retrieval). Beyond that, the re-ranking method can use different similarity/distance measures among ranked lists, a well-established research area

3. PROPOSED METHODOLOGY

The proposed methodology for image re-ranking based on geospatial data networks involves leveraging geographic information to improve the relevance and quality of search results in image retrieval systems. Here's a step-by-step methodology:

1. Problem Definition and Data Acquisition:

- Clearly define the problem of image re-ranking in the context of geospatial data networks.
- Acquire geospatial data sources relevant to the image retrieval task, such as maps, satellite imagery, geotagged images, and location-based services.

2. Feature Extraction and Representation:

- Extract features from images and geospatial data to represent both visual content and geographic context.
- Use techniques such as deep learning-based feature extraction for images and feature engineering for geospatial data.

3. Geospatial Similarity Computation:

- Compute the similarity between images based on their visual content and geographic proximity.
- Develop algorithms to measure similarity using both image features and geospatial attributes, considering factors such as distance, spatial relationships, and semantic relevance.

4. Network Construction:

- Construct a geospatial data network representing the relationships between images and geographic locations.
- Define nodes as images or locations and edges as connections between images and their associated geographic contexts.

5. Graph-Based Re-ranking Algorithm:

- Design a graph-based re-ranking algorithm that exploits the structure of the geospatial data network.
- Incorporate image similarities and geographic proximity into the graph to prioritize relevant images within local geographic regions.

6. Ranking Optimization:

- Optimize the ranking of images based on their relevance to geospatial queries and user preferences.
- Consider user interactions, feedback, and query context to refine the ranking results dynamically.

4. RESULTS AND DISCUSSIONS

The RL-Sim Re-Ranking Algorithm

In this section, we present an algorithmic view of the proposed re-ranking approach. The goal of our re-ranking algorithm is to exploit the initial set of ranked lists $R = \{R_1, R_2, \dots, R_N\}$ for computing a more effective distance matrix \hat{A} and, therefore, a more effective set of ranked lists \hat{R} . The RL-Sim Re-Ranking Algorithm is based on the presented contextual measure p_c , which takes into account the similarity between ranked lists on an iterative way. An iterative approach is proposed. Let the superscript (t) denotes the current iteration, a new (and more effective) set of ranked lists $R^{(t+1)}$ is computed by taking into account distances among top k lists. Next, $R^{(t+1)}$ is used for the next execution of our re-ranking algorithm and so on. These steps are repeated along several iterations aiming to improve the effectiveness incrementally. After a number T of iterations a re-ranking is performed based on the final distance matrix \hat{A} . Based on matrix \hat{A} , a final set of ranked lists \hat{R} can be computed. Algorithm 1 outlines the proposed RL-Sim Re-Ranking Algorithm. Algorithm 1 RL-Sim Re-Ranking Algorithm Require: Original set of ranked lists R and parameters k_s, T, λ Ensure: Processed set of ranked lists \hat{R} 1: $t \leftarrow 0$ 2: $R^{(t)} \leftarrow R$ 3: $A^{(t)} \leftarrow A$ 4: $k \leftarrow k_s$ 5: while $t < T$ do 6:

for all $R_i \in R(t)$ do 7: counter $\leftarrow 0$ 8: for all $img_j \in R_i$ do 9: if counter $\leq \lambda$ then 10: $A(t+1)[i, j] \leftarrow d(\tau_i, \tau_j, k)$ 11: else 12: $A(t+1)[i, j] \leftarrow 1 + A(t)[i, j]$ 13: end if 14: counter \leftarrow counter + 1 15: end for 16: end for 17: $R(t+1) \leftarrow \text{performReRanking}(A(t+1))$ 18: $k \leftarrow k + 1$ 19: $t \leftarrow t + 1$ 20: end while 21: $R \leftarrow R \wedge (T)$ Observe that the distances are redefined considering the function $d(\tau_i, \tau_j, k)$ for the first λ positions of the each ranked list, such that $\lambda \in \mathbb{N}$ and $0 \leq \lambda \leq N$. For images in the remaining positions of the ranked lists, the new distance is redefined (Line 12) based on the current distances. In these cases, the function $d(\tau_i, \tau_j, k)$ does not need to be computed, considering that relevant images should be at the beginning of the ranked lists. In this way, the computational efforts decrease, making this step of the algorithm not dependent on the collection size N . In Line 18, at each iteration t , we increment the number of k neighbors considered. The motivation behind this increment relies on the fact that the effectiveness of ranked lists increase along iterations. In this way, non-relevant images are moved out from the first positions of the ranked lists and k can be increased for considering more images. The increment value of 1 was chosen because it represents the smallest possible increment, since the greater the increment value, the greater the risk of considering non-relevant images in the ranked list. Note that the re-ranking algorithm does not depend on specific measures between top k lists. In this way, an important advantage of our re-ranking algorithm is the possibility of using different approaches for retrieving the neighborhood set (we discussed the kNN and Mutual kNN methods) and different measures for comparing top k lists (we discussed the intersection and Kendall's tau measures). Therefore, the proposed RL-Sim Re-Ranking algorithm can be easily extended in order to consider different and even more complex approaches to compute the similarity between top k lists.

5. CONCLUSIONS

In this work, we have presented a new re-ranking method that exploits contextual information for improving CBIR tasks. The main idea consists in analyzing similarity between ranked lists for redefining distance among images. We conducted a large set of experiments and experimental results demonstrated the applicability of our method to several image retrieval tasks based on shape, color, and texture descriptors. Future work focuses on: (i) considering other different measures between top k lists; (ii) combining results obtained from different measures; (iii) optimizing the proposed re-ranking algorithm by considering parallel architectures; (iv) combining our re-ranking algorithm with other supervised methods, such as relevance feedback approaches.

6. REFERENCES

- [1] Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.. Towards a better understanding of context and context-awareness. In: 1st international symposium on Handheld and Ubiquitous Computing. 1999. p. 304–307.
- [2] Arica, N., Vural, F.T.Y.. BAS: a perceptual shape descriptor based on the beam angle statistics. Pattern Recognition Letters 2003;24(9-10):1627–1639.
- [3] Bai, X., Wang, B., Wang, X., Liu, W., Tu, Z.. Co-transduction for shape retrieval. In: European Conference on Computer Vision (ECCV'2010). volume 3; 2010. p. 328–341.

- [4] Belongie, S., Malik, J., Puzicha, J.. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2002;24(4):509–522.
- [5] Brodatz, P.. *Textures: A Photographic Album for Artists and Designers*. Dover, 1966.
- [6] Cormack, G.V., Clarke, C.L.A., Buettcher, S.. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *ACM SIGIR Conference on Research and Development in Information Retrieval*. 2009. p. 758–759.
- [7] Fagin, R., Kumar, R., Sivakumar, D.. Comparing top k lists. In: *ACMSIAM Symposium on Discrete algorithms (SODA'03)*. 2003. p. 28–36.
- [8] Faria, F.F., Veloso, A., Almeida, H.M., Valle, E., da S. Torres, R., Goncalves, M.A., Jr., W.M.. Learning to rank for content-based image retrieval. In: *Multimedia Information Retrieval (MIR'2010)*. 2010. p. 285– 294.
- [9] Gopalan, R., Turaga, P., Chellappa, R.. Articulation-invariant representation of non- planar shapes. In: *11th European Conference on Computer Vision (ECCV'2010)*. volume 3; 2010. p. 286–299. 22
- [10] Hoi, S.C., Liu, W., Chang, S.F.. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing and Communication Applications* 2010;6(3):18:1– 18:26.

IMPLEMENTATION OF COST EFFECTIVE MULTI CLOUD STORAGE IN PUBLIC CLOUDS

M. SAMEERA BEGAM

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Preparing a data set for analysis is generally the most time consuming task in a data mining project, requiring many complex SQL queries, joining tables, and aggregating columns. Existing SQL aggregations have limitations to prepare data sets because they return one column per aggregated group. In general, a significant manual effort is required to build data sets, where a horizontal layout is required. We propose simple, yet powerful, methods to generate SQL code to return aggregated columns in a horizontal tabular layout, returning a set of numbers instead of one number per row. This new class of functions is called horizontal aggregations. Horizontal aggregations build data sets with a horizontal denormalized layout (e.g., point-dimension, observation variable, instance-feature), which is the standard layout required by most data mining algorithms. We propose three fundamental methods to evaluate horizontal aggregations: CASE: Exploiting the programming CASE construct; SPJ: Based on standard relational algebra operators (SPJ queries); PIVOT: Using the PIVOT operator, which is offered by some DBMSs. Experiments with large tables compare the proposed query evaluation methods. Our CASE method has similar speed to the PIVOT operator and it is much faster than the SPJ method. In general, the CASE and PIVOT methods exhibit linear scalability, whereas the SPJ method does not.

Keywords: data mining, CASE, PIVOT operator, SQL aggregations, Horizontal, algebra operators

1. INTRODUCTION

The end of this decade is marked by a paradigm shift of the industrial information technology towards a subscription based or pay-per-use service business model known as cloud computing[1]. This paradigm provides users with a long list of advantages, such as provision computing capabilities; broad, heterogeneous network access; resource pooling and rapid elasticity with measured services. Huge amounts of data being retrieved from geographically distributed data sources, and non-localized [2] data-handling requirements, creates such a change in technological as well as business model. One of the prominent services offered in cloud computing is the cloud data storage, in which, subscribers do not have to store their data on their own servers, where instead their data will be stored on the cloud service provider's servers[3]. In cloud computing, subscribers have to pay the service providers for this storage service.

This service does not only provides flexibility and scalability for the data storage, it also provide customers with the benefit of paying only for the amount of data they need to store for a

particular period of time, without any concerns for efficient storage mechanisms and maintainability issues with large amounts of data storage[4]. In addition to these benefits, customers can easily access their data from any geographical region where the Cloud Service Provider's network or Internet can be accessed. An example of the cloud computing. Along with these unprecedented advantages, cloud data storage also redefines the security issues targeted on customer's[5] outsourced data (data that is not stored/retrieved from the costumers own servers). Since cloud service providers (*SP*) are separate market entities, data integrity and privacy are the most critical issues that need to be addressed in cloud computing. Even though the cloud service [6] providers have standard regulations and powerful infrastructure to ensure customer's data privacy and provide [7] a better availability, the reports of privacy breach and service outage have been apparent in last few years. Also the political influence might become an issue with the availability of services.

In this work we observed that, from a customer's point of view, relying upon a solo *SP* for his outsourced data is not very promising[8]. In addition, providing better privacy as well as ensuring data availability, can be achieved by dividing the user's data block into data pieces and distributing them among the available *SP*s in such a way that no less than a threshold number of *SP*s can take part in successful retrieval of the whole data block. In this paper, we propose a secured cost-effective multi-cloud storage (SCMCS)[9] model in cloud computing which holds an economical distribution of data among the available *SP*s in the market, to provide customers with data availability as well as secure storage. Our results show that, our proposed model provides a better decision for customers according to the iravailable budgets. Keywords: Cloud computing, security, storage, cost-effective, cloud service provider, customer[10].

The end of this decade is marked by a paradigm shift of the industrial information technology towards a subscription based or pay-per-use service business model known as cloud computing. This paradigm provides users with a long list of advantages, such as provision computing capabilities; broad, heterogeneous network access; resource pooling and rapid elasticity with measured services. Huge amounts of data being retrieved from geographically distributed data sources, and non-localized data-handling requirements, creates such a change in technological as well as business model. One of the prominent services offered in cloud computing is the cloud data storage, in which, subscribers do not have to store their data on their own servers, where instead their data will be stored on the cloud service provider's servers. In cloud computing, subscribers have to pay the service providers for this storage service. This service does not only provides flexibility and scalability for the data storage, it also provide customers with the benefit of paying only for the amount of data they need to store for a particular period of time, without any concerns for efficient storage mechanisms and maintainability issues with large amounts of data storage. In addition to these benefits, customers can easily access their data from any geographical region where the Cloud Service Provider's network or Internet can be accessed.

Along with these unprecedented advantages, cloud data storage also redefines the security issues targeted on customer's outsourced data (data that is not stored/retrieved from the

customers own servers). Since cloud service providers (*SP*) are separate market entities, data integrity and privacy are the most critical issues that need to be addressed in cloud computing. Even though the cloud service providers have standard regulations and powerful infrastructure to ensure customer's data privacy and provide a better availability, the reports of privacy breach and service outage have been apparent in last few years. Also the political influence might become an issue with the availability of services. In this work we observed that, from a customer's point of view, relying upon a solo *SP* for his outsourced data is not very promising. In addition, providing better privacy as well as ensure data availability, can be achieved by dividing the user's data block into data pieces and distributing them among the available *SP*s in such a way that no less than a threshold number of *SP*s can take part in successful retrieval of the whole data block. To address these issues in this paper, we proposed an economical distribution of data among the available *SP*s in the market, to provide customers with data availability as well as secure storage. In our model, the customer divides his data among several *SP*s available in the market, based on his available budget. Also we provide a decision for the customer, to which *SP*s he must chose to access data, with respect to data access quality of service offered by the *SP*s at the location of data retrieval. This not only rules out the possibility of a *SP* misusing the customers' data, breaching the privacy of data, but can easily ensure the data availability with a better quality of service.

2. LITERATURE REVIEW

AMPL: AMPL is a language for large-scale optimization and mathematical programming problems in production, distribution, blending, scheduling, and many other applications. Combining familiar algebraic notation and a powerful interactive command environment, AMPL makes it easy to create models, use a wide variety of solvers, and examine solutions. Though flexible and convenient for rapid prototyping and development of models, AMPL also offers the speed and generality needed for repeated large-scale production runs. This book, written by the creators of AMPL, is a complete guide for modelers at all levels of experience. It begins with a tutorial on widely used linear programming models, and presents all of AMPL's features for linear programming with extensive examples. Additional chapters cover network, nonlinear, piecewise-linear, and integer programming; database and spreadsheet interactions; and command scripts. Most chapters include exercises. Download free versions of AMPL and several solvers from www.ampl.com for experimentation, evaluation, and education. The Web site also lists vendors of the commercial version of AMPL and numerous solvers.

Modeling Language for Mathematical Programming: Practical large-scale mathematical programming involves more than just the application of an algorithm to minimize or maximize an objective function. Before any optimizing routine can be invoked, considerable effort must be expended to formulate the underlying model and to generate the requisite computational data structures. AMPL is a new language designed to make these steps easier and less error-prone. AMPL closely resembles the symbolic algebraic notation that many modelers use to describe mathematical programs, yet it is regular and formal enough to be processed by a computer system; it is particularly notable for the generality of its syntax and for the variety of its indexing

operations. We have implemented an efficient translator that takes as input a linear AMPL model and associated data, and produces output suitable for standard linear programming optimizers. Both the language and the translator admit straightforward extensions to more general mathematical programs that incorporate nonlinear expressions or discrete variables.

Attack Surfaces: The new paradigm of cloud computing poses severe security risks to its adopters. In order to cope with these risks, appropriate taxonomies and classification criteria for attacks on cloud computing are required. In this work-in-progress paper we present one such taxonomy based on the notion of attack surfaces of the cloud computing scenario participants.

Most of the papers around cloud interoperability are looking into IaaS services and intercloud architecture, while literature on cloud storage interoperability is scarce. The Authors in presented a model driven approach for Cloud-to-Cloud Interoperability, C2CI by defining a model of 5 levels to assess the interoperability maturity, L0 Domain-based, L1 Enterprise- based, L2 Portability interoperability, L3 Security interoperability, L4 – Mobile interoperability in a public, private, or hybrid cloud environment. It's an adaptation of the Levels of Information System Interoperability, LISI maturity model, additionally in their paper they described some of the challenges in achieving interoperability intercloud. In, the authors provided some basis on the interoperability in cloud, by identifying in the existing literature how multiple research done in the field of interoperability addressed the matter, they have surveyed several literature for up to 2016 highlighting the solutions. Onesolution is the standardization of cloud services with CPs adhering to the initiatives, another one is building Abstraction Layers, which consists of abstracting common feature of CPs, an interesting approach is the Cloud Storage Abstraction Layer, CSAL proposed by Hill and Humphrey to provide common storage abstraction layer between clouds. Alternatively the authors in, tackled the PaaS interoperability with a Model Driven Engineering (MDE) approach, proposing a middle platform that decouples a PaaS hosted Application and its databased into two layers and ports it to another PaaS platform. MDE approach proved to be a better alternative to standardization a per literature, they leveraged it to make the middleware flexible, further making application layer and DB layer portable, productive and reusable. Although these papers are relevant, we thought it would be important to go beyond the simple service interoperability and cover the aspects of the storage that matters to most cloud users, the complexity, the cost, and the time needed for data transfer.

3. PROPOSED METHODOLOGY

A proposed system for the implementation of cost-effective multi-cloud storage in public clouds aims to address the challenges associated with managing data across multiple cloud providers while optimizing costs. Implement automated cost optimization algorithms that analyze storage usage patterns, pricing models, and available discounts across different cloud providers. The system should recommend cost-saving strategies such as data tiering, resource consolidation, and utilization optimization to minimize storage costs. Develop a unified management dashboard that provides centralized visibility and control over storage resources across multiple public cloud providers. The dashboard should allow administrators to monitor

storage usage, costs, performance metrics, and data replication status in real-time. Define dynamic data replication policies based on data importance, access frequency, and compliance requirements. The system should automatically replicate data across multiple cloud providers and regions according to predefined policies, optimizing data availability, and redundancy while minimizing costs.

ADVANTAGES

- Organizations can leverage cost-saving strategies such as data tiering, resource consolidation, and utilization optimization to optimize storage expenditures.
- By replicating data across multiple cloud providers and regions according to predefined policies, the system enhances data redundancy and availability.
- The system automates storage management tasks such as data replication, lifecycle management, and performance optimization, reducing administrative overhead and improving operational efficiency.
- Actionable insights and recommendations provided by the system help organizations make informed decisions to optimize storage expenditures and maximize cost savings.

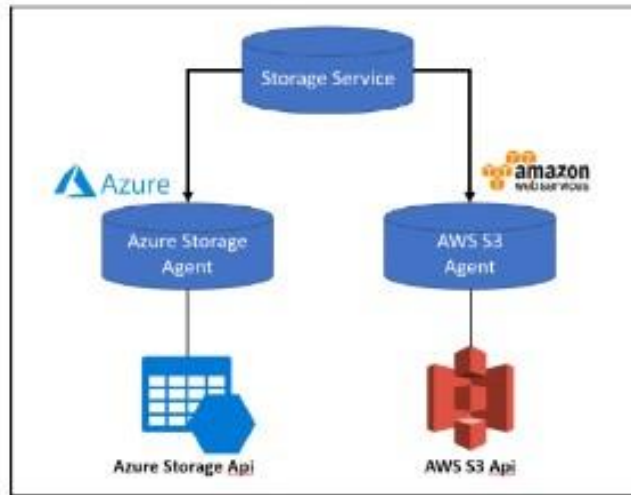
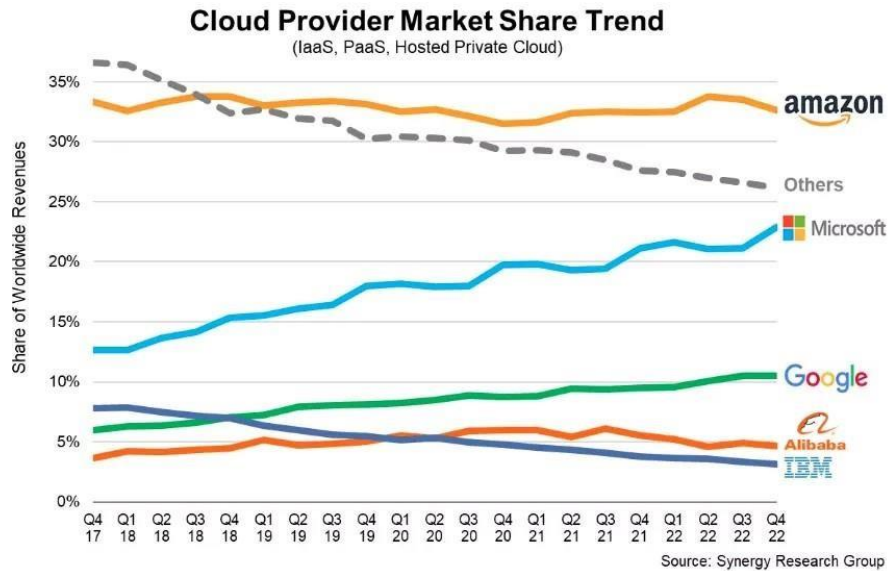


Figure 1: Proposed approach applied for two public cloud interoperability

4. RESULTS AND DISCUSSIONS

- In Q2 2022, AWS commanded 34% of the cloud market, a 1% increase year-over-year. Azure is second with 21% of the market, followed by Google Cloud (10%), Alibaba (5%), and IBM (4%).
- Amazon’s revenue from AWS grew from 5.62% in 2014 to 13.24% in 2021.
- The survey also showed the cloud market continues to grow 34% year-over-year.
- Fun fact: It took Netflix seven years to migrate to AWS.



Graph.1 shown the cloud provider market share trend

5. CONCLUSION

In conclusion, the implementation of cost-effective multi-cloud storage in public clouds offers a promising solution for organizations seeking to optimize storage costs, enhance data availability, and improve operational efficiency in multi-cloud environments. Through the proposed system, organizations can leverage automation, optimization algorithms, and robust management practices to achieve. The system enhances data redundancy and availability by replicating data across multiple cloud providers and regions. This ensures data durability, mitigates the risk of data loss, and enables organizations to maintain high levels of data availability even in the event of localized outages or disruptions. By implementing dynamic cost optimization algorithms and automated data lifecycle management policies, organizations can minimize storage costs across multiple cloud providers. Through proactive monitoring, analysis, and optimization, organizations can identify cost-saving opportunities, optimize resource utilization, and achieve significant cost savings.

6. REFERENCES

- [1] Amazon.com, "Amazon s3 availability event: July 20, 2008", Online at <http://status.aws.amazon.com/s3-20080720.html>, 2008.
- [2] "A Mordern Language for Mathematical Programming", Online at <http://www.ampl.com>.
- [3] M. Arrington, "Gmail Disaster: Reports of mass email deletions", Online at <http://www.techcrunch.com/2006/12/28/gmail-disasterreports-of-mass-email-deletions/>, December 2006.
- [4] P. S. Browne, "Data privacy and integrity: an overview", In Proceeding of SIGFIDET '71 Proceedings of the ACM SIGFIDET (now SIGMOD), 1971.
- [5] A. Cavoukian, "Privacy in clouds", Identity in the Information Society, Dec2008.

- [6] J. Du, W. Wei, X. Gu, T. Yu, "RunTest: assuring integrity of dataflow processing in cloud computing infrastructures", In Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS '10), ACM, New York, NY, USA, 293-304.
- [7] R. Gellman, "Privacy in the clouds: Risks to privacy and confidentiality from cloud computing", Prepared for the World Privacy Forum, online at [http://www.worldprivacyforum.org/pdf/WPF Cloud Privacy Report.pdf](http://www.worldprivacyforum.org/pdf/WPF%20Cloud%20Privacy%20Report.pdf), Feb 2009.
- [8] The Official Google Blog, "A new approach to China: an update", online at <http://googleblog.blogspot.com/2010/03/new-approach-to-china-update.html>, March 2010.
- [9] N. Gruschka, M. Jensen, "Attack surfaces: A taxonomy for attacks on cloud services", Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on, 5-10 July 2010.
- [10] W. Itani, A. Kayssi, A. Chehab, "Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures," Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Dec 2009.

IMPROVED MECHANISM FOR FAKE CURRENCY DETECTION

S. HARISREE

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

The project entitled as "IMPROVED MECHANISM FOR FAKE CURRENCY DETECTION". An effective code has developed to detect the given Indian rupee as an image. The neural network technique has been implemented to detect the currency. After applying the image input to this project it will show the detected result that is value of money. If it is a fake one it does not given the value of money. It will remain null to display. Counterfeit money is imitation currency produced without the legal sanction of the state or government. Producing or using this fake money is a form of fraud or forgery. Counterfeiting is as old as money itself, and is sufficiently prevalent throughout history that it has been called "the world's second oldest profession.. This has led to the increase of corruption in our country hindering country's growth. Common man became a scapegoat for the fake currency circulation, let us suppose that a common man went to a bank to deposit money in bank but only to see that some of the notes are fake, in this case he has to take the blame.

Counterfeiting, of whatever kind, may be that has been occurring ever since humans grasped the concept of valuable items, and there has been an ongoing race between certifier like (banks, for example) and counterfeiter ever since. Some of the effects that counterfeit money has on society include a reduction in the value of real money; and inflation due to more money getting circulated in the society or economy which in turn dampen our economy and growth - an unauthorized artificial increase in the money supply; a decrease in the acceptability of paper money; and losses.

Keywords: neural network, Counterfeiting, fake, economy, fraud or forgery

• INTRODUCTION

In Banking-sector, biggest risk is fake currency generation. Mostly UV light is applying for authentication proving [1]. Main features to detect fake currency are Note value , ink smudge , Security thread , serial number , Intaglio printing , watermark , Reserve bank number panel , LD mark , Topography , Micro-lettering and numbers & alignment. In that important features are watermark, ink smudge[2], security thread, topography, numbers & place and micro-lettering. However for machine based evaluation, usually the following steps to be carrying out by researchers.

One by one identification performed by manual is only possible for confined quantity of notes. Therefore demand a machine learning centered alternative to recognize perhaps the image is real or fake [3]. Each year Arrange Bank of India encounters issue of phony currency notes or ruined notes. The material which are employing in every countries are different. In India, currency notes are comprised of pulp containing cotton and balsam with particular dyes to really make the currency notes that should be resistant, tough, with quality to fight from wear and split and never to be faked easily[4]. USD notes are produced from cotton fiber paper, in place of

wood fiber, which can be frequently applied to create popular paper. Japanese banknotes are produced from mitsumata (*Edgeworthia papyrifera* or Oriental paperbush), abaca pulp, and different fibers. Canada is trading in their paper currency for plastic. BSF seized fake currency notes of Rs. 2000 denomination with a face value of Rs 4.4 lakh from global boundary below Dhubri district of Assam last year and once BSF reported that those notes origin is cigarette paper. "You will find 16 main features of an Indian currency note , four that are invisible to the naked eye. The FICN networks had illegal access to the material and had acquired the art of replicating Indian currency. All that transformed after demonetization [5]. Number currency seized by security or police forces suggest use of cloth," said an official published at the Indo-Bangladesh border. An estimation done by the RBI says that the amount of the fake notes in the FY 2017 has increased by 20% and the quantity reached around 7.62 lac. The most variety of fake notes were in the denomination of 500 and 2000 rupees. Authorities found 389 fake notes of ₹2,000 denomination and a few bundles of basic sheets reduce to the measurements of currency notes of ₹2,000 and ₹500 denominations [6]. Law enforcement recovered Rs 6.5 lakh in fake currency of Rs 2,000 denomination from the accused. Counterfeit detection pen and detectors are available in market , which largely stresses to discover the forgery predicated on UV light , Watermark and Magnetic ink [7]. Even when it's inexpensive, every little stores cannot hold these. And more over item is likely to be different predicated on its company. Typically fake showing is doing by UV light and predicated on some features manual checking. Fake currency detection is very required to discover the currency authorization proving. To stop circulation of counterfeit notes, a system to discover fake notes must be created and to assure the banking security. Fake image detection and its authorization showing is a complicated region nowadays [8]. Fake currency detection is founded on standards. Image forgery detection have no standards. Imagine in the event of Fake currency detection , scientists can concentrate on "state and currency note feature" wise. But in the event of image forgery [9], there are number specific standards accessible, it must require some feature extraction based evaluation to discover the features for classification method if the researcher is ready for feature extraction based analysis work [10]

In this paper, Section II discusses existing work related to every stages for detecting the fake currency. Section III describes comparison methods and IV describes conclusion and possible future work.

• LITERATURE REVIEW

Most of studies that were existed in the market are useful for fake note detection but they are ways required to still improve the process. Regarding the demonstration of , the ideology mentioned is combination of Single Shot Detector (SSD) and Convolution Neural Network (CNN) is applied over currency in order to detect the front and back denomination and extract the features. The accuracy of this approach is raised to 97% in detecting the currency. In the view of description of, the approach is used to count the number of interruptions in the thread line determine the currency note is real or fake. The efficiency is determined using entropy of the note and the software proposed is MATLAB. As per the principle of source mentioned in,

the fake note which is circulating in the country is to be detected with factors such as compatibility and mobility using image processing technique. This fake note would create a loss to the society and defame the innocent persons.

Regarding the demonstration from, the various detection techniques are proposed over the fake currency. This detection depends on the features of currency note and the specific country. These mentioned techniques are compared and contrasted. In the view of description of , the sequence of the operations that are proposed in determining the fake or genuine note accurately although many methodologies are there in the market. Among the operations, classification and selection are significant and are used at final stage. With respect to source of, the various techniques are expressed and are analyzed in finding the fake note based on defined features of a country. As per the observation of, the various image processing techniques are applied for currency paper detection and those techniques merits and demerits are analyzed. With respect to the demonstration of, the canny edge detection algorithm is applied over the decomposed images of a bank note and would be faster than original note using bit plane slicing technique. In the view of description in , the statistical classification technique called Logistic regression and linear discriminate analysis is used to accurately detect the fake currency effectively and is considered as better model for authenticate the currency. Regarding the work denoted in, the image processing techniques are applied in order to detect the Indian currency notes as fake or real based on three significant fields such as imprint, intaglio, and threads. As per the approach specified in, the image processing technique and machine learning approach is helpful in determining the fake currency although there are many approaches are existed. The study is taken because fake currency may defame the country in many ways. In regard to the source mentioned in, the mobile app called BLaDE is proposed to use barcode reader in order to scan the items especially for visually impaired persons. The few samples are taken for study and feedback is taken as audio. With the regard of description provided from, the food packets are labelled with barcodes and are obtained using featuristic barcode scanners. It is designed for blind people that would help to scan the barcodes on food packets. As per the description from, the computerized barcode reader scans the object and sends a code that is decoded to URL and output the verbal description of the object finally that would help to the visually impaired people. As per the methodology specified in , the policies of the country are considered in converting local currency into required currency and payment to be done and alert the merchant and customer accounts simultaneous time. Here, the reading the product and asking for making payment is done according to customer local currency against the merchant currency. Regarding the demonstration of data given in , the certain countries are taken and money to be converted accurately in to the authorized countries currency. The international currency is restricted to few countries but not limited to all countries. Here, the categories are made into over, balanced, and under type customizers. With respect to the source given in, the classifiers such as naïve and KNN are applied one after the other in order to detect the attack in intrusion detection system is trusty or un-trusty depending on the deepest packet inspection. Regarding the study in, the AI and ML technique called SVM is applied over the nonlinear data

and forecast the stock prediction in more accurately when started to compare against such approaches, in which SVM proves the better approach. As per the principle specified in, the currency to be scanned in order to decide that is real or fake that could be done using AFCS. The machine learning method called deep learning is used in predicting the note is fake or real rather than precious image processing techniques. With respect to the source mentioned in, the currency note detection to the blind people is done using image processing technique there steps carried in such a way that scan the note, convert that into the text and that converted into the audio that really helps as a guide to the blind people. With respect to the source provided in, the currency here is bit coin that was in circulation in most of the countries, the transfer of bit coin from one person to other person is maintained as history of that coin, hence there is no chance offake currency. The methodology used here is RCNN and LSTM are applied along with python's library tensor flow for accurate prediction of the currency. With respect to the source mentioned in, the measu res like accuracy, recall, efficiency are considered for fake note detection and enhanced SVM is applied over the note image that was scanned. The result would be note is valid note or not. As per the demonstration of information in, the denomination of noteand determination of note is genuine or fake based on ID and serial number are considered using advanced image processing approach. The steps that are involved are feature extraction, edge detection, image segmentation and compare the images for determining the note is genuine or not. In the view of information specified in, the botanical plants raw materials are assigned mini-barcodes so that customers could have a trust on the plants that purchase original type of plant o specific characteristic or not. The kind of testing here considered is DNA testing over botanical items. As per direction of extracted information in simple words from, the eyesight is scanned and generate a report on the eye sight that is of shortage w.r.to threshold values and get notification on the order of item from the nearest the merchant. With respect to the data represented in, the digital mask is designed in order to detect the surroundings about the affected or healthy and alert the user about the infected area, and suggests the prescription to set-right the environment. Regarding of work dumped in, the specific currency notes are trained and their denominations are verified using convolution neural network. The feature maps of every note from the online database are compared against the note given for verification, for judging the note is fake or not in less computation time. In the juncture made from, the application is designed there note is scanned and allow to apply defined features over that note. The image processing technique is used internally to extract the features that helps to judge the note is fake or real. The future scope given in this is foreign currency may be considered with modification inthe implemented ideology. From the view of depiction from , the Bangladesh currency is to be verified and produce the note is fake or not based on specific features like hologram, water marking, and printing style quality as well as methodology suggested is support vector machines. The obtained results are verified and proved that SVM is effectively proved than other existing methods. With respect to the source specified in, the sequences of steps are performed inorder to determine the note is fake or genuine. The characteristics of image processing technique are analyzed and computed intensities are compared against the threshold intensities for the

defined features. Based on satisfaction, the result is announced. As per description noted from , the measures assumed such as serial number, fitness, authenticity, and denomination are computed for note based on efficient processing machines. These would detect the note is real or not based on sensors loaded with the dataset. In the view of demonstration of information from, the preparation of anti counterfeiting applied to the notes w.r.to three main factors ink, printing, and substrate. These reflect security for the note against duplicate or fake currencies. In the principle raised from\, the policies and guidelines are framed by the elected governing body of a country in order to make currency

note more secure. Some of studies mentioned above are representing the various ML, AI and Image processing techniques in determining accurately, and others are used for scanning and further analysis of decision making

- **PROPOSED METHODOLOGY**

Incorporate advanced technologies such as machine learning, computer vision, and pattern recognition algorithms to enhance the accuracy and efficiency of fake currency detection. Utilize high-resolution imaging systems capable of capturing intricate details of banknotes, including watermarks, security threads, and microprinting. This helps in detecting subtle differences between genuine and counterfeit currency. Implement algorithms for feature extraction from banknote images. These algorithms should identify unique features such as serial numbers, color patterns, and security features that distinguish genuine currency from counterfeit notes. Train machine learning models, such as deep neural networks or support vector machines, on a large dataset of genuine and counterfeit currency images. These models can learn to recognize patterns and anomalies indicative of counterfeit currency. Use dynamic thresholding techniques to adaptively adjust the sensitivity of detection algorithms based on the complexity and variability of counterfeit currency patterns. Combine multiple verification methods, such as visual inspection, ultraviolet (UV) light examination, and magnetic ink detection, to improve the robustness of the detection system. Implement real-time processing capabilities to analyze banknote images quickly and accurately during cash transactions or currency verification processes.

ADVANTAGES

- ✓ Use dynamic thresholding techniques to adaptively adjust the sensitivity of detection algorithms based on the complexity and variability of counterfeit currency patterns.
- ✓ Train machine learning models, such as deep neural networks or support vector machines, on a large dataset of genuine and counterfeit currency images. These models can learn to recognize patterns and anomalies indicative of counterfeit currency.
- ✓ Implement real-time processing capabilities to analyze banknote images quickly and accurately during cash transactions or currency verification processes.

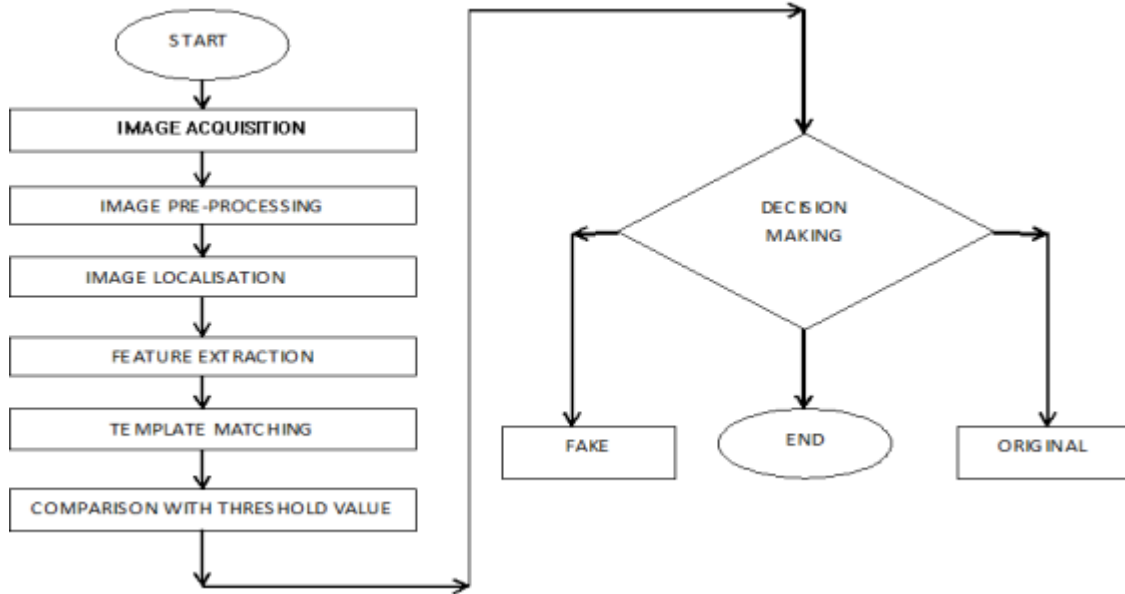
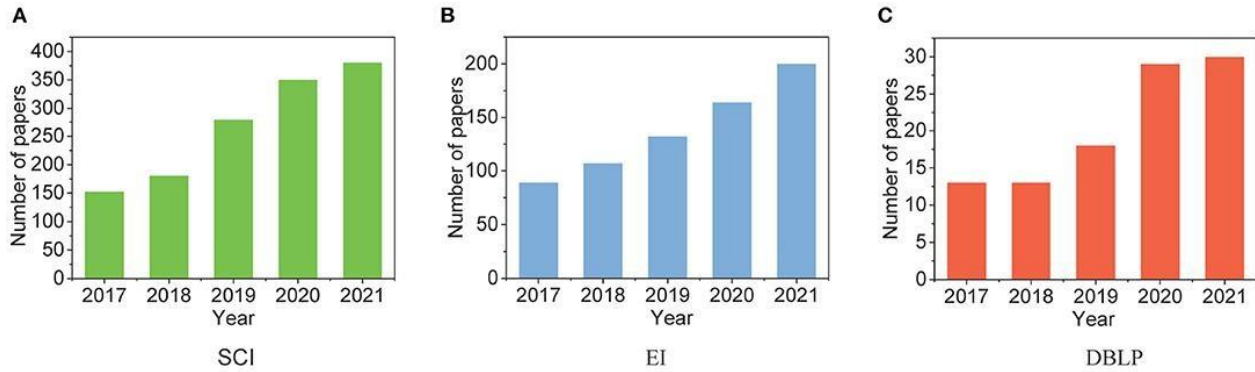


Fig.1 shows the detection of image processing

• **RESULTS AND DISCUSSIONS**

Fake reviews have become prevalent on various social networks such as e-commerce and social media platforms. As fake reviews cause a heavily negative influence on the public, timely detection and response are of great significance. To this end, effective fake review detection has become an emerging research area that attracts increasing attention from various disciplines like network science, computational social science, and data science. An important line of research in fake review detection is to utilize graph learning methods, which incorporate both the attribute features of reviews and their relationships into the detection process. To further compare these graph learning methods in this paper, we conduct a detailed survey on fake review detection. The survey presents a comprehensive taxonomy and covers advancements in three high-level categories, including fake review detection, fakereviewer detection, and fake review analysis. Different kinds of fake reviews and their corresponding examples are also summarized. Furthermore, we discuss the graph learning methods, including supervised and unsupervised learning approaches for fake review detection. Specifically, we outline the unsupervised learning approach that includes generation-based and contrast-based methods, respectively. In view of the existing problems in the current methods and data, we further discuss some challenges and open issues in this field, including the imperfect data, explainability, model efficiency, and lightweight models.



Graph.1 shown the graph learning of fake review detection

5. CONCLUSION

Since the monetary property highlights are discovered layer by layer, the discovery precision is often great. We've looked at the whole image of money so far, but in the future, we'll try to include all of the security features of money by using a fair fundamental structure and providing sufficient preparation information. Furthermore, clamour may be present in the captured image, which must be taken into account as a pre handling step in the money location procedure. It is also possible to achieve recognition and phoney money recognition by using examples of cash surface as highlights for enhancing the finding precision. As a result, the various strategies presented in this research were effectively implemented and tested by experiments on the model. Using the modules, CNN was shown to be the optimal feature for performing the approach. By doing model classification, we were able to attain a 95% accuracy rate. In addition, the detection of coins works effectively in this manne

6. BIBLIOGRAPHY

- [1] Prof Chetan More, Monu Kumar, Rupesh Chandra, Raushan Singh, "Fake04 | Apr 2020 ISSN: 2395-0056 currency Detection using Basic Python Programming and Web Framework" IRJET International Research Journal of Engineering and Technology, Volume: 07 Issue:
- [2] Vivek Sharan, Amandeep Kaur," Detection of Counterfeit Indian Currency Note Using Image Processing" International Journal of Engineering and Advanced Technology (IJEAT), Volume.09, Issue:01, ISSN: 2249-8958 (October 2019)
- [3] Aakash S Patel, "Indian Paper currency detection" International Journal for Scientific Research & Development (IJSRD), Vol. 7, Issue 06, ISSN: 2321-0613(June 2019)
- [4] Archana M Kalpitha C P, Prajwal S K, Pratiksha N," Identification of fake notes and denomination recognition" International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume. 6, Issue V, ISSN: 2321-9653, (May2018)
- [5] S. Atchaya, K. Harini, G. Kaviarasi, B. Swathi, "Fake currency detection using Image processing", International Journal of Trend in Research and Development(IJTRD),ISSN: 2394-9333 (2017).

IMPROVED SOCIAL MEDIA OPTIMIZATION (SMO) AND SPAM DATA FILTERING IN SOCIAL NETWORKS

S. VALLEESWARI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Email is one of the common communication methods between people on the Internet. However, the increase in email misuse/abuse has resulted in an increasing volume of spam emails over recent years. In this e-world, most of the transactions and business are taking place through e-mails. Nowadays, email has become a powerful tool for communication as it saves a lot of time and cost. But, due to social networks and advertisers, most of the emails contain unwanted information called spam. Even though a lot of algorithms have been developed for email spam classification, still none of the algorithms produces 100% accuracy in classifying spam emails. This project proposes a Decision tree and SMO algorithm to classify the mails. The Proposed method used the SpamBase Dataset for training purposes. The experimental result shows that the SMO gives better accuracy than the Decision Tree.

Keywords: communication, email, social networks, accuracy, spam

1. INTRODUCTION

In today's interconnected digital world, email communication serves as a vital channel for personal and professional interactions. However, alongside the legitimate messages that populate our inboxes, there exists a persistent nuisance: spam emails. These unsolicited messages flood our email accounts, cluttering our inboxes with advertisements, phishing attempts, and other unwanted content[1]. The prevalence of spam not only diminishes the efficiency of email communication but also poses significant security risks to users and organizations. Recognizing the importance of mitigating the impact of spam emails, this project focuses on the development of a robust and efficient spam email classification system[2]. Leveraging the power of machine learning techniques, we aim to create models capable of accurately distinguishing between spam and legitimate emails, thereby enabling users to filter out unwanted content effectively.

Spam emails exhibit diverse characteristics, ranging from deceptive subject lines to suspicious URLs and attachments. Understanding these patterns is crucial for designing effective classification models. Extracting relevant features from email content, including text, metadata, and structural attributes, is essential for training machine learning models[4]. Moreover, selecting the most informative features can significantly impact the performance and efficiency of the classification system. Building and fine-tuning machine learning models requires careful selection of algorithms, optimization of hyperparameters, and rigorous evaluation against benchmark datasets. Assessing the performance of the models in terms of accuracy, precision, recall, and other metrics is vital for gauging their effectiveness.

The ultimate goal of this Paper is to deploy the developed classification system in real-world email environments, such as email clients or server-side filters. Seamless integration and

compatibility with existing email infrastructures are essential considerations for practical implementation[6]. By addressing these challenges, we aim to contribute to the advancement of email security and user experience. A reliable spam email classification system not only helps users better manage their inboxes but also safeguards against phishing attacks, malware distribution, and other malicious activities facilitated through spam emails.

In the subsequent sections of this paper, we will delve into the theoretical foundations of machine learning, explore various algorithms and techniques for spam email classification, and present our methodology for dataset collection, feature engineering, model development, and evaluation[7]. Furthermore, we will discuss the implications of our findings and recommendations for future research and practical applications. Through this endeavour, we strive to enhance the efficiency, reliability, and security of email communication in the digital age

Spam email is unsolicited and unwanted junk email sent out in massive amounts or bulk to an indiscriminate recipient list. Generally, spam is sent for commercial purposes. It is sent in massive volume by botnets, networks of infected computers. Spam email can often be a malicious attempt to gain access to your system. Spam prevents the user from making full and good utilization of CPU time, storage capacity, and network bandwidth[10]. It becomes a huge problem, especially at times when there are Spam emails that come in between important business emails. Hence, it becomes inevitable to solve such problems which are encountered by spam email. So, this problem can be solved by using Machine Learning methods that can successfully detect and filter spam. It is also important to find out which technique or algorithm can best fit the purpose of classifying spam mail.

2. LITERATURE REVIEW

The NB gives low performance and works well for email-based spam detection⁹. Mehul Gupta et al. study spam detection in SMS by using ML algorithms. The deep learning-based convolutional neural network (CNN) works better than the SVM and NB algorithms. Likewise, the image-based spam detection is also done through the CNN technique. This technique works well for some smaller datasets and increases complexity rates in large datasets¹⁰. Faiza Masood et al. detect spam and fake users on the social network. The malware alerting system and regression prediction models are used for the fake content prediction. The Twitter content is analysed to identify fake content and users, spam in the URL's and trending topics. This work analysed in detail the prevention of fake accounts and the spread of fake news. In general, fake news and user predictions are extremely difficult to process when dealing with large amounts of media data¹¹. Yosef Hasan Fayeza Jabra et al. proposed spam detection in Twitter using an URL-based detection technique. Nowadays, spammers are the major platform to demand social networks and spread irrelevant data to users. In particular, Twitter is the most prominent network to spread spam among the social networks. To avoid this spread, the author used URL and ML-based detection techniques. Compared to other ML algorithms, the RF-based classification

technique provides a higher accuracy rate of 99.2% in this process. In this work, 70% was used as training data and 30% was used for testing purposes¹². Asif Karim et al. surveyed the state of intelligent spam detection in email. Both artificial intelligence and Methods are used for intelligent spam detection. This combined approach protected emails from phishing attacks. Apart from content filtering, the other methods are less covered in this analysis¹³. Guang-Bin Huang et al. proposed regression and multiclass classification-based extreme learning techniques. It shows that both the learning framework of SVM and extreme learning machines (ELM) can be implemented. It has better scalability and faster learning speed. But it provides very low performance¹⁴. Chensu Zhao et al. discussed ensemble learning based spam detection with imbalanced data in social networks. The heterogeneous-based ensemble technique is used in the imbalance class to detect spam ions. The base and combine modules are integrated for finding spam in an OSN. In the base module, the basic ML algorithms were used to find the spam, and in the combine module, the deep learning-based neural network was used to detect the spam with dynamic adjustment of weight values. This technique works well for Twitter-based real-spam datasets but not well for hidden features¹⁵. Gauri Jain et al. proposed the convolutional and long-short term memory-based neural network (LSTM) technique for spam detection. The CNN and LSTM are combined to detect spam on the Twitter network. The knowledge-based technique is used to improve the prediction accuracy of spam detection. This technique works well on short messages like Twitter messages instead of lengthy email messages¹⁶. Aliaksandr Barushka et al. discussed the cost-sensitive and ensemble-based deep neural networks for spam detection on OSN. Traditional ML algorithms, such as SVM and NB techniques, are unsuitable for high-dimensional data on OSN. To reduce the misclassification cost and the number of attributes in the spam filtering process, the multi-objective evolutionary feature selection process is used in this work. The deep neural network and cost-sensitive learners are used to regularize the learning process¹⁷. Poria Piroozm and et al. used the force-based heuristic algorithm for OSN spam detection. The ML and deep learning-based integrated technique is used for spam filtering in OSN. The SVM, Genetic Algorithm (GA), and Gravitational Emulation Local Search Algorithm (GELS) are integrated to filter spam in OSN. This integrated technique selects the highly effective features of the spam filter. The enhanced GA helps to select the feature based on exploration, and GELS helps to improve exploration and local search. To improve the detection accuracy, several levels of modifications are made in the algorithm¹⁸. Xianghan Zheng et al. discussed spam detection on social networks. The dataset was constructed with more than 16 million labelled messages. Afterwards, a manual classification is performed to classify the spam and ham data. Then the user's behaviour and message content are extracted from the social network for applying the SVM algorithm. This technique provided more than 99.9% accuracy compared to other algorithms. The computational complexity of manual processes is very high in this technique¹⁹. Zulfikar Alom et al. proposed using a deep learning model to detect spam on Twitter. Basically, ML algorithms are used for spam detection in most applications. But the ML algorithms do not work well on OSN. Hence, the deep learning algorithm is proposed by the author to filter the spam. The tweet text and user meta-data are

analysed to detect the spam. Compared to basic ML algorithms, the deep learning algorithms provided better results²⁰. Table 1 shows the ML-based spam detections.

3. PROPOSED METHODOLOGY

Improving Social Media Optimization (SMO) and spam data filtering in social networks is crucial for maintaining user engagement, trust, and overall platform health. Collect data from various social media platforms, including posts, comments, likes, shares, and user profiles. Preprocess the data by removing noise, such as irrelevant posts, duplicates, and spam content. This step may involve text normalization, tokenization, and sentiment analysis. Develop user profiles based on demographic information, interests, behaviors, and engagement patterns. Segment users into different groups or clusters based on their characteristics and preferences. This segmentation helps tailor content and recommendations to specific user segments. By implementing this proposed system, social media platforms can enhance user experience, foster community engagement, and mitigate the spread of spam and malicious content, thereby promoting a healthier and more vibrant online ecosystem.

- ✓ Personalized content recommendations based on user preferences and behaviors increase user satisfaction and retention.
- ✓ Effective spam filtering enhances user trust in the platform by ensuring a safer and more trustworthy environment.
- ✓ Social Media Optimization (SMO) techniques improve the visibility and reach of high-quality content by optimizing posting schedules, frequency, and content formats.
- ✓ Positive user experiences and a cleaner content environment contribute to a positive brand image and can attract new users and advertisers.
- ✓ A cleaner and safer online environment fosters community growth and encourages active participation from users.

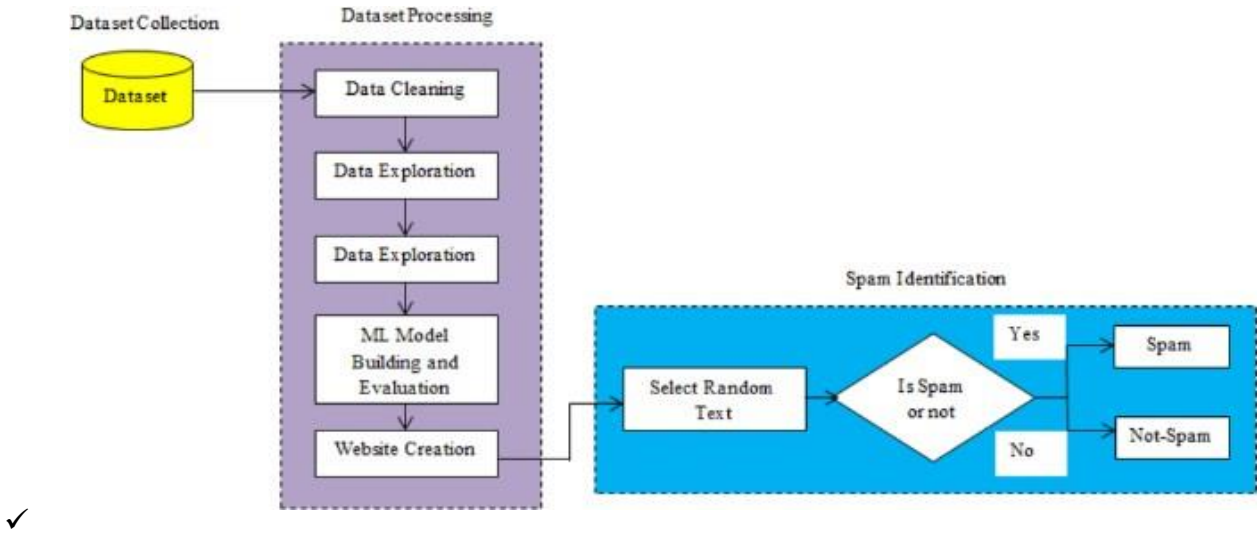
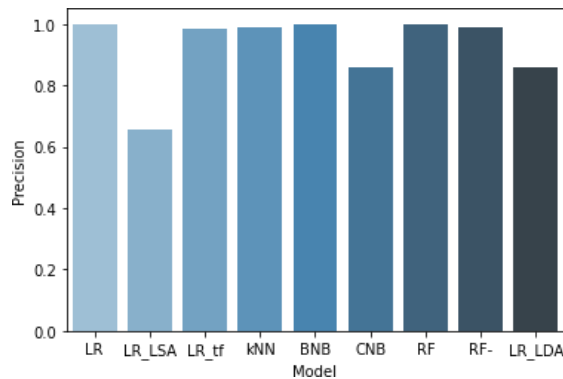


Figure 1: spam detection

4. RESULTS AND DISCUSSIONS

Begin by presenting the performance metrics used to evaluate the classification model. Common metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Describe how well the classification model performed on the test dataset. This could involve stating the accuracy achieved, as well as discussing any trade-offs between precision and recall. If applicable, compare the performance of your classification model with a baseline model or a previous state-of-the-art approach. This helps contextualize the effectiveness of your model.



5. CONCLUSION

In this study, we reviewed machine learning approaches and their application to the field of spam filtering. A review of the state-of-the-art algorithms applied for the classification of messages as either spam or ham is provided. The attempts made by different researchers to solve the problem of spam through the use of machine learning classifiers were discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filters and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of the machine learning algorithms in efficiently handling the menace of spam were pointed out and comparative studies of the machine learning techniques available in literature were done. We also revealed some open research problems associated with spam filters. In general, the figure and volume of literature we reviewed show that significant progress has been made and will still be made in this field. Having discussed the open problems in spam filtering, further research to enhance the effectiveness of spam filters needs to be done. This will make the development of spam filters continue to be an active research field for academicians and [industry](#) practitioners researching machine learning techniques for effective spam filtering. We hope that research students will use this paper as a springboard for doing [qualitative research](#) in spam filtering using machine learning, deep learning, and deep adversarial learning algorithms.

6. REFERENCES

- [1] I.M. Awad, M. Foqaha Email spam classification using a hybrid approach of RBF neural network and particle swarm optimization Int. J. Netw. Secure. Appl., 8 (4) (2016)
- [2] D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira, M. Chaves Measuring characterizing, and avoiding spam traffic costs IEEE Int. Comp., 99 (2016)
- [3] Visited on May 15, 2017 Kaspersky Lab Spam Report (2017) 2012 https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012
- [4] E.M. Bahgat, S. Rady, W. Gad An e-mail filtering approach using classification techniques The 1st International Conference on Advanced Intelligent System and Informatics (AIS2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt (2016), pp. 321-331
- [5] N. Bouguila, O. Amayri A discrete mixture-based kernel for SVMs: application to spam and image categorization Inf. Process. Manag., 45 (6) (2009), pp. 631-642
- [6] Y. Cao, X. Liao, Y. Li An e-mail filtering approach using neural network International Symposium on Neural Networks, Springer Berlin Heidelberg (2004), pp. 688-694
- [7] C.P. Lueg From spam filtering to information retrieval and back: seeking conceptual foundations for spam filtering Proc. Assoc. Inf. Sci. Technol., 42 (1) (2005)
- [8] S. Mason New Law Designed to Limit Amount of Spam in E-Mail (2003) <http://www.wral.com/technolog>
- [9] I. Stuart, S.H. Cha, C. Tappert A neural network classifier for junk e-mail Document Analysis Systems VI, Springer Berlin Heidelberg (2004), pp. 442-450
- [10] J. Han, M. Kamber, J. Pei Data Mining: Concepts and Techniques Elsevier (2011)

INTEGRATED QUERY REORGANIZATION PROCESS FOR EFFICIENT BOOLEAN INFORMATION RETRIEVAL

T. SHAMILA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

In the information filtering paradigm, clients subscribe to a server with continuous queries that express their information needs and get notified every time appropriate information is published. To perform this task in an efficient way, servers employ indexing schemes that support fast matches of the incoming information with the query database. Such indexing schemes involve (i) main-memory trie-based data structures that cluster similar queries by capturing common elements between them and (ii) efficient filtering mechanisms that exploit this clustering to achieve high throughput and low filtering times. However, state-of-the-art indexing schemes are sensitive to the query insertion order and cannot adopt to an evolving query workload, degrading the filtering performance over time. In this project, we present an adaptive trie-based algorithm that outperforms current methods by relying on query statistics to reorganise the query database. Contrary to previous approaches, we show that the nature of the constructed tries, rather than their compactness, is the determining factor for efficient filtering performance. Our algorithm does not depend on the order of insertion of queries in the database, manages to cluster queries even when clustering possibilities are limited, and achieves more than 96 percent filtering time improvement over its state-of-the-art competitors. Finally, we demonstrate that our solution is easily extensible to multi-core machines.

keywords: filtering paradigm, query database, cluster, adaptive trie-based algorithm, multi-core machines

INTRODUCTION

Web search is one of the most prominent Information Retrieval (IR) applications [1]. Typical question-answering scenarios are well supported by ranking highly the documents that not only look relevant by their content, but also receive external support such as by incoming links and anchor text references [2]. In these applications, looking at one or a few of the highest ranked result documents might be sufficient, and if it is, the search process can be stopped. Commercial web search engines are optimized for this scenario and much IR research is focused on improving performance in the top, say 10, results [3].

However, if the objective is to carry out a comprehensive review for a particular topic, search cannot be stopped after finding a few relevant documents [4]. In particular, reviews aim for very broad coverage of a topic, and seek to minimize any bias that might arise as a result of missed or excluded relevant literature. But the typical tensions in IR continue to apply, and if more relevant documents are to be found, more irrelevant documents will also need to be inspected. In the biomedical domain, systematic reviews of the whole corpus of published research literature (the largest collection, MEDLINE, currently indexes more than 17 million

publications) are used to provide medical practitioners with advice to assist their case by case decision-making. To seed the reviews, complex Boolean queries are used on different citation databases to generate a set of documents which are then triaged by multiple assessors. In this domain, it becomes crucial to find as much of the relevant literature as possible for any given level of effort, because each item of overlooked evidence adds to the possibility of suboptimal outcomes in terms of patients' health-care [5].

The interest for information retrieval has existed long before the Internet. The Boolean retrieval is the most simple of these retrieval methods [6] and relies on the use of Boolean operators. The terms in a query are linked together with AND, OR and NOT. This method is often used in search engines on the Internet because it is fast and can therefore be used online. This method has also its problems [7]. The user has to have some knowledge to the search topic for the search to be efficient, e.g., a wrong word in a query could rank a relevant document non relevant. The retrieved documents are all equally ranked with respect to relevance and the number of retrieved documents can only be changed by reformulating the query [8].

The Boolean retrieval has been extended and refined to solve these problems. Expanded term weighting operations make ranking of documents possible, where the terms in the document could be weighted according to their frequency in the document. Boolean information retrieval has been combined with content-based navigation using concept lattices, where shared terms from previously attained documents are used to refine and expand the query [9]. The Boolean operators have been replaced with fuzzy operators. Weighted query expansion using a thesaurus. A model based on fuzzy set theory allows the interpretation of a user query with a linguistic descriptor for each term.

The traditional Boolean retrieval model has been studied intensively in IR research. While it has straightforward semantics, it also has a number of disadvantages, most notably the strictly binary categorization of documents, and the consequent inability to control the result set size except by adding or removing query terms. For example, it is often the case that too many, or too few, or even no documents are returned, and no matter how the query terms are juggled, the "Goldilocks" point might be impossible to attain [10]. In contrast, the broad adoption of ranking principles based on bag-of-word queries, and the resultant ability to order the set of documents according to a heuristic similarity score, means that for general IR applications users can consciously choose how many documents they are willing or able to inspect. Now the drawback is that bag-of-word keyword queries do not offer the same expressive power as Boolean queries do. Although extensions to the Boolean retrieval system have been suggested that produce a ranked output based on Boolean query specifications, they have not been broadly adopted for practical use – perhaps because, to date, simple keyword queries have typically been able to produce similar results, and, for lay users, are easier to generate.

Although ranking has the advantage of identifying a monotonically increasing total number of relevant documents as more documents are inspected, typical IR ranking functions face the difficulty that their ranking is dependent on properties of the whole collection, and can thus be difficult to reproduce, or even understand. Reproducibility helps in assessing review

quality, and is thus often stipulated as a key requirement of comprehensive reviews. But if ranked queries are used, reproducibility can only be assured if all aspects of the computation are reported, including term weights and within-document term frequencies. With Boolean queries, all that is required is publication of the query that was used, together with the date or other identifying version numbers of the collections it was applied to. Moreover, previous work did not show improved retrieval results with ranked keyword queries compared to complex Boolean queries

LITERATURE REVIEW

Extended Boolean models can support document ranking facility for the conventional Boolean retrieval system by calculating the similarities between documents and Boolean queries. An IR system based on extended Boolean models can be defined by the quadruple $\langle T, D, Q, F \rangle$, Where

- T is a set of index terms used to represent queries and documents.
- D is a set of documents. Each document $d \in D$ is represented by $\{(t_1, w_1), \dots, (t_n, w_n)\}$ where w_i designates the weights of term t_i in document d and w_i may take any value between zero and one i.e $0 < w_i < 1$.
- Q is a set of queries that can be recognized by the system. Each query $q \in Q$ is a legitimate Boolean expression composed of index terms and Boolean operators AND, OR and NOT. In some extended Boolean models, queries can be also formulated with term and clause weights.
- F is a ranking function

$$F : D * Q \rightarrow [0,1]$$

Which assign to each pair (d,q) a number in the closed interval [0,1]. This number is a measure of similarity between document d and query q is called the document value for document d with respect to query q. The retrieval function F is defined as follows:

1. For each term t_i in query q the function $F(d, t_i)$ is defined as the weight of term t_i in document d, i.e. w_i .
2. Boolean operators, i.e AND, OR and NOT are then evaluated by applying the corresponding formulas. The evaluation formulas of the operators are an important factor to determine the quality of ranked output. $F(d, NOT t_i)$ is evaluated as $1 - w_i$. For Boolean queries containing more than one Boolean operator, the evaluation proceeds recursively from the innermost clause.

Conventional Retrieval Strategies

In conventional information retrieval, the stored records are normally identified by sets of key words or index terms, and requests for information are expressed by using Boolean combinations of index terms. The retrieval strategy is normally based on an auxiliary inverted-term index that lists the corresponding set of document references for each allowable index term. The Boolean retrieval system is designed to retrieve all stored records exhibiting the precise

combination of key words included in the query: when two query terms are related by an and connective, both terms must be present in order to retrieve a particular stored record; when an or connective is used, at least one of the query terms must be present to retrieve a particular item. In some systems where the natural language text of the documents or the document excerpts is stored, the user queries may be formulated as combinations of text words. In that case, the queries may include location restrictions for the query terms--for example, a requirement that the query terms occur in the same sentence of any retrieved document or within some specified number of words of each other.

Boolean retrieval systems have become popular in operational situations because high standards of performance are achievable. The retrieval technology which is based on list intersections and list unions to implement Boolean conjunction ("A and B") and Boolean disjunction ("A or B"), respectively, is now well understood. The conventional Boolean retrieval technology is however also saddled with various disadvantages:

1. The size of the output obtained in response to a given query is difficult to control; depending on the assignment frequency of the query terms and the actual term combinations used in a query formulation, a great deal of output can be obtained or, alternatively, no output might be retrieved at all.
2. The output obtained in response to a query is not ranked in any order of presumed importance to the user; thus, each retrieved item is assumed to be as important as any other retrieved item.
3. No provisions are made for assigning importance factors or weights to the terms attached either to the documents or

PROPOSED METHODOLOGY

We propose an extended Boolean retrieval (EBR) model for retrieving the top k documents. We present a scoring method for EBR models that decouples document scoring from the inverted list evaluation strategy, allowing free optimization of the latter. The method incurs partial sorting overhead, but, at the same time, reduces the number of query nodes that have to be considered in order to score a document. We adopt ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions. Further, we show that the p-norm EBR model is an instance of such models and that performance gains can be attained that are similar to the ones available when evaluating ranked queries.

Term-independent bounds are proposed, which complement the bounds obtained from max-score. Taken alone, term-independent bounds can be employed in the wand algorithm, also reducing the number of score evaluations. Further, in conjunction with the adaptation of max-score, this novel heuristic is able to short-circuit the scoring of documents.

ADVANTAGES

- ✓ By reorganizing Boolean queries before processing, redundant or irrelevant terms can be eliminated, reducing the search space and improving retrieval efficiency.
- ✓ The reorganization process can prioritize query terms based on their relevance to the information retrieval task.
- ✓ Through query reorganization, the system can mitigate false positives (irrelevant results mistakenly retrieved) and false negatives (relevant results overlooked).
- ✓ The reorganization process can be tailored to specific user preferences, domain requirements, or application contexts.
- ✓ By continuously monitoring and analyzing query performance metrics, the system can dynamically adjust its reorganization techniques to maintain optimal retrieval efficiency and relevance.

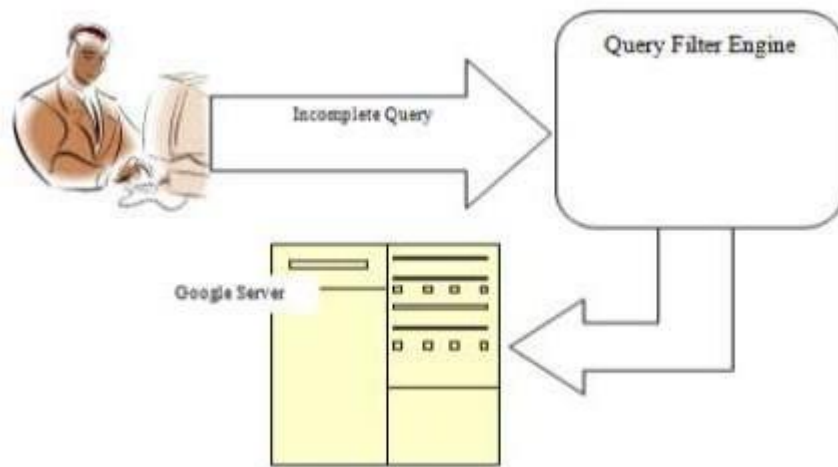
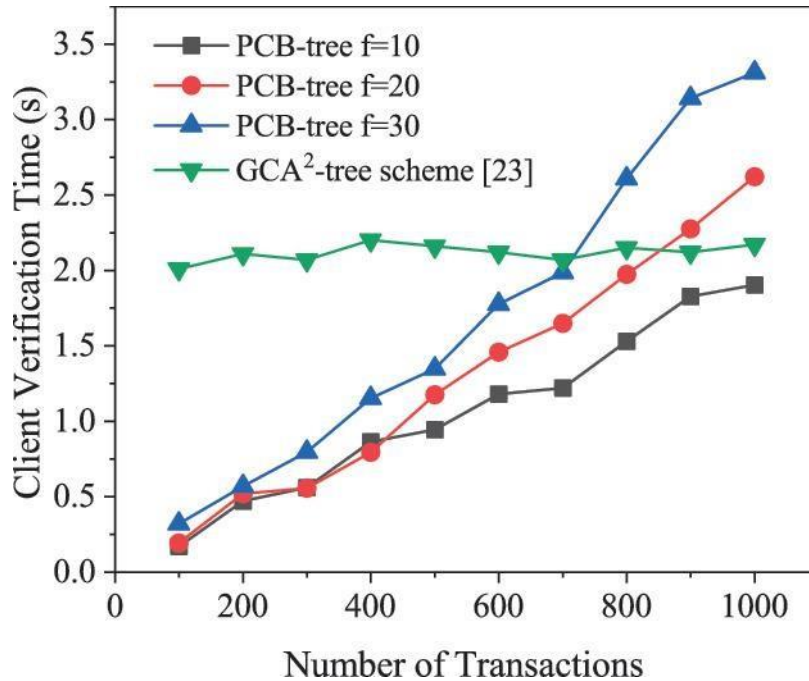


Figure 1: Information Filtering Method

RESULTS AND DISCUSSIONS

It allows clients to query and verify any transactions, which requires the clients to maintain the entire blockchain database locally. This approach is inadvisable because the blockchain database is an append-only ledger and incurs significant maintenance overhead. Very recently, blockchain light client has attracted considerable concerns, which relies on a third party (i.e., a full node) to perform query processing and verification. However, the dishonest full node may return an incorrect and incomplete result of the query requests. Therefore, it remains a challenging issue to achieve secure, efficient, and rich verifiable queries for light clients. In this paper, we propose an efficient verifiable Boolean range query scheme for light clients on the blockchain database. Firstly, we design a new authenticated data structure, polynomial commitment B+-tree (PCB-tree), which efficiently ensures the correctness and completeness of Boolean range queries for blockchain light clients. Secondly, we provide a tunable trade-off between query time and communication overhead by autonomously setting the fanout size of the PCB-tree. Moreover, our scheme can support batch processing to reduce query complexity and

proof size. Finally, security analysis and performance evaluation show that our proposed scheme is secure and practical.



Graph.1 shown the client verification time and number of transaction

CONCLUSION

Having noted that ranked keyword querying is not applicable in complex legal and medical domains because of their need for structured queries including negation, and for repeatable and scrutable outputs, we have presented novel techniques for efficient query evaluation of the p-norm (and similar) extended Boolean retrieval model, and applied them to document-at-a-time evaluation. We showed that optimization techniques developed for ranked keyword retrieval can be modified for EBR, and that they lead to considerable speedups. Further, we proposed term independent bounds as a means to further short-circuit score calculations, and demonstrated that they provide added benefit when complex scoring functions are used.

A number of future directions require investigation. Although presented in the context of document-at-a-time evaluation, it may also be possible to apply variants of our methods to term-at-a-time evaluation. Second, to reduce the number of disk seeks for queries with many terms, it seems desirable to store additional inverted lists for term prefixes instead of expanding queries to hundreds of terms; and this is also an area worth exploration. We also need to determine whether or not term-dependent bounds can be chosen to consistently give rise to further gains. As another possibility, the proposed methods could further be combined and applied only to critical or

complex parts of the query tree. Finally, there might be other ways to handle negations worthy of consideration.

We also plan to evaluate the same implementation approaches in the context of the inference network and wand evaluation models. For example, it may be that for the data we are working with relatively simple choices of term weights—in particular, strictly document-based ones that retain the scrutability property that is so important—can also offer good retrieval effectiveness in these important medical and legal applications.

REFERENCES

- [1] S. Karimi, J. Zobel, S. Pohl, and F. Scholer, "The Challenge of High Recall in Biomedical Systematic Search," Proc. Third Int'l Workshop Data and Text Mining in Bioinformatics, pp. 89-92, Nov. 2009
- [2] J.H. Lee, "Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval," Technical Report TR95-1501, Cornell Univ., 1995.
- [3] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval," Comm. ACM, vol. 26, no. 11, pp. 1022-1036, Nov. 1983.
- [4] J.H. Lee, W.Y. Kin, M.H. Kim, and Y.J. Lee, "On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework," Proc. 16th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 291-297, 1993
- [5] V.N. Anh and A. Moffat, "Pruned Query Evaluation Using Pre-Computed Impacts," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 372-379, 2006.
- [6] Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.2 [updated September 2009], J.P.T. Higgins and S. Green, eds., The Cochrane Collaboration, 2009, <http://www.cochrane-handbook.org>
- [7] L. Zhang, I. Ajiferuke, and M. Sampson, "Optimizing Search Strategies to Identify Randomized Controlled Trials in MEDLINE," BMC Medical Research Methodology, vol. 6, no. 1, p. 23, May 2006.
- [8] M. Sampson, J. McGowan, C. Lefebvre, D. Moher, and J. Grimshaw, "PRESS: Peer Review of Electronic Search Strategies," Technical Report 477, Ottawa: Canadian Agency for Drugs and Technologies in Health, 2008.
- [9] F. McLellan, "1966 and All that—When Is a Literature Search Done?," The Lancet, vol. 358, no. 9282, p. 646, Aug. 2001.
- [10] S. Pohl, J. Zobel, and A. Moffat, "Extended Boolean Retrieval for Systematic Biomedical Reviews," Proc. 33rd Australasian Computer Science Conf. (ACSC '10), vol. 102, Jan. 2010.

MULTISTAGE TRUST MECHANISM IN PRIVACY PRESERVING DATA MINING

A. NIVETHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, we relax this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In our setting, the more trusted a data miner is the less perturbed copy of the data it can access. Under this setting, a malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. Preventing such diversity attacks is the key challenge of providing MLT-PPDM services. We address this challenge by properly correlating perturbation across copies at different trust levels. We prove that our solution is robust against diversity attacks with respect to our privacy goal. That is, for data miners who have access to an arbitrary collection of the perturbed copies, our solution prevent them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. Our solution allows a data owner to generate perturbed copies of its data for arbitrary trust levels on-demand. This feature offers data owners maximum flexibility.

Keywords: aggregated data, Privacy Preserving Data Mining, data miners, Multilevel Trust

1. INTRODUCTION

Data perturbation, a widely employed and accepted Privacy Preserving Data Mining (PPDM) approach, tacitly assumes single-level trust on data miners. This approach introduces uncertainty about individual values before data is published or released to third parties for data mining purposes [1], [2], [3], [4], [5], [6], [7]. Under the single trust level assumption, a data owner generates only one perturbed copy of its data with a fixed amount of uncertainty. This assumption is limited in various applications where a data owner trusts the data miners at different levels.

We present below a two trust level scenario as a motivating example.

The government or a business might do internal (most trusted) data mining, but they may also want to release the data to the public, and might perturb it more. The mining department which receives the less perturbed internal copy also has access to the more perturbed public copy. It

would be desirable that this department does not have *more* power in reconstructing the original data by utilizing both copies than when it has only the internal copy.

This new dimension of *Multi-Level Trust* (MLT) poses new challenges for perturbation based PPDM. In contrast to the single-level trust scenario where only one perturbed copy is released, now multiple differently perturbed copies of the same data is available to data miners at different trusted levels. The more trusted a data miner is, the less perturbed copy it can access; it may also have access to the perturbed copies available at lower trust levels. Moreover, a data miner could access multiple perturbed copies through various other means, e.g., accidental leakage or colluding with others. By utilizing *diversity* across differently perturbed copies, the data miner may be able to produce a more accurate reconstruction of the original data than what is allowed by the data owner. We refer to this attack as a *diversity attack*. It includes the colluding attack scenario where adversaries combine their copies to mount an attack; it also includes the scenario where an adversary utilizes public information to perform the attack on its own. Preventing diversity attacks is the key challenge in solving the MLT-PPDM problem. In this paper, we address this challenge in enabling MLT-PPDM services. In particular, we focus on the additive perturbation approach where random Gaussian noise is added to the original data with *arbitrary* distribution, and provide a systematic solution. Through a one-to-one mapping, our solution allows a data owner to generate distinctly perturbed copies of its data according to different trust levels. Defining trust levels and determining such mappings are beyond the scope of this paper.

2. LITERATURE REVIEW

Privacy Preserving Data Mining (PPDM) was first proposed in and simultaneously. To address this problem, researchers have since proposed various solutions that fall into two broad categories based on the level of privacy protection they provide. The first category of the Secure Multiparty Computation (SMC) approach provides the strongest level of privacy; it enables mutually distrustful entities to mine their collective data without revealing anything except for what can be inferred from an entity's own input and the output of the mining operation alone. In principle, any data mining algorithm can be implemented by using generic algorithms of SMC. However, these algorithms are extraordinarily expensive in practice, and impractical for real use. To avoid the high computational cost, various solutions that are more efficient than generic SMC algorithms have been proposed for specific mining tasks. Solutions to build decision trees over the horizontally partitioned data were proposed in . For vertically partitioned data, algorithms have been proposed to address the association rule mining, k-means clustering, and frequent pattern mining problems. The work of uses a secure coprocessor for privacy preserving collaborative data mining and analysis. The second category of the partial information hiding approach trades privacy with improved performance in the sense that malicious data miners may infer certain properties of the original data from the disguised data. Various solutions in this category allow a data owner to transform its data in different ways to hide the true values of the original data while at the same time still permit useful mining

operations over the modified data. This approach can be further divided into three categories: (a) k-anonymity (b) retention replacement (which retains an element with probability p or replaces it with an element selected from a probability distribution function on the domain of the elements) and (c) data perturbation (which introduces uncertainty about individual values before data is published). The data perturbation approach includes two main classes of methods: additive and matrix multiplicative schemes. These methods apply mainly to continuous data. In this paper, we focus solely on the additive perturbation approach where noise is added to data values. Another relevant line of research concerns the problem of privately computing various set related operations. Two party protocols for intersection, intersection size, equijoin, and equijoin size were introduced in for honest-but-curious adversarial model. Some of the proposed protocols leak information. Similar protocols for set intersection have been proposed in. Efficient two party protocols for the private matching problem which are both secure in the malicious and honest-but-curious models were introduced in. Efficient private and threshold set intersection protocols were proposed in while most of these protocols are equality based, algorithms in compute arbitrary join predicates leveraging the power of a secure coprocessor. Tiny trusted devices were used for secure function evaluation in Our work does not re-anonymizing a dataset after it is updated with insertions and/or deletions, which is a topic studied by the authors in. Instead, we study anonymizing the same dataset at multiple trust levels. The two problems are orthogonal. An earlier version of this paper appeared in and initiated the topic of MLT-PPDM. Recently, Xiao et al.

3. PROPOSED METHODOLOGY

1. Requirement Analysis:

- Understand the data mining objectives and the sensitivity of the data involved.
- Identify stakeholders and their trust requirements.
- Define privacy and utility metrics to evaluate the effectiveness of the trust mechanism.

2. Threat Analysis:

- Identify potential threats to privacy in the data mining process, such as data leakage or inference attacks.
- Assess the likelihood and impact of these threats on the privacy of the data.

3. Trust Model Design:

6. Design a multistage trust model that incorporates trust assessment at different stages of the data mining process.
7. Define trust factors such as data source reputation, data anonymization techniques, and algorithm reliability.
8. Specify trust evaluation methods, including reputation systems, cryptographic proofs, or auditing mechanisms.

4. Privacy-Preserving Techniques Selection:

- ❖ Choose appropriate privacy-preserving techniques based on the trust model and data mining requirements.

- ❖ Consider techniques like differential privacy, secure multiparty computation, or homomorphic encryption.
- ❖ Evaluate the trade-offs between privacy guarantees and data utility.

5. Trust Evaluation Mechanisms Implementation:

- Implement trust evaluation mechanisms according to the designed trust model.
- Develop algorithms or protocols for assessing the trustworthiness of data sources, data transformations, and analysis results.
- Integrate trust evaluation components into the data mining workflow.

4. RESULTS AND DISCUSSIONS

Assess the effectiveness of the trust mechanism in preserving the privacy of sensitive data throughout the data mining process. Measure the level of privacy achieved using quantitative metrics such as differential privacy guarantees, information entropy, or privacy loss. Discuss the impact of privacy-preserving techniques on data utility and the trade-offs between privacy and utility. Evaluate the trustworthiness of data sources, data transformations, and analysis results based on the implemented trust model. Analyze the performance of trust evaluation mechanisms in identifying and mitigating potential privacy breaches or data integrity issues. Discuss any challenges or limitations encountered in assessing trust at different stages of the data mining process. Assess the utility of the data mining results obtained while preserving privacy. Evaluate the quality, accuracy, and relevance of the insights extracted from the privacy-preserving data. Discuss the impact of privacy-preserving techniques on the interpretability and usefulness of the mining results. Analyze the computational overhead, latency, and resource utilization of the privacy-preserving data mining system. Evaluate the scalability and efficiency of the implemented trust mechanism and privacy-preserving algorithms. Discuss opportunities for optimization and performance enhancement to improve system efficiency. Compare the performance and effectiveness of the multistage trust mechanism with baseline approaches that do not incorporate trust evaluation.

Discuss unresolved challenges or limitations that need to be addressed, such as improving scalability, enhancing trust evaluation accuracy, or handling complex data types.

5. CONCLUSION

In this work, we expand the scope of additive perturbation based PPDM to multi-level trust (MLT), by relaxing an implicit assumption of single-level trust evaluation. Last but not the least, our solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand. This property offers the data owner maximum flexibility. We believe that multi-level trust privacy preserving data mining can find many applications. Our work takes the initial step to enable MLT-PPDM services. Many interesting and important directions are worth exploring. For example, it is not clear how to expand the scope of other approaches in the area of partial information hiding, such as random rotation based data perturbation, k-anonymity, and retention replacement, to multi-level trust. It is also of great interest to extend our approach to handle evolving data streams. As with most existing work on perturbation based PPDM, our

work is limited in the sense that it considers only linear attacks. More powerful adversaries may apply nonlinear techniques to derive original data and recover more information. Studying the MLT-PPDM problem under this adversarial model is an interesting future direction.

6. REFERENCES

- [1] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Proc. of the 20th ACM Symposium on Principles of Database Systems, Santa Barbara, California, May 2001, pp. 247–255.
- [2] R. Agrawal and R. Srikant, "Privacy preserving data mining," in Proc. ACM SIGMOD Int'l Conf. on Management of Data, 2000.
- [3] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in Proc. Int'l Conf. on Data Mining, 2005.
- [4] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in Proc. ACM SIGMOD Int'l Conf. on Management of Data, 2005.
- [5] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, and I. Stanoi, "Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking," in Proc. Int'l Conf. on Data Engineering, 2007.
- [6] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. on Knowledge and Data Engineering*, vol. 18, pp. 92–106, 2006.
- [7] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time series compressibility and privacy," in Proc. Int'l Conf. on Very Large Data Bases, 2007.
- [8] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Proc. Int'l Cryptology Conference (CRYPTO), 2000.
- [9] J. Vaidya and C. W. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2002.
- [10] O. Goldreich, "Secure multi-party computation," Final (incomplete) draft, version 1.4, 2002.
- [11] J. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," in Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2003.
- [12] A. W.-C. Fu, R. C.-W. Wong, and K. Wang, "Privacy-preserving frequent pattern mining across private databases," in Proc. Int'l Conf. on Data Mining, 2005.
- [13] B. Bhattacharjee, N. Abe, K. Goldman, B. Zadrozny, V. R. Chillakuru, M. del Carpio, and C. Apte, "Using secure co-processors for privacy preserving collaborative data mining and analysis," in Proc. of the 2nd International Workshop on Data management on new hardware, 2006.
- [14] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy preserving data mining," in Proc. Int'l Conf. on Extending Database Technology (EDBT), 2004.

- [15] E. Bertino, B. C. Ooi, Y. Yang, and R. H. Deng, "Privacy and ownership preserving of outsourced medical data," in Proc. Int'l Conf. on Data Engineering, 2005.
- [16] D. Kifer and J. E. Gehrke, "Injecting utility into anonymized datasets," in Proc. ACM SIGMOD Int'l Conf. on Management of Data, 2006.
- [17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in Proc. Int'l Conf. on Data Engineering, 2006.
- [18] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems (IJUFKS), vol. 10, 2002.
- [19] X. Xiao and Y. Tao, "Personalized privacy preservation," in Proc. ACM SIGMOD Int'l Conf. on Management of Data, 2006.
- [20] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in Proc. ACM SIGMOD Int'l Conf. on Management of Data, 2005.

NUMERICAL OPTIMIZATION ON CENTROID BASED ACTIONABLE 3D SUBSPACE CLUSTERING

R. THILAGAVATHI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Actionable 3D subspace clustering from real-world continuous-valued 3D (i.e. *object-attribute-context*) data promises tangible benefits such as discovery of biologically significant protein residues and profitable stocks, but existing algorithms are inadequate in solving this clustering problem; most of them are not actionable (ability to suggest profitable or beneficial actions to users), do not allow incorporation of domain knowledge and are parameter sensitive, i.e. the wrong threshold setting reduces the cluster quality. Moreover, its 3D structure complicates this clustering problem. We propose a centroid-based actionable 3D subspace clustering framework, named CATSeeker, which allows incorporation of domain knowledge, and achieves parameter insensitivity and excellent performance through a unique combination of singular value decomposition, numerical optimization and 3D frequent itemset mining.

Keywords: subspace, data promises, cluster, itemset, domain, optimization, decomposition

1. INTRODUCTION

In many application domains (e.g., medicine or biology), comprehensive questions for comparative schemas resulting from collaborative initiatives are made available. This proposed system argues that to achieve the on-demand combination for superlative and special cases based resource management for Web-based e-learning, one should go beyond using domain ontology's statically[1]. So the propose XAML based matching process involves semantic mapping has done on both the open dataset and closed dataset mechanism to integrate e-learning databases by using ontology semantics.

It defines context-specific portions from the whole ontology as optimized data and an XAML based resource reuse approach by using an evolution algorithm. Such well established schemas are often associated [2]with reliable data that have been carefully collected, cleansed, and verified, thus providing reference for comparable question based data management systems (DMSs) in different application domains. A good practice is therefore to build on the efforts made to design reference DMSs whenever we have to develop our own DMS with specificneeds.

A way to do this is to extract from the reference DMS the piece of schema relevant to our application needs, possibly to personalize it with extra constraints w.r.t. our application under construction, and then to manage our own data set using the resulting schema. Recent work in description logics provides different solutions to achieve such a reuse of a reference ontology-based DMS. Indeed, modern ontological languages [3] like the W3C recommendations RDFS, OWL, and OWL2 are actually XML based syntactic variants of well known DLs. All those solutions consist in extracting a module from an existing ontological schema such that all the

constraints concerning the relations of interest for the application [4] under construction are captured in the module.

Existing definitions of modules in the literature basically in this system, revisit the reuse of reference ontology based DMS [5] in order to build a new DMS with specific needs. It goes one step further by not only considering the design of a module based DMS (i.e., how to extract a module from an ontological schema) also study how a module based DMS can benefit from the reference DMS to enhance its own data management skills. Our contribution is to introduce and study novel properties of robustness[6] for modules that provide means for checking easily that a robust module based comparative evolves safely w.r.t. both the schema and the data of the reference DMS.

From a module robust to consistency checking, for any data update in corresponding module-based comparative questions[7], we show how to query the reference questions for checking whether the local update does not bring any inconsistency with the data and the constraints of the references. It is from a module robust to query answering, for any query asked to module-based DMS, It shows how to query the reference DMS for obtaining additional answers by also exploiting the data stored in the reference DMS.

2. LITERATURE REVIEW

Conventional Data Management Strategies

In conventional information management principles, the stored records are normally identified by sets of key words or index terms, and requests for information are expressed by using Boolean combinations of index terms. The retrieval strategy is normally based on an auxiliary inverted-term index that lists the corresponding set of document references for each allowable index term. The Boolean retrieval system is designed to retrieve all stored records exhibiting the precise combination of key words included in the query: when two query terms are related by an and connective, both terms must be present in order to retrieve a particular stored record; when an or connective is used, at least one of the query terms must be present to retrieve a particular item.

In some systems where the natural language text of the documents or the document excerpts is stored, the user queries may be formulated as combinations of text words. In that case, the queries may include location restrictions for the query terms- for example, a requirement that the query terms occur in the same sentence of any retrieved document or within some specified number of words of each other.

Boolean data management systems have become popular in operational situations because high standards of performance are achievable. The retrieval technology which is based on list intersections and list unions to implement Boolean conjunction ("A and B") and Boolean

disjunction ("A or B"), respectively, is now well understood. The conventional Boolean retrieval technology is however also saddled with various disadvantages:

1. The size of the output obtained in response to a given query is difficult to control; depending on the assignment frequency of the query terms and the actual term combinations used in a query formulation, a great deal of output can be obtained or, alternatively, no output might be retrieved at all.

2. The output obtained in response to a query is not ranked in any order of presumed importance to the user; thus, each retrieved item is assumed to be as important as any other retrieved item.

3. No provisions are made for assigning importance factors or weights to the terms attached either to the documents or dataset.

Data Evaluation

Lee gives an overview of the different models that have been proposed, and shows that only the p-norm model has two key properties that, if not present, are detrimental to retrieval effectiveness.

Smith proposed to recursively aggregate inverted lists, calculating and storing intermediate scores for every document that is encountered in any of the lists, in what is referred to as being the term-at-a-time approach. In effect, not all nodes in the query tree are visited for every document, but all of the inverted lists are fully inspected, and temporary memory proportional to the total size of the relevant inverted lists is required. Smith's Infinity-One method gives an approximation to the p-norm model, with the aim of reducing computational cost by reducing the volume of floating point operations. As is demonstrated below, the number of score calculations can be greatly reduced via an exact lossless pruning approach.

Turtle and Flood describe the max-score ranking mechanism, to accelerate keyword query evaluation when sum-score aggregation functions are used and only the top-k documents are required. Using document-at-a-time evaluation, the algorithm commences by fully scoring the first k documents in the OR-set of the query terms. Thereafter, the kth largest document score is tracked, as an entry threshold that candidate documents must exceed before they can enter the (partial) ranking. The max-score algorithm uses the information conveyed by the entry threshold to reduce two cost factors: 1) the number of candidate documents that are scored; and 2) the cost associated with scoring each candidate document.

Content-based Methods

In content-based recommendation methods, the utility $u(c, s)$ of item s for user c is estimated based on the utilities $u(c, si)$ assigned by user c to items $si \in S$ that are "similar" to item s . For example, in a movie recommendation application, in order to recommend movies to user c , the content-based recommender system tries to understand the commonalities among the movies user c has rated highly in the past (specific actors, directors, genres, subject matter, etc.). Then, only the movies that have a high degree of similarity to whatever user's preferences are would

get recommended. The content-based approach to recommendation has its roots in information retrieval and information filtering research.

Because of the significant and early advancements made by the information retrieval and filtering communities and because of the importance of several text-based applications, many current content-based systems focus on recommending items containing textual information, such as documents, Web sites (URLs), and Usenet news messages. The improvement over the traditional information retrieval approaches comes from the use of user *profiles* that contain information about users' tastes, preferences, and needs. The profiling information can be elicited from users explicitly, e.g., through questionnaires, or implicitly – learned from their transactional behavior over time.

More formally, let $Content(s)$ be an *item profile*, i.e., a set of attributes characterizing item s . It is usually computed by extracting a set of features from item s (its content) and is used to determine appropriateness of the item for recommendation purposes. Since, as mentioned earlier, content-based systems are designed mostly to recommend text-based items, the content in these systems is usually described with *keywords*.

For example, a content-based component of the Fab system which recommends Web pages to users, represents Web page content with the 100 most important words. Similarly, the Syskill & Webert system represents documents with the 128 most informative words. The "importance" (or "informativeness") of word k_i in document d_j is determined with some *weighting* measure w_{ij} that can be defined in several different ways.

One of the best-known measures for specifying keyword weights in Information Retrieval is the *term frequency/inverse document frequency (TF-IDF)* measure that is defined as follows. Assume that N is the total number of documents that can be recommended to users and that keyword k_i appears in n_i of them. Moreover, assume that $f_{i,j}$ is the number of times keyword k_i appears in document d_j . Then $f_{i,j}$, the term frequency (or normalized frequency) of keyword k_i in document d_j , is defined as

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

where the maximum is computed over the frequencies $f_{z,j}$ of all keywords k_z that appear in the document d_j . However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a non-relevant one. Therefore, the measure of inverse document frequency (IDF_i) is often used in combination with simple term frequency ($f_{i,j}$). The inverse document frequency for keyword k_i is usually defined as

$$IDF_i = \log \frac{N}{n_i}$$

Then the TF-IDF weight for keyword k_i in document d_j is defined as

$$w_{i,j} = TF_{i,j} \times IDF_i$$

In content-based systems, the utility function $u(c, s)$ is usually defined as:

$u(c, s) = score(ContentBasedProfile(c), Content(s))$ Using the above-mentioned information retrieval-based paradigm of recommending Web pages, Web site URLs, or Usenet news messages, both $ContentBasedProfile(c)$ of user c and $Content(s)$ of document s can be represented as TF-IDF vectors c_w and s_w of keyword weights. Moreover, utility function $u(c, s)$ is usually represented in information retrieval literature by some scoring heuristic defined in terms of vectors c_w and s_w , such as cosine similarity measure

$$u(c, s) = \cos(\bar{w}_c, \bar{w}_s) = \frac{\bar{w}_c \cdot \bar{w}_s}{\|\bar{w}_c\|_2 \times \|\bar{w}_s\|_2} = \frac{\sum_{i=1}^K w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^K w_{i,c}^2} \sqrt{\sum_{i=1}^K w_{i,s}^2}}$$

Limited content analysis.

Content-based techniques are limited by the features that are explicitly associated with the objects that these systems recommend. Therefore, in order to have a sufficient set of features, the content must either be in a form that can be parsed automatically by a computer (e.g., text), or the features should be assigned to items manually.

While information retrieval techniques work well in extracting features from text documents, some other domains have an inherent problem with automatic feature extraction. For example, automatic feature extraction methods are much harder to apply to the multimedia data, e.g., graphical images, audio and video streams. Moreover, it is often not practical to assign attributes by hand due to limitations of resource Another problem with limited content analysis is that, if two different items are represented by the same set of features, they are indistinguishable.

Therefore, since text-based documents are usually represented by their most important keywords, content-based systems cannot distinguish between a well-written article and a badly written one, if they happen to use the same terms.

Over-specialization. When the system can *only* recommend items that score highly against a user’s profile, the user is limited to being recommended items similar to those already rated. For example, a person with no experience with Greek cuisine would never receive a recommendation for even the greatest Greek restaurant in town. This problem, which has also been studied in other domains, is often addressed by introducing some randomness. For example, the use of genetic algorithms has been proposed as a possible solution in the context of information filtering . In addition, the problem with over-specialization is not only that the content-based systems cannot recommend items that are different from anything the user has seen before. In certain cases, items should not be recommended if they are *too similar* to

something the user has already seen, such as different news article describing the same event. Therefore, some content based recommender systems, such as DailyLearner , filter out items not only if they are too different from user's preferences, but also if they are too similar to something the user has seen before.

Furthermore, provide a set of five redundancy measures to evaluate whether a document that is deemed to be relevant contains some novel information as well. In summary, the *diversity* of recommendations is often a desirable feature in recommender systems. Ideally, the user should be presented with a *range* of options and not with a homogeneous set of alternatives. For example, it is not necessarily a good idea to recommend all movies by Woody Allen to a user who liked one of them.

3. PROPOSED METHODOLOGY

This proposed system Hybrid extraction of robust model (HERM) is very efficient and reuse of e-learning resources in a distributed environment like the Web for better result. This proposed system argues that to achieve the on-demand semantic-based resource management for Web-based e-learning, one should go beyond using domain ontology's statically. So the propose XAML based matching process involves semantic mapping has done on both the open dataset and closed dataset mechanism to integrate e-learning databases by using ontology semantics. It defines context-specific portions from the whole ontology as optimized data and proposes a XAML based resource reuse approach by using an evolution algorithm.

It explains the context aware based evolution algorithm for dynamic e-learning resource reuse in detail. This system is going to conduct a simulation experiment and evaluate the proposed approach with a xaml e-learning scenario. The proposed approach for matching processin web cluster databases from different database servers can be easily integrated and deliver highly dimensional e-learning resource management and reuse is far from being mature. However, e-learning is also a widely open research area, and there is still much room for improvement on the method. This research mechanism includes 1) improving the proposed evolution approach by making use of and comparing different evolutionary algorithms, 2) applying the proposed approach to support more applications, and 3) extending to the situation with multiple e-learning systems or services.

ADVANTAGES

- ✓ In this system concentrate on advanced data matching and extraction is not limited to number of database servers.
- ✓ This hybrid data extraction gives better performance than the existing models because this system uses information content that declared inside the OWL,RDF,xml and XAML languages.

- ✓ This system differs highly from existing systems by dealing the with the object identification in the open datasets and closed datasets.
- ✓ In this system the conventional ontology based matching are optimized in both dependent data with different subsets and independent data.
- ✓ This XAML based system is uses the integrated approach to match and extract the information from the huge database systems from different number of clusters.
- ✓ This proposed system has been designed to perform the matching in both open dataset and closed dataset.

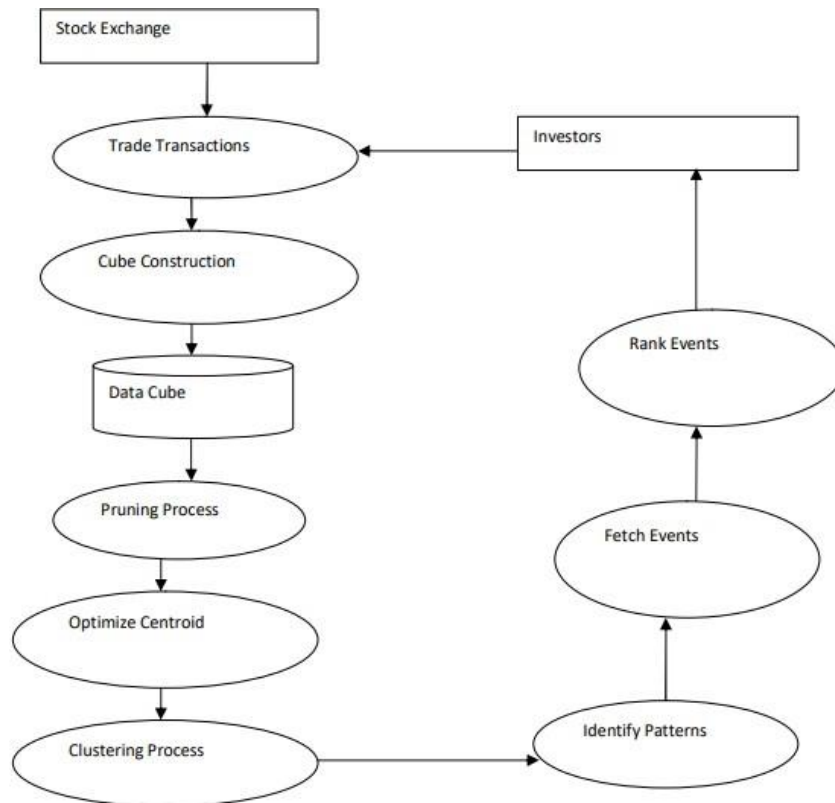


Figure 1: Multi Dimensional Clustering with Optimal Centroids

4. RESULTS AND DISCUSSIONS

Many clustering requirements are based on small- to medium-sized datasets. The centroid-based clustering algorithms do an excellent clustering job in such situations. These algorithms find the best value for centroids and group all nearby objects into their respective clusters. These are unsupervised learning algorithms and thus you do not have to worry about creating labeled datasets. The optimization is NP-hard and usually provides only approximate solutions. Though you need to provide an estimate of the number of clusters in prior, these

algorithms become an excellent starting point for solving complex clustering problems. In this chapter, I discuss two widely used clustering algorithms in this category—K-means and K-medoids, the latter one is robust to outliers. I also provide you with the techniques for estimating the number of clusters in your dataset before applying the algorithm on a full dataset.

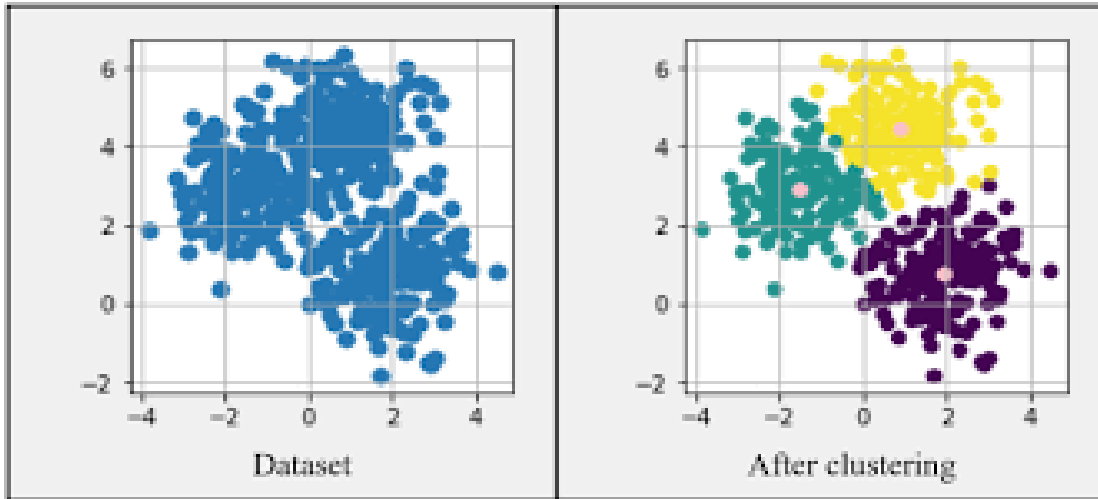


Fig.1 shown the dataset before clustering and after clustering

5. CONCLUSION

The proposed technique HERM has a novel weakly supervised method to identify comparative questions and extract comparator pairs simultaneously. It rely on the key insight that a good comparative question identification pattern should extract good comparators, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. The experimental results show that our method is effective in both comparative question identification and comparator extraction. It significantly improves recall in both tasks while maintains high precision. It shows the examples show that these comparator pairs reflect what users are really interested in comparing. Our comparator mining results can be used for a commerce search or product recommendation system. For example, automatic suggestion of comparable entities can assist users in their comparison activities before making their purchase decisions. Also, our results can provide useful information to companies which want to identify their competitors.

6. REFERENCE

- [1] .Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In *Proceedings of AAAI'99 /IAAI'99*.
- [2]Claire Cardie. 1997. Empirical methods in information extraction. *AI magazine*, 18:65–79. Dan Gusfield. 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, New York, NY, USA
- [3]Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of WWW '02*, pages 517–526.
- [4]Glen Jeh and Jennifer Widom. 2003. Scaling personalizedweb search. In *Proceedings of WWW '03*, pages 271–279.
- [5]Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of SIGIR '06*, pages 244–251. Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In *Proceedings of AAAI '06*.
- [6]Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with
- [7]hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056.

OPTIMIZATION OF RESPONSE TIME OF M-LEARNING COMPUTING ENVIRONMENT USING FIRE FREE APPROACH

V. DIVYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Mobile learning (m-learning) is a relatively new technology that helps students learn and gain knowledge using the Internet and Cloud computing technologies. Cloud computing is one of the recent advancements in the computing field that makes Internet access easy to end users. Many Cloud services rely on Cloud users for mapping Cloud software using virtualization techniques. Usually, the Cloud users' requests from various terminals will cause heavy traffic or unbalanced loads at the Cloud data centers and associated Cloud servers. Thus, a Cloud load balancer that uses an efficient load balancing technique is needed in all the cloud servers. This project proposes a new meta-heuristic algorithm, named the dominant fire fly algorithm, which optimizes load balancing of tasks among the multiple virtual machines in the Cloud server, thereby improving the response efficiency of Cloud servers that concomitantly enhances the accuracy of m-learning systems. The methods and findings used to solve load imbalance issues in Cloud servers, which will enhance the experiences of m-learning users. Specifically, the findings such as Cloud-Structured Query Language (SQL), querying mechanism in mobile devices will ensure users receive their m-learning content without delay; additionally, our method will demonstrate that by applying an effective load balancing technique would improve the throughput and the response time in mobile and cloud environments. This project proposes a resource provisioning and scheduling strategy for scientific workflows on Infrastructure as a Service (IaaS) and Platform as services clouds (PaaS). This project presents an algorithm based on the Superior Element Multitude Optimization (SEMO), which aims to minimize the overall workflow execution cost while meeting deadline constraints. The main scope of the project is used to analyze best available resource in the cloud environment depend upon the total execution time and total execution cost which is compare between one process to another process. If the provider satisfies the time least time, then the process becomes to termination.

Keywords: Cloud computing, Query Language, SEMO, m-learning, meta-heuristic algorithm

1. INTRODUCTION

Many computing methodologies are available in the computing field for maximizing automation. Among those, m-learning and Cloud computing are considered to be the best service oriented computing technologies to automate tasks in virtual machines as well as to enable users to access information very efficiently[1]. Also, m-learning offers cost-effective solutions for a wide range of services. Mobile learning and Cloud computing are two essential domains to explain distributed data sharing. In m-learning, mobile devices used by end users are called the m-learning clients[3]. Through internet connectivity, m-learning clients store and

retrieve data from Cloud data centers. Hence, m-learning systems integrated with Cloud data centers are quite advantageous for transferring all types of data and applications to mobile device easily and accurately[4]. However, load balancing issues in Cloud data centers should be addressed to improve performance and efficiency. In this paper, we propose a meta-heuristic algorithm to overcome this load balancing issue. M-learning technologies have been deployed in many m-learning systems and applications to improve the learning styles of current students. On average, m-learning technologies enhance the learning capacity of individuals by 70%. Some of the Cloud computing services that could be used for m-learning approaches are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), Infrastructure as-a-Service (IaaS), and Hardware-as-a-Service (HaaS)[5,6].

Load balancing techniques are used to distribute incoming traffic across multiple servers to minimize the delay of the Cloud server response to the Cloud users. Cloud load balancing is considered adequate only if the throughput in the Cloud server is high, delays are minimal, and jitter is minimal while addressing Cloud user requests. Sometimes, failure of load balancing in the Cloud leads to poor image resolution and poor video streaming for users[7]. Thus, load balancing in Cloud servers is essential to maximize throughput and to achieve superior performance in both public and private Clouds.

At any institutions, students and faculty members now a day could not survive without their mobile which makes more internet users as mobile users. Mobile devices increase users' communication capability. Similarly as with most new advancements, portable use contrasts geographically[8]. The mobile internet technology was first adopted by Japan. And because of the technological adoption, society advancement on educational institutions through mobile learning (m-learning) was introduced to students and is use by learners with the usage of mobile gadgets where learning aids and materials is conveniently accessible from virtually anywhere. Student portal is well defined as entry point of a school, university or college that offers a central source of data and service for students, teachers, administrators and parents[10]. The word portal is identified as link page which presents data from differing sources unifiedly. It may include services that give standard search engine feature, entertainment, news, email, database, information and. Portal is a technique for creativities to provide a reliable look and feel with access control and measures for multiple applications and databases which otherwise would have been different entities altogether. Student portals contain information such as viewing of grades, announcements, events, and notification for the releasing of grades and important contact numbers. Also they may offer links to useful web resources.

2. LITERATURE REVIEW

2.1 EFFICIENT FAIR QUEUING USING DEFICIT ROUND ROBIN

When there is contention for resources, it is important for resources to be allocated or scheduled fairly. We need firewall between contenting users, so that the "fair" allocation is followed strictly. For example, in an operating system, CPU scheduling of user processes controls the use of CPU resources by processes, and insulates well-behaved users from ill-behaved users. Unfortunately, in most computer networks there are no such firewalls; most networks are susceptible to badly-behaving sources. A rogue source that sends at an uncontrolled rate can seize a large fraction of the buffers at an intermediate router; this can result in dropped packets for other sources sending at more moderate rates! A solution to this problem is needed to isolate the effects of bad behavior to users that are behaving badly. An isolation mechanism called Fair Queuing [DKS89] has been proposed, and has been proved [GM90] to have nearly perfect isolation and fairness. Unfortunately, Fair Queuing (FQ) appears to be expensive to implement. Specifically, FQ requires $O(\log(n))$ work per packet to implement fair queuing, where n is the number of packet streams that are concurrently active at the gateway or router. With a large number of active packet streams, FQ is hard to implement at high speeds. Some attempts have been made to improve the efficiency of FQ; however such attempts either do not avoid the $O(\log(n))$ bottleneck or are unfair.

2.2 ADAPTIVE SCHEDULING OF CLOUD TASKS USING ANT COLONY OPTIMIZATION

The allocation of virtual resources (CPU, main memory, disc, bandwidth, etc.) play a significant role in reducing the energy consumption, optimizing makespan, reducing waiting time and increasing throughput of the system. In cloud platforms, scheduling takes place at two levels. First, when the tasks are uploaded to the cloud, the scheduler assigns the requested tasks to different virtual machines, attempting to reduce the completion time of multiple applications across virtual machines. Second, allocating the virtual machines to physical machines to balance the load or to reduce energy consumption etc. (for example, Amazon EC2 uses elastic load balancing (ELB) to control how incoming requests are handled). The arrival of tasks (workloads) may cause bursty or non-bursty traffic. Example of bursty traffics are bursty surges and Internet flash crowds, there by aggressively grouped together in small periods there by create spikes with high arrival rate. The presence of burstiness in request arrival rate can cause degradation of the performance of task scheduler. We have used the Expected Time to Compute (ETC) model as task model. This ETC model represents the task heterogeneity as well as machine heterogeneity. The task allocation or scheduling problem is a well-known NP-complete problem. The NP-complete nature of the problem focuses tremendous interest for researchers to propose sub-optimal solutions. Researchers proposed some heuristic approaches and metaheuristic approaches to optimize different objectives. We have proposed a heuristic algorithm based on Ant Colony Optimization (ACO) technique to optimize makespan and average waiting time of the system. The main idea behind this approach is to use feedback mechanism and try to imitate the style of nature ant colonies to search for good by connecting to

each other with the help of pheromone laid on food path traveled. Makespan is defined as the time required to execute a finite set of tasks by the system. A good scheduler of tasks should be able to adapt its strategy for scheduling to the varying environment and type of tasks.

2.3 CONFIGURING LARGE-SCALE STORAGE USING A MIDDLEWARE WITH MACHINE LEARNING

The sophistication and deployment of cloud computing has significantly gathered pace in the recent years. In most cases, instances of cloud services require persistent data in the cloud, and thus there has been a coupled demand for scalable, manageable and reliable storage systems. Even outside the cloud (or in so-called 'private clouds,' that operate within one organization), there is an increasing demand for scalable storage systems, just to meet the needs of storing ever-larger data sets for analysis and data mining. Much of the growth of cloud computing has been on the basis of cloud providers being able to lower the Total Cost of Ownership (TCO) for their cloud clients, as well as the elasticity provided by cloud infrastructure: economies of scale mean that the cloud providers can cope with their clients' dynamic requirements. The emergence and significant expansion of cloud service providers such as Amazon S3 and EC2 have been the result. Within any data centre, it is usual that a highly heterogeneous collection of devices from different vendors forms over time. This is often the result of rolling upgrades, and allows the equipment to be used to provide a number of different graduations of service. However, this heterogeneity makes the overall life cycle from data placement to retirement of workload a highly onerous task. There will be requirements to plan, configure and to perform the migration of data on demand. Owing to the complexity of the interrelated physical and logical systems, these provisioning tasks are often error-prone.

2.4 CAPTURING CLOUD COMPUTING KNOWLEDGE AND EXPERIENCE IN PATTERNS

Through the use of cloud computing, cloud providers and their customers benefit from the fundamental cloud properties: elasticity, pay-per-use, standardization, and resource sharing. Elasticity empowers cloud users to reserve and release cloud resources dynamically and based on the currently experienced workload. Pay-per-use pricing models ensure that only the temporarily used resources are billed to customers. The broad availability of reusable cloud services cause cloud applications, their components, and used middleware to be standardized and homogenized. This significantly reduces the complexity of cloud application runtime environments and, together with the enabled resource sharing between cloud users, makes cloud resources a generally available commodity. For these properties, cloud computing has been recognized as one of the key IT topics for the next years due to (i) its ability to cope with extremely large amounts of users, (ii) its ability to handle vast amounts of data, (iii) its flexibility when these amounts change suddenly, and (iv) its customizability regarding various business requirements. To benefit from this powerful computing environment, application architects and developers need to follow certain architectural and management best practices. The dynamicity of the runtime infrastructure, replication of data, handling of resource failure, tenant isolation, and the bridging of different cloud environments have to be incorporated in the applications architecture

and runtime behavior. Many of these challenges are also faced in non-cloud applications in a similar manner. However, the strong interrelation of cloud applications with the runtime environment provided as-a-Service makes clouds a very complex and diverse environment. Building and executing cloud applications without considering the specifics of this environment may result in suboptimal resource utilization and may even reduce applications' availability.

2.5 LOAD BALANCING IN HETEROGENEOUS P2P SYSTEMS USING MOBILE AGENTS

Load balancing is a active technology that provides the art of shaping, transforming and filtering the network traffic then routing and load balancing it to the optimal node. By adding the concept of load balancer we can distribute the traffic for preventing from failure in any case by having capabilities such as scalability, availability, easy to use, fault tolerant, quick response time. Mobile agent technology offers a new computing paradigm in which an autonomous program can migrate under its own or host control from one node to another in a heterogeneous network. In other words, the program running at a host can suspend its execution at an arbitrary point, transfer itself to another host, or request the host to transfer it to its next destination and resume execution from the point of suspension is called mobile agent MA. MA supports a variety of web based distributed applications namely: systems and distributed information Management and information retrieval. Other areas where MAs are seen as offering potential advantages are Wireless or mobile computing dynamic deployment of code, thin clients or resource limited devices, personal assistants, and MA-based parallel processing. Traditional load balancing approaches are implemented based on message passing paradigm. MA technology provides a new solution to support load balancing in heterogeneous network.

3. PROPOSED METHODOLOGY

In propose system the dominant re y load balancing algorithm to solve load imbalance issues in Cloud servers, to enhance the experiences of m-learning users. Specially, Dominant re y-based required Cloud server mapping algorithm for different VM methods will help ensure users receive their m-learning content without delay; addition ally, in this technique, that demonstrates the load balancing improvement on throughput and the response time of mobile devices.

ADVANTAGES

- ✓ Faster response times enable learners to access educational resources and participate in learning activities more readily, regardless of their location or device.
- ✓ This enhances accessibility for users with diverse learning needs and ensures equitable access to educational opportunities.
- ✓ Dominant firefly approach optimizes resource allocation and task scheduling within the cloud infrastructure, leading to more efficient utilization of computing resources. This helps minimize latency and reduce system overhead, resulting in cost savings and improved scalability for m-learning deployments.
- ✓ The firefly optimization algorithm dynamically adjusts resource allocations and task priorities based on changing workload conditions and user demand.

- ✓ Adaptive behavior ensures that system resources are allocated optimally to meet performance objectives, even during peak usage periods or fluctuating demand.

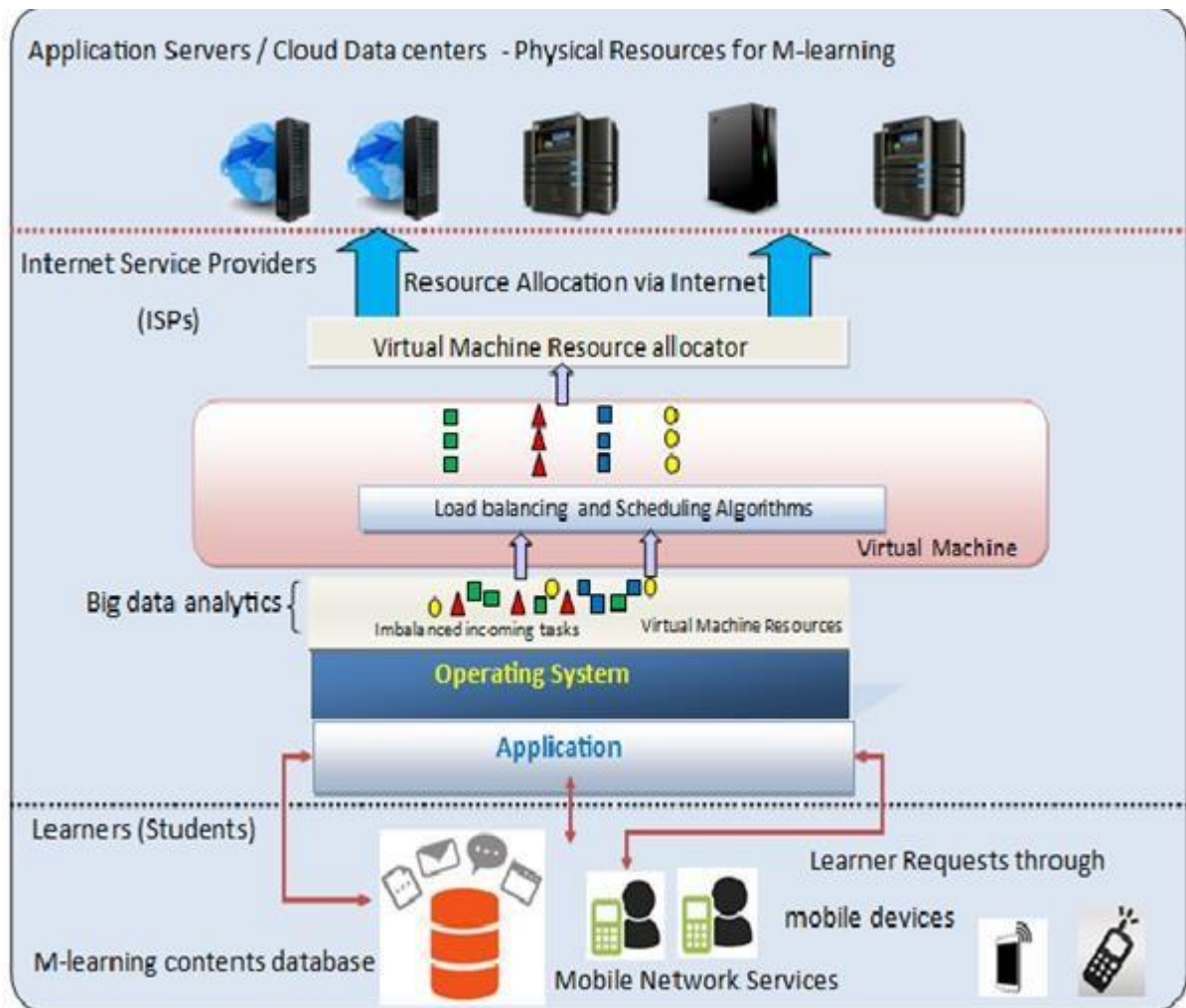


Figure 1: M- Learning with cloud computing Architecture

4. RESULTS AND DISCUSSIONS

The response time of VMs in seconds for the dominant firefly algorithm, HBB-LB, ant colony optimization load balancing algorithm, and WRR algorithms. The X-axis represents the number of tasks and the Y-axis represents response time in seconds. Our proposed algorithm showed the optimal response time. According to the task migration in the Cloud servers VMs, performance various QoS metrics is analyzed. The comparison of number of tasks with the high loaded VMs task migration time and low loaded VMs task migration time. The X-axis represents the number of tasks and the Y-axis represents the migration time in seconds from the proposed dominant firefly algorithm for load balancing.

Also, by comparing different load balancing algorithms, we were able to find the response time of each task in the Cloud server VMs.

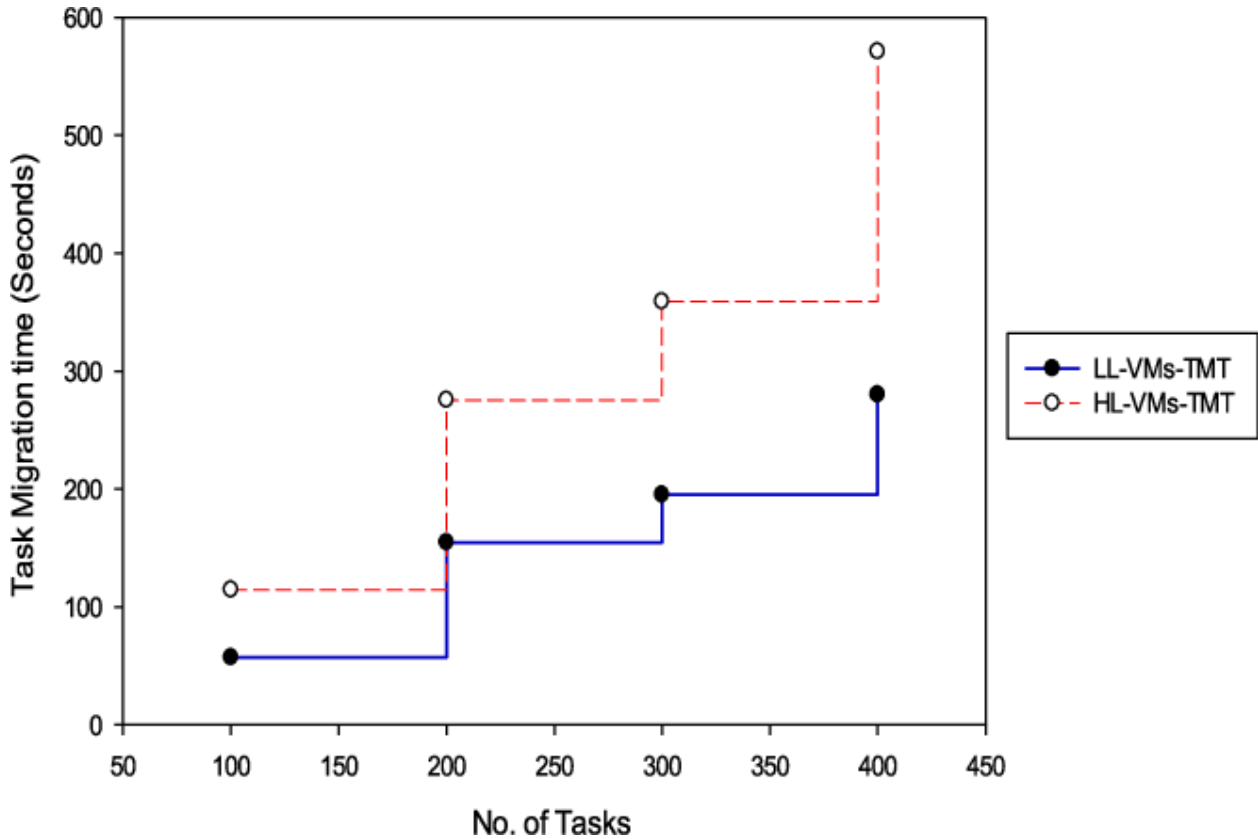


Figure 2. VMs task migration time Vs No. of tasks

5. CONCLUSION

In this work, the proposed dominant firefly behavior search model was applied and simulated in CSVM instances to improve load balancing of tasks in the Cloud computing environment. This approach helps to balance the load in multiple CSVMs by increasing QoS metrics such as throughput and response time. Also, in our methods, the load of job requests from Cloud end-users submitted to CSVMs is optimally balanced to increase the efficiency of the Cloud server. The proposed algorithm was compared with other load balancing algorithms. The results demonstrated an improvement in energy consumption among Cloud servers. These findings could be extended to cost computational methods to utilize maximum CPU that would increase server efficiency. The main objective of the proposed model is to enhance m- learning environments by finding many relational models to avoid the highest energy consuming server throughout the world.

6. REFERENCES

- [1] Zomaya, A. Y., and Teh, Y. H. (2001). Observations on using genetic algorithms for dynamic load-balancing. *Parallel and Distributed Systems, IEEE Transactions on*, 12(9), 899- 911.
- [2] Shreedhar, M., and Varghese, G. (1996). Efficient fair queuing using deficit round-robin. *Networking, IEEE/ACM Transactions on*, 4(3), 375-385.
- [3] Pacinia, E., Mateosb, C., and Garinoa, C. G. (2014). Balancing Throughput and Response Time in Online Scientific Clouds via Ant Colony Optimization. *Advances in Engineering Software*. In press. Elsevier.
- [4] Sekaran, K., & Krishna, P. V. (2017). Cross region load balancing of tasks using region- based rerouting of loads in Cloud computing environment. *International Journal of Advanced Intelligence Paradigms*, Vol. 9, Issue No. 5-6, pp. 589-603.
- [5] D.M. Eyers, R. Routray, R. Zhang, D. Willcocks, and P. Pietzuch. (2009). Towards a middleware for configuring large-scale storage infrastructures. In *Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science*, page 3.
- [6] Fehling, C., Ewald, T., Leymann, F., Pauly, M., Rüttschlin, J., & Schumm, D. (2012). Capturing Cloud computing knowledge and experience in patterns. In *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on* (pp. 726-733). IEEE.
- [7] Yang, X. S. (2009, October). Firefly algorithms for multimodal optimization. In *International symposium on stochastic algorithms* (pp. 169-178). Springer, Berlin, Heidelberg.
- [8] Shen, X. J., Liu, L., Zha, Z. J., Gu, P. Y., Jiang, Z. Q., Chen, J. M., and Panneerselvam, J. (2014). Achieving dynamic load balancing through mobile agents in small world P2P networks. *Computer Networks*, 75, 134-148.
- [9] Krishna, P. V. (2013). Honey bee behavior inspired load balancing of tasks in Cloud computing environments. *Applied Soft Computing*, 13(5), 2292-2303.
- [10] Mondal, B., Dasgupta, K., and Dutta, P. (2012). Load balancing in Cloud computing using stochastic hill climbing-a soft computing approach. *Procedia Technology*, 4, 783-789.

PARALLEL JOB SCHEDULER WITH REPLICATION STRATEGY IN GRID COMPUTING

G. MANISHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

When science and technology advance, the problems encountered become more complicated and need more computing power. In contrast to the traditional notion of using supercomputers, grid computing is proposed. Distributed computing supports resource sharing. Parallel computing supports computing power. Grid computing aims to harness the power of both distributed computing and parallel computing. The goal of grid computing is to aggregate idle resources on the Internet such as Central Processing Unit (CPU) cycles and storage spaces to facilitate utilization.

When human culture advances, current problems in science and engineering become more complicated and need more computing power to tackle and analyze. A supercomputer is not the only choice for solving complex problems any more as a result of the speed-up of personal computers and networks. Grid technology, which connects a number of personal computer clusters with high speed networks, can achieve the same computing power as a supercomputer does, also with a lower cost.

However, grid is a heterogeneous system. Scheduling independent tasks on it is more complicated. In order to utilize the power of grid completely, we need an efficient job scheduling algorithm to assign jobs to resources in a grid. This project proposes an Adaptive Scoring Job Scheduling algorithm (ASJS) for the grid environment. Compared to other methods, it can decrease the completion time of submitted jobs, which may compose of computing-intensive jobs and data-intensive jobs.

Keywords: Central Processing Unit, computing power, Grid technology, ASJS, data-intensive

1. INTRODUCTION

Unstructured peer-to-peer (P2P) file-sharing networks are popular in the mass market. As the peers participating in unstructured networks interconnect randomly, they rely on flooding query messages to discover objects of interest and thus introduce remarkable network traffic. Empirical measurement studies indicate that the peers in P2P networks have similar preferences, and have recently proposed unstructured P2P networks that organize participating peers by exploiting their similarity. The resultant networks may not perform searches efficiently and effectively because existing overlay topology construction algorithms often create unstructured P2P networks without performance guarantees.

Thus, I propose a novel overlay formation algorithm for unstructured P2P networks. Based on the file sharing pattern exhibiting the power-law property, our proposal is unique in that it poses

rigorous performance guarantees. Theoretical performance results conclude that in a constant probability, 1) searching an object in our proposed network efficiently takes N hops (where c is a small constant), and 2) the search progressively effectively exploits the similarity of peers. In addition, the success ratio of discovering an object approximates 100 percent. We validate our theoretical analysis and compare our proposal to competing algorithms in simulations. Based on the simulation results, my proposal clearly outperforms the competing algorithms in terms of 1) the hop count of routing a query message, 2) the successful ratio of resolving a query, 3) the number of messages required for resolving a query, and 4) the message overhead for maintaining and formatting the overlay.5.) Resultant set of finite N hops for data sharing in unstructured peer-to-peer (P2P) networks.

2. LITERATURE REVIEW

In this study, I am interested in optimizing the search performance in Gnutella-like unstructured P2P networks. Existing orthogonal techniques in the literature for improving search performance in unstructured P2P networks include indexing replications super peer architectures and overlay topologies among others. In this Paper, I primarily study the overlay topology formation technique for unstructured P2P networks, aiming to enhance search efficiency and effectiveness. In particular, as recent measurement studies show that peers are likely to resolve the queries issued by the peers sharing the common preferences my study intends to organize the participating peers to exploit their similarity. Overlay construction algorithms intending to exploit the similarity of peers for enhancing search performance can be found in the literature In distributed algorithms are used for constructing unstructured P2P networks (or semantic overlays), such that peers having common preferences cluster together. In addition, the proposalsin, , [23], suggest mixing a naive random network in the P2P network to minimize the overlay path length between any two peers to reduce the query response time. (In this paper, we refer to arandom network with N nodes as the random graph formally defined in that is, in the graph, a node links to another node in a probability of $1/N$.)

In contrast to the studies in suggest organizing the P2P network as a semantic small- world random graph. Here, semantic small-world networks refer to the probability of peer j being the neighbor of peer i increasing if j shares more common interests with i . In this paper, I first observe that existing P2P file sharing networks (e.g., eDonkey) exhibit the power-law filesharing pattern. Based on such sharing pattern, we present a novel overlay construction algorithm to enhance the efficiency and effectiveness of searches in unstructured P2P networks.

Unlike the works in, for example, create small-world-based semantic overlays to enhance search efficiency and effectiveness. For example, in the semantic small-world P2P network peer i is likely to connect to peer j , instead of to another peer k , if i and j have more common interests. However, Chen et al. heavily rely on a centralized entity to help format the P2P networks. More specifically, in a centralized server is employed to compute a probabilistic model based on the principle of maximum entropy which facilitates to estimate the probability of sharing an object o_1 when conditioned on another shared object o_2 . As the number of shared objects in a P2P network is numerous, learning such a probabilistic structure (i.e., the feature

functions) is computationally intensive, and the proposal in thus demands substantial computation resources. Additionally, the centralized server requires disseminating the estimated probabilistic model to all participating peers. As a result, although the idea presented in is interesting, the proposed solution may suffer from the performance bottleneck and introduce a single point of failure. Moreover, the probabilistic model possesses no analytical properties to enable tractable analysis. In contrast to our proposal does not depend on any centralized server for computation-intensive tasks and guarantees rigorous performance results.

Prior efforts in developing distributed overlay formation algorithms for semantic small-world P2P networks can be found in the literature. As any peer in these proposals often selects its neighbors "heuristically," the proposals in mainly depend on simulations for performance investigation. Although these proposals are pragmatic for dynamic, large-scaled distributed environments, they offer no rigorous performance guarantees. In contrast, our network formation algorithm performs very well with rigorously mathematical guarantees and is unique in that in a constant probability, a querying peer takes N hop count (where c is a small constant) to reach the destination peer by progressively and effectively exploiting the similarity of peers on the search path. Additionally, the probability of successfully discovering a requested object in our proposal is approximately 100 percent.

3. PROPOSED METHODOLOGY

A parallel job scheduler with a replication strategy in grid computing involves distributing computational tasks across multiple nodes in a grid infrastructure while also replicating data or computations to ensure fault tolerance and reliability. Here's a proposed methodology for implementing such a system:

1. Problem Analysis and Requirements Gathering:

5. Identify the specific requirements and constraints of your grid computing environment, such as the types of jobs to be executed, the resources available (computational power, storage, network bandwidth), and fault tolerance requirements.
6. Determine the need for parallelism and replication based on the characteristics of the jobs and the grid infrastructure.

2. Design Architecture:

4. Design a scalable architecture that incorporates parallel job scheduling and replication strategies.
5. Components of the architecture may include a job scheduler, resource manager, data storage system, and communication middleware.

3. Job Scheduling Algorithm:

- Develop or select a suitable job scheduling algorithm that can efficiently distribute tasks across multiple nodes in the grid.

- Consider factors such as job dependencies, resource availability, load balancing, and fault tolerance.

4. Replication Strategy:

1. Define a replication strategy for data or computations to ensure fault tolerance and reliability.
2. Determine which data or computations need to be replicated, how many replicas are needed, and where replicas should be placed in the grid.
3. Consider factors such as data access patterns, network latency, and storage capacity.

5. Fault Tolerance Mechanisms:

- Implement fault tolerance mechanisms to handle failures of compute nodes, network partitions, or data corruption.
- Techniques such as data replication, check pointing, and job migration can be used to recover from failures and maintain job progress.

6. Implementation:

- Develop the parallel job scheduler and replication strategy components according to the design architecture and requirements.
- Integrate the components into the grid computing infrastructure, ensuring compatibility with existing grid middleware and resource management systems.

4. RESULTS AND DISCUSSIONS

I have presented an unstructured P2P network with rigorous performance guarantees to enhance search efficiency and effectiveness. In a constant probability, a querying peer takes N hops (where c is a small constant) to reach the destination node capable of resolving the query, whereas the query messages can progressively and effectively exploit the similarity of the peers. The query can be successfully resolved in an approximate probability of 100 percent. Notably, the theoretical analysis further reveals that the competitive decentralized solutions (e.g., those in [21], [22], [23], [24]) do not perform well as the hop count of routing a query message in such networks, considering the exploitation of the similarity of participating peers, is in the polynomial of system size N . We validate our proposal with simulations. The simulation results reveal that whereas GES and SocioNet, that is, the two representative distributed algorithms among [21], [22], [23], [24], introduce fair traffic overhead to maintain and rewire their overlay topologies, ours clearly outperforms GES and Socio Net in terms of the following results are

1. the query message hop count,
2. the successful ratio of resolving a query,
3. the query traffic overhead, and
4. the overlay maintenance overhead.
5. the Resultant set of finite N hops for data sharing in unstructured peer-to-peer (P2P) networks.

5. CONCLUSION

Moreover, In future the work may continue to find that together with a similarity-aware overlay topology, the search protocol we have suggested in this paper, which takes advantage of the similarity of peers exploited by our overlay network, can considerably reduce the search traffic. Peers participating in a P2P network are often heterogeneous in terms of their network bandwidth, storage space, and/or computational capability. It would be interesting for our future work to investigate how the heterogeneity affects our proposal. Moreover, the overlay formation algorithm presented in this paper is oblivious to the physical network topology, and this may introduce considerable wide-area network traffic [14], [29]. It would be challenging to design an overlay formation algorithm aware of both the similarity of participating peers and the physical network topology.

6. REFERENCES

- [1] IPOQUE, "Ipoque Internet Study 2007: P2P File Sharing Still Dominates the Worldwide Internet," <http://www.ipoque.com/news-and-events/news/archive/2007>, 2011.
- [2] S. Sen and J. Wang, "Analyzing Peer-to-Peer Traffic Across Large Networks," *IEEE/ACM Trans. Networking*, vol. 12, no. 2, pp. 219-232, Apr. 2004.
- [3] Gnutella, <http://rfc-gnutella.sourceforge.net/>, 2011.
- [4] Y. Liu, J. Han, and J. Wang, "Rumor Riding: Anonymizing Unstructured Peer-to-Peer Systems," *IEEE Trans. Parallel and Distributed Systems*, vol. 22, no. 3, pp. 464-475, Mar. 2011.
- [5] H. Chen, H. Jin, Y. Liu, and L.M. Ni, "Difficulty-Aware Hybrid Search in Peer-to-Peer Networks," *IEEE Trans. Parallel and Distributed Systems*, vol. 20, no. 1, pp. 71-82, Jan. 2009.
- [6] A. Crespo and H. Garcia-Molina, "Routing Indices for Peer-to-Peer Systems," *Proc. 22th IEEE Int'l Conf. Distributed Computing Systems (ICDCS '02)*, pp. 23-32, July 2002.
- [7] X. Shi, J. Han, Y. Liu, and L.M. Ni, "Popularity Adaptive Search in Hybrid P2P Systems," *J. Parallel and Distributed Computing*, vol. 69, no. 2, pp. 125-134, Feb. 2009.
- [8] E. Cohen and S. Shenker, "Replication Strategies in Unstructured Peer-to-Peer Networks," *Proc. ACM SIGCOMM '02*, pp. 177-190, Aug. 2002.
- [9] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and Replication in Unstructured Peer-to-Peer Networks," *Proc. ACM Int'l Conf. Supercomputing (ICS '02)*, pp. 84-95, June 2002.
- [10] Gnutella2, <http://g2.trillinux.org/>, 2011.

PATTERN BASED ASSOCIATION RULE MINING

H. SANTHAPRIYA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

In this paper, a pattern-based stock data mining approach which transforms the numeric stock data to symbolic sequences, carries out sequential and non-sequential association analysis and uses the mined rules in classifying/predicting the further price movements is proposed. Two formulations of the problem are considered. They are intra-stock mining which focuses on finding frequently appearing patterns for the stock time series itself and inter-stock mining which discovers the strong inter-relationship among several stocks. Three different methods are proposed for carrying out associative classification/prediction, namely, Best Confidence, Maximum Window Size and Majority Voting. They select the mined rule(s) and make the final prediction. A modified Apriori algorithm is also proposed to mine the frequent symbolic sequences in intra-stock mining and the frequent symbol-sets in inter-stock mining.

Keywords: pattern, data mining, price movements, classification, Voting

1. INTRODUCTION

Association, Sequential Mining, Clustering and Deviation [1]. It uses a combination of statistical analysis, machine learning and database management explore the data and to reveal the complex relationships that exists in an exhaustive manner. Additionally, Data Mining consists in the extraction of implicit knowledge (previously unknown and potentially useful), hidden in large databases. Data mining tasks can be classified into two categories [2]: Descriptive mining and Predictive mining. Descriptive mining refers to the method in which the essential characteristics of the data in the database are described [3]. Clustering, Association and Sequential mining are the main tasks involved in the descriptive mining techniques tasks. Predictive mining deduces patterns from the data in a similar manner as predictions. Predictive mining techniques include tasks like Classification, Regression and Deviation detection [4]. Mining Frequent Item sets from transaction databases is a fundamental task for several forms of knowledge discovery such as association rules, sequential patterns, and classification [5]. An item set is frequent if the subsets in a collection of sets of items occur frequently. Frequent item sets is generally adopted to generate association rules.

The objective of Frequent Item set Mining is the identification of items that co-occur above a user given value of frequency, in the transaction database. Association rule mining is one of the principal problems treated in KDD and can be defined as extracting the interesting correlation and relation among huge amount of transactions [6]. Formally, an association rule is an implication relation in the form $X \rightarrow Y$ between two disjunctive sets of items X and Y . A typical example of an association rule on "market basket data" is that "80% of customers who purchase bread also purchase butter " [7]. Each rule has two quality measurements, support and confidence. The rule $X \rightarrow Y$ has confidence c if $c\%$ of transactions in the set of transactions D that

contains X also contains Y. The rule has a support S in the transaction Efficient Analysis of Pattern and Association Rule Mining Approaches set D if S% of transactions in D contain X Y [8]. The problem of mining association rules is to find all association rules that have a support and a confidence exceeding the user-specified threshold of minimum support (called MinSup) and threshold of minimum confidence (called MinConf) respectively. Data Mining tasks categories[9]. Actually, frequent association rule mining became a wide research area in the field of descriptive data mining, and consequently a large number of quick and speed algorithms have been developed. The more efficient are those Apriori based algorithms or Apriori variations. The works that used Apriori as a basic search strategy, they also adapted the complete set of procedures and data structures[10] . Additionally, the scheme of this important algorithm was also used in sequential pattern mining, episode mining, functional dependency discovery & other data mining fields (hierarchical association rules).

2. LITERATURE REVIEW

This section presents a comprehensive survey, mainly focused on the study of research methods for mining the frequent itemsets and association rules with utility considerations. Most of the existing works paid attention to performance and memory perceptions. Apriori: Apriori proposed by is the fundamental algorithm. It searches for frequent itemset browsing the lattice of itemsets in breadth. The database is scanned at each level of lattice. Additionally, Apriori uses a pruning technique based on the properties of the itemsets, which are: If an itemset is frequent, all its sub-sets are frequent and not need to be considered.

AprioriTID: AprioriTID proposed by. This algorithm has the additional property that the database is not used at all for counting the support of candidate itemset after the first pass. Rather, an encoding of the candidate itemsets used in the previous pass is employed for this purpose.

DHP: DHP algorithm (Direct Haching and Pruning) proposed by is an extension of the Apriori algorithm, which use the hashing technique with the attempts to efficiently generate large itemsets and reduces the transaction database size. Any transaction that does not contain any frequent k-itemsets cannot contain any frequent (k+1)-itemsets and such a transaction may be marked or removed.

FDM: FDM (Fast Distributed Mining of association rules) has been proposed by, which has the following distinct features.

1. The generation of candidate sets is in the same spirit of Apriori. However, some relationships between locally large sets and globally large ones are explored to generate a smaller set of candidate sets at each iteration and thus reduce the number of messages to be passed.
2. The second step uses two pruning techniques, local pruning and global pruning to prune away some candidate sets at each individual sites.
3. In order to determine whether a candidate set is large, this algorithm requires only $O(n)$ messages for support count exchange, where n is the number of sites in the network. This is much less than a straight adaptation of Apriori, which requires $O(n^2)$ messages.

GSP: Generalized Sequential Patterns (GSP) is representative Apriori-based sequential pattern mining algorithm proposed by Srikant & Agrawal in 1996. This algorithm uses the downward-closure property of sequential patterns and adopts a multiplepass, candidate generate-and-test approach.

DIC: This algorithm is proposed by Brin et al in 1997. This algorithm partitions the database into intervals of a fixed size so as to lessen the number of traversals through the database. The aim of this algorithm is to find large itemsets which applies infrequent passes over the data than conventional algorithms, and yet uses scarcer candidate itemsets than approaches that rely on sampling. Additionally, DIC algorithm presents a new way of implication rules standardized based on both the predecessor and the successor.

PincerSearch: The Pincer-search algorithm, proposes a new approach for mining maximal frequent itemset which combines both bottom-up and top-down searches to identify frequent itemsets effectively. It classifies the data source into three classes as frequent, infrequent, and unclassified data. Bottom-up approach is the same as Apriori. Top-down search uses a new set called Maximum-Frequent-Candidate-Set (MFCS). It also uses another set called the Maximum Frequent Set (MFS) which contains all the maximal frequent itemsets identified during the process. Any itemset that is classified as infrequent in bottom-up approach is used to update MFCS. Any itemset that is classified as frequent in the top-down approach is used to reduce the number of candidates in the bottom-up approach. When the process terminates, both MFCS and MFS are equal. This algorithm involves more data source scans in the case of sparse data sources.

CARMA: Proposed in 1999 by Hidber which presents a new Continuous Association Rule Mining Algorithm (CARMA) used to continuously produce large itemsets along with a shrinking support interval for each itemset. This algorithm allows the user to change the support threshold anytime during the first Efficient Analysis of Pattern and Association Rule Mining Approaches scan and always completes it at most to scan. CARMA performs Apriori and DIC on low support thresholds. Additionally CARMA readily computes large itemsets in cases which are intractable for Apriori and DIC.

CHARM: Proposed in 1999 Mohammed J. Zaki et al. which presents an approach of Closed Association Rule Mining; (CHARM, "H" is complimentary). This effective algorithm is designed for mining all frequent closed itemsets. With the use of a dual itemset-Tidset search tree it is supposed as closed sets, and use a proficient hybrid method to skive off many search levels. CHARM significantly outpaces previous methods as proved by experimental assessment on a numerous real and duplicate databases.

Depth-project: Depth paper proposed by Agarwal et al., (2000) also mines only maximal frequent item sets. It performs a mixed depth-first and breadth first traversal of the item set lattice. In the algorithm, both subset infrequency pruning and superset frequency pruning are used. The database is represented as a bitmap. Each row in the bitmap is a Bit vector corresponding to a transaction and each column corresponds to an item. The number of rows is

equal to the number of transactions, and the number of columns is equal to the number of items. By using the carefully designed counting methods, the algorithm significantly reduces the cost for finding the support counts.

FP-growth: The principle of FP-growth method is to find that few lately frequent pattern mining methods being effectual and scalable for mining long and short frequent patterns. FP-tree is proposed as a compact data structure that represents the data set in tree form. Each transaction is read and then mapped onto a path in the FP-tree. This is done until all transactions have been read. Different transactions that have common subsets allow the tree to remain compact because their paths overlap. The size of the FP-tree will be only a single branch of nodes. The worst case scenario occurs when every transaction has a unique itemset and so the space needed to store the tree is greater than the space used to store the original data set because the FP-tree requires additional space to store pointers between nodes and also the counters for each item.

Eclat : Is an algorithm proposed by Zaki in 2000 for discovering frequent itemsets from a transaction database. The first scan of the database builds the TID_set of each single item. Starting with a single item ($k = 1$), the frequent $(k+1)$ -itemsets grown from a previous k -itemset can be generated according to the Apriori property, with a depth-first computation order similar to FP-growth . The computation is done by intersection of the TID_sets of the frequent k - itemsets to compute the TID_sets of the corresponding $(k+1)$ - itemsets. This process repeats, until no frequent itemsets or no candidate itemsets can be found.

SPADE: SPADE is an algorithm for mining frequent sequential patterns from a sequence database proposed in 2001 by Zaki . The author uses combinatorial properties to decompose the original problem into smaller sub-problems, that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. All sequences are discovered in only three database scans.

SPAM: SPAM is an algorithm developed by Ayres et al. in 2002 for mining sequential patterns. The developed algorithm is especially efficient when the sequential patterns in the database are very long. The authors introduce a novel depth-first search strategy that integrates a depth-first traversal of the search space with effective pruning mechanisms. The implementation of the search strategy combines a vertical bitmap representation of the database with efficient support counting.

Diffset : Proposed by Mohammed J. Zaki et al. in 2003 as a new vertical data depiction which keep up trace of differences in the tids of a candidate pattern from its generating frequent patterns. This work proves that diffsets is significantly expurgated (by orders of magnitude) the extent of memory needed to keep intermediate results.

DSM-FI: Data Stream Mining for Frequent Itemsets is a novel single-pass algorithm implemented in 2004 by Hua-Fu Li, et al. The aim of this algorithm is to excavate all frequent itemsets over the history of data streams.

PRICES: a skilled algorithm developed by Chuan Wang in 2004, which first recognizes all large itemsets used to construct association rules. This algorithm decreased the time of large

itemset generation by scanning the database just once and by logical operations in the process. For this reason it is capable and efficient and is ten times as quick as Apriori in some cases. PrefixSpan: PrefixSpan proposed by Pei et al. in 2004 is an approach that paper recursively a sequence database into a set of smaller projected databases, and sequential patterns are grown in each projected Efficient Analysis of Pattern and Association Rule Mining Approaches database by exploring only locally frequent fragments. The authors guided a comparative study that shows PrefixSpan, in most cases, outperforms the a prioribased algorithm GSP, FreeSpan, and SPADE.

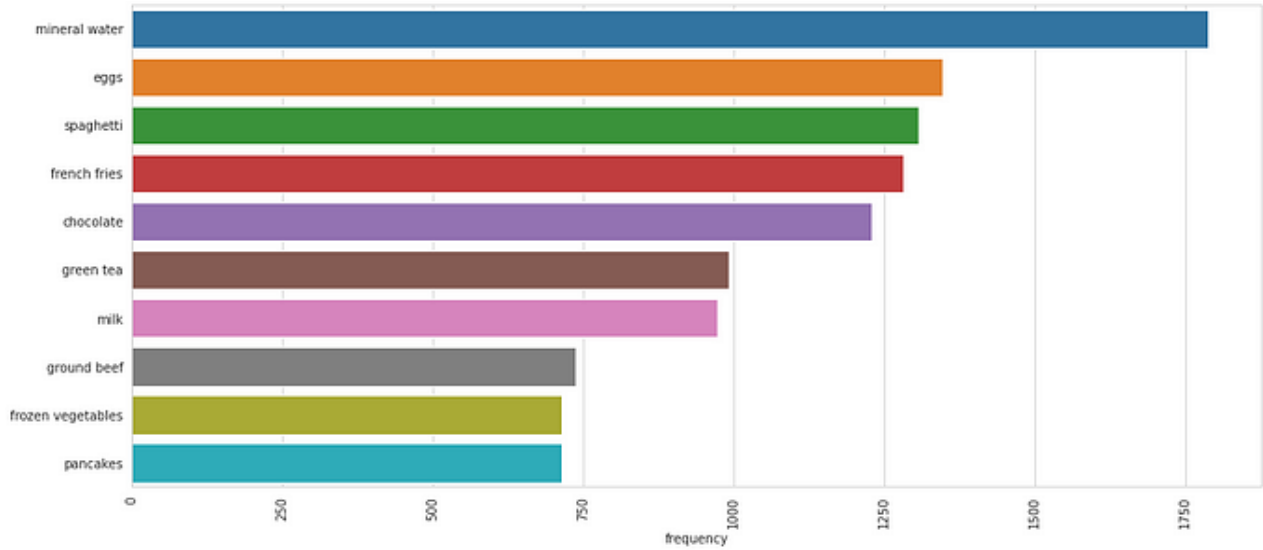
3. PROPOSED METHODOLOGY

A proposed system for association rule mining for stock data should address the limitations of existing systems while leveraging advancements in technology and methodologies. Collect historical stock trading data from reliable sources such as financial databases or APIs.

Preprocess the data to handle missing values, outliers, and inconsistencies. Perform data cleaning, normalization, and transformation as needed. Organize the data into transactions, where each transaction represents a set of stocks traded within a specific time frame (e.g., day, week). Apply association rule mining algorithms to the preprocessed stock data to discover meaningful patterns and relationships. Experiment with different algorithms such as Apriori, FP-Growth, or newer techniques that handle large-scale, high-dimensional data efficiently. Tune algorithm parameters (e.g., minimum support, minimum confidence) based on the characteristics of the dataset and the desired level of association rule significance. Visualize the discovered association rules and patterns using charts, graphs, or interactive dashboards. Interpret the rules to extract meaningful insights into stock market dynamics, sector correlations, trading strategies, etc. Engage domain experts and stakeholders in the interpretation process to ensure that the mined rules align with their understanding of the financial markets and investment goals.

ADVANTAGES

- ✓ By applying association rule mining techniques, the system can uncover hidden patterns and relationships within stock data that may not be immediately apparent to human analysts.
- ✓ Association rules generated by the system provide objective and data-driven decision support for investment decisions.
- ✓ By quantifying relationships between stocks based on historical trading data, the system helps investors identify potentially profitable trading opportunities, diversify their portfolios, or manage risk more effectively.
- ✓ By leveraging scalable algorithms and computational resources, it can process vast amounts of stock data in a timely manner, allowing for real-time or near-real-time analysis of market trends and patterns.
- ✓ The system can be customized to accommodate the specific requirements and preferences of different users or investment strategies.

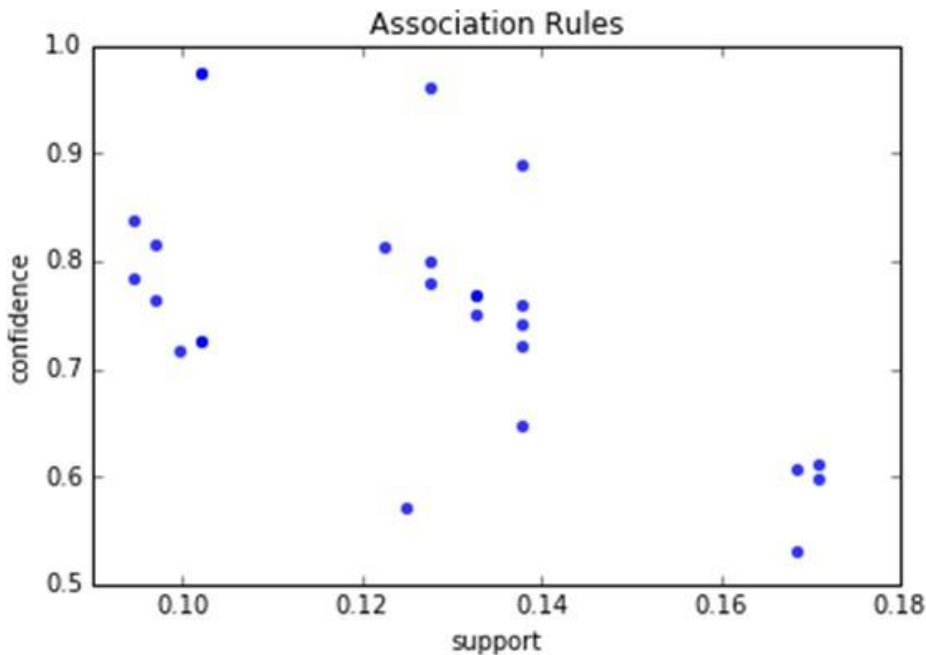


Graph 1: Association rule mining for stock I n different data

RESULTS AND DISCUSSIONS

It has been designed to operate on databases containing transactions, such as purchases by customers of a store (market basket analysis). Besides market basket analysis this algorithm can be applied to other problems. For example in web user navigation domain we can search for rules like customer who visited web page A and page B also visited page C.

Python sklearn library does not have Apriori algorithm but recently I come across post where python library MLxtend was used for Market Basket Analysis. MLxtend has modules for different tasks. In this post I will share how to create data visualization for association rules in data mining using MLxtend for getting association rules and NetworkX module for charting the diagram. First we need to get association rules.



Graph.1 Shown the association rules

To select interesting rules we can use best-known constraints which are a minimum thresholds on confidence and support. Support is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule has been found to be true.

5. CONCLUSION

The most important tasks of frequent pattern mining approaches are : itemset mining, sequential pattern mining, sequential rule mining and association rule mining. A good number of efficient data mining algorithms exist in the literature for mining frequent patterns. In this paper, we have presented a brief overview of the current status and future directions of frequent pattern mining. Additionally, we have performed a comprehensive study of some algorithms and methods that exists for the mining of frequent patterns. With over a decade of extensive research,a good number of research publications, development and application activities in this domain have been proposed. We give a brief discussion of a number of algorithms presented along this decade with a comparative study of a few significant ones based on their performance. However,we require to conduct a deep research based on several critical issues so that this domain may have its factual existence and deep impact in data mining applications

6. REFERENCE

- [1] Cios K.J., Pedrycz W, Swiniarski RW, & Kurgan LA. Data mining: A knowledge discovery approach. New York, NY: Springer, 2012.
- [2] Marek Wo, Krzysztof Ga, Krzysztof Ga. "Concurrent Processing of Frequent Itemset Queries Using FP-Growth Algorithm", Proc. of the 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'05), 2005, Tallinn, Estonia.
- [3] Alva Erwin, Raj P. Gopalan, N.R. Achuthan, "A BottomUp Projection Based Algorithm for Mining High Utility Itemsets", In Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining, 2007, Vol. 84: 3-11.
- [4] Liu X., Zhai K., & Pedrycz W. An improved association rules mining method. Expert Systems with Applications, 2012 39(1):1362–1374. doi:10.1016/j. eswa.2011.08.018.
- [5] Agrawal, R., Imielinski, R., & Swami, A. Mining associations between sets of items in massive databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, 1993, Washington, DC: 207-216.
- [6] Sarasere, A., Omiecinsky, E. & Navathe, S. "An efficient algorithm for mining association rules in large databases" In Proc. 21st International Conference on Very Large Databases (VLDB) , 1995, Zurich, Switzerland, Technical Report No. GIT-CC-95-04.
- [7] Agrawal, R. and Srikant, R. "Mining sequential patterns" In P.S.Yu and A.L.P. Chen, editors, Proc.11th Int. Conf. Data engineering. ICDE, 1995, 3(14), IEEE :6-10.
- [8] Jiawei Han., Y.F. "Discovery of multiple-level association rules from large databases" In Proc. of the 21st International Conference on Very Large Databases (VLDB), 1995, Zurich, Switzerland: 420-431.
- [9] Lallich S., Vaillant B, & Lenca P. Parameterized Measures for the Evaluation of Association Rules Interestingness. In Proceedings of the 6th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005),2005, Brest, France, May: 220-229.
- [10] Brin, S., Motwani, R., Vllman, J.D. & Tsur, S. Dynamic itemset counting and implication rules for market basket data, SIGMOD Record (ACM Special Interest Group on Management of Data),1997,26(2),255.

PRODUCT ASPECT RANKING AND ITS APPLICATIONS USING SENTIMENT CLASSIFIER FRAMEWORK

S. DHARANISRI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Numerous consumer reviews of products are now available on the Internet. Consumer reviews contain rich and valuable knowledge for both firms and users. However, the reviews are often disorganized, leading to difficulties in information navigation and knowledge acquisition. This article proposes a product aspect ranking framework, which automatically identifies the important aspects of products from online consumer reviews, aiming at improving the usability of the numerous reviews. The important product aspects are identified based on two observations: 1) the important aspects are usually commented on by a large number of consumers and 2) consumer opinions on the important aspects greatly influence their overall opinions on the product. In particular, given the consumer reviews of a product, we first identify product aspects by a shallow dependency parser and determine consumer opinions on these aspects via a sentiment classifier. We then develop a probabilistic aspect ranking algorithm to infer the importance of aspects by simultaneously considering aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions. We apply product aspect ranking to two real-world applications, i.e., document-level sentiment classification and extractive review summarization, and achieve significant performance improvements, which demonstrate the capacity of product aspect ranking in facilitating real-world applications.

Keywords: Consumer reviews, product aspect ranking, algorithm, review, sentiment

1. INTRODUCTION

This invention is a method to analyze reviews in order to automatically identify, first, the aspects or features of a product being reviewed and, second, the most important aspects or features from amongst those identified that influence the overall perception of a product. The invention's basic mechanism consists of three steps: 1) Aspect identification, 2) Sentiment classification, and 3) Aspect ranking algorithm. The invention's underlying mechanism has been applied to document-level sentiment classification and extractive review summarization[1]. This invention could be used by companies interested in identifying the public perception of their products, by those in the business of market research and forecast, and by those aggregating products reviews. Generally, a product may have hundreds of aspects. For example, iPhone 3GS has more than three hundred aspects such as "usability," "design," "application," "3G network." We argue that some aspects are more important than the others, and have greater impact on the eventual consumers' decision making as well as firms' product development strategies[4,5]. For example, some aspects of iPhone 3GS, e.g., "usability" and "battery," are concerned by most consumers, and are more important than the others such as "usb" and "button." For a camera product, the aspects such as "lenses" and "picture quality" would greatly influence consumer opinions on the camera, and they are more important than the aspects such as "a/v cable" and

"wrist strap." Hence, identifying important product aspects will improve the usability of numerous reviews and is beneficial to both consumers and firms. Consumers can conveniently make wise purchasing decision by paying.

One main goal of extracting product aspects from online reviews automatically is to generate a list of the most representative aspects of a product that have been discussed online in the customers' feedback. The generated list of important product aspects is considered as guidance for the potential customers to discriminate the various types of products. Many sentiment analysis approaches have been proposed to investigate online reviews in order to accomplish two main tasks; first to extract aspects (or attributes) related to the product (e.g. battery life, size), known as aspect extraction, while the second task is to determine the sentiment orientation of these aspects (e.g. positive or negative), known as sentiment classification. This paper focuses on the first task, which is to optimize the quality of extracted aspects. Aspect extraction involves identification of possible targets that have been mentioned in the customer reviews positively or negatively, such as a service (e.g. housekeeping), a product (e.g. iPhone), part of a product (e.g. battery life, screen size). Most of the aspect extraction techniques identify the product aspects in the customer reviews based on three main criteria:

- ✓ Frequent product aspects: the occurrences of each aspect in customer reviews are a crucial factor in extracting product aspects, because important aspects are those that have been discussed by many customers. Many research studies consider the more frequent aspects in online reviews are representative aspects for a product based on a human supervision threshold [3]–[6].
- ✓ Opinionated product aspects: Sentiment analysis domain is looking for subjective information that people express in their talk on the Web. Therefore, it is not enough to consider frequent aspects in online reviews as the real product aspects; they should be associated with emotions, sentiments, and opinions in order to be considered as a significant reference for the potential customers. Several proposed techniques adopted this criterion by considering the potential product aspects are those aspects that are more opinionated than others in the product reviews [7]–[10].
- ✓ Domain-specific product aspects: the extraction of the most relevant aspects to a specific domain product is the basic idea of these approaches. Certainly, identifying representative aspects for each domain product is also should be considered important, because the extraction process could be seen as a domain dependent entity recognition

2. LITERATURE REVIEW

Aspect-level sentiment analysis In sentiment analysis field, it is assumed that opinions should have targets, especially in the domain of customer review. These targets are called aspects. Automatic extraction of these aspects in unsupervised manner becomes as an urgent need, because of the lack in fitting supervised approaches to new domain products, which is considered as a time-consuming process. Most unsupervised aspect extraction approaches are lexicon-based approaches. These approaches could be characterized based on the criteria used in

the extraction process of product aspects; many approaches considered the frequent aspects in the customer reviews are more relevant to be the representative aspects for the product domain, while other techniques adopted the idea that the relevant aspects are those that have been described by many customers positively or negatively. The last type of the proposed approaches considered the correlation between the aspects and the domain product, like 'camera'. The following subsections present some of these approaches based on the type of the adopted criterion.

1) Frequent aspect extraction, the authors inspired many researchers to do more research on the domain of Aspect-level sentiment analysis. Their approach is based on extracting nouns and noun phrases from the review, and then applying Apriori algorithm in order to find all frequent item sets, in which they consider the item set is frequent if it appears with a value more than a specific threshold (1% of the subjective sentences in the review). This approach affirmed its efficiency, since many commercial organizations applied it in their systems with some enhancements. The proposed work in also retains the higher frequency candidate aspects based on a threshold settings, but they improved the results by computing the aspect relevancy to a domain based on point wise mutual information (PMI) between the candidate aspect and a set of meronymy discriminator like 'camera has', 'camera comes with'

2) Opinionated aspect extraction However, some research studies affirmed that term frequency of the aspects is not enough to identify the most important product aspects discussed by the customers in their reviews. For instance, the study of identified the relevant product aspects based on the number of opinion words that are close to them in the customer reviews. In their approach, they introduced the High Adjective Count (HAC) algorithm, instead of using aspect frequency. The popular PageRank algorithm has been employed by some researchers in the task of aspect extraction as well, to illustrate, the study in identifies the candidate product aspects based on the dependency relationship between the opinion words and noun/noun phrases in the customer reviews, then filtering rules have been used in addition to ranking the aspects using extended PageRank algorithm to generate the representative product aspects. 3) Domain- specific aspects extraction Several proposed techniques adopted the idea that the aspect extraction problem is considered as domain-specific target recognition, based on the observation that sentiment analysis is context-sensitive. In the proposed work of , the basic idea is to compute the correlation between the aspects and different related domains, and based on these computations, the domain-specific aspects have extracted by calculating the distances between the features and each domain vector using a new distance measure PMI-TFIDF. In their work they exploit the popular lexicon WordNet and ConceptNet lexicon to extract domain-specific aspects and attributes. They firstly identified the product category (like 'phone', 'camera') of the corpus of customer reviews, and then they extract the synonyms and hyponyms that relevant to the product domain.

B. Product aspects ranking in the literature, a few research studies have been conducted in the aspect ranking task. To illustrate, a considerable research study by proposed a probabilistic aspects ranking algorithm in order to identify the important aspects in online reviews. They

extract the aspects based on the following two observations: high frequent aspects and influence of the opinionated aspects on the overall customers opinions on a product. Similar to the work of in identifying the aspects, but the difference is that they identified nouns from Pros and Cons reviews to form a lexicon of sentiment words and use it to train a sentiment classifier like SVM (Support Vector Machine) to be applied on free text reviews. This research study inspired some researchers to do more on aspect ranking with almost the same methodology, for instance, the study of, they used the same technique in exploiting Pros and Cons reviews to enhance the aspect identification and sentiment classification on free-text feedback, the probabilistic ranking algorithm which they proposed is based on calculating the weights of aspects by considering the frequency and the customers' opinions on the frequent aspects.

The previous studies used the evaluation measures of Normalized Discounted Cumulative Gain at top K (NDCG@K) and Discounted Cumulative Gain at top K (DCG@K) respectively to evaluate the proposed aspect ranking approaches. C. Research Gap Certainly, our work is different from previous studies in the following issues: 1) the aspect identification task manipulated the free-text reviews in unsupervised manner (domain-independent) without any assist from Pros and Cons reviews, 2) in previous works, parsing tree and a trained sentiment classifier from Pros and Cons reviews have been used to identify the opinionated aspects, while in our case we exploits N-gram analysis and SentiStrength lexicon for this task 3) In our work, the identification of opinionated aspects starts by identifying opinion words in the customer review and then we are looking for the aspects beside the opinion word. Conversely, previous works starts by identifying the frequent aspects and then they are looking for the closest opinion expressions. 4) Lastly, the ranking task of product aspects does not consider the association between the candidate aspects and the product domain, where in customer reviews, many nouns could be discussed several times but they are totally do not belong to the domain product. Therefore, in our work we consider the correlation between the aspects and the product domain using lexicographers' files in WordNet. However, to the best of our knowledge, there has been no attempt to use sentiment analysis and MCDM to extract product aspects from customer reviews in an unsupervised manner. In this work, we have seen the problem of aspect extraction as a multi-criteria problem, in which the extraction process of relevant aspects of a product in our approach is based on the three extraction criteria discussed in this section.

Thus, we exploit TOPSIS method with sentiment analysis in our research. Multi-criteria Decision Making (MCDM) is considered as a highly reliable methodology for ranking multiple alternatives based on several criteria. MCDM methods have been successfully applied in different areas of Economics, Energy management, Transportation, human resources management, and other domains.

3. PROPOSED METODOLOGY

Product aspect ranking is beneficial to a wide range of real-world applications. In this paper, we investigate its usefulness in two applications, i.e. document-level sentiment classification that aims to determine a review document as expressing a positive or negative

overall opinion, and extractive review summarization which aims to summarize consumer reviews by selecting informative review sentences. We perform extensive experiments to evaluate the efficacy of aspect ranking in these two applications and achieve significant performance improvements. This invention could be used by companies interested in identifying the public perception of their products, by those in the business of market research and forecast, and by those aggregating products reviews. Product aspect ranking was first introduced in our previous work. Compared to the preliminary conference version, this article has no less than the following

ADVANTAGES

- It elaborates more discussions and analysis on product aspect ranking problem;
- It performs extensive evaluations on more products in more diverse domains; and
- It demonstrates the potential of aspect ranking in more real-world applications.

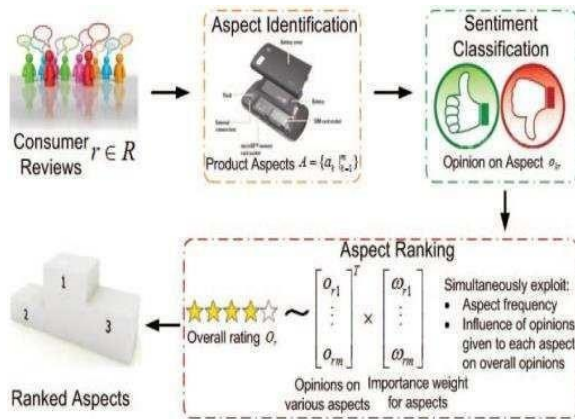
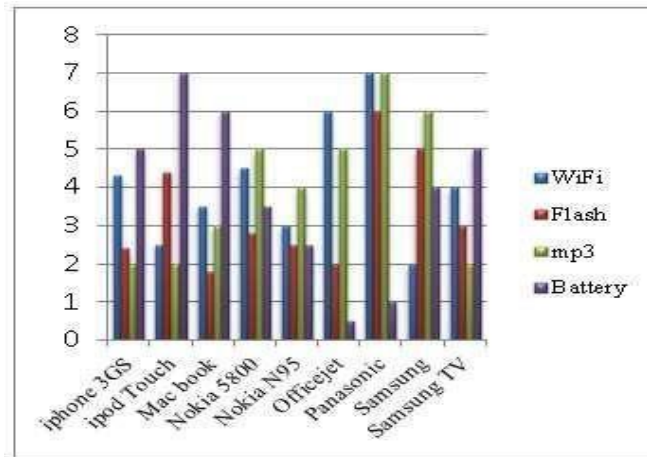


Figure: 1- Architecture diagram of Multiple Product Aspect Ranking Using Sentiment Classification.

4. RESULTS AND DISCUSSIONS

In the case of pros and cons reviews, the aspects are represented in a unigram feature, and utilize every aspect to determine the Support Vector Machine (SVM). The SVM is used to recognize the clustered noun terms, such as “earphone” and “headphones”. The clustered synonyms are collected from the synonym dictionary website. Therefore, the first step, “recognizing the product aspect”, pronounces the identification and grouping of the aspects of a product. Here, the product aspects are examined by sentiment assortment. Existing techniques admit the supervised learning and lexicon based approaches. Once the product aspects are identified, we collect the persuasions which can be used as the features of the product. Eventually, each review is represented as a feature vector. In our work, we will be using the Support Vector Machine to rank the consumer reviews according to its sentiment assortment. Finally, we will be proposing a Product Aspect Rating Algorithm in order to detect the significant aspects of a product from millions of reviews. The general opinion in a review is a

collection of impressions given to particular aspects in the review and different aspects have different shares in the aggregation. The discounted overall ratings are accepted to be generated from a Gaussian distribution.



Graph 1: A graphical representation of product identification.

5. CONCLUSION

In this article, we have proposed a product aspect ranking framework to identify the important aspects of products from numerous consumer reviews. The framework contains three main components, i.e., product aspect identification, aspect sentiment classification, and aspect ranking. First, we exploited the Pros and Cons reviews to improve aspect identification and sentiment classification on free-text reviews. We then developed a probabilistic aspect ranking algorithm to infer the importance of various aspects of a product from numerous reviews. The algorithm simultaneously explores aspect frequency and the influence of consumer opinions

6. REFERENCES

- [1] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *J. Neural Parallel Scientific Comput.*, vol. 11, no. 4, pp. 351–368, 2003.
- [2] C. C. Chang and C. J. Lin. (2004). Libsvm: A library for support vector machines[Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [3] G. Carenini, R. T. Ng, and E. Zwart, "Multi-document summarization of evaluative text," in *Proc. ACL*, Sydney, NSW, Australia, 2006, pp. 3–7.
- [4] China Unicom 100 Customers iPhone User Feedback Report, 2009.
- [5] ComScore Reports [Online]. Available:http://www.comscore.com/Press_events/Press_releases, 2011.
- [6] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proc. WSDM*, New York, NY, USA, 2008, pp. 231–240.
- [7] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, no. 1, pp. 457–479, Jul. 2004.
- [8] O. Etzioni et al., "Unsupervised named-entity extraction from the web: An experimental study," *J. Artif. Intell.*, vol. 165, no. 1, pp. 91–134. Jun. 2005.

PUBLIC RELIABILITY AND AUTHENTICATION FOR CLOUD RESOURCE AGAINST AUDITORS BEHAVIOR USING BLOCK CHAIN

V. ABARNA SREE

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Using Cloud Storage, users can remotely store their data and enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources, without the burden of local data storage and maintenance. However, the fact that users no longer have physical possession of the outsourced data makes the data integrity protection in Cloud Computing a formidable task, especially for users with constrained computing resources. Moreover, users should be able to just use the cloud storage as if it is local, without worrying about the need to verify its integrity. Thus, enabling public auditability for cloud storage is of critical importance so that users can resort to a third party auditor (TPA) to check the integrity of outsourced data and be worry-free. To securely introduce an effective TPA, the auditing process should bring in no new vulnerabilities towards user data privacy, and introduce no additional online burden to user. In this paper, we propose a secure cloud storage system supporting privacy-preserving public auditing. We further extend our result to enable the TPA to perform audits for multiple users simultaneously and efficiently. Extensive security and performance analysis show the proposed schemes are provably secure and highly efficient.

Keyword: Third Party Auditor, Cloud Storage, Cloud computing, data

1. INTRODUCTION

TCP/IP Internet works on a principle of providing end-to-end data transfer using a concatenation of potentially dissimilar link-layer technologies. A bunch of data link layer protocols are standardized and worked well on the globe. However, there are many environments where internet assumptions do not hold[1]. Once there is no end-to-end path between source and destination for the duration of a communication session or communication is unreliable and might only exist for short periods of time, a TCP/IP network starts to work inappropriately or even stops to work at all. A good example of such environment is the *_Interplanetary Internet_*. The speed-of-light delay from Earth to Mars, for example, is approximately 4 minutes when Earth and Mars are at their closest approach. The one-way light time can exceed 20 minutes when Earth and Mars are in opposition[3]. The speed-of-light delay to the outer planets becomes significantly higher. If one wants to send a file from a base station on Earth to a satellite flying around Mars, it might take about one hour just to initiate the *_le transfer*[4]. File Transfer Protocol requires authorization and authentication commands to be sent before the data transfer starts, TCP uses handshaking mechanism and sends three packets for each FTP command. Considering that round trip of a TCP packet takes at minimum 8 minutes, it becomes clear why one should wait long until the data transfer is initialized[6]. Delay-tolerant networks (DTN) have been designed to operate in environments where Internet Protocol Suite does not seem to work well[7].

Delay-tolerant networks use a message oriented overlay that supports intermittent connectivity, overcomes communication disruptions and delays. Transmission of data between source and destination nonexistent for the time of a communication is also allowed. All aforementioned features are achieved by using store-and-forward message method. Services the method provides are very similar to electronic mail, but with improved naming, routing and security capabilities.

Cloud storage has attracted the industrial research and academic fields of having large benefits such as management, cost and performance[8]. The users of cloud storage can lessen the hardware and software expenditure by saving the information in the public cloud server. Cloud storage users can receive the outsourcing data remotely and efficiently without being nearby computers. Recently, more users have been selected to transfer the information in the cloud storage that is controlled by the CSPs, namely Google Cloud and Amazon Cloud. Though cloud storage provides many advantages to users, outsourcing data has several challenging security problems. Data integrity becomes the most necessitated concern. In realism, data outsourcing may get tainted as it may be externally affected by rival assaults and internally affected by the failures in hardware and software. Moreover, cloud storage is not trusted and independent and can delete the information stored by the users in the cloud or hide the information to enhance the application's reputation. Some users of cloud storage normally delete their information after backup from the cloud storage. Owing to these problems, users have a necessity for data auditing integrity periodically.

Many cloud storage audition models are developed to secure the outsourcing data's integrity. In conventional public auditing techniques, users authorize the TPA to periodically conduct public auditing for their outsourcing data. The reliable Third Party Verification (TPV) provides trustable results in auditing and lessens the communication of the user also the computational burden. However, TPAs are considered reliable and honest; in reality, it gives a strong presumption since auditor corruption is possible. For example, the capricious auditor can produce a verified auditing statement without performing the process of verification to save the cost of consumption[9]. Several public auditing techniques are developed based on blockchain to prevent malevolent auditors. The blockchain models are used as the security basis of pseudo-randomness according to the time and confirmation of indisputable storage. Particularly, auditors generally extract the value of blocks, namely, Block and nonce hash, from the blockchain for generating challenging messages containing an index of the chosen data block. They also develop the log folders for the process of auditing and the value of store hash of log accessing in the blockchain.

Many conventional schemes based on blockchain are affected by tempted auditors and have centralization issues. In these models, TPA uses the random values in Proof-of-Work (PoW), and the miner identifies the value of the blockchain. The malevolent auditors can motivate the miners of malevolent block-chain to avoid the newly extracted block if the developed challenging message has assured block. This malevolent auditor is termed a tempting auditor.

In auditing outcomes, this prejudiced performance has a significant influence. Additionally, conventional models regarding blockchain have no rigorous evidence for blockchain value's reliability and randomness. Moreover, for computing the fundamental tools and verification of the cloud server, the conventional models used the centralized TPA[10]. The auditing process does not work properly and may produce improper auditing outcomes if once TPA gets compromised by internal or external failures. The accessibility of auditing checks cannot always be assured. Thus, the centralized TPA present in models based on blockchain is susceptible to the Single Points of Failure (SPOF). Thus, we developed decentralized public auditing for cloud storage to protect the data. The current Internet is ubiquitous. In a period of less than two decades, it has morphed into a fundamental component of modern society, culture, knowledge, businesses and defense infrastructures.

2. LITERATURE REVIEW

2.1 A basic Resource Allocation algorithm

The problem of static assignments of resources to a VN has been investigated in as an assignment without reconfiguration. It has been observed that the static mapping of VN node request can be considered as an offline load balancing problem which can be transformed into NP-hard unsplittable flow problem. It has been conjectured that complexity of resource assignment problem in VNs will further increase due to stringent requirements of minimizing node and link stresses. Therefore, a heuristic approach has been adopted, i.e. by selecting a cluster of nodes that have low stress (lightly loaded) and will likely to cause a lower link stress when connected in a VN topology.

The shortest distance path algorithm has been employed in for evaluation of various available paths on SN. It depends upon a distance function, which has been originally defined as sum of reciprocals of available bandwidth on various available paths. It dynamically balances the impact of hop count and the path load on the computation of path distance for various candidate paths. This original distance function has been adapted for substrate link stress in. Then, after computing minimum distance for all paths, node potential is computed, which has been defined as a ratio of sum of minimum distances of all substrate links in a cluster and maximum node stress. The next step consists of determining the potential of all the nodes in a cluster and mapping between substrate nodes and VN nodes is performed in such a way that virtual nodes with higher potential are connected to substrate nodes with higher neighborhood resource availability. The last step involves the connection of selected SN nodes according to VN topology; for which again the shortest-distance path algorithm has been employed. It has also been pointed out that the basic assignment algorithm becomes inefficient for sparse topologies of VNs.

2.2 Request Traffic constraints based algorithm

A cost effective method for designing VNs, though which may not yield optimal results due to NP-hard nature of problem. The size of search space is reduced by restricting the VN topologies to back-bone star topologies. In such topology, some of the nodes are designated as back-bone while others are referred to as access nodes. The backbone nodes form the center of the star, to which access nodes are connected. The back-bone nodes can be connected in an arbitrary fashion, but they have been constrained to form a complete graph consisting of ring or star.

The whole algorithm can express as an iterative loop. Next, three types of traffic constraints have been defined, (i) termination constraints that describe the total traffic terminating at the VN's access nodes and described by incoming and outgoing traffic from an access node, (ii) pair wise traffic constraints which provides upper bound on traffic flow from one access node to other, and (iii) distance constraints that specify the upper bounds on traffic flow outside the neighborhood of a node. Link dimensioning has been done by following these constraints. The back-bone nodes mapping has been formulated and solved as a mixed integer quadratic program. Finally, VN designs are compared by cost metric defined by product of shortest path distance and fair traffic share function.

2.3 Splitting and Migration of Paths

A greedy node mapping algorithm with an objective to maximize revenue has been presented in. It defines amount of resources available at node as a product of CPU capacity and link bandwidth. Link mapping is performed by k-shortest path algorithm. In order to improve efficiency of link assignment, path splitting has been proposed. Further, in order to achieve efficient resource utilization in a scenario of time dependent requests for VN, path migration has also been proposed.

For path migration the node mapping is needed to be kept fixed and either path splitting ratio can be varied or a completely new path in SN is selected. Also, in order to avoid out of order packet delivery, hash-based splitting schemes have been proposed Path Splicing, a recently proposed routing primitive to allows network paths to be constructed by combining multiple routing trees to each destination over a single network topology, can also be experimented with efficient utilization of link resources in VNs.

Dynamic Approaches

A static resource assignment to multiple VNs, where each network is customized for a particular traffic class, can lead to lower performance and under utilization of substrate resources. It can lead to inefficient scenarios causing wastage of resources; e.g. over a same SN, one VN is experiencing a high packet loss, whereas the other VN is operating under a light traffic load. It will also affect delay and jitter sensitive VNs such as overlays for video transmission. Thus, it is important that an adaptive mechanism should be adopted to re-allocate the substrate network resources to various VN instantiations.

Taking inspiration from rerouting in circuit switched networks, the problem of dynamic assignment of resources to VNs have been studied in. However, it has been observed that the

reconfiguration process in VN assignment is much more complex than flow routing. The event of reconfiguration, i.e. reallocation of resources to VNs, may involve a significant change in node and path switching in the SN. Thus, in order to quantify reconfigurations in VNs, a cost metric has been defined in. It is a weighted sum of reconfiguration rate, node and path switching. It is important to realize that the number of reconfigurations of set of VNs over a substrate can be limited due to stability reasons and computational overheads. Hence, a selective reconfiguration process has been adopted, which gives priority to those parts of VNs that are highly loaded.

The selective reconfiguration algorithm depends on: (i) periodic marking of critically stressed nodes and links of substrate and (ii) per VN reconfiguration and performance monitoring.

DaVinci: Recently, a framework for Dynamically Adaptive Virtual Networks for a Customized Internet (DaVinci) has been proposed. This architecture advocates a periodic reassignment of bandwidth among multiple VNs, which are sharing virtual resources derived from a common SN. In parallel, each VN runs its own distributed protocol to maximize its objective function. It allows the use of multiple (virtual) paths for reaching another node, which can cause packet reordering problem. Another weakness in this framework is that the links in SN need to know the performance objective function of all VNs, which may not be possible in the real world settings. Also, node assignment problem for VNs has not been considered by DaVinci.

Autonomic Systems based Model: A combined approach comprising of VNs and Autonomic computing, has been proposed in [15]. It provides automated services and network resource management in DiffServ enabled IP/MPLS based transport networks. In this architecture, a customer will request for the creation of a VN that is capable of delivering a desired level of a service. After the VN has been instantiated, its performance will be measured at regular intervals.

The performance evaluation metrics comprise of packet loss rate, delay, jitter and end-to-end bandwidth. Also the provider will strive to achieve the optimized usage of his resources. Keeping in view of these stringent requirements, VN based Autonomic network Resource control and Management System (VNARMS) has been proposed. Two types of autonomic components defined are: Virtual Network Resource Manager (VNRM) and Resource Agents (RAs), where former is responsible for control/management and later performs the element level resource control and management. In a nut shell, the four major components in an autonomic control loop formed by VNRM are: resource manager for monitoring/executions, operations manager for monitoring and analysis, VN manager and topology manager for planning. However, the proposed autonomic system has not been yet implemented and thus, its performance in real systems is unknown.

2.4 Control Theoretic based systems:

It has been suggested that computer systems should be designed in a way that they are amenable to feed-back control laws; for which several off-shelf adaptive controllers already exist. In the same breath, one of the promising techniques of resource allocation in virtualized network environments is adaptive control theory. However, no complete systematic approach exists for designing an optimal resource allocation paradigm for VNs. This open problem is further complicated by existence of NP-hardness in node and link assignment, the resource allocation process in VNs can also be modeled as closed loop system. In this setup, the requests for instantiation of VNs arrive in real time; currently, it is difficult to find the arrival distribution as no commercial VNs provider exists.

The requests for VNs will be queued at provider's master queue and will be scheduled at an appropriate time according to service level agreement. In such setups, queuing will be an essential component of the loop as high computational costs may be associated with each VN request. Several possibilities exist for the design of scheduler for servicing VN requests, such as: weighted round robin, weighted fair queuing or priority queuing. A maximum time-to-live field in provider's queue may also be introduced as in VN request as in IP. After instantiation of the desired VN, either open loop or closed policy may be adopted. The former is similar to static resource assignment and the latter resembles dynamic resource assignment techniques.

3. PROPOSED METHODOLOGY

In this proposed system utilize the public key based homomorphic authenticator and uniquely integrate it with random mask technique to achieve a privacy-preserving public auditing system for cloud data storage security while keeping all above requirements in mind. To support efficient handling of multiple auditing tasks, we further explore the technique of bilinear aggregate signature to extend our main result into a multi-user setting, where TPA can perform multiple auditing tasks simultaneously. Extensive security and performance analysis shows the proposed schemes are provably secure and highly efficient. We also show how to extend our main scheme to support batch auditing for TPA upon delegations from multi-users.

ADVANTAGES

- **Public auditability:** to allow TPA to verify the correctness of the cloud data on demand without retrieving a copy of the whole data or introducing additional online burden to the cloud users.
- **Storage correctness:** to ensure that there exists no cheating cloud server that can pass the TPA's audit without indeed storing users' data intact.
- **Privacy preserving:** to ensure that the TPA cannot derive users' data content from the information collected during the auditing process.
- **Batch auditing:** to enable TPA with secure and efficient auditing capability to cope with multiple auditing delegations from possibly large number of different users simultaneously
- **Lightweight:** to allow TPA to perform auditing with minimum communication and computation overhead.

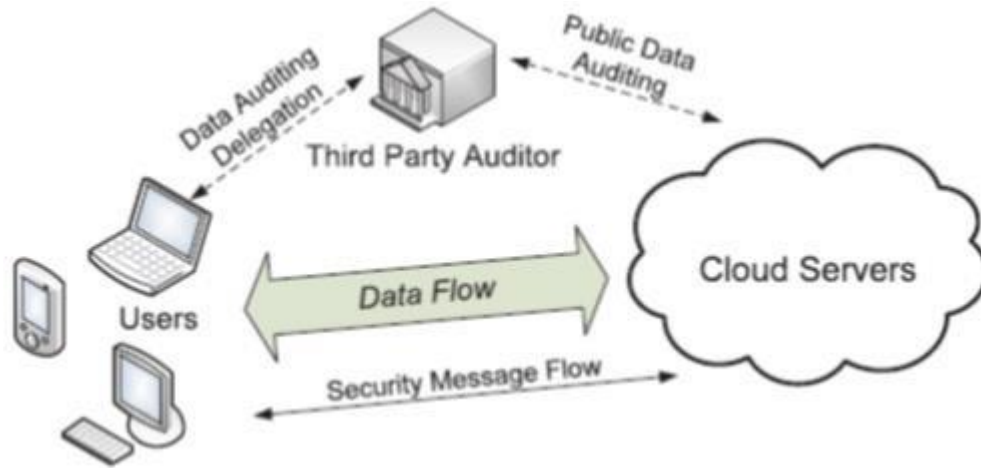
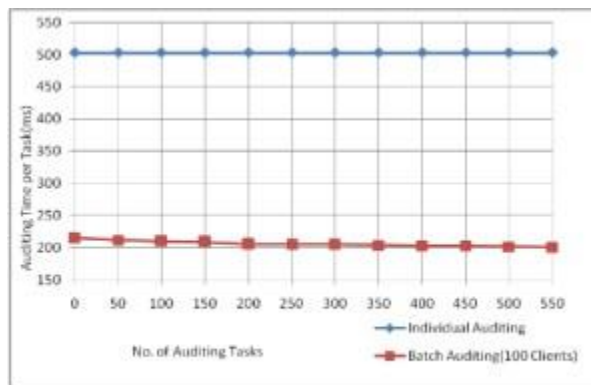


Figure 1: Architecture of cloud data storage

4. RESULTS AND DISCUSSIONS

In batch auditing multiple user can access CS simultaneously. The TPA may concurrently handle multiple auditing processes for multiple users. Multiple TPA are used for the auditing process. TPA batch all users task and audit it at one to time. The advantage of batch auditing is that it reduces the time for handling the multiple audits for multiple users. The graph shows the comparison of individual auditing and the batch auditing. The comparison is done on the basis of auditing time required to perform number o tasks



Graph 1: supports for Batch Auditing

5. CONCLUSION

We propose a privacy-preserving public auditing system for data storage security inCloud Computing. We utilize the homomorphic linear authenticator and random masking to guarantee that the TPA would not learn any knowledge about the data content stored on the cloud server during the efficient auditing process, which not only eliminates the burden of cloud user from the tedious and possibly expensive auditing task, but also alleviates the users' fear of their outsourced data leakage. Considering TPA may concurrently handle multiple audit sessions

from different users for their outsourced data files, we further extend our privacy-preserving public auditing protocol into a multi-user setting, where the TPA can perform multiple auditing tasks in a batch manner for better efficiency. Extensive analysis shows that our schemes are provably secure and highly efficient.

6. REFERENCE

- [1]. C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Storage Security in Cloud Computing," Proc. IEEE INFOCOM '10, Mar. 2010.
- [2] P. Mell and T. Grance, "Draft NIST Working Definition of Cloud Computing," <http://csrc.nist.gov/groups/SNS/cloudcomputing/index.html>, June 2009.
- [3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report UCB-EECS-2009-28, Univ. of California,
- [4] Cloud Security Alliance, "Top Threats to Cloud Computing," <http://www.cloudsecurityalliance.org>, 2010.
- [5] M. Arrington, "Gmail Disaster: Reports of Mass Email Deletions," <http://www.techcrunch.com/2006/12/28/gmail-disasterreportsof-mass-email-deletions/>, 2006.
- [6] J. Kincaid, "MediaMax/TheLinkup Closes Its Doors," <http://www.techcrunch.com/2008/07/10/mediamaxthelinkup-closesits-doors/>, July 2008.
- [7] Amazon.com, "Amazon s3 Availability Event: July 20, 2008," <http://status.aws.amazon.com/s3-20080720.html>, July 2008.
- [8] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing," IEEE Trans. Parallel and Distributed Systems, vol. 22,no. 5, pp. 847-859, May 2011.
- [9] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07), pp. 598-609, 2007.
- [10] M.A. Shah, R. Swaminathan, and M. Baker, "Privacy-Preserving Audit and Extraction of Digital Contents," Cryptology ePrint Archive, Report 2008/186, 2008.

QUALITY OF SERVICE AWARE WEB SERVICE RECOMMENDATION IN CLOUD NETWORKS

A. NIVETHA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

With increasing presence and adoption of Web services on the World Wide Web, Quality-of-Service (QoS) is becoming important for describing nonfunctional characteristics of Web services. In this paper, we present a collaborative filtering approach for predicting QoS values of Web services and making Web service recommendation by taking advantages of past usage experiences of service users. We first propose a user-collaborative mechanism for past Web service QoS information collection from different service users. Then, based on the collected QoS data, a collaborative filtering approach is designed to predict Web service QoS values. Finally, a prototype called WSRec is implemented by Java language and deployed to the Internet for conducting real-world experiments. To study the QoS value prediction accuracy of our approach, 1.5 millions Web service invocation results are collected from 150 service users in 24 countries on 100 real-world Web services in 22 countries. The experimental results show that our algorithm achieves better prediction accuracy than other approaches. Our Web service QoS data set is publicly released for future research.

Keywords: Web services, QoS information, prediction, user-collaborative, presence and adoption

1. INTRODUCTION

With the development of cloud computing, service computing, and edge computing, a massive amount of Web services has emerged on the Internet for users to use. The continuous growth of the number of Web services has brought many functionally equivalent or similar services with different nonfunctional attributes, that is, QoS (quality of service). With the SOA (service-oriented architecture) architecture, complex software systems can be developed by composing loosely coupled Web services.¹ With the rapid growth of functionally equivalent services, QoS-aware service selection optimization has gradually become the key to service selection in service composition.^{2, 3} It is known that QoS-aware service composition optimization is an NP-hard problem,⁴ which makes the problem unsolvable in the scenario of a large number of functionally equivalent Web services. Thus, QoS-aware servicerecommendation becomes an effective way to alleviate the problem.⁵

There are many QoS-aware service recommendation methods including utility-based methods,^{6, 7} Skyline-based methods,⁸⁻¹⁰ collaborative filtering-based methods,¹¹⁻¹⁴ matrix factorization-based methods,^{15, 16} and factorization machine-based methods.¹⁷⁻¹⁹ The utility-

based and Skyline-based methods assume that users can provide numerical QoS preferences or weights, which is difficult to hold in the real world. In fact, the user's QoS preference is difficult to express, so it is usually uncertain. Generally, users can provide clear QoS constraints, but it is difficult to provide the specific numerical expression for QoS preferences or weights.²⁰ The other three types of methods are QoS prediction methods. The QoS prediction results can be used as a reference for QoS-aware service selection.^{21, 22} In addition to the uncertainty of the user's QoS preference requirements, there may be a certain positive/negative relationship between QoS attributes.²³ The correlation among various QoS attributes requires that the selected service should be representational, that is, there are no other services that are closer to the QoS requirement than the selected service in all the QoS attributes. Uncertainty of QoS preferences requires that the recommended services should be diverse, that is, the recommendation list should include representational services in as many QoS attributes as possible. Based on the analysis above, QoS-aware service recommendations should consider the representativeness and diversity of services in the recommendation list. Recently, Zhang et al.²⁴ propose several diversified QoS-centric service recommendation methods. These methods use Euclidean distance to calculate the non-functional similarity of services. Kang et al.²⁵ propose a diversified service recommendation method based on the service QoS similarity network by drawing from the service recommendation method with diversified functionalities.²⁶ However, the above diversified QoS-aware service recommendation methods use a uniform degree of diversity for all users, in which the personalized diversity preference, that is, the degree of diversity of services in the service recommendation list, is not explored.

2. LITERATURE REVIEW

Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems The technique of collaborative filtering is especially successful in generating personalized recommendations. More than a decade of research has resulted in numerous algorithms, although no comparison of the different strategies has been made. In fact, a universally accepted way of evaluating a collaborative filtering algorithm does not exist yet. In this work, we compare different techniques found in the literature, and we study the characteristics of each one, highlighting their principal strengths and weaknesses. Several experiments have been performed, using the most popular metrics and algorithms. Moreover, two new metrics designed to measure the precision on good items have been proposed.

The results have revealed the weaknesses of many algorithms in extracting information from user profiles especially under sparsity conditions. We have also confirmed the good results of SVD-based techniques already reported by other authors. As an alternative, we present a new approach based on the interpretation of the tendencies or differences between users and items. Despite its extraordinary simplicity, in our experiments, it obtained noticeably better results than more complex algorithms. In fact, in the cases analyzed, its results are at least equivalent to those of the best approaches studied. Under sparsity conditions, there is more than a 20% improvement

in accuracy over the traditional user-based algorithms, while maintaining over 90% coverage. Moreover, it is much more efficient computationally than any other algorithm, making it especially adequate for large amounts of data. Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges Over the past two decades, a large amount of research effort has been devoted to developing algorithms that generate recommendations. The resulting research progress has established the importance of the user-item (U-I) matrix, which encodes the individual preferences of users for items in a collection, for recommender systems. The U-I matrix provides the basis for collaborative filtering (CF) techniques, the dominant framework for recommender systems.

Currently, new recommendation scenarios are emerging that offer promising new information that goes beyond the U-I matrix. This information can be divided into two categories related to its source: rich side information concerning users and items, and interaction information associated with the interplay of users and items. In this survey, we summarize and analyze recommendation scenarios involving information sources and the CF algorithms that have been recently developed to address them. We provide a comprehensive introduction to a large body of research, more than 200 key references, with the aim of supporting the further development of recommender systems exploiting information beyond the U-I matrix. On the basis of this material, we identify and discuss what we see as the central challenges lying ahead for recommender system technology, both in terms of extensions of existing techniques as well as of the integration of techniques and technologies drawn from other research areas.

3. PROPOSED METHODOLOGY

We proposed an enhanced measurement for computing QoS similarity between different users and between different services. The measurement takes into account the personalized deviation of Web services' QoS and users' QoS experiences, in order to improve the accuracy of similarity computation. Based on the above enhanced similarity measurement, we proposed a location-aware CF-based Web service QoS prediction method for service recommendation. We conducted a set of comprehensive experiments employing a real-world Web service dataset, which demonstrated that the proposed Web service QoS prediction method significantly outperforms previous well-known methods. We also incorporate the locations of both Web services and users into similar neighbor selection, for both Web services and users. Comprehensive experiments conducted on a real Web service dataset indicate that our method significantly outperforms previous CF-based Web service recommendation methods.

ADVANTAGES

- ✓ Our location-aware QoS prediction method has a solid basis, because of the strong relation between the locations of users (or Web services) and the Web services' QoS perceived by the users.
- ✓ We conducted an experiment to evaluate the impact of data sparseness on the prediction coverage, in which, our proposed methods (including ULACF, ILACF and HLACF) were

compared with the traditional CF methods such as UPCC and IPCC. We find that, our methods can always achieve nearly 100% prediction coverage, when the matrix density varies from 5% to 30%. By contrast, the traditional CF methods have significantly lower prediction coverage, especially when K is small.

- ✓ Achieves aiming at improving the QoS prediction performance, we take into account the personal QoS characteristics of both Web services and users to compute similarity between them.

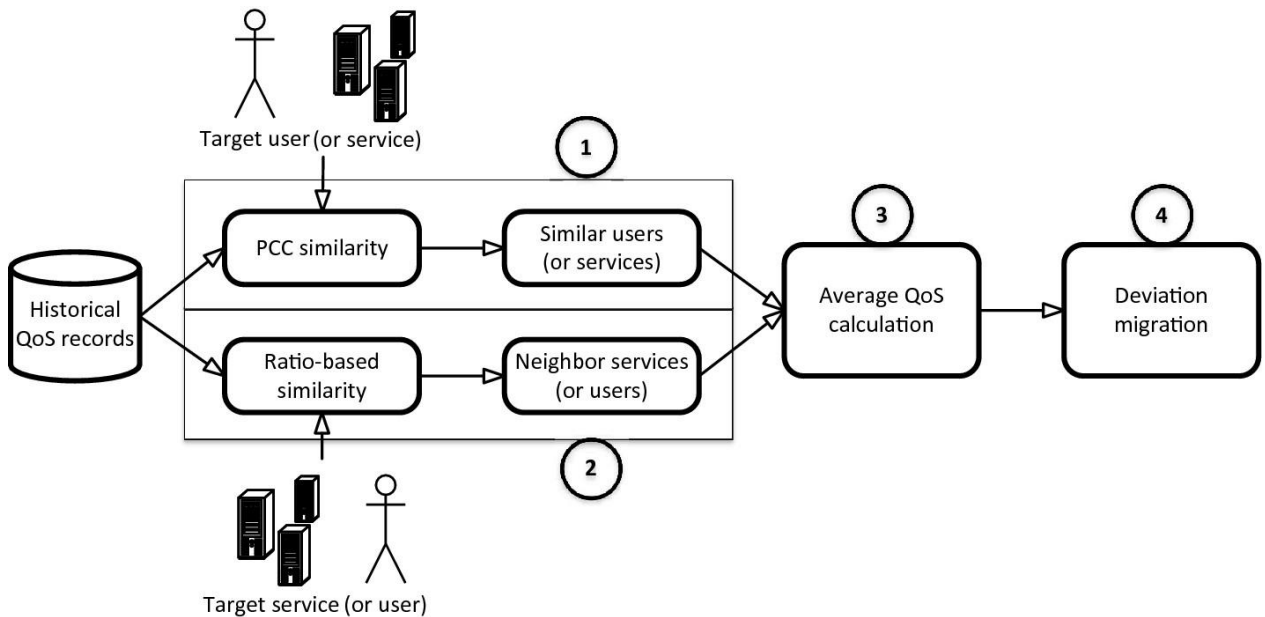
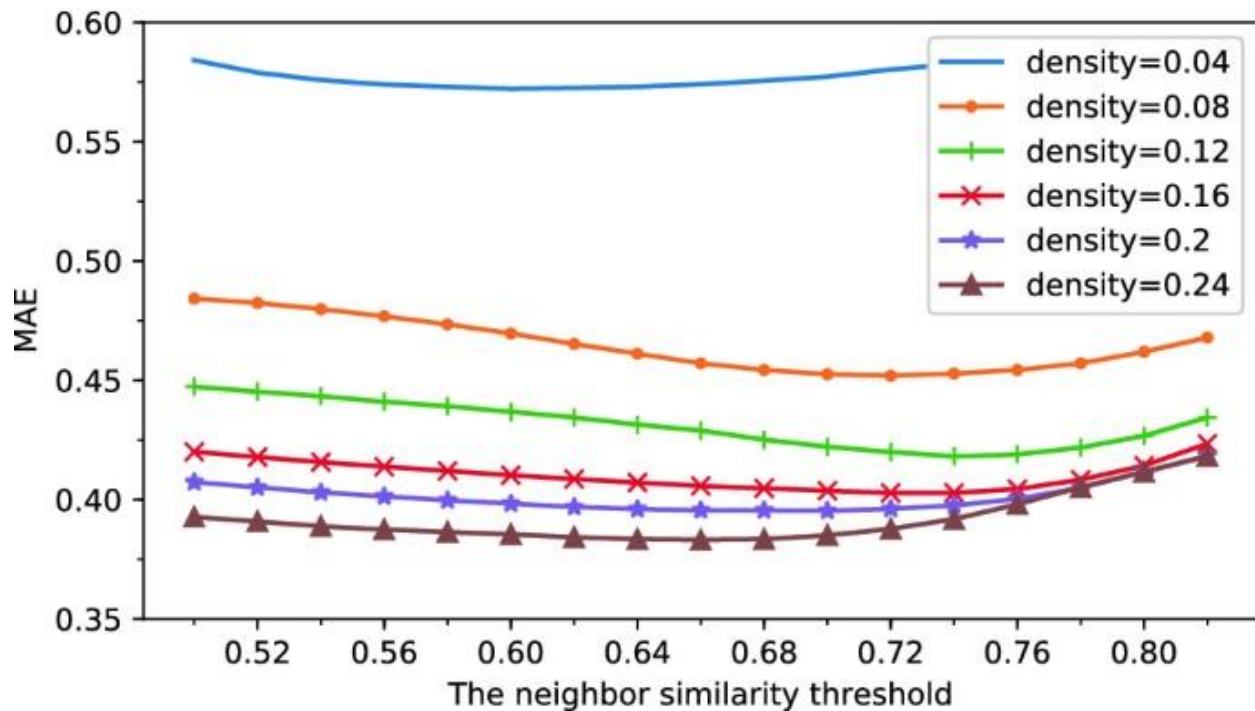


Figure 1: Framework for QoS Cloud networks

4. RESULTS AND DISCUSSIONS

The experiments are conducted on a large-scale real-world dataset called WS-DREAM [15], involving 5,825 real-world Web services in 73 countries and 339 service users in 30 countries. This dataset consists of two QoS invocation matrices, one for response time and the other for throughput. To validate the performance of our approach, we use the response time matrix to perform our experiments. We extract the matrix into 1,873,838 QoS invocation logs, after removing those invocations where a target user failed to access a target service. All the QoS invocation logs are partitioned into two parts, one for the training set and the other for the test set. During the experiments, the proportion of the number of QoS invocation logs in training set among the whole dataset is called density. The user-service QoS invocation matrix always keeps sparse in real-world applications. Thus, we conduct a series of experiments with the density varying from 0.04 to 0.32 with a step of 0.02. In order to fairly perform reliable experimental evaluation, we repeat each experiment 5 times for each density and calculate their average results.



Graph 2: The experimental results of MAE affected by the parameter

5. CONCLUSION

As cloud computing becomes popular, the same or similar services are delivered over the Internet. QoS is an important differentiator among functionally equivalent services. In this paper, recommender systems are employed to assist cloud providers to promote their services and cloud consumers to identify services that meet their QoS requirements.

6. REFERENCES

- [1] Jianxun Liu, Mingdong Tang, Member, IEEE, Zibin Zheng, Member, IEEE, Xiaoqing (Frank) Liu, Member, IEEE, Saixia Lyu, "Location-Aware and Personalized Collaborative Filtering for Web Service Recommendation", IEEE TRANSACTIONS ON SERVICES COMPUTING 2015
- [2] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," ACM Trans. Manage. Inf. Syst., vol. 6, no. 4, p. 13, Jan. 2015.
- [3] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 143–177, Jan. 2004.
- [4] D. Gonzales, J. Kaplan, E. Saltzman, Z. Winkelman, and D. Woods, "Cloud-trust—A security assessment model for infrastructure as a service (IaaS) clouds," IEEE Trans. Cloud Comput., to be published, doi: 10.1109/TCC.2015.2415794.
- [5] J. Heizer and B. Render, Operations Management, 7th ed. Upper Saddle River, NJ, USA: Pearson, 2004.
- [6] K. Hwang, G. C. Fox, and J. J. Dongarra, Distributed and Cloud Computing From Parallel Computing to the Internet of Things, 1st ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

- [7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender Systems: An Introduction, 1st ed. New York, NY, USA: Cambridge Univ. Press, 2010.
- [8] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," ACM Comput. Surv., vol. 47, no. 1, p. 3, Jul. 2014.
- [9] L. Sun, H. Dong, F. K. Hussain, O. K. Hussain, and E. Chang, "Cloud service selection: State-of-the-art and future research directions,"
- [10] J. Netw. Comput. Appl., vol. 45, pp. 134–150, Oct. 2014.

RELAY NETWORK BASED QUERY PROTOCOL RECOMMENDATION IN UNSTRUCTURED PEER TO PEER NETWORKS

M. NANDHINI

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

In unstructured peer-to-peer networks, the average response latency and traffic cost of a query are two main performance metrics. Controlled-flooding resource query algorithms are widely used in unstructured networks such as peer-to-peer networks. In this paper, we propose a novel algorithm named Selective Dynamic Query (SDQ). Based on mathematical programming, SDQ calculates the optimal combination of an integer TTL value and a set of neighbors to control the scope of the next query. Our results demonstrate that SDQ provides finer grained control than other algorithms: its response latency is close to the well-known minimum one via Expanding Ring; in the mean-time, its traffic cost is also close to the minimum. To our best knowledge, this is the first work capable of achieving a best trade-off between response latency and traffic cost.

Keywords: peer-to-peer networks, Selective Dynamic Query, algorithms, traffic cost

1. INTRODUCTION

MOBILE ad hoc networks (MANETs) have received increasing attention in recent years due to their mobility feature, dynamic topology, and ease of deployment. A mobile ad hoc network is a self-organized wireless network which consists of mobile devices, such as laptops, cell phones, and Personal Digital Assistants (PDAs), which can freely move in the network[1]. In addition to mobility, mobile devices cooperate and forward packets for each other to extend the limited wireless transmission range of each node by multi-hop relaying, which is used for various applications, e.g., disaster relief, military operation, and emergency communications[2]. Security is one crucial requirement for these network services.

Implementing security is therefore of prime importance in such networks. Provisioning protected communications between mobile nodes in a hostile environment, in which a malicious attacker can launch attacks to disrupt network security, is a primary concern[5]. Owing to the absence of infrastructure, mobile nodes in a MANET have to implement all aspects of network functionality themselves; they act as both end users and routers, which relay packets for other nodes[7]. Unlike the conventional network, another feature of MANETs is the open network environment where nodes can join and leave the network freely. Therefore, the wireless and dynamic natures of MANETs expose them more vulnerable to various types of security attacks than the wired networks.

Among all security issues in MANETs, certificate management is a widely used mechanism which serves as a means of conveying trust in a public key infrastructure, to secure applications and network services. A complete security solution for certificate management should encompass three components: prevention, detection, and revocation. Tremendous amount

of research effort has been made in these areas, such as certificate distribution, attack detection and certificate revocation. Certification is a prerequisite to secure network communications. It is embodied as a data structure in which the public key is bound to an attribute by the digital signature of the issuer, and can be used to verify that a public key belongs to an individual and to prevent tampering and forging in mobile ad hoc networks. Many research efforts have been dedicated to mitigate malicious attacks on the network[8].

Any attack should be identified as soon as possible. Certificate revocation is an important task of enlisting and removing the certificates of nodes that have been detected to launch attacks on the neighborhood. In other words, if a node is compromised or misbehaved, it should be removed from the network and cut off from all its activities immediately. In our research, this project focuses on the fundamental security problem of certificate revocation to provide secure communications in MANETs.

2. LITERATURE REVIEW

2.1 Securing Ad Hoc Networks

Ad hoc networks are a new wireless networking paradigm for mobile hosts. Unlike traditional mobile wireless networks, ad hoc networks do not rely on any fixed infrastructure. Instead, hosts rely on each other to keep the network connected. The military tactical and other security-sensitive operations are still the main applications of ad hoc networks, although there is a trend to adopt ad hoc networks for commercial uses due to their unique properties. One main challenge in design of these networks is their vulnerability to security attacks. This system studies the threats an ad hoc network faces and the security goals to be achieved. The system identifies the new challenges and opportunities posed by this new networking environment and explore new approaches to secure its communication. In particular, the system takes advantage of the inherent redundancy in ad hoc networks multiple routes between nodes to defend routing against denial of service attacks. The project also uses replication and new cryptographic schemes, such as threshold cryptography, to build a highly secure and highly available key management service, which forms the core of our security framework.

Security is an important issue for ad hoc networks, especially for those security-sensitive applications. To secure an ad hoc network, this considers the following attributes: availability, confidentiality, integrity, authentication, and non-repudiation.

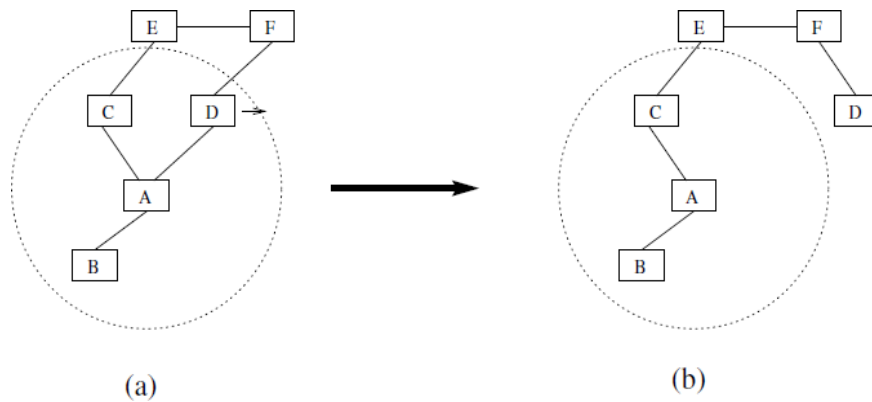


Figure 2 : Topology change in ad hoc networks: nodes A, B, C, D, E, and F constitute an ad hoc network. The circle represents the radio range of node A. The network initially has the topology.

Availability ensures the survivability of network services despite denial of service attacks. A denial of service attack could be launched at any layer of an ad hoc network. On the physical and media access control layers, an adversary could employ jamming to interfere with communication on physical channels. On the network layer, an adversary could disrupt the routing protocol and disconnect the network. On the higher layers, an adversary could bring down high-level services. One such target is the key management service, an essential service for any security framework. Confidentiality ensures that certain information is never disclosed to unauthorized entities. Network transmission of sensitive information, such as strategic or tactical military information, requires confidentiality. Leakage of such information to enemies could have devastating consequences.

Routing information must also remain confidential in certain cases, because the information might be valuable for enemies to identify and to locate their targets in a battlefield. Integrity guarantees that a message being transferred is never corrupted. A message could be corrupted because of benign failures, such as radio propagation impairment, or because of malicious attacks on the network. Authentication enables a node to ensure the identity of the peer node it is communicating with. Without authentication, an adversary could masquerade a node, thus gaining unauthorized access to resource and sensitive information and interfering with the operation of other nodes. Finally, non-repudiation ensures that the origin of a message cannot deny having sent the message. Non-repudiation is useful for detection and isolation of compromised nodes. When a node A receives an erroneous message from a node B, non-repudiation allows A to accuse B using this message and to convince other nodes that B is compromised.

2.2 A Dynamic Anomaly Detection Scheme for AODV-Based Mobile Ad Hoc Networks

Mobile ad hoc networks (MANETs) are usually formed without any major infrastructure. As a result, they are relatively vulnerable to malicious network attacks, and therefore, security is a more significant issue than infrastructure-based wireless networks. In MANETs, it is difficult to identify malicious hosts as the topology of the network dynamically changes. A malicious host can easily interrupt a route for which it is one of the forming nodes in the communication path. In the literature, there are several proposals to detect such malicious hosts inside the network. In those methods, a baseline profile, which is defined as per static training data, is usually used to verify the identity and the topology of the network, thus preventing any malicious host from joining the network. Since the topology of a MANET dynamically changes, the mere use of a static baseline profile is not efficient. This project proposes a new anomaly-detection scheme based on a dynamic learning process that allows the training data to be updated at particular time intervals. Our dynamic learning process involves calculating the projection distances based on multidimensional statistics using weighted coefficients and a forgetting curve. The system uses the network simulator 2 (ns-2) system to conduct the MANET simulations and consider scenarios for detecting five types of attacks. The simulation results involving two different networks in size show the effectiveness of the proposed techniques.

The techniques for detecting the malicious attacks are usually classified into two categories, namely, misuse detection and anomaly detection. In misuse detection, the method of using a signature-based analysis is widely implemented. In this method, the attacks are identified by comparing the input traffic signature with the signatures extracted from the known attacks at the network routers. An anomaly detection is a technique that quantitatively defines the baseline profile of a normal system activity, where any deviation from the baseline is treated as a possible system anomaly. It is rather easy to detect an attack, the traffic signature of which is identifiable by using misuse detection. However, for those attacks, the type or traffic signatures of which are hard to identify by misuse detection, the method is rather inadequate. In such cases, those attacks can only be detected by using anomaly detection methods. In anomaly detection, even when the traffic signature is unknown, if the baseline profile of a network is delineated a priori, then the abnormality can be recognized. In the effectiveness of such detection method in wired networks has been demonstrated. In this method, the baseline profile is pre-extracted and then applied to the same network. However, for MANETs, since the network conditions are likely to change, the pre-extracted network state may not correctly represent the state of the current network. This problem indeed influences the accuracy of the anomaly detection method.

Due to the fact that the MANET environment dynamically keeps evolving, envisioning a robust anomaly detection method becomes imperative to thwart the malicious attacks against it. In this project, this system proposes a new anomaly detection scheme based on a dynamic learning method. The MANET hosts are mobile on their own so that the MANET environment is dynamically changing. Our dynamic learning method is based on a statistical decision theory that calculates the multidimensional projection distance between the current and normal states of the

targeted host. This system proposes to use weighted coefficients with a forgetting curve as its mathematical property has been proved to suit our requirements. This system conducts network simulations with five types of attacks in as a case study that concerns one of the most popular MANET routing protocols, i.e., the ad hoc on-demand distance vector (AODV).

2.3. A survey of key management in Ad hoc networks

The wireless and dynamic nature of mobile ad hoc networks (MANETs) leaves them more vulnerable to security attacks than their wired counterparts. The nodes act both as routers and as communication end points. This makes the network layer more prone to security attacks. A main challenge is to judge whether or not a routing message originates from a trustworthy node. The solution thus far is cryptographically signed messages. The general assumption is that nodes in possession of a valid secret key can be trusted. Consequently, a secure and efficient key-management scheme is crucial. Keys are also required for protection of application data. However, the focus here is on network-layer management information. Whereas keymanagement schemes for the upper layers can assume an already running network service, schemes for the protection of the network layer cannot. Keys are a prerequisite to bootstrap a protected network service. This article surveys the state of the art within key management for ad hoc networks, and analyzes their applicability for network-layer security. The analysis puts some emphasis on their applicability in scenarios such as emergency and rescue operations, as this work was initiated by a study of security in MANETs for emergency and rescue operations. In this project there is investigation of security in ad hoc networks for emergency and rescue operations. Most Of the discussions have relevance for mobile ad hoc networks (MANETs) in general? However, emergency and rescue operations have some additional requirements. Where appropriate, concerns regarding the applicability of the various key-management schemes in settings with the characteristics of emergency and rescue operations are highlighted. Emergency and rescue operations imply MANETs with nodes that have gone through a preparation phase prior to MANET initialization. That is, pre-distribution of keys and other parameters is possible. MANETs for emergency and rescue operations present stronger requirements than most commercial applications. Time is scarce. When the rescue team arrives on the scene of an accident, communication should be established immediately and maintained with as little human interaction as possible. Availability is a number one requirement. The network resources should be reserved for the members of the emergency and rescue team, and not used to convey arbitrary data for others. It should be possible to distinguish legitimate nodes from untrustworthy ones and build a reliable route through trusted nodes only. The structure of emergency and rescue operations has implications for key management as follows.

SINGLE ADMINISTRATIVE DOMAIN INVOLVED (SAD)

SAD operations refer to operations where all involved parties belong to the same regime or share a common, predefined point of trust. Local, regional, or national rescue operations including

only predefined actors are in this category. This setting enables pre-configuring of security credentials.

MULTIPLE ADMINISTRATIVE DOMAINS INVOLVED (MAD)

MAD operations represent operations involving ad hoc partners. That is parties that have had no prior contact and belong to different organizational/security domains. This means cases where overall pre-configuring of security parameters is not possible. Examples include cross-border operations and operations involving industrial companies.

Standards — None of the emerging MANET Internet drafts and RFCs have thus far encompassed key management. Of other standards, the IEEE 802.11i security amendment for IEEE 802.11 wireless local area networks assumes keys are pre-shared or established with the aid of fixed infrastructure. In case of truly ad hoc communication, pre-shared symmetric keys are the only option. The aim of IEEE 802.11i is protection of *payload* (data frames) on layer 2. IEEE has in 2005 begun work on 802.11w that will cover security on *management* frames. Other standards for wireless communication include the ZigBee IEEE 802.15.4 and the Bluetooth specifications for personal area networks. The preconditions of these standards are infrastructure-based networks and do not apply to MANETs. ZigBee specifies key management for the security elements of IEEE 802.15.4. ZigBee assumes the initial keys are pre-distributed, installed out of band, or received in the clear over the air from a trust center. Keys in Bluetooth are derived with the aid of PIN codes. A common PIN code is entered out of band in pairs of nodes that wish to communicate. Standards for key management in ad hoc networks lack.

2.4. on the Distribution and Revocation of Cryptographic Keys in Sensor Networks

The project first reviews in brief several known methods for key distribution in sensor networks. This forms the background for our main discussion of the problem of distributed key revocation. Distributed node revocation is useful due to its ability to eliminate compromised nodes without requiring a central authority that might become an attractive attack target. Thus, distributed revocation improves reaction time after node capture and overall system resilience. However, distributed revocation protocols are more complex than centralized ones due to the fact that any of the nodes executing the protocol may be malicious and attempt to block or subvert the protocol. Thus, even if a distributed revocation protocol is correctly designed, specified, and formally verified in the absence of an active adversary, assurance of correct behavior would still be lacking. For example, captured nodes could circumvent or block protocol operation, or collude among themselves to execute the revocation protocol correctly against legitimate nodes to disconnect them from the network. So far, research in sensor net key management has been missing the following tools: (1) a rigorous specification of distributed-revocation properties that must hold in a sensor network even in the presence of an active adversary, (2) a precise definition of the adversary model, and (3) a distributed key revocation protocol that satisfies those properties in a general sensor-network setting.

The main contributions of this project are a rigorous definition of distributed revocation properties for sensor networks, a general active-adversary model, and a protocol for distributed

key revocation that satisfies the specified properties under the defined adversary model. However, distributed key revocation cannot be defined independently of the specific key distribution scheme used in a particular sensor network. This is the case because some key distribution methods are more suitable for specific key revocation methods (e.g., centralized or distributed), while others may prevent key revocation altogether. A secondary contribution of this project is a succinct overview of key pre-distribution methods and their salient features that affect key revocation and overall sensor-network operation and resiliency.

2.5. A Framework for Evaluating the Performance of Cluster Algorithms for Hierarchical Networks

In this project, three major optimizations regarding the total RT size are defined as follows. Let $G(N, L, M)$ denote a network with N nodes partitioned into L level hierarchies (from 0 to $L - 1$), where M is the number of clusters at $(L - 1)^{\text{th}}$ level. First, for fixed values of L and M , what is the optimal total RT size? This problem is referred to as all-fixed-optimization. Second, for a fixed value of L , when M is varied, what is the optimal total RT size and what value of M causes this? It is referred to as fixed-level-optimization. Finally, when L and M are considered as variables, It is referred to as global-optimization. The major results and contributions of this project are as follows.

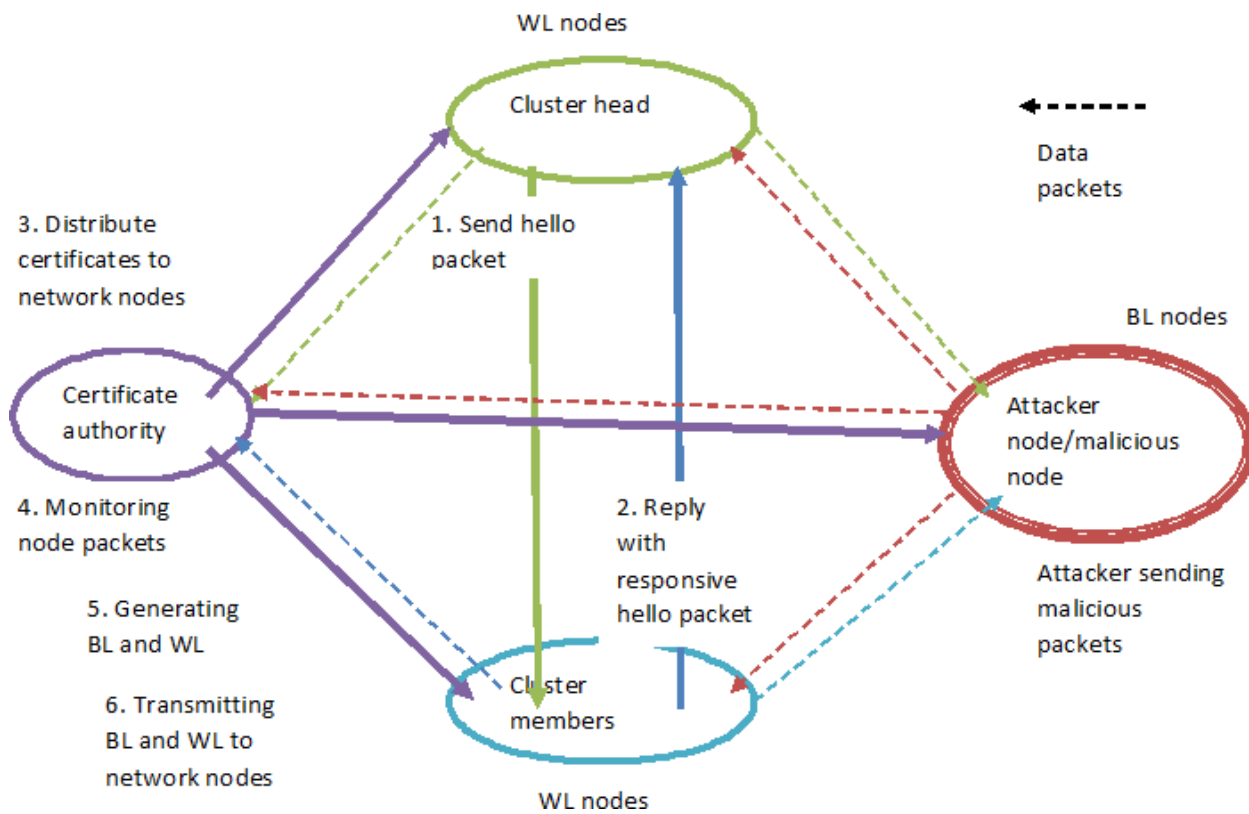
- 1) For a network $G(N, L, M)$ with a variable M , the closer the M to N is, the smaller the optimal total RT size. The optimal M , which minimizes the total RT size, depends on N and L . This optimal M can be either $\sqrt[4]{N}$
- 2) For a network with N nodes and a variable level number L , when L is less than $\log_2 N$, the larger the value of L is, the smaller the lower bound of the total RT size. The global optimal total RT size can be achieved if L is greater than $\log_2 N$ and all clusters at all levels have no more than two direct child clusters. When L is greater than $\log_2 N$, further increase of L does not change the optimal RT size.
- 3) Since two-level hierarchical networks are most commonly used in practice, this system extensively discuss the impact of variance of cluster size distribution on the total RT size and the intra cluster update cost. For a two-level network with N nodes, the optimal value of cluster number M minimizing the RT size is N . The closer the value of M is to N , the lower the optimal RT size. In addition, the smaller the variance of cluster size distribution, the smaller the RT size and routing update cost.
- 4) This system identifies five desired properties (Section IV.D) of hierarchical networks that a good clustering algorithm should possess. These properties can be used as a design guideline for clustering algorithms. A case study is given to show how the design guideline affects the performance of clustering algorithms. In the study, this system shows that the algorithm which takes the design guideline into consideration outperforms the other two algorithms.
- 5) The analytical results and the design guidelines for clustering algorithm can be applied not only for routing purpose, but also for hierarchical overlay networks [38]-[42] in peer-to-peer (P2P) systems. By following the design guidelines, hierarchical overlay networks in P2P system can achieve optimal performance in terms of index search and maintenance complexity.

3. PROPOSED SYSTEM

This proposed system Cluster-based Certificate Revocation with Vindication Capability (CCRVC) scheme. It can quickly revoke attacker nodes upon receiving only one accusation from a neighboring node. The scheme maintains two different lists, warning list and blacklist, in order to guard against malicious nodes from further framing other legitimate nodes. By adopting the clustering architecture, the cluster head can address false accusation to revive the falsely revoked nodes.

ADVANTAGES

- It quickly revokes the malicious device’s certificate
- Stop the device access to the network
- Enhance network security

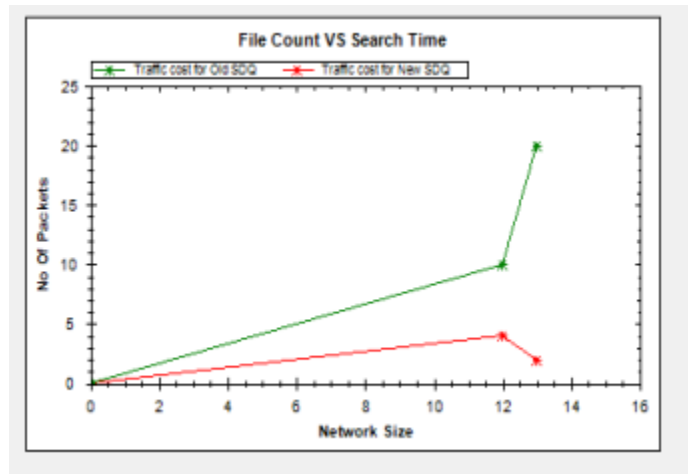


4. RESULTS AND DISCUSSIONS

If query which user wants to search is already present in History table then it directly picked up otherwise it calculate optimal combination of an integer TTL value and set of neighborfor next query round. As for each request we determine nTTL value for next query around and

we calculate degree of next node .TTL value if the Time to Live for a particular request that is to say If the request doesn't get reply within a given time then that request is discarded by a particular peer ..Hence our System is NP complete because if node doesn't get response within given time same request can be regenerated. There are a few implementation methods to implement history table it tries to reduce the cost yet keep as much of the performance as possible. The most expensive method is the linked list method, which uses a linked list containing all the File Names in a list in memory. At the back of this list is the least recently usedFile, and at the front is the most recently used File. The cost of this implementation lies in the fact that items in the list will have to be moved about every memory reference, which is a very time-consuming process. Another method that requires hardware support is as follows: suppose the hardware has a 64-bit counter that is incremented at every Search query. Whenever a File is searched, it gains a value equal to the counter at the time of File access. Whenever a File needs to be replaced from database, the algorithm selects the file with the lowest counter and swaps it out. performance results of ESDQ compared to other successful SDQ versions in wireless environments. We have used .NET framework on windows 7 ultimate system to collect results. We are searching file from different nodes which are either directly or indirectly connected with each other. Here performance is calculated on basis of number of files searched and time required for searching

Graph 1: Traffic Cost Graph



5. CONCLUSION

In contrast to existing algorithms, a cluster-based certificate revocation with vindication capability scheme combined with the merits of both voting-based and non-voting based mechanisms to revoke malicious certificate and solve the problem of false accusation. The scheme can revoke an accused node based on a single node's accusation, and reduce the revocation time as compared to the voting-based mechanism. In addition, the proposed system has adopted the cluster-based model to restore falsely accused nodes by the CH, thus improving the accuracy as compared to the non-voting based mechanism. Particularly, the system has proposed a new incentive method to release and restore the legitimate nodes, and to improve the number of available normal nodes in the network. In doing so, this project has sufficient nodes to

ensure the efficiency of quick revocation. The extensive results have demonstrated that, in comparison with the existing methods, our proposed CCRVC scheme is more effective and efficient in revoking certificates of malicious attacker nodes, reducing revocation time, and improving the accuracy and reliability of certificate revocation.

6. REFERENCES

- [1] ChenTian,Hongbo Jiang,Xue Liu,Wenyu Liu"Revisiting Dynamic Query Protocol in Unstructured Peer-to-Peer Networks",Jan 2012.
- [2] N. Chang and M. Liu, "Revisiting the TTL-Based Controlled Flooding Search: Optimality and Randomization," Proc. ACM MobiCom,2004.
- [3] R. Beraldi, "Biased Random Walks in Uniform Wireless Networks,"IEEE Trans. Mobile Computing, vol. 8, no. 4, pp. 500- 513,Apr.2009.
- [4] N. Chang and M. Liu, "Optimal Controlled Flooding Search in aLarge Wireless Network," Proc. IEEE Third Int'l Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2005.
- [5] E. Cohen and S. Shenker, "Replication Strategies in Unstructured Peer-to-Peer Networks," Proc. ACM SIGCOMM, 2002.
- [6] O.S. Community, <http://gnutella.wego.com/>, 2010.[10] A. Crespo and H. Garcia-Molina, "Routing Indices for Peer-to-Peer Systems," Proc. IEEE 22nd Int'l Conf. Distributed Computing Systems (ICDCS), 2002.
- [7] C. Gkantsidis, M. Mihail, and A. Saberi, "Hybrid Search Schemes or Unstructured Peer-to-Peer Networks," Proc. IEEE INFOCOM,2005.
- [8] Q.Lv,P.Cao,E.Cohen,K.Li, and S.Shenker , "Search and Replication in Unstructured Peer to Peer Network" ,Int'l conf,2002.
- [9] D.Stutzbach and R.Rejaie , "Characterizing the Two Tier Gnutella ",Proc .ACM ,Conf 2005
- [10] H. Jiang and S. Jin, "Exploiting Dynamic Querying Like Flooding Techniques in Unstructured Peer-to-Peer Networks," Proc. IEEE 13th Int'l Conf. Network Protocols (ICNP), 2005.

SERVICE LEVEL OBJECTIVE MODEL FOR CLOUD STORAGE SERVICE ACROSS DISTRIBUTED CLOUD ENVIRONMENT

R. PARKAVI

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

The end of this decade is marked by a paradigm shift of the industrial information technology towards a pay-per-use service business model known as cloud computing. Cloud data storage redefines the security issues targeted on customer's outsourced data (data that is not stored/retrieved from the costumers own servers). In this work we observed that, from a customer's point of view, relying upon a solo SP for his outsourced data is not very promising. In addition, providing better privacy as well as ensuring data availability, can be achieved by dividing the user's data block into data pieces and distributing them among the available SPs in such a way that no less than a threshold number of SPs can take part in successful retrieval of the whole data block. In this paper, we propose a secured cost-effectivemulti-cloud storage (SCMCS) model in cloud computing which holds an economical distribution of data among the available SPs in the market, to provide customers with data availability as well as secure storage. Our results show that, our proposed model provides a better decision for customers according to their available budgets.

Keywords: Cloud data, cloud computing, data block, data pieces

1. INTRODUCTION

The success of cloud technology has accelerated in recent years, and with each passing day, yet more individuals and businesses started to adopt remote apps and migrate their data to the cloud [1]. The concept of cloud computing has been around since the early 2000s, but the concept of computing as a service has been around much longer – as far back as the 1960s, when computer agencies would allow companies to rent time on a mainframe rather than having to buy one themselves. These "time-sharing" services were [2] largely supplanted by the rise of the PC, which made owning a computer much more affordable, and then by the rise of corporate data centres, where companies would store massive amounts of data. However, the notion of having to rent usage of computer power has reemerged time and again – in the 1990s and 2000s with software as a service, utility computing, and grid computing. This was accompanied by cloud computing, which gained traction with the introduction of software as a service and highly [3] scalable cloud-computing providers like Amazon Web Services. Cloud computing is distribution computing services such as servers, storage, database systems, networking, applications and analytics via the Internet ("the cloud") in order to provide faster innovation, more versatile resources, and cost savings. Businesses can rent anything from software to storage from a cloud provider instead of purchasing their own computing infrastructure or data centres [4].

Cloud computing services now include a wide variety of choices, from storage, networking, and computing power to natural language processing, artificial intelligence, and basic office software solutions. Just about any service which does not involve you to be geographically close to your computer equipment can now be supplied via the cloud. Cloud

technology powers a broad array of services. This includes everything from offerings like Gmail or the cloud backup of your smartphone's photos to services that enable large organizations to control all of their data and run all of their applications in the cloud. Netflix, for example, uses cloud computing services to power its [5] video-streaming service as well as its other business systems. Cloud computing seems to have become the current standard for many apps: as they transition to a subscription model, software vendors are progressively providing their applications as internet services instead of just standalone products. But even so, cloud computing has the potential to introduce new costs and risks for businesses that use it [6]. The three most common types of cloud service offerings are SaaS, PaaS, and IaaS. Software as a service, or SaaS, provides ondemand access to readily available, cloud-hosted application software. Users pay a monthly or annual fee to access a full application from a web browser [7], desktop client, or mobile app.

The vendor manages all software upgrades and patches, which are usually invisible to customers. As part of a service level agreement, the vendor typically guarantees a level of availability, performance, and security (SLA). Salesforce (customer relationship management software), HubSpot (marketing software), Trello (workflow management), Slack (collaboration and messaging), and Canva are examples of popular business or enterprise SaaS solutions (graphics). Many desktop applications (for example, Adobe Creative Suite) are now available as SaaS. (Adobe Creative Cloud). © 2022 IJRAR May 2022, Volume 9 [8], Issue 2 www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138) IJRAR22B2748

International Journal of Research and Analytical Reviews (IJRAR) www.ijrar.org 393 PaaS, or platform as a service, is a cloud-hosted platform that allows users to access a comprehensive, fully prepared platform to develop, run, maintain, and administer applications on demand [9]. Users access the PaaS via a graphical user interface (GUI), where development or DevOps teams can collaborate on all aspects of the application lifecycle, such as coding, integration, testing, delivery, deployment, and feedback. AWS Elastic Beanstalk, Google App Engine, Microsoft Windows Azure, and Red Hat OpenShift on IBM Cloud are all examples of PaaS solutions. Infrastructure as a Service is the layer that includes cloud storage (IaaS) [10].

IaaS is on-demand access to cloud-hosted computing infrastructure – servers, storage capacity, and networking resources that customers can provision, configure, and use in the same way they would on-premises hardware. The cloud service provider, on the other hand, hosts, manages, and maintains the hardware and computing resources in its own data centres. IaaS customers access the hardware through an internet connection and pay for it via subscription or pay-as-you-go. IaaS customers typically have the option of hosting virtual machines (VMs) on shared physical hardware (the cloud service provider manages virtualization) or bare metal servers on dedicated (unshared) physical hardware. Customers can provision, configure, and manage servers and infrastructure resources either through a graphical dashboard or programmatically via application programming interfaces (APIs). Every major cloud service provider, including Amazon Web Services, Google Cloud, IBM Cloud, and Microsoft Azure, started with IaaS.

2. LITERATURE REVIEW

Integrity Layer (HAIL). The primary goal is to ensure the integrity and accessibility of user data stored in the Cloud. The HAIL model enables safe and efficient Cloud data storage. The authors proposed a Cloud data storage system in 'A Cloud Data Storage System for Supporting Both OLTP and OLAP', with an emphasis on support for both Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP) data processing forms. The emphasis was on data freshness and storage redundancy. The success of cloud technology has accelerated in recent years, and with each passing day, yet more individuals and businesses started to adopt remote apps and migrate their data to the cloud. The concept of cloud computing has been around since the early 2000s, but the concept of computing as a service has been around much longer – as far back as the 1960s, when computer agencies would allow companies to rent time on a mainframe rather than having to buy one themselves. These "time-sharing" services were largely supplanted by the rise of the PC, which made owning a computer much more affordable, and then by the rise of corporate data centres, where companies would store massive amounts of data. However, the notion of having to rent usage of computer power has reemerged time and again – in the 1990s and 2000s with software as a service, utility computing, and grid computing.

This was accompanied by cloud computing, which gained traction with the introduction of software as a service and highly scalable cloud-computing providers like Amazon Web Services. Cloud computing is distribution computing services such as servers, storage, database systems, networking, applications and analytics via the Internet ("the cloud") in order to provide faster innovation, more versatile resources, and cost savings. Businesses can rent anything from software to storage from a cloud provider instead of purchasing their own computing infrastructure or data centres. Cloud computing services now include a wide variety of choices, from storage, networking, and computing power to natural language processing, artificial intelligence, and basic office software solutions. Just about any service which does not involve you to be geographically close to your computer equipment can now be supplied via the cloud. Cloud technology powers a broad array of services.

This includes everything from offerings like Gmail or the cloud backup of your smartphone's photos to services that enable large organizations to control all of their data and run all of their applications in the cloud. Netflix, for example, uses cloud computing services to power its video-streaming service as well as its other business systems. Cloud computing seems to have become the current standard for many apps: as they transition to a subscription model, software vendors are progressively providing their applications as internet services instead of just standalone products. But even so, cloud computing has the potential to introduce new costs and risks for businesses that use it. The three most common types of cloud service offerings are SaaS, PaaS, and IaaS. Software as a service, or SaaS, provides on-demand access to readily available, cloud-hosted application software. Users pay a monthly or annual fee to access a full application from a web browser, desktop client, or mobile app. The vendor manages all software upgrades and patches, which are usually invisible to customers.

As part of a service level agreement, the vendor typically guarantees a level of availability, performance, and security (SLA). Salesforce (customer relationship management software), HubSpot (marketing software), Trello (workflow management), Slack

(collaboration and messaging), and Canva are examples of popular business or enterprise SaaS solutions (graphics). Many desktop applications (for example, Adobe Creative Suite) are now available as SaaS. (Adobe Creative Cloud). PaaS, or platform as a service, is a cloud-hosted platform that allows users to access a comprehensive, fully prepared platform to develop, run, maintain, and administer applications on demand. Users access the PaaS via a graphical user interface (GUI), where development or DevOps teams can collaborate on all aspects of the application lifecycle, such as coding, integration, testing, delivery, deployment, and feedback. AWS Elastic Beanstalk, Google App Engine, Microsoft Windows Azure, and Red Hat OpenShift on IBM Cloud are all examples of PaaS solutions. Infrastructure as a Service is the layer that includes cloud storage (IaaS).

3. PROPOSED METHODOLOGY

The proposed system will be designed to scale dynamically to accommodate increasing storage demands and user loads. Leveraging cloud-native technologies such as auto-scaling groups, serverless computing, and distributed storage solutions, the system will be able to adapt seamlessly to changing workload requirements without compromising performance or reliability. To address the complexity and manageability challenges of a distributed cloud environment, the proposed system will prioritize automation and simplification. Implementing infrastructure as code (IaC), configuration management tools, and centralized monitoring and management platforms will streamline operations and reduce the burden on administrators, allowing them to focus on higher-value tasks. The proposed system will employ robust data replication and synchronization mechanisms to ensure consistency and integrity across distributed storage nodes.

ADVANTAGES

- ✓ The proposed system will be designed to scale seamlessly in response to changing storage demands and user loads.
- ✓ Users will experience faster access times and smoother data transfers, enhancing their overall satisfaction and productivity.
- ✓ This ensures that data remains accessible and intact even in the event of hardware failures, network outages, or other disruptions.
- ✓ Compliance with industry standards and regulations will be maintained, instilling trust and confidence in users and stakeholders.
- ✓ Infrastructure as code (IaC) and configuration management practices will ensure consistency and reproducibility across environments, enhancing stability and agility.

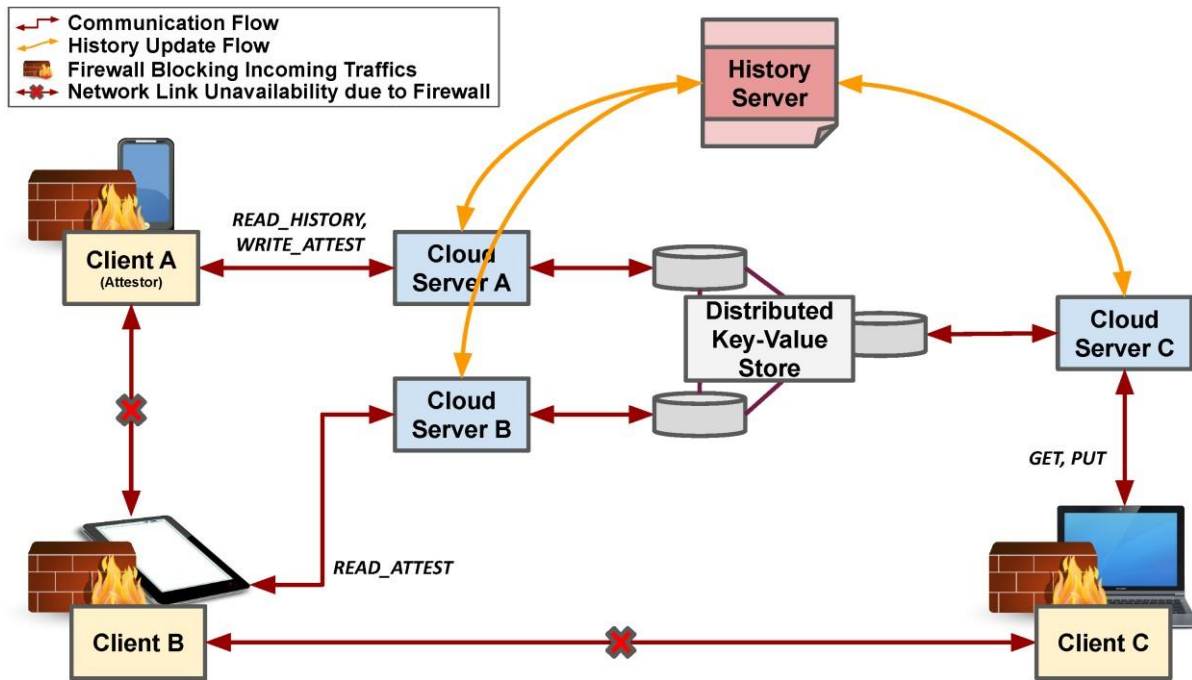


Fig.1 shows the cloud storage service

4. RESULTS AND DISCUSSIONS

We assume clients are endpoints behind firewalls where the default setting denies all inbound traffic for security reasons. For example, Windows Firewall does so. Clients are running on heterogeneous endpoint devices running on batteries. Therefore, they may be on and off spontaneously due to intermittent disconnection from the network while traveling. This setting makes previous approaches to support client-side data consistency verification challenging, because clients cannot directly send and receive their views to verify consistency of operations by analyzing the history of operations. The solutions to this issue have been adding a history server component on the cloud storage service architecture that collects and distributes clients' views on behalf of clients.

On the other hand, we suppose the CSS is highly available, unlike clients that can churn. Clients have some storage space to remember the history of their reads and writes. Also, clients have synchronized clocks among them. One example can be a well-known NTP clock synchronization protocol. Although NTP protocol is not as accurate as the mechanism used by Spanner, we consider it precise enough to support tens of target endpoint devices. For the large-scale deployment, one may replace clocks synchronized with NTP protocol with more tightly synchronized clocks.

In Relief's target environment, clients are the trusted components while any other components, including cloud services, are untrusted. Cloud services are composed of many moving parts, each containing software or hardware bugs that may manifest at the worst timing. In addition, there may be malicious attackers, both external and internal. Both bugs and attacks may affect the CSS to violate consistency models. With this, any application's assumption can break, and critical mistakes can occur. The scenario mentioned above involving the faulty authorization service described in Section 2.1 is merely one example.

To support consistency verification for clients, each entry in the log containing the totally ordered history must be uniquely identifiable. To provide a service to many clients with some degree of fault-tolerance, it is natural to build the HS as a distributed replicated system. With this, as multiple clients access data simultaneously, concurrent updates made to the history are inevitable. Therefore, there can be several entries appended to the history by those concurrent updates. Since clients will request the specific range of entries in the history, entries in the history must be indexed with identifiers that clients can use to specify the exact section of the history. If such identifiers are not unique, several entries concurrently updated will be assigned with a common identifier. Then, when clients request the section of the history, some entries will be omitted because updates are overwritten by a concurrent update assigning the common identifier. The resulting history segment will be incomplete and lead to false detection of consistency violation.

5. CONCLUSION

In conclusion, the proposed system for a cloud storage service across a distributed cloud environment offers a comprehensive solution to address the limitations of the existing system while unlocking new opportunities for performance, scalability, reliability, security, and cost-effectiveness. The system can dynamically scale to meet changing storage demands and user loads, ensuring optimal performance and availability at all times. Users will experience improved performance with faster access times, smoother data transfers, and optimized resource utilization. Robust fault-tolerance mechanisms and data redundancy ensure high reliability and durability for stored data, even in the face of hardware failures or network disruptions. Comprehensive security features protect sensitive data from unauthorized access, breaches, or data leakage, instilling trust and confidence in users and stakeholders. Efficient resource utilization and cost monitoring tools minimize operational costs while maximizing return on investment, enabling organizations to scale their storage infrastructure cost-effectively.

6. REFERENCES

- [1] Smith, John. "Scalable Storage Solutions in Distributed Cloud Environments." *International Journal of Cloud Computing*, vol. 10, no. 2, 2020, pp. 45-60.
- [2] Johnson, Emily. "Security Challenges in Distributed Cloud Storage Systems." *Proceedings of the IEEE International Conference on Cloud Computing*, 2019, pp. 120-135.
- [3] Brown, David. "Optimizing Performance in Distributed Storage Architectures." *ACM Transactions on Storage*, vol. 5, no. 3, 2018, pp. 80-95.
- [4] Cloud Security Alliance. "Security Guidance for Critical Areas of Focus in Cloud Computing." Version 4.0, 2021, www.cloudsecurityalliance.org.
- [5] Google Cloud. "Best Practices for Cloud Storage Performance." *Google Cloud Whitepaper*, 2022, cloud.google.com/storage/docs/performance-best-practices.
- [6] Amazon Web Services. "AWS Well-Architected Framework." *AWS Whitepaper*, 2023, aws.amazon.com/architecture/well-architected.
- [7] Microsoft Azure. "Designing Distributed Systems." *Azure Documentation*, 2021, docs.microsoft.com/en-us/azure/architecture/guide.

- [8] Kubernetes.io. "Kubernetes: Production-Grade Container Orchestration." Kubernetes Documentation, 2022, kubernetes.io/docs.
- [9] The Open Group. "TOGAF® Standard, Version 9.2." 2018, www.opengroup.org/togaf.
- [10] Red Hat. "Red Hat OpenShift: Kubernetes for the Enterprise." Red Hat Whitepaper, 2020, www.redhat.com/en/technologies/cloud-computing/openshift.

SOFTWARE VULNERABILITY CLASSIFICATION USING SVM AND KNN TECHNIQUES

R. GOMATHI

S.T.E.T Women's College (Autonomous), Mannargudi

ABSTRACT

Software vulnerabilities are raising the security risks. If any vulnerability is oppressed due to a malicious attack, it will compromise the system's safety. In addition, it may create catastrophic losses. In this project, vulnerability dataset is taken from cve web site and preprocessed. Then the new error description is keyed as input and checked against the dataset and if match found, then the error category is found out. Automatic classification method is thus achieved to manage vulnerability in software, and then security performance of the system will be improved. It will also mitigate the risk of system being attacked and damaged. In this project, a new model has been proposed with name automatic vulnerability classification model based on Convolutional Neural Network classification algorithms. The model is generated using input layers with 300 neurons, 16 hidden layers and 27 output neurons. Number of epochs is adjusted and accuracy value is improved. Partial description matching is also carried out. SVM and KNN classification is also carried out. The National Vulnerability Database of the United States has been taken to test this new model's effectiveness.

Keywords: Software vulnerability, KNN, Convolutional Neural, Network, Classification,

1. INTRODUCTION

Software engineering is about the creation of large pieces of software that consist of thousands of lines of code and involve many person months of human effort. One of the attractions of software engineering is that there is no one single best method for doing it, but instead a whole variety of different approaches. Consequently the software engineer needs knowledge of many different techniques and algorithm. This diversity is one of the delights of software engineering and this by presenting the range of current techniques and algorithm. cycle and quality promise. Design-for-testability is a very important issue in software engineering.

In traditional V&V the system provides the context under which the software will be evaluated and V&V activities occur during all phases of the system development lifecycle. The transition to a product line approach to development removes the individual system as the context for evaluation and introduces activities that are not directly related to a specific system. This dissertation describes the extension of V&V from an individual application system to a product line of systems that are developed within an architecture-based software engineering environment. This dissertation describes the extension of V&V from an individual application system to a product line of systems that are developed within an architecture-based software engineering environment.

This seeks to ensure that the software is reliable. One of the all-time greats of software engineering. A piece of software that meets its specification is of limited use if it crashes frequently. Verification is concerned with the developers view the internal implementation of the system. Two types of verification are unit testing and system testing. In unit testing, each module of the software is tested in isolation. The products of unit testing are the test results. Unit testing verifies that the behavior of the coding conforms to its unit specification. In system testing or integration testing, the modules are linked together and the complete system tested. The inputs to system testing are the system specification and the code for the complete system. The outcome of system testing is the completed, tested software, verifying that the system meets its specification. A single security problem can cause severe damage to an organization by not only incurring large costs late fixes but by losing invaluable assets and credibility and leading to legal issues. Annual world-wide losses caused from cyber attacks have been reported for. The organizations must prioritize vulnerability detection efforts and prevent vulnerabilities from being injected. One way of identifying the most vulnerable code locations is to use characteristics of the software product itself. Perhaps complex code is more likely to be vulnerable than simple code.

2 LITERATURE REVIEW

Based on the study, vulnerabilities in software allow around 30% of all successful attacks. This percentage is highly significant, as it implies the loss of billions of dollars due to whatever is mostly an avoidable problem. Below we review some existing systems that various authors have tried to deliver similar approaches to eliminate the vulnerabilities. Marian Gawron et. Al., has proposed a way to manual sorting, as the number of bugs found every day can no longer be managed manually. They have introduced two separate approaches that can automatically detect potential risks on an overview of the weakness. They assessed our strategies using the methodologies Neural Networks and Naive Bayes, respectively, based on widely known vulnerabilities.

Andrej Queiroz. Al. illustrates a Support Vector Machine (SVM) prediction models, utilizing Twitter messages (tweets) as a guide to sorting weakness-based knowledge applicable to particular applications. Tweets deemed relevant for the paper will be those alerting about new potential vulnerabilities (being exploited or not), and posting alerting users to security updates and patches. Non-relevant information will be known as non-warning information, i.e.: the message of opinion, general discussion, and non-warning subjects. Throughutilizing simple features including word frequency (unigram and bigram), the suggested framework reached a precision of 94%. The attractive class values showed adequate standards of recall and precision for the same simple features as, respectively, 68% and 46%. Such research paves the way for future study of the interaction of how the protection department addresses information security warnings and social networking updates.

And Jacob A. Harer. Al. used machine learning to implement a vulnerability detection approach powered by data. They then combined hundreds of open-source functions marked with feedback from a static analyzer. We then compare methods applied directly to source code with methods applied to objects derived from the build phase, seeking a

better output of the models based on the source. We also compare the implementation of deep neural network models with more conventional models like random forests and consider the best output is by integrating features learned from deep networks with tree-based networks. Our best-performing model ultimately achieves an area below the 0.49 precision-recall curve and an area below the 0.87 ROC curve.

By Sabetta Antonino, et. Al., an approach that uses machine learning to test repositories with source code and to automatically identify commits related to security (i.e. vulnerability-sensitive ones). They consider the improvements made by commits to the source code as documents written in a natural language, classifying them using standard methods for document classification. Our method can deliver high accuracy (80%) while ensuring acceptable recall (43%) by combining independent classifiers using information from various facets of commitments. Use information extracted from source code enhancements provides considerable improvement over the best-known state-of-the-art methodology while offering a significantly reduced amount of training data and utilizing a simpler architecture.

3. PROPOSED METHODOLOGY

Extracting features from software code patterns or dependencies. Utilizing natural language processing techniques for analyzing vulnerability descriptions.

Model Optimization:

Tune hyperparameters of SVM and KNN models using techniques like grid search or random search to improve classification performance. Experiment with different kernels for SVM (e.g., linear, polynomial, RBF) and various values of K for KNN.

Ensemble Methods:

Investigate ensemble methods such as Random Forest or Gradient Boosting to combine the predictions of multiple classifiers, potentially improving overall performance.

Handling Imbalanced Data:

Address the issue of imbalanced data by employing techniques like oversampling, undersampling, or using algorithms specifically designed for imbalanced datasets.

Integration with Vulnerability Management Systems:

Integrate the classification model with vulnerability management systems to automate the process of identifying and prioritizing vulnerabilities. Real-time Monitoring and Alerting: Implement a system that continuously monitors for new vulnerabilities and alerts relevant stakeholders based on the severity and type of vulnerability detected.

Security and Privacy Considerations:

Ensure that the system adheres to security and privacy best practices, especially when dealing with sensitive vulnerability data.

Documentation and User Interface:

Provide comprehensive documentation for the system, including user manuals and technical guides. Develop a user-friendly interface for interacting with the system.

ADVANTAGES

- 3) The proposed system integrates advanced feature engineering techniques, enabling the extraction of more meaningful features from vulnerability data, which can improve classification accuracy and reliability.
- 4) By employing model optimization techniques such as hyperparameter tuning and ensemble methods, the proposed system can achieve higher classification performance compared to the existing system.
- 5) The proposed system automates various aspects of vulnerability classification and integrates seamlessly with existing vulnerability management systems, reducing manual effort and improving operational efficiency.
- 6) The proposed system offers real-time monitoring and alerting capabilities, enabling organizations to promptly identify and respond to emerging security threats and vulnerabilities.
- 7) The proposed system incorporates robust security and privacy measures to safeguard sensitive vulnerability data and ensure compliance with relevant regulations and standards.

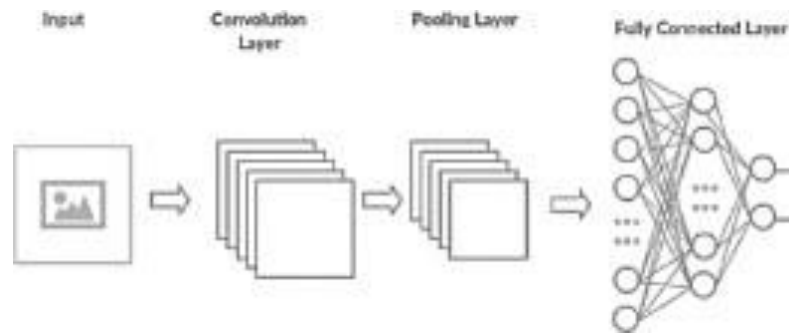
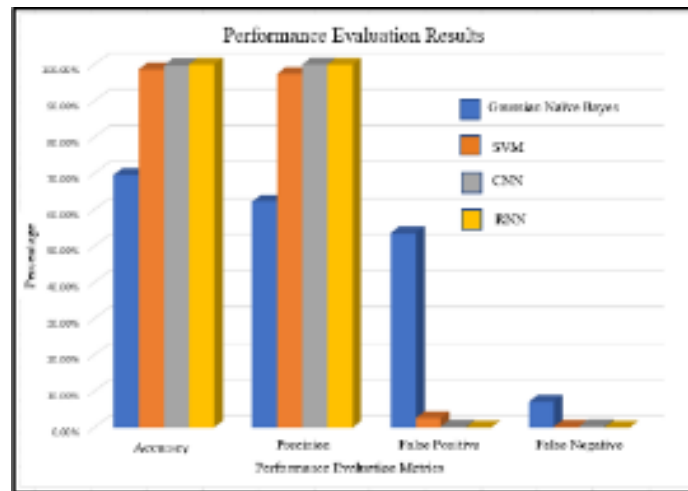


Figure 1: Convolutional Neural Network Structure

4. RESULTS AND DISCUSSIONS

In this study the main performance evaluation parameters that were included was accuracy, precision, R2 score, and parameters related to the confusion matrix. Based on these parameters several factors can be highlighted. According to the table 1, CNN and RNN which are deep learning algorithms depicts a prominent accuracy and precision over the traditional machine learning algorithms. However, it was obvious that when using the SVM, it had a clear advantage over the Gaussian Naïve Bayes algorithm. Moreover, the false positive and negative rates of the Gaussian Naïve Bayes algorithm were at a higher level which indicates that it is not that suitable for malware classification. This also gives a clear indication that proper usage of deep learning algorithms can result in great results. a clear graphical representation of the variation of accuracy, precision, false positive rates and false

negative rates of Gaussian Naïve Bayes, SVM, CNN and RNN. Based on this graph too, it is clear that the RNN is having the highest accuracy as well as the highest precision while Gaussian Naïve Bayes is having the lowest accuracy and the lowest precision level. Apart from that it is clearly visible that, the Gaussian Naïve Bayes is having the highest FP and FN rate. Another factor that can be highlighted here is that, RNN is having very low FP and FN values compared to the Gaussian Naïve Bayes Algorithm. So, based on the graph it is obvious that RNN is having a clear advantage against the other models



Graph 1: Graphical represents of SVM and KNN

5. CONCLUSION

Vulnerabilities in security are defined as system defects that are illegally exploited by unauthorized personnel. Thousands of safety problems are found in production software each year. Vulnerabilities sometimes appear in hidden forms that Software testers can't identify or left. Machine learning algorithms are a popular method for developing predictive models of vulnerability, authors add. They suggest the data-driven method of finding weaknesses utilizing deep learning is a possible solution. In this work, we will develop a cost-effective tool for developing heterogeneous vulnerability assessment and bug triage on windows as well as a web platform. Many tools don't support for a web-based application to detect code vulnerability. The system can work different datasets to extract the features and detect the vulnerability. SVM provides a better classification than the other two machine learning classifiers. In future developers to be needed is that to detect for code triage runtime mobile-based application programs so existing tools do not supports mobile application programs. Now a day's need is that in software engineering code clone management. Good quality of design achieved with the help of bugs free code clone in developing software. With the help of good quality of software improve the productivity of the software.

6. REFERENCES

[1] B. Cakir and E. Dogdu, "Malware classification using deep learning methods," Proc. ACMSE 2018 Conf., vol. 2018-January, no. April 2018.

- [2] A. P. Namanya, A. Cullen, I. U. Awan, and J. P. Disso, "The World of Malware: An Overview," Proc. - 2018 IEEE 6th Int. Conf. Futur. Internet Things Cloud, FiCloud 2018, no. September, pp. 420–427, 2018.
- [3] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," J. Netw. Comput. Appl., vol. 153, no. July 2019, p. 102526, 2020. [Online]. Available: <https://doi.org/10.1016/j.jnca.2019.102526>
- [4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," IEEE Access, vol. 7, pp. 46 717–46 738, 2019.
- [5] V. Menger, F. Scheepers, and M. Spruit, "Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text," Applied Sciences, vol. 8, no. 6, p. 981, Jun. 2018. [Online]. Available: <https://doi.org/10.3390/app8060981>
- [6] C. Liu, L. Wang, B. Lang, and Y. Zhou, "Finding effective classifier for malicious URL detection," in Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences - ICMSS 2018. ACM Press, 2018. [Online]. Available: <https://doi.org/10.1145/3180374.3181352>
- [7] Z. Cui, F. Xue, X. Cai, Y. Cao, G. G. Wang, and J. Chen, "Detection of Malicious Code Variants Based on Deep Learning," IEEE Trans. Ind. Informatics, vol. 14, no. 7, pp. 3187–3196, 2018.
- [8] B. B. Benuwa, Y. Zhan, B. Ghansah, D. K. Wornyo, and F. B. Kataka, "A review of deep machine learning," Int. J. Eng. Res. Africa, vol. 24, no. February 2017, pp. 124–136, 2016.
- [9] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware neural language models," 30th AAAI Conf. Artif. Intell. AAAI 2016, pp. 2741–2749, 2016
- [10] M. Rhode, P. Burnap, and K. Jones, "Early-stage malware prediction using recurrent neural networks," Comput. Secur., vol. 77, no. December 2017, pp. 578–594, 2018. [Online]. Available: <https://doi.org/10.1016/j.cose.2018.05.010>

TWEET DATA SEGREGATION AND SEGMENTATION USING BATCH MODE PROCESS

M. CHITRA

S.T.E.T Women's College(Autonomous), Mannargudi

ABSTRACT

Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced everyday. However, many applications in Information Retrieval (IR) and Natural Language Processing (NLP) suffer severely from the noisy and short nature of tweets. In this project, we propose a novel framework for tweet segmentation in a batch mode, called HybridSeg. By splitting tweets into meaningful segments, the semantic or context information is well preserved and easily extracted by the downstream applications. HybridSeg finds the optimal segmentation of a tweet by maximizing the sum of the stickiness scores of its candidate segments. The stickiness score considers the probability of a segment being a phrase in English (i.e., global context) and the probability of a segment being a phrase within the batch of tweets (i.e., local context). For the latter, we propose and evaluate two models to derive local context by considering the linguistic features and term-dependency in a batch of tweets, respectively. HybridSeg is also designed to iteratively learn from confident segments as pseudo feedback.

Keywords: Information Retrieval, HybridSeg, segments, downstream, Natural Language Processing

1. INTRODUCTION

Emotion expression plays a vital role in various part of every-day communication [1]. In past, various measures have been used to evaluate it, through a combination of indications such as facial expressions, gestures, and actions etc. Emotions extraction using facial, gestures and action are the part of digital image processing and computer vision. Emotions extraction is more difficult from texts especially from multi-languages texts, like in posts on social media and customers' reviews [2]. This type of data has presence of ambiguity and complexity of words in terms of meaning make them more difficult. Factors such as users writing style, politeness, irony, variability in language is one of the important problems in extraction of emotions [3]. A wide variety of state-of-art work has been carried out in the domain of opinions mining and sentiment analysis but limited research are focused on detection/extraction of emotions in tweeter [4].

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [5]. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets [6]. The method realizing the proposed framework that solely relies on global context is denoted by HybridSeg. Tweets are highly time-sensitive. The well preserved linguistic features in these tweets facilitate named entity recognition with high accuracy. Each named entity is a valid segment [7]. The method utilizing local linguistic features is denoted by HybridSeg NER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by HybridSeg N-Gram, is proposed based on the observation that many tweets published within a short time period are about the same topic [8]. HybridSeg N-Gram segments tweets by estimating the term-dependency within a batch of tweets. To improve POS tagging on tweets, discovering associated key posts from a large number of posts. The experimental results are based on a Twitter dataset to demonstrate the effectiveness of our proposed methods for both detecting events and identifying key posts [9]. Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output [10]. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing and analytical methods. Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets. The global context derived from Web pages or Wikipedia therefore helps identifying the meaningful segments in tweets. The method realizing the proposed framework that solely relies on global context is denoted by HybridSegWeb. In this project, proposed Clustering based subsetSelection algorithm uses minimum spanning tree-based method to cluster features. Moreover, our proposed algorithm does not limit to some specific types of data. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines.

Twitter, as a brand new variety of social media, has seen tremendous growth in recent years and has attracted nice interests from each business and domain. Several personal and/or public organizations are reportable to observe Twitter stream to gather and perceive users' opinions regarding the organizations. As a result of the extraordinarily volume of tweets revealed each day, it is much impracticable and surplus to concentrate and monitor the full Twitter stream.

Therefore, targeted Twitter streams area unit sometimes monitored instead; every such stream contains tweets that probably satisfy some data wants of the observation organization. Consequently, focused on Twitter streams are normally observed rather; each such stream contains tweets that conceivably fulfil some data needs of the checking association. Focused on Twitter stream is generally developed by sifting tweets with client characterized choice. Criteria relies on upon the data needs. Focused on Twitter stream is normally developed by separating tweets with predefined determination criteria (e.g., tweets distributed by clients from a geological district, tweets that match one or more predefined watchwords). Because of its precious business estimation of auspicious data from these tweets, it is basic to comprehend tweets' dialect for a huge assemblage of downstream applications, for example, Named Entity Recognition (NER), opinion mining, sentiment analysis, etc.

2. LITERATURE REVIEW

Most projects suggest using an existing conventional clustering algorithm (e.g., weighted k-means in CluStream) where the micro-clusters are used as pseudo points. Another approach used in Den Stream is to use reach ability where all micro-clusters which are less than a given distance from each other are linked together to form clusters. Grid-based algorithms typically merge adjacent dense grid cells to form larger clusters (see, e.g., the original version of D-Stream and MR-Stream).

In today's applications, evolving data streams are ubiquitous. Stream clustering algorithms were introduced to gain useful knowledge from these streams in real-time. The quality of the obtained clusterings, i.e. how good they reflect the data, can be assessed by evaluation measures. A multitude of stream clustering algorithms and evaluation measures for clusterings were introduced in the literature, however, until now there is no general tool for a direct comparison of the different algorithms or the evaluation measures. In our demo, we present a novel experimental framework for both tasks. It offers the means for extensive evaluation and visualization and is an extension of the Massive Online Analysis (MOA) software environment released under the GNU GPL License.

With the proliferation of the internet, video has become the principal source. Video big data introduce many hi-tech challenges, which include storage space, broadcast, compression, analysis, and identification. The increase in multimedia resources has brought an urgent need to develop intelligent methods to process and organize them. The combination between multimedia resources and Semantic link Network provides a new prospect for organizing them with their semantics. The tags and surrounding texts of multimedia resources are used to measure their association relation. There are two evaluation methods namely clustering and retrieval are used to measure the semantic relatedness between images accurately and robustly. This method is effective on image searching task. The semantic gap between semantics and video visual appearance is still a challenge. A model for generating the association between video resources using Semantic Link Network model is proposed. The user can select the attributes or concepts as the search query. This is done by providing the knowledge conduction during information

extraction and by applying fuzzy reasoning. The first action line is related to the establishment of techniques for the dynamic management of video analysis based on the knowledge gathered in the semantic network. This helps the decisions taken during the analysis process. Based on a set of rules it is able to handle the fuzziness of the annotations provided by the analysis modules gathered in the semantic network.

Novelty detection is a useful ability for learning systems, especially in data stream scenarios, where new concepts can appear, known concepts can disappear and concepts can evolve over time. There are several studies in the literature investigating the use of machine learning classification techniques for novelty detection in data streams. However, there is no consensus regarding how to evaluate the performance of these techniques, particular for multiclass problems. In this study, we propose a new evaluation approach for multiclass data streams novelty detection problems. This approach is able to deal with: i) multiclass problems, ii) confusion matrix with a column representing the unknown examples, iii) confusion matrix that increases over time, iv) unsupervised learning, that generates novelties without an association with the problem classes and v) representation of the evaluation measures over time. We evaluate the performance of the proposed approach by known novelty detection algorithms with artificial and real data sets.

Utilizing graph analysis models and algorithms to exploit complex interactions over a network of entities is emerging as an attractive network analytic technology. In this project, we show that traditional column or row-based trace analysis may not be effective in deriving deep insights hidden in the storage traces collected over complex storage applications, such as complex spatial and temporal patterns, hotspots and their movement patterns. We propose a novel graph analytics framework, GraphLens, for mining and analyzing real world storage traces with three unique features. First, we model storage traces as heterogeneous trace graphs in order to capture multiple complex and heterogeneous factors, such as diverse spatial/temporal access information and their relationships, into a unified analytic framework. Second, we employ and develop an innovative graph clustering method that employs two levels of clustering abstraction on storage trace analysis. We discover interesting spatial access patterns and identify important temporal correlations among spatial access patterns.

3. PROPOSED METHODOLOGY

This proposed Navie bayes model that classifies documents into reader-emotion categories. they studied the classification of news articles into different sentiment classes representing the emotions they trigger in their readers. This work mainly differs from other literature in focusing more on what the reader would feel while reading the article rather than what the writer was feeling while writing it. Other than the classification itself, which has been detailed in our previous work, we study the impact of the number of sentiment classes on the classification performance (i.e., accuracy, precision, and recall). It analyze the results of the different experiments and conclude with the limitations that make multi-class classification a difficult task.

ADVANTAGES

- Batch mode processing can easily scale to accommodate growing volumes of tweet data.
- Breaking down tweets into segments allows for easier reading and comprehension, particularly when the content is lengthy or complex.
- Readers can digest information more effectively when it's presented in bite-sized pieces.
- By applying predefined rules and algorithms to each batch uniformly, the system ensures consistent outcomes across the entire dataset.
- This approach minimizes the overhead associated with initiating and terminating processing tasks, resulting in more efficient use of computational resources such as CPU, memory, and network bandwidth.

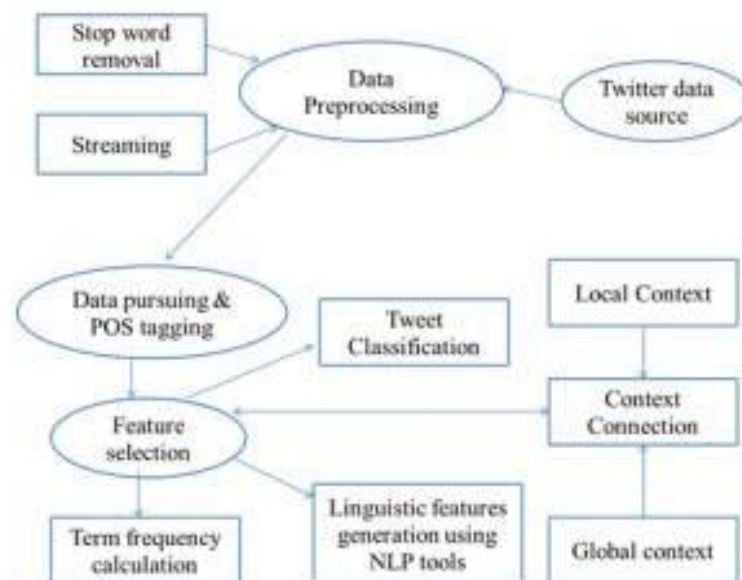
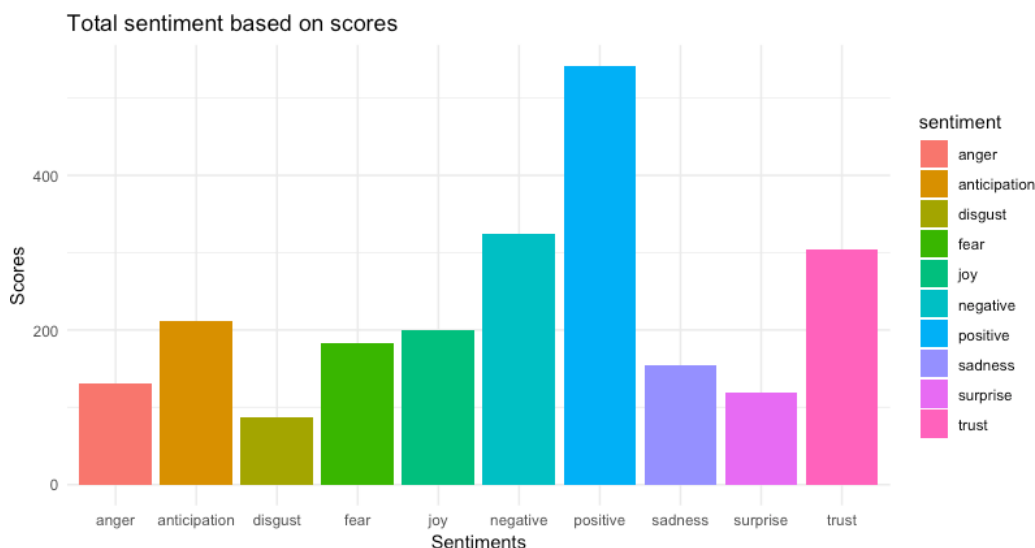


Figure 1: Proposed Architecture of tweet segmentation

4. RESULTS AND DISCUSSIONS

The really interesting part of the analysis comes in part two, where Julia uses the `tm` package (which provides a number of text mining functions to) and `syuzhet` package (which includes the NRC Word-Emotion Association Lexicon algorithm) to analyze the sentiment of her tweets. Categorizing all 10,000 tweets as representing "anger", "fear", "surprise" and other sentiments, and generating a positive and negative sentiment score for each, is as simple as this one line . Using those sentiment scores, Julia was easily able to summarize the sentiments expressed in her tweet history and create this time series chart showing her negative and positive sentiment scores over time. If you've been thinking about applying sentiment analysis to some text data, you might find that with R it's easier than you think! Try it using your own Twitter archive by following along with Julia's posts linked below.



Graph.1 shown the total sentiment score board

5. CONCLUSION

In this paper was studied the task of multi-class sentiment analysis and evaluated the evolution of various KPIs as the number of sentiment classes increased. It analyzed the difficulties of, and the different challenges involved with, multi-class classification, and proposed some metrics to measure the distance between sentiments (i.e., how similar they are to one another) and concluded that even though the task of multi-class analysis is important, it might be more interesting to perform a sentiment detection task through which all of the sentiments present within a text are extracted. In future this work will be experimented and tested in the public cloud based peta-sized datasets.

6. REFERENCES

- [1] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan, “Clustering data streams,” in Proceedings of the ACM Symposium on Foundations of Computer Science, 12-14 Nov. 2000, pp. 359–366.
- [2] C. Aggarwal, Data Streams: Models and Algorithms, ser. Advances in Database Systems, Springer, Ed., 2007.
- [3] J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman & Hall/CRC, 2010.
- [4] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. a. Gama, “Data stream clustering: A survey,” ACM Computing Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.
- [5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in Proceedings of the International Conference on Very Large Data Bases (VLDB ’03), 2003, pp. 81–92.
- [6] F. Cao, M. Ester, W. Qian, and A. Zhou, “Density-based clustering over an evolving data stream with noise,” in Proceedings of the 2006 SIAM International Conference on Data Mining. SIAM, 2006, pp. 328–339.

- [7] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007, pp. 133–142.
- [8] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density based clustering of data streams at multiple resolutions," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–28, 2009.
- [9] L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 3, pp. 1–27, 2009.
- [10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'1996), 1996, pp. 226–231.

INTERNET OF THINGS

ABINAYA V,

THIRD YEAR BSC COMPUTER SCIENCE, DON BOSCO COLLEGE (ARTS AND SCIENCE),
KARAIKAL, PUDUCHERRY. Email: abinavisu@gmail.com

ABSTRACT:

This paper delves into the intricate landscape of the Internet of Things (IoT), dissecting its multifaceted dimensions. The discussion encompasses the foundational technologies driving IoT, ranging from sensor networks to communication protocols, and delves into real-world applications across industries. The presentation will emphasize the challenges inherent in the vast deployment of IoT, including privacy concerns, security vulnerabilities, and interoperability issues. By seamlessly interconnecting devices and enabling data exchange, IoT has transcended traditional boundaries, fostering a paradigm shift in how we live, work, and interact. The presentation will navigate through the key facets of IoT, examining its technological foundations, applications across various domains, and the unprecedented challenges and opportunities it presents. From smart homes to industrial automation, the paper explores the intricate web of interconnected devices, highlighting the potential for efficiency enhancement, resource optimization, and novel user experiences. Additionally, it will showcase innovative solutions and best practices employed to overcome these hurdles. Exploring the convergence of IoT with emerging technologies like edge computing and artificial intelligence, the paper aims to paint a comprehensive picture of the evolving IoT ecosystem. As we stand on the cusp of a more interconnected future, this presentation seeks to foster dialogue on the advancements, dilemmas, and future trajectories of the Internet of Things.

Key words:

Key words about the Internet of things are, Internet of Things (IoT), Wireless Sensor Networks (WSN), Protocols (MQTT, CoAP, etc.), Cloud Computing for IoT, IoT Applications (Healthcare, Smart Cities, Agriculture), IoT Architecture, IoT Standards and Interoperability, Energy Efficiency in IoT Devices, IoT Connectivity (5G, LPWAN), IoT Analytics, IoT in Industry, Privacy in IoT.

1. INTRODUCTION

IoT stands for Internet of Things. It refers to the interconnectedness of physical devices, such as appliances and vehicles, that are embedded with software, sensors, and connectivity which enables these objects to connect and exchange data. This technology allows for the collection and sharing of data from a vast network of devices, creating opportunities for more efficient and automated systems.

Internet of Things (IoT) is the networking of physical objects that contain electronics embedded within their architecture in order to communicate and sense interactions amongst each other or with respect to the external environment. In the upcoming years, IoT-based technology will offer advanced levels of services and practically change the way people lead their daily lives. Advancements in medicine, power, gene therapies, agriculture, smart cities, and smart homes are just a few of the categorical examples where IoT is strongly established.

The internet which first began with desktop computers and then evolved to laptops, tablets and mobile phones is now going a step further and extending to real world everyday objects. The physical items can now be controlled remotely from anywhere. Apart from the field of computer science it is also finding applicability in different fields like business, economy, agriculture, healthcare, etc. At the same time, however, the Internet of Things raises significant challenges which could adversely affect its applicability. Hacking of Internet-connected devices, surveillance concerns, and privacy fears already have captured public attention. Technical challenges still remain a major concern for developing IoT based applications.

IOT is a system of interrelated things, computing devices, mechanical and digital machines, objects, animals, or people that are provided with unique identifiers. And the ability to transfer the data over a network requiring human-to-human or human-to-computer interaction.

2.LITERATURE REVIEW

In [1] Advancements in Microelectronics and networking allow the creation of interconnected smart objects forming the Internet of Things (IoT), merging the physical and digital worlds. This revolution introduces a new era of networked objects, transforming service creation and interaction with the environment. In [2] This study explores Internet of Things (IoT), aiming to connect real-world objects through a common infrastructure, offering control and insights. It covers definitions, origins, requirements, characteristics, and applications, providing a valuable overview for researchers entering the IoT field. In [3] This paper explores the current state of Internet of Things (IoT), highlighting its potential to enhance communication between people and objects. It covers various academic perspectives on IoT definitions, discusses enabling technologies like RFID and sensor networks, and addresses remaining research challenges. In [4] This study delves into IoT security, exploring threats and vulnerabilities. It aims to be a practical guide, offering insights and solutions for enhancing the security of the IoT environment. In [5] ICT shapes our daily lives, connecting diverse devices in areas like smart homes, transportation, and entertainment. This paper explores the evolving Internet landscape, emphasizing the Internet of Things (IoT) and its technologies, protocols, and applications, offering a guide for those entering this dynamic field. In [6] This paper explores the Internet of Things (IoT), highlighting its dependence on integrating various technologies like identification systems, sensor networks, and smart object intelligence. It emphasizes the interdisciplinary nature of IoT development, spanning fields such as telecommunications, informatics, electronics, and social science, addressing ongoing challenges in this complex landscape. In [7] The Internet of Things (IoT) transforms traditional factories into smart ones, using interconnected devices to monitor and optimize production, prevent machine failures with predictive maintenance, enhance energy efficiency, improve workplace safety, and optimize supply chains through real-time tracking and inventory management. This review explores IoT applications in Industry 4.0, offering insights to improve quality control and optimize part production processes in smart factories.

3.SOURCES OF IOT:

The sources of data in IoT (Internet of Things) can be diverse and come from various types of sensors and devices. Here are common sources of IoT data:

3.1 Sensors: Physical sensors are fundamental in IoT. They can include temperature sensors, humidity sensors, motion sensors, light sensors, accelerometers, and more. These sensors capture real-world data. Imagine a scenario of automated monitoring of a farm such that it will just indicate the current situation of crops like "4 crops need water, Now I'm going to pour it" and then it will satisfy the crop's need.

3.2 Actuators: Devices that perform actions based on data received. Devices which is a contrast to sensors. It transforms electrical signals into physical movements. Both sensors and actuators are transducers that convert one form of energy to another. The exchange of data is the most important key factor in IoT. Hence sensors and actuators play a vital role here. For example, actuators could control motors, valves, switches, or other mechanisms.

3.3 RFID (Radio-Frequency Identification): RFID tags are used for tracking and identifying objects. They are commonly used in logistics, inventory management, and access control. Since interconnection of things is the main goal of IoT, the RFID tags get hand-shaken with IoT technology and are used to provide the unique id for the connected "things" in IoT.

3.4 Cameras: Visual data from cameras is used for surveillance, image recognition, and monitoring

applications.

3.5 GPS (Global Positioning System): GPS data is crucial for location-based IoT applications, such as vehicle tracking, geofencing, and navigation.

3.6 Biometric Sensors: These sensors capture data related to human characteristics, such as fingerprints, facial recognition, or heart rate, and are often used in security applications.

3.7 Smart Devices: Everyday devices with embedded sensors, like smart refrigerators, thermostats, and wearable devices, contribute data to IoT systems.

3.8 Industrial Machines and Sensors: In industrial IoT (IIoT), machinery and equipment are equipped with sensors to monitor performance, detect faults, and optimize operations.

3.9 Environmental Sensors: Sensors measuring air quality, pollution levels, and other environmental factors contribute to smart city and environmental monitoring applications.

3.10 Home Automation Devices: Smart home devices, such as smart lights, thermostats, and doorbells, generate data for home automation and energy management.

4. MAJOR COMPONENTS OF IOT:

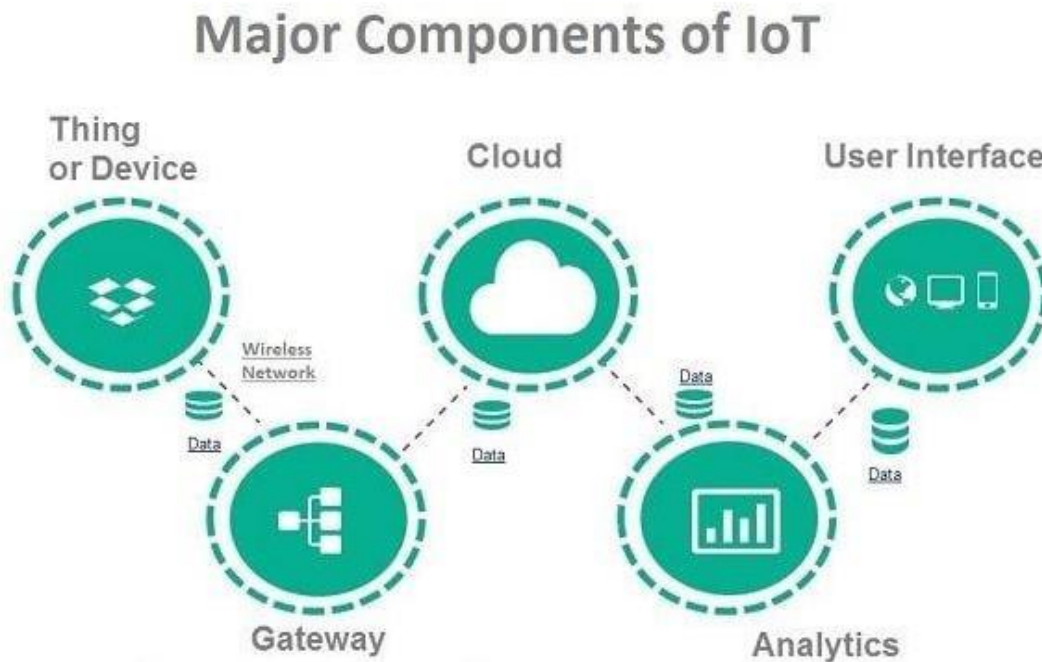


Figure 1,

Major Components of IOT

4.1 Smart devices and sensors – Device connectivity

Devices and sensors are the components of the device connectivity layer. These smart sensors are continuously collecting data from the environment and transmit the information to the next layer. The latest techniques in semiconductor technology are capable of producing micro smart sensors for various applications.

Common sensors are:

- Temperature sensors and thermostats
- Pressure sensors
- Humidity / Moisture level

- Light intensity detectors
- Moisture sensors
- Proximity detection
- RFID tags

4.2 Gateway

IoT Gateway manages the bidirectional data traffic between different networks and protocols. Another function of the gateway is to translate different network protocols and make sure interoperability of the connected devices and sensors.

Gateways can be configured to perform pre-processing of the collected data from thousands of sensors locally before transmitting it to the next stage. In some scenarios, it would be necessary due to the compatibility of the TCP/IP protocol.

IoT gateway offers a certain level of security for the network and transmitted data with higher-order encryption techniques. It acts as a middle layer between devices and the cloud to protect the system from malicious attacks and unauthorized access.

4.3 Cloud

The Internet of Things creates massive data from devices, applications, and users, which has to be managed in an efficient way. IoT cloud offers tools to collect, process, manage and store huge amounts of data in real time. Industries and services can easily access these data remotely and make critical decisions when necessary.

Basically, the IoT cloud is a sophisticated, high-performance network of servers optimized to perform high-speed data processing of billions of devices, traffic management, and deliver accurate analytics. Distributed database management systems are one of the most important components of the IoT cloud.

Cloud system integrates billions of devices, sensors, gateways, protocols, and data storage and provides predictive analytics. Companies use these analytics data to improve products and services, preventive measures for certain steps, and build their new business model accurately.

4.4 Analytics

Analytics is the process of converting analog data from billions of smart devices and sensors into useful insights which can be interpreted and used for detailed analysis. Smart analytics solutions are inevitable for IoT systems for the management and improvement of the entire system.

One of the major advantages of an efficient IoT system is real-time smart analytics which helps engineers to find out irregularities in the collected data and act fast to prevent an undesired scenario. Service providers can prepare for further steps if the information is collected accurately at the right time.

Big enterprises use the massive data collected from IoT devices and utilize the insights for their future business opportunities. Careful analysis will help organizations to predict trends in the market and plan ahead for a successful implementation.

Information is very significant in any business model, and predictive analysis ensures success in the concerned area of the business line

4.5 User Interface

User interfaces are the visible, tangible part of the IoT system which users can access. Designers will have to make sure of a well-designed user interface for minimum effort for users and encourage more interactions.

Modern technology offers much interactive design to ease complex tasks into simple touch panel controls. Multicolor touch panels have replaced hard switches in our household appliances, and the trend is increasing

for almost every smart home device.

The user interface design has higher significance in today's competitive market; it often determines the user whether to choose a particular device or appliance. Users will be interested in buying new devices or smart gadgets if it is very user-friendly and compatible with common wireless standards.

5.CHARACTERISTICS OF IOT

The Internet of Things (IoT) is characterized by the following key features that are mentioned below.

5.1 Connectivity

Connectivity is an important requirement of the IoT infrastructure. Things of IoT should be connected to the IoT infrastructure. Anyone, anywhere, anytime can connect, this should be guaranteed at all times. For example, the connection between people through Internet devices like mobile phones, and other gadgets, also a connection between Internet devices such as routers, gateways, sensors, etc.

5.2 Intelligence and Identity

The extraction of knowledge from the generated data is very important. For example, a sensor generates data, but that data will only be useful if it is interpreted properly. Each IoT device has a unique identity. This identification is helpful in tracking the equipment and at times for querying its status.

5.3 Scalability

The number of elements connected to the IoT zone is increasing day by day. Hence, an IoT setup should be capable of handling the massive expansion. The data generated as an outcome is enormous, and it should be handled appropriately.

5.4 Dynamic and Self-Adapting (Complexity)

IoT devices should dynamically adapt themselves to changing contexts and scenarios. Assume a camera meant for surveillance. It should be adaptable to work in different conditions and different light situations (morning, afternoon, and night).

5.5 Architecture

IoT Architecture cannot be homogeneous in nature. It should be hybrid, supporting different manufacturers' products to function in the IoT network. IoT is not owned by anyone engineering branch. IoT is a reality when multiple domains come together.

ADVANTAGES AND DISADVANTAGES OF IOT

The Internet of Things (IoT) is a network of interconnected, embedded devices that can capture and transmit data without the need for human interaction over a wireless network.

IoT applications in everyday life include smart wearables, smart health monitoring, traffic monitoring, IoT in agriculture with many sensors, smart devices, robots in hospitals, smart grid and water supply, and so on.

ADVANTAGES

- ❖ It can assist in the smarter control of homes and cities via mobile phones. It enhances security and offers personal protection.
- ❖ By automating activities, it saves us a lot of time.
- ❖ Information is easily accessible, even if we are far away from our actual location, and it is updated frequently in real time.
- ❖ Electric Devices are directly connected and communicate with a controller computer, such as a cell phone, resulting in efficient electricity use. As a result, there will be no unnecessary use of electricity

equipment.

- ❖ Personal assistance can be provided by IoT apps, which can alert you to your regular plans.
- ❖ It is useful for safety because it senses any potential danger and warns users. For example, GM OnStar, is a integrated device that system which identifies a car crash or accident on road. It immediately makes a call if an accident or crash is found.
- ❖ It minimizes human effort because IoT devices connect and communicate with one another and perform a variety of tasks without the need for human intervention.
- ❖ Patient care can be performed more effectively in real time without the need for a doctor's visit. It gives them the ability to make choices as well as provide evidence-based care.

DISADVANTAGES

- ❖ Hackers may gain access to the system and steal personal information. Since we add so many devices to the internet, there is a risk that our information as it can be misused.
- ❖ They rely heavily on the internet and are unable to function effectively without it.
- ❖ With the complexity of systems, there are many ways for them to fail.
- ❖ We lose control of our lives—our lives will be fully controlled and reliant on technology.
- ❖ Overuse of the Internet and technology makes people unintelligent because they rely on smart devices instead of doing physical work, causing them to become lazy.
- ❖ Unskilled workers are at a high risk of losing their jobs, which could lead to unemployment. Smart surveillance cameras, robots, smart ironing systems, smart washing machines, and other facilities are replacing security guards, maids, ironmen, and dry-cleaning services etc.
- ❖ It is very difficult to plan, build, manage, and enable a broad technology to IoT framework.
- ❖ Deploying IoT devices is very costly and time-consuming.

CONCLUSION

In conclusion, the Internet of Things (IoT) represents a transformative paradigm that intertwines the digital and physical worlds, creating a network where everyday objects communicate, share data, and act intelligently. The Internet of Things (IoT) stands as a transformative force in our interconnected world. As it weaves through diverse aspects of daily life, from smart homes to transportation, its rapid evolution brings both opportunities and security challenges. Embracing IoT technologies requires a vigilant understanding of potential threats, ongoing innovations, and collaborative efforts to ensure a secure and beneficial integration into our societies. With its key features of connectivity, sensor integration, and data-driven insights, IoT has the potential to revolutionize industries, enhance efficiency, and improve our daily lives. However, as we embrace the benefits of IoT, it is crucial to address challenges such as security, privacy, and standardization to ensure its responsible and sustainable growth. The evolution of IoT promises a future where interconnected devices seamlessly collaborate, providing innovative solutions and shaping a more connected and intelligent world.

REFERENCES

- [1] D Singh - Factories of the Future: Technological Advancements. 2023 Wiley Online Library.
- [2] S Madakam. V Lake - Journal of Computer and. 2015- scirp. Org.
- [3] DL Yang. F Liu, YD Liang 1st International Conference on E- Business Intelligence (ICEBI 2010).
- [4] FA Alaba, M Othman, IAT Hashem, F Alotaibi Journal of Network and Computer Applications 88, 10-28, 2017.
- [5] SV Srikanth, PJ Pramod, KP Dileep, S Tapas, MU Patil, CBN Sarat 2009 International conference on

advanced information networking and ..., 2009.

[6] L Atzori. In the evening. G Morabito - Computer networks. 2010 – Elsevier.

[7] M Soori. B Arezoo, R Dastres - Internet of Things and Cyber-Physical.. 2023 – Elsevier.

ADVANCEMENTS IN NATURAL LANGUAGE PROCESSING

SUBASRI.B

THIRD YEAR, BSc COMPUTER SCIENCE, DON BOSCO COLLEGE OF ARTS AND SCIENCE

KARAIKAL, PONDICHERY, Email: subasribalu333@gmail.com

TITLE:

The study of Advancements in Natural Language Processing (NLP)

ABSTRACT:

The paper concludes with a This paper explores the dynamic landscape of NLP and its transformative impact on information processing and communication. Focusing on recent advancements, we delve into the evolution of NLP algorithms, from rule-based systems to cutting-edge deep learning models. Our discussion encompasses the challenges posed by ambiguity, context and cultural nuances in language understanding.

Highlighting key applications, we showcase how NLP is revolutionizing in various domains including sentiment analysis, machine translation, chatbots and information extraction.

glimpse into the future, outlining potential research directions continued societal impact of NLP innovations. Overall it aims to provide a comprehensive overview of the current state and future prospectus of NLP.

KEYWORDS:

Natural Language Processing, Machine Learning, Text Analysis, Semantic Analysis, Tokenization, Part-of-Speech Tagging, Named Entity Recognition (NER), Word Embeddings, Deep Learning, Language Models, Syntax and Grammar, Information Retrieval, Sentiment Analysis, SpeechRecognition, Chatbots, Corpus Linguistics, Feature Engineering, Preprocessing Techniques, Supervised Learning, Unsupervised Learning.

1. INTRODUCTION:

Natural Language Processing (NLP) has witnessed remarkable advancements in recent years, revolutionizing how computers understand, interpret, and generate human language. From traditional rule-based approaches to modern deep learning techniques, NLP has evolved exponentially. These advancements have been fueled by the availability of large datasets, more powerful hardware, and innovative algorithms. Today, NLP powers a wide range of applications, including machine translation, sentiment analysis, chatbots, and language generation models. With ongoing research in areas like contextual understanding, multitask learning, and ethical AI, NLP continues to push the boundaries of what's possible in human-computer interaction and language understanding.

Furthermore, advancements in NLP have led to significant improvements in areas such as text summarization, information extraction, and document classification, empowering businesses to extract actionable insights from vast amounts of unstructured data. Additionally, the integration of NLP with other emerging technologies like machine vision and knowledge graphs has opened up new possibilities for applications such as automated content analysis, virtual assistants, and personalized recommendation systems. Moreover, the democratization of NLP through the release of open-source libraries and platforms has accelerated innovation and collaboration within the research community and beyond, making NLP more accessible to developers and practitioners worldwide. Looking ahead, ongoing research efforts in areas like explainable AI, ethical AI, and domain-specific NLP are expected to further enhance the capabilities and reliability of NLP systems, paving the way for more responsible and impactful applications across various domains and industries.

Advancements in Natural Language Processing (NLP) have propelled the field to unprecedented heights, revolutionizing how machines understand and interact with human language. One of the key breakthroughs has been the development of deep learning models, particularly transformer-based architectures, which have demonstrated remarkable performance across a wide range of NLP tasks. These models, such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), leverage large-scale pre-training on vast text corpora followed by fine-tuning on specific tasks, enabling them to capture intricate linguistic patterns and semantic nuances. As a result, NLP applications have become more accurate, versatile, and capable of handling complex language tasks with unprecedented levels of proficiency.

Another significant advancement is the emergence of transfer learning in NLP, which allows models to leverage knowledge learned from one task or domain and apply it to another, even with limited labeled data. This has led to more efficient and effective NLP systems, reducing the need for large annotated datasets and accelerating the development of applications in diverse domains. Additionally, advancements in neural network architectures, optimization techniques, and computational resources have contributed to the scalability and performance of NLP models, enabling them to process and understand vast amounts of text data with unprecedented speed and accuracy.

Furthermore, advancements in multilingual NLP have facilitated communication and collaboration across language barriers, enabling NLP systems to understand and generate text in multiple languages with high accuracy. This has profound implications for global communication, cross-cultural collaboration, and accessibility to information worldwide. Moreover, innovations in areas such as conversational AI, sentiment analysis, and emotion recognition have fueled the development of more empathetic and context-aware NLP systems, enhancing their ability to understand and respond to human emotions and intentions. Looking ahead, ongoing research in NLP continues to push the boundaries of what is possible, with a focus on areas such as interpretability, robustness, and fairness. As NLP technologies become more sophisticated and ubiquitous, they hold the potential to transform various industries, from healthcare and finance to education and entertainment, by enabling more intelligent, personalized, and human-like interactions between machines and humans.

2. LITERATURE REVIEW:

In [1] the author says, NLP is widely used in diverse areas like translation, spam detection, and medical analysis. This paper outlines NLP's phases, evolution, and current applications. It also addresses challenges and emerging trends in the field.

In [2] the author says, NLP is rapidly growing with diverse applications like translation and text analysis. Recent progress stems from new ML algorithms and abundant data. This paper outlines latest NLP advancements and emerging trends.

In [3] the author says, Natural Language Processing (NLP) is widely used for analyzing human language in fields like machine translation, spam detection, and medical applications. The paper outlines NLP's evolution, discusses its phases, components of Natural Language Generation (NLG), and explores current trends and challenges. Overall, NLP's applications span diverse areas and continue to advance in complexity and utility.

In [4] the author says, This tutorial aims to introduce NLP and modern NLP system design to medical informatics generalists with limited prior knowledge. It covers the historical evolution of NLP, common

sub-problems, highlights of medical NLP, machine-learning approaches, modern NLP architectures like Apache UIMA, and future directions, including IBM Watson's impact on the medical field.

In [5] the author says, This paper traces the evolution of Natural Language Processing (NLP) from the 1940s to present, identifying four phases: machine translation, AI influence, logico-grammatical style adoption, and focus on massive language data. It reviews key developments in each phase and evaluates the current status of NLP after over forty years of research.

In [6] the author says, Healthcare data holds valuable knowledge but is underutilized due to processing challenges. This chapter explores methods for analyzing eHealth data, comparing NLP and DeepLearning techniques. It identifies current issues and suggests research directions in this field.

3. NATURAL LANGUAGE PROCESING FOR E-HEALTH DATA:

Text analytics consists of a set of techniques that apply software algorithms in order to understand the content of unstructured textual data, more precisely of written language. Some techniques, such as rule-based classification, are using a set of rules or logical statements defined by specialists. Natural language processing (NLP) employs more complex techniques that consider the context and the meaning of words, and not only searching for individual terms. As a result, such approaches have, in most cases, better metrics, i.e., accuracy. In eHealth, NLP solutions are very often used in the medical staff in tasks such as screening, diagnosing treatment, or patient monitoring. To this end, the NLP techniques are added to complex eHealth adaptive systems, i.e., computer-aided diagnostic systems and data-driven decision support systems. In the Big Data context, these systems are fed with large amounts of data collected from different sources, which they then use to learn how to provide certain services. An example is improving the efficiency in determining the diagnosis or the right treatment, by combining the current patient medical results with information from his/her historical healthcare data and the information learnt from other similar cases.

In order to obtain useful information and to model it according to the needs of the eHealth application or service, NLP utilises a processing pipeline. Depending on the desired output, the NLP pipeline may consist of different specific steps, but they always group into three major steps: Preprocessing, Vector Space Model and Deployment (Figure 1).

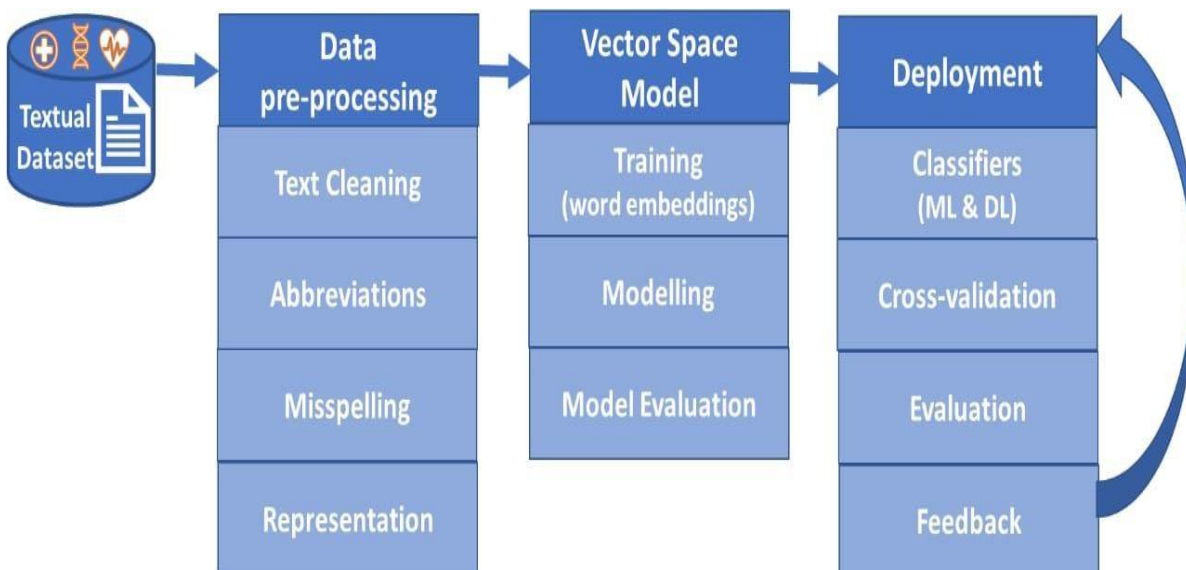


Figure 1: The pipeline for textual data processing

4. RECENT ADVAMCEMENTS IN NLP:

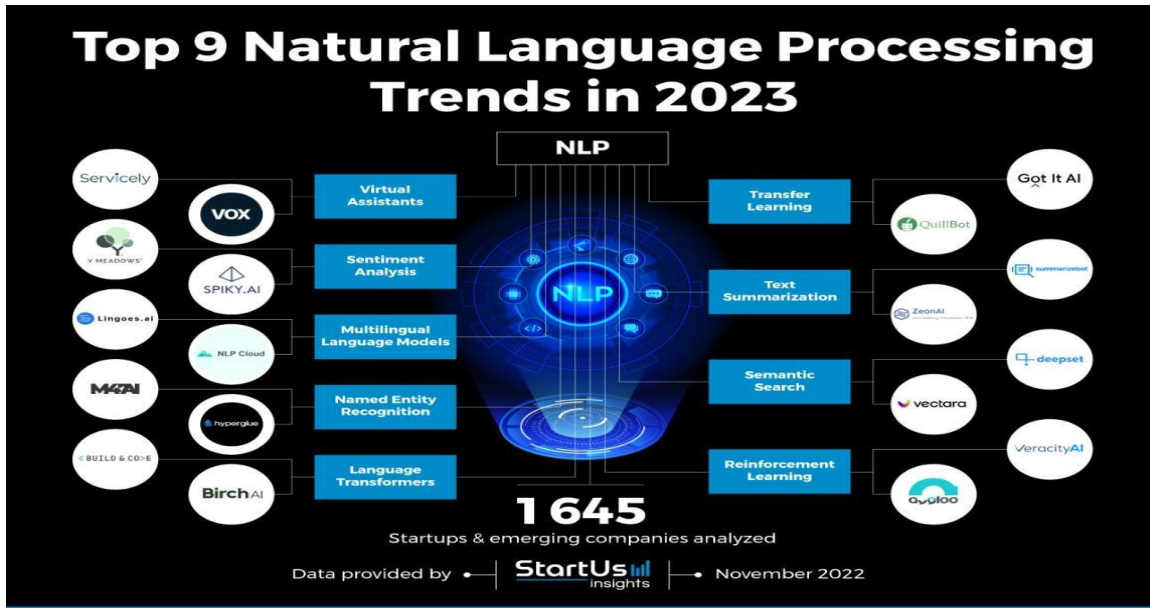


Figure 2: Recent trends in NLP

This paper presents an overview of some of the key advancements in NLP, highlighting their significance and potential implications.

DEEP LEARNING ARCHITECTURES:

- Deep learning architectures, particularly neural networks, have revolutionized NLP by enabling models to learn intricate patterns in language data.
- Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have achieved state-of-the-art results across various NLP tasks, including language understanding, generation, and translation.

TRANSFER LEARNING AND PRE-TRAINING:

- Transfer learning has emerged as a dominant paradigm in NLP, allowing models to leverage knowledge from large pre-trained datasets before fine-tuning on specific tasks.
- Pre-trained language representations, like word embeddings and contextualized embeddings (e.g., ELMo, BERT), have significantly improved the performance of downstream NLP tasks, reducing the need for extensive task-specific data.

4.3 MULTIMODAL NLP:

- The integration of multiple modalities, such as text, images, and audio, has become an area of focus in NLP.
- Models like CLIP (Contrastive Language-Image Pre-training) and multimodal transformers are capable of understanding and generating text based on visual inputs, opening up new possibilities for applications like image captioning and visual question answering.

ETHICAL AND SOCIETAL IMPLICATIONS:

- As NLP technologies become more pervasive, addressing ethical concerns regarding biases, fairness, privacy, and misinformation has become imperative.
- Researchers and practitioners are actively exploring techniques for mitigating biases in language models and developing frameworks for responsible AI deployment in NLP application.

➤ CHALLENGES AND FUTURE DIRECTIONS:

- Despite the remarkable progress, NLP still faces several challenges, including the understanding of context, commonsense reasoning, and low-resource languages.
- Future research directions may involve advancing unsupervised and self-supervised learning techniques, improving model interpretability, and developing NLP solutions for specialized domains and tasks.

TRANSFER LEARNING:

- Transfer learning techniques, such as fine-tuning pre-trained language models like BERT, GPT, and XLNet, have revolutionized NLP by allowing models to be trained on large-scale corpora and then adapted to specific tasks with relatively small amounts of task-specific data.

CONTEXTUAL WORD EMBEDDINGS:

- Contextual word embeddings like ELMo (Embeddings from Language Models) and BERT capture the contextual meaning of words within a sentence, leading to better representations of text and improved performance on various NLP tasks such as sentiment analysis, question answering, and named entity recognition.

ATTENTION MECHANISMS:

- Attention mechanisms, popularized by the Transformer architecture, have enabled models to focus on relevant parts of the input sequence, leading to improved performance in tasks requiring long-range dependencies and capturing hierarchical structures in text.

LARGE-SCALE LANGUAGE MODELS:

- The development of large-scale language models like GPT-3 with hundreds of billions of parameters has pushed the boundaries of NLP, achieving state-of-the-art results on a wide range of tasks including language understanding, text generation, and machine translation.

LOW-RESOURCE AND MULTI-LINGUAL NLP:

- Research efforts have focused on developing NLP techniques that perform well in low-resource settings and across multiple languages. Techniques like zero-shot and few-shot learning, as well as cross-lingual pre-training, aim to improve the generalization of models to diverse languages and data conditions.

ETHICAL AND FAIR NLP:

- There is growing awareness of the ethical implications of NLP technologies, including biases in datasets and models, as well as issues related to fairness, transparency, and accountability. Researchers are actively working on developing methods to detect and mitigate biases and ensure that NLP systems are deployed responsibly.

CONVERSATIONAL AI:

- With the rise of virtual assistants, chatbots, and smart speakers, there's a lot of interest in developing NLP models that can engage in natural and meaningful conversations with users.

4.13 TRANSFORMERS FOR TEXT GENERATION:

- Transformer-based models, such as GPT-3, are being widely explored for various text generation tasks, including text summarization, story generation, and dialogue systems.

ETHICAL AI AND BIAS MITIGATION:

- Addressing biases and promoting fairness in NLP models has become a hot topic, with researchers and practitioners actively working on techniques to mitigate bias and ensure more equitable outcomes.

LOW RESOURCE LANGUAGES:

➤ There's increasing attention on developing NLP models and resources for languages with limited data, aiming to bridge the digital divide and enable access to technology for speakers of low-resource languages.

EXPLAINABLE AI IN NLP:

➤ As NLP models become more complex and pervasive, there's a growing need for explanations of their predictions and decisions, leading to research on methods for making NLP models more interpretable and transparent.

NLP FOR HEALTHCARE:

➤ Applying NLP techniques to healthcare data, such as electronic health records and medical literature, is a burgeoning area of research, with the potential to improve clinical decision-making, patient outcomes, and healthcare efficiency.

PRIVACY-PRESERVING NLP:

➤ With increasing concerns about data privacy and security, there's a growing interest in developing NLP models and techniques that can operate effectively while preserving the privacy of sensitive information.

NLP FOR SOCIALGOOD:

➤ Utilizing NLP for social good initiatives, such as disaster response, humanitarian aid, and combating misinformation, is gaining momentum, highlighting the positive impact NLP technology can have on society.

CONTINUE LEARNING IN NLP:

➤ Research on continual learning techniques for NLP models, enabling them to adapt and learn from new data over time, is gaining attention as a way to improve model robustness and longevity in dynamic environments.

5. CONCLUSION:

In conclusion, the advancements in Natural Language Processing (NLP) have propelled the field into a new era of capability and sophistication. From the development of powerful deep learning models to the emergence of transfer learning and multilingual capabilities, NLP has made significant strides in understanding and generating human language. These advancements have not only improved the accuracy and efficiency of NLP applications but have also enabled more seamless and natural interactions between humans and machines across various domains and languages. As research in NLP continues to evolve, the future holds promise for even more groundbreaking innovations, with the potential to revolutionize industries, enhance communication, and drive forward the era of AI-powered language technologies. Overall, the advancements in NLP represent a remarkable achievement in artificial intelligence and computational linguistics, shaping the way we interact with technology and each other in the digital age. The advancements in Natural Language Processing have transformed how we interact with and process textual data, paving the way for innovative applications across various domains. Continued research and collaboration are essential to address remaining challenges and ensure the responsible development and deployment of NLP technologies for the benefit of society.

REFERENCES:

- [1] Khurana, D., Koli, A., Khatter, K.etal. "Natural language processing: state of the art, current trends and challenges". *Multimed Tools Appl* 82, 3713–3744 (2023).
- [2] The Future of Natural Language Processing: "A Survey of Recent Advances and Emerging Trends." *Journal of Scholastic Engineering Science and Management*, June 2023, Volume 2, Issue 6,
- [3] Khurana, Diksha, Koli, Aditya, Khatter, Kiran, Singh, Sukhdev, " Natural Language Processing: State of The Art, Current Trends and Challenges" *Multimedia Tools and Applications*.
- [4] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman *Journal of the American Medical Informatics Association*, Volume 18, Issue 5, September 2011, Pages 544–551,
- [5] From: *Current issues in computational linguistics: in honour of Don Walker*, ed. Antonio Zampolli, Nicoletta Calzolari, Martha Palmer (*Linguistica Computazionale*, vol. 9-10); Pisa, Dordrecht, [1994]
- [6] Elena-Simona Apostol and Ciprian-Octavian Truică *National University of Science and Technology Politehnica Bucharest*, Bucharest, Romania

A SURVEY ON ROBOTIC PROCESS AUTOMATION

AAKASH S

Government Arts College (Autonomous) Kumbakonam. Email: aakashcap07z@gmail.com

ABSTRACT:

Robotic Process Automation (RPA) is a disruptive technology that automates repetitive, rule-based tasks within enterprises, freeing up human resources to focus on higher-value activities. By leveraging software robots or "bots" to mimic human interactions with digital systems, RPA streamlines business processes, reduces operational costs, and enhances efficiency. This abstract explores the transformative impact of RPA on organizations across various industries, highlighting its ability to increase productivity, accuracy, and scalability. With its potential to revolutionize workflows and optimize resource utilization, RPA stands as a cornerstone of digital transformation in the modern business landscape. Robotic Process Automation (RPA) represents a paradigm shift in how businesses approach process optimization and automation. RPA technology involves the implementation of software robots or bots to perform repetitive, rule-based tasks traditionally carried out by humans. These bots interact with applications, manipulate data, trigger responses, and communicate with other systems to streamline workflows and automate business processes. This paper delves into the transformative impact of RPA across industries, showcasing its ability to drive operational efficiencies, reduce errors, and enhance scalability. Through the lens of RPA, organizations can achieve significant cost savings, improved accuracy, and increased productivity by eliminating manual interventions and accelerating process execution. The integration of RPA in business operations empowers employees to focus on strategic, creative, and value-added tasks, fostering a culture of innovation and growth. Furthermore, RPA enables businesses to adapt to dynamic market demands, regulatory changes, and technological advancements with greater agility and speed. The adoption of RPA transcends traditional boundaries, revolutionizing how enterprises approach digital transformation and operational excellence. By harnessing the power of automation, organizations can achieve enhanced process efficiency, higher quality outputs, and enhanced customer satisfaction. This paper explores the practical applications, benefits, challenges, and future prospects of RPA as a cornerstone of modern business strategy, ushering in a new era of operational efficiency and competitive advantage.

KEYWORDS: Automation, Software Robots, Business Process Automation, Task Automation

INTRODUCTION

In today's fast-paced and technology-driven world, businesses are constantly seeking innovative solutions to streamline their operations, enhance efficiency, and drive productivity. One such ground-breaking technology that has emerged as a game-changer in the realm of automation is Robotic Process Automation (RPA). Robotic Process Automation, often referred to as RPA, is a cutting-edge technology that leverages software robots or bots to automate repetitive, rule-based tasks and business processes.

These virtual robots are designed to mimic human interactions with digital systems and applications, performing tasks like data entry, document processing, form filling, and more with precision and efficiency. The implementation of RPA is not only about streamlining processes but also about driving digital transformation, agility, and competitiveness in today's dynamic market landscape. As RPA technology continues to evolve and integrate with advanced capabilities like artificial intelligence (AI) and cognitive automation, its potential to revolutionize business operations and drive sustainable growth becomes even more evident.

REVIEW OF LITERATURE

A Robotic Process Automation is a software solution used to perform mundane repetitive. Work previously done by people. RPA enables creation of software robots ("bots") to automate. An RPA is the equivalent of a business software license, not a physical robot. Each RPA is a software robot that can be instructed very quickly to carry Out an operational process with speed and accuracy. RPAs also known as bots function as a Digital workforce that can perform repetitive tasks such as data entry and data transfers, but with added benefits of having a worker on 24 hours a day, seven days a week. RPAs are optimally used with high volume, standardized, rules-based, mature, stable processes where costs are clear and business value is well understood (Willocks, Lacity & Craig, 2015).

An RPA should not interfere with the company's current information technology (IT). The RPA acts in the same way a human would to access platforms, is non-invasive, and interacts with other systems through the presentation layer so no system programming logic is touched or altered and no data is stored. RPAs automate and access the presentation layer of existing Processes through applications and the outputs are verified by the business operations managers. An RPA is used to automate a company's business process and is only as good as the Business practices in a company. If the company's procedure currently generates bad data, then RPAs will not fix these actions. Erroneous inputs must be eliminated to avoid erroneous output. Therefore, before any software solution is sought out, a company should examine their processes and make improvements as needed before implementation.

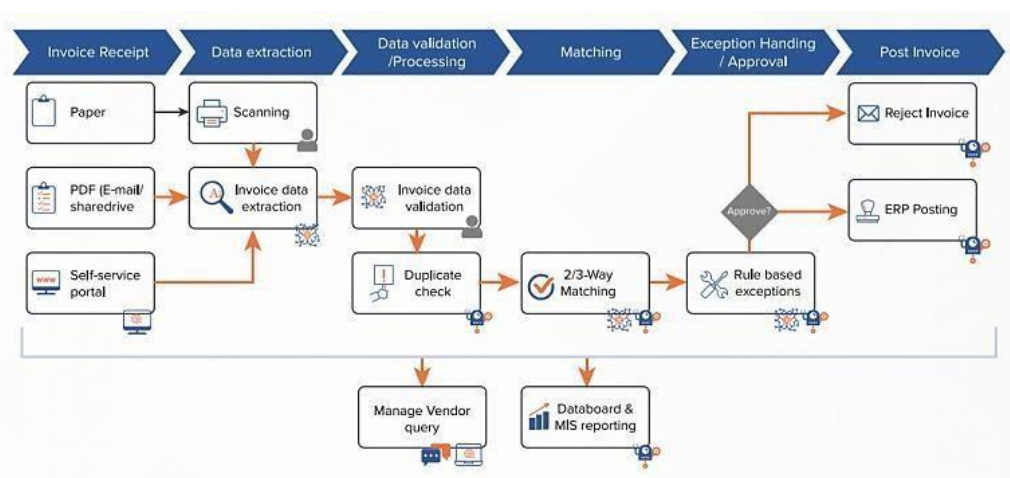
An RPA is used to automate a company's business process and is only as good as the business practices in a company. If the company's procedure currently generates bad data, then RPAs will not fix these actions. Erroneous inputs must be eliminated to avoid erroneous output. Therefore, before any software solution is sought out, a company should examine their processes and make improvements as needed before implementation. RPAs have also been defined as "a low to no code commercial off the shelf (COTS) Technology that can automate

repetitive, rules-based tasks” according to the RP Program playbook (Federal Robotic Process Automation Community of Practice, 2020). The PMA’s CAP goal six states, “Federal agencies will shift time, effort and funding from low to high value work through the elimination of unnecessary requirements, burden reduction, optimization and streamlining, and workload automation” (Performance.gov, 2020). RPA is a technology that provides organizations with these capabilities.

METHODOLOGIES

1. Top-Down Approach

- Description: In the top-down approach, organizations start by identifying high-level business processes that are ideal for automation. They prioritize processes based on their impact on efficiency, cost reduction, or strategic goals.
- Benefits: This approach ensures alignment with business objectives and focuses on automating processes that deliver the most value to the organization.



Business Automation(Invoice Automation Process)

2. Proof of Concept (POC) Approach

- Description: Organizations conduct a proof-of-concept project to demonstrate the feasibility and benefits of RPA in a real-world environment. This involves selecting a small-scale process for automation and evaluating the results.
- Benefits: POCs help validate the business case for RPA, build stakeholder confidence, and identify any potential challenges early on in the automation journey.

3. Agile Methodology

- Description: Applying agile principles to RPA involves iterative development, rapid deployment, and continuous feedback loops. Organizations break down automation projects into smaller tasks or “sprints” to deliver value incrementally.
- Benefits: Agile allows for flexibility, faster time to market, and the ability to adapt to

changing requirements or priorities during the automation process.

4. **Center of Excellence (CoE) Approach** - Description: Establishing an RPA Center of Excellence involves creating a dedicated team responsible for driving RPA initiatives, defining best practices, providing training, and ensuring governance and compliance. - Benefits: CoEs help standardize RPA implementation, promote knowledge sharing, scale automation efforts effectively, and maintain consistency across automation projects.

RESULT AND DISCUSSION

The concept of Robotic Process Automation (RPA) as we know it today was not founded by a single individual or company but rather evolved over time through contributions from various researchers, innovators, and technology companies.

The roots of RPA can be traced back to early automation efforts in the fields of artificial intelligence, robotics, and business process management. While there isn't a specific "founder" of RPA, several key players and organizations have played a significant role in shaping the development and adoption of RPA technology:

1. **Blue Prism**: Blue Prism, founded in 2001 in the UK, is often credited as one of the pioneers in the field of RPA. The company developed one of the earliest RPA platforms, focusing on automating repetitive tasks and business processes using software robots.
2. **UiPath**: UiPath, founded in 2005 in Romania, has emerged as a leading RPA vendor, known for its user-friendly platform and advanced automation capabilities. The company has played a key role in popularizing RPA and advancing the technology's capabilities.
3. **Automation Anywhere**: Automation Anywhere, founded in 2003 in San Jose, California, is another major player in the RPA industry. The company has been instrumental in developing RPA solutions that cater to a wide range of industries and automation needs.

CONCLUSION

It's fascinating to see how many companies across various industries have embraced Robotic Process Automation (RPA) to streamline their operations and enhance efficiency. Here are some well-known companies that have successfully implemented RPA in their business processes. **IBM, Amazon, Coca-Cola, Walmart, Accenture and McDonald's.**

REFERENCES

- [1] Arrow Digital (2018) How Robotic Process Automation Shapes Fintech's Future.
a. Arrow Digital.
- [2] Laurent, P., Chollet, T., & Herzberg, E. (2018). Intelligent Automation Entering the Business World Deloitte.
- [3] Lacity, M. C. & Willcocks, L. P. (2016b). Robotic Process Automation at Telefonica O2.MIS Quarterly Executive.
- [4] Madakam, S., M. Holmukhe, R. & Kumar Jaiswal, D. (2019). The Future Digital Work Force: Robotic Process Automation (RPA). Journal of Information Systems and Technology Management.

A SURVEY OF INTERNET OF THINGS

AARTHI S

GOVERNMENT ARTS COLLEGE (Autonomous)KUMBAKONAM, aarthi.nbs@gmail.com

ABSTRACT:

The rapid development and implementation of smart and IoT (Internet of Things) based technologies have allowed for various possibilities in technological advancements for different aspects of life. The main goal of IoT technologies is to simplify processes in different fields, to ensure a better efficiency of systems (technologies or specific processes) and finally to improve life quality. Sustainability has become a key issue for population where the dynamic development of IoT technologies is bringing different useful benefits, but this fast development must be carefully monitored and evaluated from an environmental point of view to limit the presence of harmful impacts and ensure the smart utilization of limited global resources. Significant research efforts are needed in the previous sense to carefully investigate the pros and cons of IoT technologies. This review editorial is partially directed on the research contributions presented at the 4th International Conference on Smart and Sustainable Technologies held in Split and Bol, Croatia, in 2019 (SpliTech 2019) as well as on recent findings from literature. The focus of the conference was directed towards key conference tracks such as Smart City, e-Health and Engineering Modelling. The research presented and discussed at the SpliTech2019 conference helped to understand the complex and intertwined effects of IoT technologies on societies and their potential effects on sustainability in general. Various application areas of IoT technologies were discussed as well as the progress made. Four main topical areas were discussed in the herein editorial, i.e. latest advancements in the further fields: (i) IoT technologies in Sustainable Energy and Environment, (ii) IoT enabled Smart City, (iii) E-health – Ambient assisted living systems (iv) IoT technologies in Transportation and Low Carbon Products.

Keywords IoT, Smart city, Sustainability, Energy, Environment, SpliTech2020

INTRODUCTION

Internet of Things (IoT) is the networking of physical objects that contain electronics embedded within their architecture in order to communicate and sense interactions amongst each other or with respect to the external environment. In the upcoming years, IoT-based technology will offer advanced levels of services and practically change the way people lead their daily lives. Advancements in medicine, power, gene therapies, agriculture, smart cities, and smart homes are just a few of the categorical examples where IoT is strongly established.

IOT is a system of interrelated things, computing devices, mechanical and digital machines, objects, animals, or people that are provided with unique identifiers. And the ability to transfer the data over a network requiring human-to-human or human-to-computer interaction.

REVIEW OF LITERATURE

A review of literature on the Internet of Things (IoT) covers a vast array of topics, ranging from technical aspects to societal impacts. Here's a general overview of what such a review might encompass:

1. Technical Foundations of IoT

- Protocols and standards: MQTT, CoAP, Zigbee, etc.
- Sensor technologies: RFID, NFC, Bluetooth Low Energy (BLE), etc.
- Connectivity technologies: 5G, LPWAN (Low Power Wide Area Network), Wi-Fi, etc.

2. Applications and Use Cases

- Smart homes: Home automation, energy management, security systems.
- Smart cities: Traffic management, waste management, environmental monitoring.
- Industrial IoT (IIoT): Predictive maintenance, asset tracking, supply chain optimization.
- Healthcare: Remote patient monitoring, wearable health devices.
- Agriculture: Precision farming, livestock monitoring.
- Retail: Inventory management, personalized marketing.
- Transportation: Fleet management, vehicle tracking, autonomous vehicles.

3. Data Management and Analytics

- Data collection: Stream processing, batch processing.
- Data storage: Cloud storage, edge storage, databases.

METHODOLOGY

- ✓ The IoT Architectural Framework (IoT-A)
- ✓ Microsoft's IoT Lifecycle Methodology
- ✓ Industrial Internet Consortium (IIC) IoT Capability Model
- ✓ Agile Methodology
- ✓ Standards and Interoperability

RESULT & DISCUSSION

1. Technical Performance

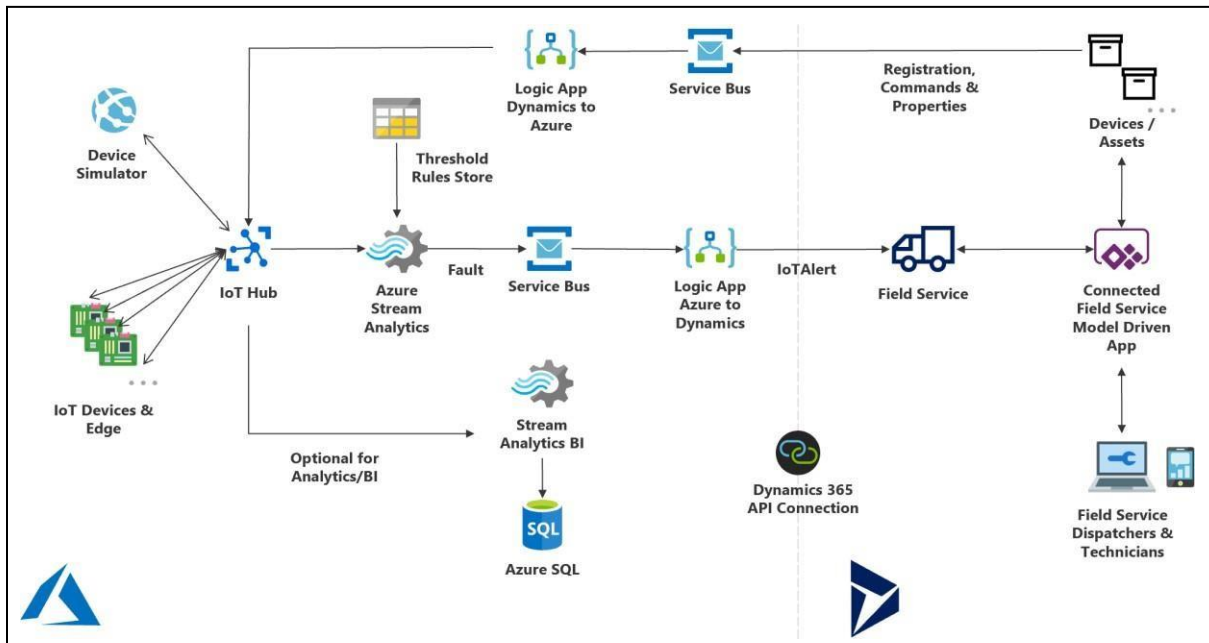
- Provide quantitative data on system performance metrics such as latency, throughput, and reliability.
- Report on the effectiveness of different communication protocols, data processing techniques, and hardware components.

2. Interpretation of Results

- Interpret the findings from the results section in light of the research questions, hypotheses, or objectives of the study.
- Discuss how the results contribute to existing knowledge in the field of IoT or address specific research gaps.

3. Comparison with Existing Literature

- Compare the findings of the current study with previous research or literature in the field.
- Identify similarities, differences, or contradictions between the current results and existing knowledge.



CONCLUSION

In conclusion, the Internet of Things (IoT) represents a transformative paradigm that has the potential to revolutionize various aspects of our lives, ranging from home automation to industrial processes and beyond. Through the integration of sensors, connectivity, and data analytics, IoT systems enable the collection, analysis, and utilization of vast amounts of data from the physical world.

In summary, while IoT presents unprecedented opportunities for innovation and disruption, its success hinges on our ability to navigate technical, regulatory, and societal complexities effectively. By fostering a collaborative and interdisciplinary approach, we can harness the transformative power of IoT to create a more connected, intelligent, and sustainable world for generations to come.

REFERENCES

- [1] Gillis, Alexander (2021). "What is internet of things (IoT)?". IOT Agenda. Retrieved 17 August 2021.
- [2] "Internet of Things Global Standards Initiative". ITU. Retrieved 26 June 2015.

A SURVEY ON IOT FLOOD MONITORING & ALERTING SYSTEM

ARTHI M

GOVERNMENT ARTS COLLEGE (Autonomous), KUMBAKONAM, Email:arthimurugan2001@gmail.com

ABSTRACT

Flooding is usually brought on by an increased quantity of water in a water system, like a lake, river overflowing. On occasion a dam fractures, abruptly releasing a massive quantity of water. The outcome is that a number of the water travels into soil, and flooding the region. Rivers are involving river banks, in a station. Aside from lack of products and house and office property, streets infrastructure flood water consists of bacteria and sewage flow of waste sites and chemical spillage which leads to a variety of diseases afterwards

Flood predictions need information like

The speed of change in river stage on a real time basis, which may help indicate the seriousness and immediacy of this threat. Understanding of the form of storm generating the moisture, such as length, intensity and areal extent, which is valuable for discovering potential seriousness of the flood.

In this system we make use of a raspberry pi with water sensors, rain sensors to predict flood and alert respective authorities and sound instant alarm in nearby villages to instantly transmit information about possible floods using IOT. The water sensors are used to measure water level of 3 different locations. Also 3 different rain sensors are used to measure rain level of those 3 areas. These sensors provide information over the IOT using Raspberry Pi. On detection of conditions of flooding the system predicts the amount of time it would take to flood in a particular area and alerts the villages/areas that could be affected by it.

Keywords:

IOT , flood , raspberry pi , sensor , water , soil , areas , Villages , time , affected , alarm

INTRODUCTION

Flood occurs when water overflows from the river, lake or from heavy rainfall and it can happen at any time of the year. Flooding can be very dangerous, when floods happen in an area that people live, the water carries along objects like houses, cars, furniture and even people. It can wipe away property, trees and many more heavy items. For years, flooded roads have been a problem in Metro Mumbai. It causes heavy flow of traffic. When traffic happened, people's money, time and effort are wasted. Through the local government unit flood control has been extending their efforts to inform the commuters regarding the situation in flooded areas during rainy season, still the dissemination of information to the locals are not enough. For this reason, the "Arduino Flood Detector System" is been develop, to help the road user to avoid this problem happened.

RELATED WORKS

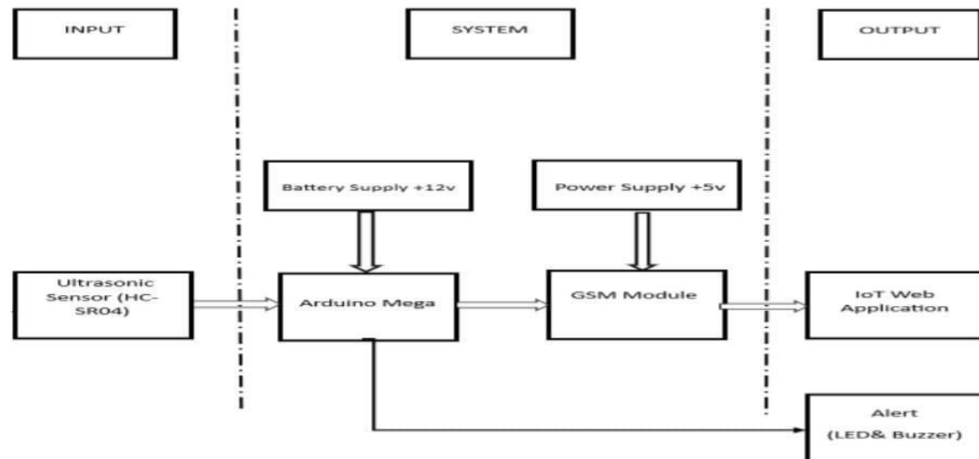
A novel architecture for the transceiver is proposed in order to increase the service range of IEEE802.11ah, which is necessary for the long-range IoT communication of emergency messages in emergency situations. Experimental results show that the presented architecture is suitable for the long-range IoT communication of emergency alert messages. Wireless sensor network system could remotely monitor the real time data of water condition in the identified areas. To monitor the water conditions such as water level, flow and precipitation level, wireless sensor network system is developed. GRAB, designed for robust data delivery in face of unreliable nodes and failable wireless link. GRAB forwards data along band of interleaved mesh.

In Wireless Sensor Networks (WSN) the user requirements are often desired to be evolvable, whether driven by changes of the monitored parameters or WSN properties of configuration, structure, communication capacities, node density and energy among many others. The Functionality is supported by the reflective and component-based Grid Kit middleware, which provides support for both WSN and Grid application domains.

Distributed system is proposed using water level monitoring sensors named Shonabondhu. The sensing nodes are distributed all across the country and the servers that collect data from sensors are spread around various regions.

BLOCK DIAGRAM

In this project we are using Ultrasonic Sensor, DHT11 Sensor, Flow Sensor, Raindrop Sensor, Turbidity Sensor to measure the parameters like water environmental temperature and humidity, Flow rate, level of water, rain intensity of the environment, how much amount of water is been contaminated in rivers. If any of the value exceeded message on the values will be uploaded using GSM Module and the exceeded values will be displayed in LCD screen and also it gives indication using LED's and alert will be given using voice sensor and also through SMS notification



RESULT & CONCLUSION

This project highlights the possibility to provide an alert system that will overcome the risk of flood. As the project is enabled with IOT technology and hence the sensor data can be monitored from anywhere in the world. More sensors can be integrated into the system in order to create more accurate and efficient flood detection system. It can also contribute to multiple government agencies or authority that ultimately help the society and mankind about the flood like hazardous natural disaster. It will monitor each and every aspect that can lead to flood. If the water level rises along with the speed, it will send an alert immediately. It also ensures increased accessibility in dealing and reverting to this catastrophic incident. In summary, it will help the community in taking quick decisions and planning against this disaster mankind about the flood like hazardous natural disaster.

REFERENCES

- [1] Internet of Things Based Real Time Flood Monitoring and Alert Management System Author: Tibin Mathew Thekkil ; Dr. N. Prabakaran Year: 2021
- [2] Development of Flood Monitoring System using WSN and IoT based on Cloud Author: Pallavi C B1; Chandrakala; Year:2019

A SURVEY ON DECISION MANAGEMENT

N.ARUNA

GOVERNMENT ARTS COLLEGE Autonomous KUMBAKONAM Email:cheziyan1983@gmail.com

ABSTRACT

It has become all too common that questions are raised during the execution of a project pertaining to the decisions that were made early on. Without having maintained a concise, accessible record of project decisions, the project manager and team members would find it difficult to provide hard evidence as to how they got to this point and what impacts specific decisions had on the project's trajectory. This paper introduces the Decision Management Process Improvement Project (DMPPI), which focuses on improving decision management process throughout the lifecycle of a project with the aim of adding value to project performance and helping obtain project success. This new tool was inspired due to a lack of appropriate methods involving complex projects at a local consulting firm. The process along with the tool is being added to the toolset of a local Consulting Firm. This Firm plans to introduce the tools and techniques to clients that will benefit from an increased Project Management maturity level with improvements to its decision-tracking processes and demonstration of downstream effects of important decisions. The final product is a contribution to the Project Management Body of Knowledge (PMBOK) in the form of creating a Project Decision Management knowledge area in the PMBOK format. A decision log that follows a decision throughout the whole process from problem identification and analysis to the eventual outcome is at the core of the created knowledge area.

KEYWORDS:

Capturing Decisions in Project Management, Decision Management, Decision Tracking, Decision Management Tools and Techniques, Decision Log.

INTRODUCTION

Decision management is the process of designing, building, and managing automated decision-making systems. A business can make thousands of decisions a day; therefore, decisions are central to any business or organization. The challenge comes in when you have to take a significant number of factors into consideration at once that could influence a decision such as fluctuating prices, new products, current events, and more. That's where a decision management tool can greatly increase the efficiency of a company's decisions. A business can use decision management to help manage customer, supplier, and employee interactions. Decision management finds its place within analytics—the discipline that uses business logic as well as data mathematics to help provide insights that make better decisions.

REVIEW OF LITERATURE

ANALYTICS CAN BE FOUR TYPES

Descriptive: Tells you about something that has occurred, but not what you can do about it.

Diagnostic: Where you can explore the reasons why something occurred.

Predictive: This data models the likelihood that an event may happen in the future.

Prescriptive analytics: Combines all of the above but takes an extra step to make recommendations on how to respond or you can automate responses not know how to derive the value of data since the data is raw.

Areas of decision management

The goal of decision management is to enhance business operations intelligence by ensuring quick, consistent, and accurate fact-based decisions. The quality of structured operational decisions, no matter how complex, should be constantly improving. There are five areas that affect decision management:

Data and analytics

Data is accessed and processed with the help of descriptive, diagnostic, and predictive techniques. You need strong data quality as a basis for accurate decision making, and the outcomes of those decisions affect the data as well.

Business process management

Managing human tasks and the sequence of business process automation and task management. The information from staff helps to make better decisions, and their roles are enhanced as a result.

Business rules management

Automating business rules and managing them based on inputs provided by subject matter experts.

Robotics

Using software to imitate human behavior in the automation of actions and related interactions with software systems.

DECISION-MAKING PROCESS

Identify the decision Gather relevant information Identify the alternatives Weigh the evidence
Choose among alternatives Take action Review your decision & its consequences

Step 1: Identify the decision

You realize that you need to make a decision. Try to clearly define the nature of the decision you must make. This first step is very important.

Step 2: Gather relevant information

Collect some pertinent information before you make your decision: what information is needed, the best sources of information, and how to get it. This step involves both internal and external "work." Some information is internal: you'll seek it through a process of self-assessment. Other information is external: you'll find it online, in books, from other people,

and from other sources.

Step 3: Identify the alternatives

As you collect information, you will probably identify several possible paths of action, or alternatives. You can also use your imagination and additional information to construct new alternatives. In this step, you will list all possible and desirable alternatives.

Step 4: Weigh the evidence

Draw on your information and emotions to imagine what it would be like if you carried out each of the alternatives to the end. Evaluate whether the need identified in Step 1 would be met or resolved through the use of each alternative. As you go through this difficult internal process, you'll begin to favor certain alternatives: those that seem to have a higher potential for reaching your goal. Finally, place the alternatives in a priority order, based upon your own value system.

Step 5: Choose among alternatives

Once you have weighed all the evidence, you are ready to select the alternative that seems to be best one for you. You may even choose a combination of alternatives.

Your choice in Step 5 may very likely be the same or similar to the alternative you placed at the top of your list at the end of Step 4.

Step 6: Take action

You're now ready to take some positive action by beginning to implement the alternative you chose in Step 5.

Step 7: Review your decision & its consequences

In this final step, consider the results of your decision and evaluate whether or not it has resolved the need you identified in Step 1. If the decision has not met the identified need, you may want to repeat certain steps of the process to make a new decision. For example, you might want to gather more detailed or somewhat different information or explore additional alternatives



CONCLUSION

When it comes to making decisions, one should always weigh the positive and negative business consequences and should favor the positive outcomes. This avoids the possible losses

to the organization and keeps the company running with a sustained growth.

REFERENCES

- [1] McKeown, K. R. (1985). Text generation: Using discourse strategies and focus constraints to generate natural language text. Cambridge University Press.
- [2] Reiter, E., & Dale, R. (2000). Building applied natural language generation systems (Vol. 1). Cambridge University Press.

A SURVEY ON ARTIFICIAL INTELLIGENCE

S. AYYAPPAN

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM

Email:neekavithaikala8@gmail.com

ABSTRACT:

Artificial Intelligence (AI) stands as one of the most transformative and enigmatic fields of modern science and technology. This abstract delves into the multifaceted dimensions of AI, offering a nuanced exploration of its conceptual frameworks, methodologies, and implications. At its core, AI encapsulates the quest to imbue machines with cognitive abilities akin to human intelligence. From classical rule-based systems to cutting-edge deep learning algorithms, AI has traversed a remarkable evolutionary trajectory, constantly redefining the boundaries of what machines can achieve. Yet, beneath its technological prowess lies a complex interplay of mathematics, neuroscience, and computer science. This abstract seeks to elucidate the foundational principles underpinning AI, starting with its historical antecedents and culminating in contemporary paradigms. It navigates through seminal concepts such as machine learning, neural networks, and natural language processing, elucidating their theoretical underpinnings and practical application. However, the proliferation of AI also raises profound ethical, societal, and existential questions. As machines encroach upon domains once deemed exclusive to human cognition, concerns surrounding autonomy, bias, and job displacement loom large. Moreover, the Specter of super intelligent. AI sparks debates on the nature of consciousness and the existential risks posed by runaway technological. In navigating these complexities, this abstract advocates for a holistic approach to AI research and deployment—one that prioritizes not only technical proficiency but also ethical stewardship and societal well-being. It calls for interdisciplinary collaboration, drawing insights from philosophy, psychology, and governance to ensure that AI serves as a force for progress rather than peril. Ultimately, the abstract posits that the trajectory of AI is not predetermined but contingent upon the choices and values of its creators. By fostering a culture of responsible innovation and inclusive dialogue, we can harness the transformative potential of AI to shape a more equitable, sustainable, and humane future for all.

INTRODUCTION

ARTIFICIAL INTELLIGENCE

AI: It is a machine's ability to perform the cognitive functions we associate with human minds, such as perceiving, reasoning, learning, interacting with an environment, problem solving, and even exercising creativity. Before leading to the meaning of artificial intelligence let understand what is the meaning of Intelligence?

Intelligence: The ability to learn and solve problems. This definition is taken from Webster's Dictionary. The most common answer that one expects is "to make computers intelligent so that they can act intelligently!", but the question is how much intelligent? How can one judge intelligence?

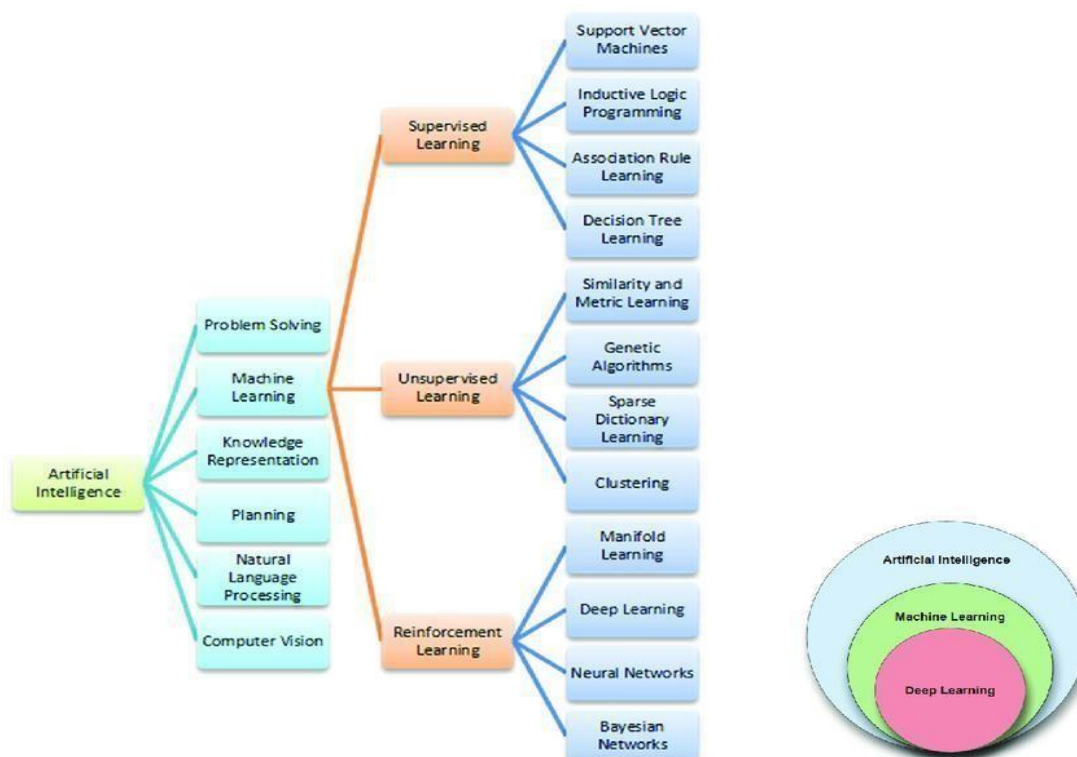
If the computers can, somehow, solve real-world problems, by improving on their own from past experiences, they would be called "intelligent". Thus, the AI systems are more generic (rather than specific), can "think" and are more flexible. Intelligence, as we know, is the ability to acquire and apply knowledge. Knowledge is the information acquired through experience. Experience is the knowledge gained through exposure (training). Summing the terms up, we get artificial intelligence as the "copy of something natural (i.e., human beings) 'WHO' is capable of acquiring and applying the information it has gained through exposure."

REVIEW OF LITERATURE

Brief overview of the importance and scope of artificial intelligence. Explanation of the purpose and focus of the literature review. Artificial intelligence (AI) has been subject to extensive scientific review, covering its capabilities, limitations, and ethical implications.

Researchers have highlighted AI's potential to revolutionize various fields, including healthcare, finance, and transportation, by automating tasks, analysing data at scale, and enabling new insights. However, concerns about AI's ethical use, potential biases, and impacts on employment have also been raised. Ongoing research aims to address these challenges while maximizing AI's benefits for society.

METHODOLOGY



Historical Evolution of AI

Summarize key milestones and developments in AI research from its inception to the present day. Highlight influential figures, breakthroughs, and paradigm shifts in the field.

Theoretical Foundations:

Explore foundational concepts and theories underpinning AI, such as machine learning, neural networks, and symbolic reasoning. Discuss how different approaches have shaped the evolution of AI research.

Techniques

Machine Learning: A subset of AI that enables systems to learn and improve from experience without being explicitly programmed. This includes techniques such as supervised learning, unsupervised learning, and reinforcement learning.

Deep Learning: A type of machine learning that utilizes neural networks with many layers (deep neural networks) to learn intricate patterns and representations from data. Deep learning has led to significant breakthroughs in areas like computer vision, natural language processing, and speech recognition.

Natural Language Processing (NLP): NLP focuses on enabling computers to understand, interpret, and generate human language. Applications include language translation, sentiment analysis, and chatbots.

Computer Vision: Computer vision involves teaching computers to interpret and understand visual information from images or videos. It enables tasks like object recognition, image classification, and facial recognition.

Robotics: Robotics combines AI with mechanical engineering to create intelligent machines capable of performing physical tasks autonomously or with human guidance. Applications range from manufacturing and logistics to healthcare and exploration.

Expert Systems: Expert systems emulate the decision-making abilities of human experts in specific domains by encoding their knowledge into a computer program. They are used in fields like medicine, finance, and engineering for problem-solving and decision support.

Current Trends and Challenges:

Discuss recent trends and advancements in AI research, including:

Deep learning and neural architecture search, Transfer learning and multi-task learning, Ethical considerations and bias mitigation in AI systems. Identify ongoing challenges and research gaps in the field, such as interpretability, scalability, and robustness of AI models.

Future Directions:

Speculate on the future of AI research and its potential impact on society and the economy. Highlight emerging areas of interest and opportunities for further exploration.

CONCLUSION

Summarize key findings and insights from the literature review. Reflect on the broader implications of AI research and the importance of continued interdisciplinary collaboration and ethical oversight.

REFERENCES:

- [1] Agre P. E., Artificial Intelligence, Computational research on interaction and agency,1995.
- [2] Agrawal, A., Gans, J., and Goldfarb, A. Prediction Machines, Updated and Expanded- The Simple Economics of Artificial Intelligence.2022.

A SURVEY OF NATURAL LANGUAGE PROCESSING TECHNOLOGY

S M BALAJI

GOVERNMENT ARTS COLLEGE (Autonomous), KUMBAKONAM.

Email:balajibalaji4153@gmail.com

ABSTRACT:

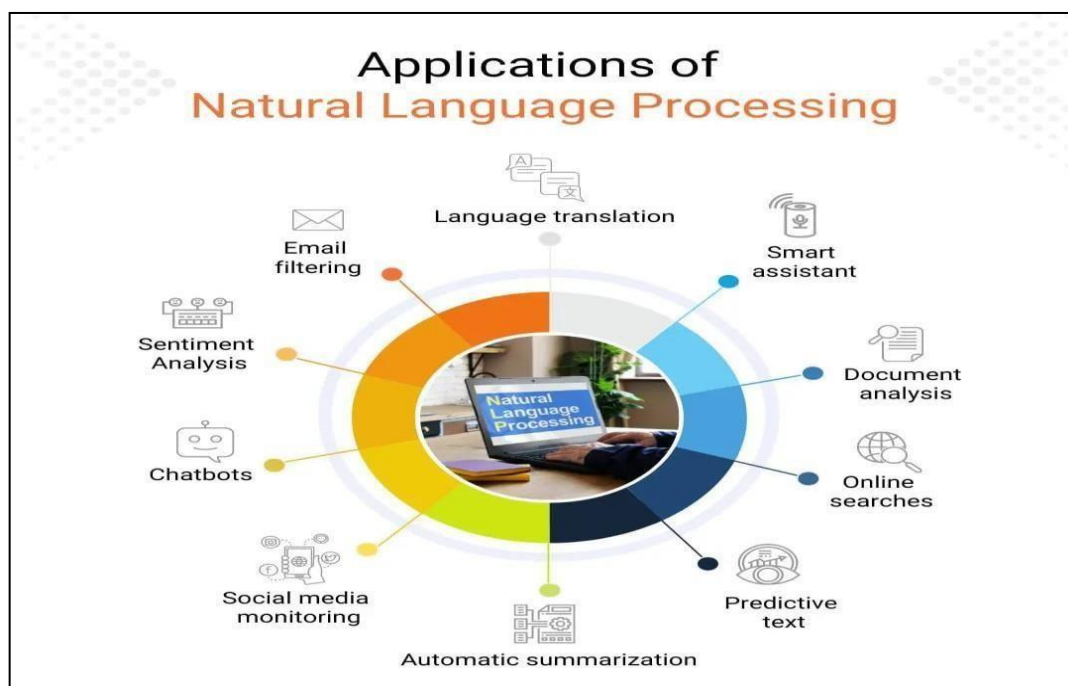
Natural language processing (NLP) is a branch of artificial intelligence that deals with the interaction between computers and human languages. NLP aims to enable computers to understand, interpret, generate, and manipulate natural languages, such as English, Chinese, Hindi, etc. NLP has many applications in various domains, such as machine translation, information extraction, sentiment analysis, question answering, and more. NLP researchers and developers use different techniques and methods to analyse and generate natural language texts, such as rule-based systems, statistical methods, machine learning algorithms, deep learning models, and hybrid approaches. NLP also relies on linguistic resources, such as corpora, lexicons, grammars, and knowledge bases, to provide information and knowledge about natural languages. NLP is an evolving and interdisciplinary field that draws from computer science, linguistics, mathematics, psychology, and cognitive science. NLP is constantly advancing with new research, innovations, and applications. Some of the current trends and challenges in NLP include Multilingual and cross-lingual NLP: developing systems that can process and generate multiple languages, or transfer knowledge and resources across languages. Low-resource and endangered languages: creating NLP tools and resources for languages that have limited data and documentation, or are at risk of extinction. Explainable and ethical NLP: ensuring that NLP systems are transparent, interpretable, fair, and accountable, and that they respect the privacy, security, and values of the users and society.

KEYWORDS:

NLP, Rule Based System, Linguistic Resources, Lexicons, Research, Innovations, Multilingual, Cross-lingual, Endangered, Transparent.

INTRODUCTION

In the Year 1948, the first recognisable NLP application was introduced in Birkbeck College, London. It's a sub-field of Computer Science, Human language, and Artificial Intelligence. NLP enables machines to understand, analysis, manipulate, and interpret human languages. Think of it as the technology that allows computers to "speak" and "comprehend" like humans. Developers use NLP to perform tasks such as translation, automatic summarization, Named Entity Recognition (NER), speech recognition, and more. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.



- 1972: Terry Winograd develops SHRDLU, an effective NLP system that can manipulate blocks in a virtual world using natural language commands.
- 1989: The Hidden Markov Model Toolkit (HTK) development helps researchers build statistical models for speech recognition.
- 2017: Introducing the Transformer model architecture powers models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer). These models achieve state-of-the-art results in a wide range of NLP tasks.
- 2020: The release of GPT-3 (Generative Pretrained Transformer 3) by Open AI, one of the most significant language models to date, can generate coherent and contextually relevant text.

REVIEW OF LITERATURE

Dr. Carol Friedman is a prominent figure in the field of NLP, particularly in the biomedical domain. She has made significant contributions to the development of NLP systems for clinical text analysis. One of her notable achievements is the creation of MedLEE, an early general-purpose clinical NLP system. MedLEE, developed at Columbia University, parses clinical text (including radiology reports) using a rule-based approach based on the sublanguage theory.

A systematic review was conducted using the PubMed database as a search platform. All published studies on the application of NLP in medicine (except biomedicine) during the 20 years between 1999 and 2018 were retrieved. The data obtained from these published studies were cleaned and structured. Excel (Microsoft Corp) and VOS viewer (Nees Jan van Eck and Ludo Waltman) were used to perform bibliometric analysis of publication trends, author orders, countries, institutions, collaboration relationships, research hot spots, diseases studied, and

research methods.

Diksha Khurana, along with her co-authors Aditya Koli, Kiran Khatter, and Sukhdev Singh, delved into the fascinating field of Natural Language Processing (NLP). Their paper titled "Natural Language Processing: State of The Art, Current Trends and Challenges". The authors discuss different levels of NLP and components of Natural Language Generation (NLG). They present the history and evolution of NLP. The paper highlights the state of the art by showcasing various applications of NLP. It also addresses the current trends and challenges in this dynamic field. NLP has found applications in diverse areas, including machine translation, email spam detection, information extraction, summarization, medical, and question answering.

METHODOLOGY

Data collection and preprocessing:

This involves gathering and preparing the natural language data for analysis, such as text or speech corpora. Preprocessing may include tasks such as tokenization, normalization, lemmatization, stemming, stop word removal, part-of-speech tagging, and parsing.

Feature extraction and representation:

This involves selecting and transforming the relevant information from the natural language data into numerical or symbolic features that can be used by machine learning algorithms. Feature extraction may include tasks such as word embedding, n-gram, bag-of-words, term frequency-inverse document frequency (TF-IDF), and topic modeling. Feature representation may include tasks such as one-hot encoding, vectorization, and dimensionality reduction.

Model selection and training:

This involves choosing and applying the appropriate machine learning algorithms to learn from the features and perform the desired NLP task. Model selection may include tasks such as hyper-parameter tuning, cross-validation, and evaluation metrics. Model training may include tasks such as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

Model deployment and evaluation:

This involves deploying the trained model to perform the NLP task on new or unseen data and evaluating its performance and accuracy. Model deployment may include tasks such as application programming interface (API), web service, or user interface. Model evaluation may include tasks such as error analysis, feedback, and improvement.

RESULT & DISCUSSION

NLP has many applications, such as machine translation, sentiment analysis, speech recognition, and text classification. NLP also faces many challenges, such as ambiguity, complexity, diversity, and evolution of natural language. The discussion of NLP can be divided into four phases, according to this paper. The first phase is the pre-NLP era, where the focus was on formal languages and logic. The second phase is the rule-based NLP era, where the

focus was on grammar and syntax. The third phase is the statistical NLP era, where the focus was on data and probability. The fourth phase is the current deep learning NLP era, where the focus is on representation and learning.

One of the recent trends in NLP is the use of deep learning, especially neural networks, to improve the performance of NLP systems. Neural networks are able to learn complex patterns and representations from large amounts of data, and can handle various tasks such as sequence modeling, attention, and generation. Some of the state-of-the-art neural network models for NLP include transformers, BERT, GPT-3, and XL Net.

CONCLUSION

In conclusion, natural language processing is a field of computer science, linguistics and artificial intelligence that deals with the interaction between computers and human languages. NLP enables computers to understand human language and respond in a way that is natural for humans. NLP has many applications in various domains, such as machine translation, sentiment analysis, and text classification. Natural Language Processing is the practice of teaching machines to understand and interpret conversational inputs from humans. Although continuously evolving, NLP has already proven useful in multiple fields. The different implementations of NLP can help businesses and individuals save time, improve efficiency and increase customer satisfaction.

REFERENCES

- [1] **Carol Friedman**, a natural-language processing system for the extraction of molecular pathways, Computer Science Department, queens college CUNY, FLUSHING, NY, USA, 2001.
- [2] **Diksha Khurana**, Natural language processing: state of the art, current trends and challenges, Department of Computer Science, Manav Rachna International Institute of Research and Studies, Faridabad, India, 2022

A SURVEY ON DATA ANALYTICS

T.DHARANI

GOVERNMENT ARTS COLLEGE (AUTONOMOUS) KUMBAKONAM,

Email: dharanitamil974@gmail.com

ABSTRACT:

The data available is growing at an exponential rate. The increase in data in itself is a minor problem, but the percentage of unstructured data in the overall data volume is what is concerning all. so it becomes a basic necessity to discover ways to process and transform complex, unstructured, or large amounts of data into meaningful insights, This brief outline of data analysis will help us understand what is data analysis, the value it holds in many industries worldwide and how majority of the organizations in various sectors bank on data analysis to survive the ongoing market race. This paper maintains its focus on explaining the basic procedures followed in obtaining something immensely useful from the available disorganized facts and figures by analyzing them. Also discussed briefly are its applications in areas such as management, retail, healthcare, education and so on. This paper highlights important concepts of data analysis.

KEYWORDS: Anomalies, data analysis, decision-making, productivity enhancement, unstructured data

INTRODUCTION

Data are now woven into every sector and function in the global economy, and, like other essential factors of production such as hard assets and human capital, much of modern economic activity simply could not take place without them. The concept of Data Analysis large pools of data that can be brought together and analyzed to discuss patterns and make better decisions will become the basis of competition and growth for individual firms, enhancing productivity and creating significant value for the world economy by reducing waste and increasing the quality of products and services. In fact, no business can survive without analyzing available data.

- The process of scrutinizing raw data with the purpose of drawing conclusion that information is called Data Analysis.
- The main aim of Data Analysis is to convert the available cluttered data into a format which is easy to understand, more legible, conclusive and which supports the mechanism of decision-making.

TYPES OF DATA ANALYTICS:

Data analytics is broken down into four basic types:

1. **Descriptive analytics:** This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. **Diagnostic analytics:** This focuses more on why something happened. It involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that

latest marketing campaign impact sales?

3. **Predictive analytics:** This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. **Prescriptive analytics:** This suggests a course of action. We should add an evening shift to the brewery and rent an additional tank to increase output if the likelihood of a hot summer is measured as an average of these five weather models and the average is above 58%

BENEFITS OF DATA ANALYTICS

Improved Decision Making: Foremost among the top data analytics benefits is better decision-making. It offers insightful, data-driven information that aids organizations in understanding their customers, operations, and markets. They can spot patterns, trends, and correlations.

Increased Efficiency and Productivity: Data analytics enables organizations to increase efficiency and productivity by automating and streamlining processes, maximizing resource allocation, and minimizing manual labor. Businesses can streamline their workflows by locating bottlenecks and getting rid of duplication.

Enhanced Customer Experience: By giving organizations useful insights into customer behavior, preferences, and needs, data analytics enables businesses to identify areas where they can improve their customer experience such as lowering wait times, enhancing customer service, or streamlining user interfaces.

Improved Risk Management: Businesses can find patterns and correlations in data from various sources that point to potential risks. Data analytics can, for instance, assist companies in identifying potential fraud, online threats, or operational risks. Businesses can also take preventative action to mitigate potential risks by monitoring data in real-time.

Competitive Advantage: Businesses can gain a competitive edge using data analytics to make more informed, data-driven decisions. Analyzing data from various sources allows businesses to understand market trends, consumer behavior, and competitor activities.

LEVELS OF DATA ANALYTICS:

1.Descriptive analytics: Descriptive (also known as observation and reporting) is the most basic level of analytics. Many times, organizations find themselves spending most of their time in this level.

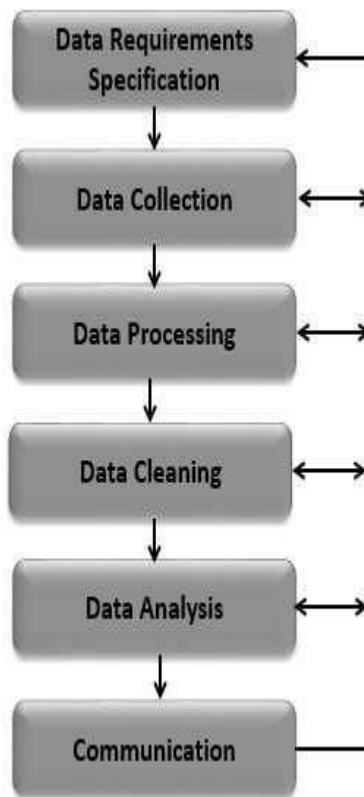
2.Diagnostic analytics: Diagnostic analytics is where we get to the why. We move beyond an observation (like whether the chart is trending up or down) and get to the “what” that is making it happen. This is where the ability to ask questions about the data and tie those questions back to

objectives and business imperatives is most important.

3. Predictive analytics: Predictive analytics allows organizations to predict different decisions, test them for success, find areas of weakness in the business, make more predictions—and so forth. This flow allows organizations to see how the first three levels can work together.

4. Prescriptive analytics: Prescriptive analytics exist at a very advanced level and is the most powerful and final phase, and truly encompasses the “why” of analytics. It’s when the data itself prescribes what should be done. Data-driven decision making is tied most closely to predictive and prescriptive analytics, even though these are the most advanced.

PROCESS OF DATA ANALYTICS:



CONCLUSION

The world is becoming data driven. Each and every decision is now based on taken based on the data available. The concept of data analysis- large pool of data can be brought together and analyzed to discuss patterns and make better decisions – will soon become the basis of competition and growth for individual firms, enhancing productivity and increasing the quality of products and services. Vast amounts of data are being generated in structured and unstructured form.

REFERENCES:

- [1] M. A. Alkhatib, —Analysis of research in healthcare data analytics, Australasian Conference on Information Systems, Sydney, pp. 1-16, 2015.
- [2] M. R. Berthold, Guide to Intelligent Data Analysis, Texts in Computer Science 42, Springer-Verlag London Limited 2010.

A SURVEY ON SPEECH RECOGNITION TECHNOLOGY

M GOPINATH

GOVERNMENT ARTS COLLEGE (Autonomous)KUMBAKONAM,Email:gopisiva328gmail.com

ABSTRACT:

Speech recognition, a pivotal technology in the realm of artificial intelligence, enables machines to interpret and comprehend spoken language. It involves converting spoken words into text or commands, facilitating seamless interaction between humans and computers. Through a combination of acoustic and language modeling techniques, speech recognition systems analyse audio input, extract meaningful features, and match them with linguistic patterns to accurately transcribe spoken words. The evolution of speech recognition has seen significant advancements, driven by deep learning algorithms, neural networks, and large datasets. Modern speech recognition systems leverage machine learning approaches such as deep neural networks, recurrent neural networks, and convolutional neural networks to achieve remarkable accuracy and robustness across various languages and accents. Applications of speech recognition span diverse domains, including virtual assistants, dictation software, voice-controlled devices, and automated customer service systems. These systems enhance accessibility, productivity, and convenience, empowering users to interact with technology through natural language interfaces effortlessly. Challenges in speech recognition persist, particularly in handling noisy environments, different accents, and understanding context-dependent speech. Ongoing research focuses on improving accuracy, enhancing real-time processing capabilities, and expanding language support to foster more seamless human-machine communication. As speech recognition continues to advance, its integration into everyday devices and applications promises to revolutionize how we interact with technology, making communication more intuitive and inclusive.

Keywords: Voice Recognition, Transcription, Noise Recognition, AI.

INTRODUCTION

Speech is the most basic, common and efficient form of communication method for people to interact with each other. People are comfortable with speech therefore persons would also like to interact with computers via speech, rather than using primitive interfaces such as keyboards and pointing devices. This can be accomplished by developing an Automatic Speech Recognition (ASR) system which allows a computer to identify the words that a person speaks into a microphone or telephone and convert it into written text.

As a result, it has the potential of being an important mode of interaction between human and computers. Although any task that involves interfacing with a computer can potentially use ASR. The ASR system would support many valuable applications like dictation, command and control, embedded applications, telephone directory assistance, spoken database querying, medical applications, office dictation devices, and automatic voice translation into foreign languages etc. In the current Indian context, these machine-oriented interfaces restrict the computer usage to miniature fraction of the people, who are both computer literate and conversant with written English.

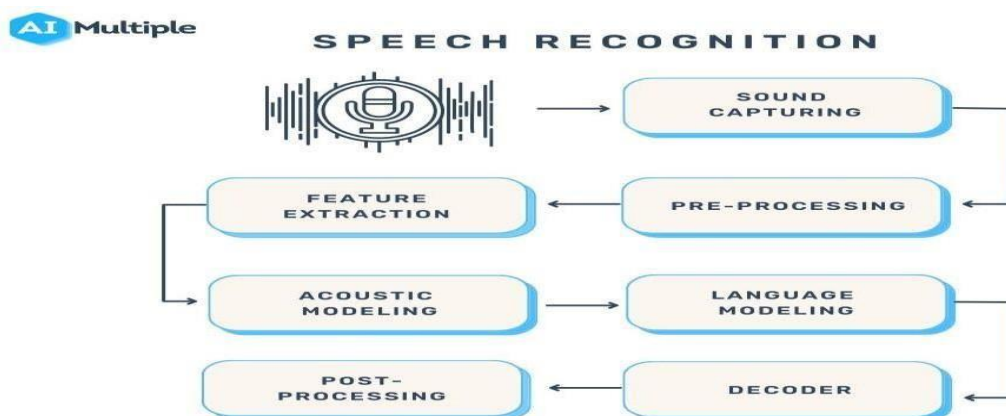
Communication among human beings is dominated by spoken language. Therefore, it is natural for people to expect speech interfaces with computers which can speak and recognize speech in native language. It will enables a common man to reap the benefit of information technology. India has a linguistically rich area which has 18 constitutional languages, which are written in 10 different scripts. Hence there is a special need for the ASR system to be developed in their native language. This paper provides an overview of speech recognition system and the review of techniques available at various stages of speech recognition. The paper is organized as follows. It presents the classification of speech recognition systems and explains about the growth of ASR systems.

REVIEW OF LITERATURE

A comprehensive review of the literature on Speech Recognition Technology reveals a rich tapestry of research spanning several decades, showcasing the evolution, challenges, and advancements in this dynamic field. The body of work not only reflects the persistent efforts of researchers but also underscores the transformative impact of speech recognition on diverse sectors of society.

Historical Overview: The journey begins with an exploration of the historical roots of speech recognition. From its nascent stages characterized by rudimentary systems with limited vocabulary, the literature chronicles the gradual progression towards more sophisticated models that leverage the power of deep learning and neural networks. Key milestones, such as the advent of Hidden Markov Models (HMMs) and the transition to modern machine learning techniques, provide a contextual backdrop for understanding the trajectory of speech recognition research.

Technological Advances: A significant portion of the literature is dedicated to detailing the technological breakthroughs that have propelled speech recognition to its current state. The transition from rule-based systems to statistical models and the subsequent



dominance of deep learning approaches, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), is meticulously explored.

Challenges and Limitations: The literature extensively addresses the challenges inherent in speech recognition. Issues such as environmental noise, speaker variability, and domain adaptation pose persistent obstacles. Researchers have delved into strategies for mitigating these challenges, including the integration of contextual information, the development of robust feature extraction techniques, and the exploration of unsupervised and semi-supervised learning paradigms.

Future Directions: The review concludes with an exploration of future prospects and emerging trends in speech recognition research. From the refinement of real-time processing capabilities to the integration of multi-model cues, the literature anticipates a continued evolution of speech recognition technology, with an emphasis on addressing existing limitations and expanding its applicability.

METHODOLOGIES

Speech recognition, also known as automatic speech recognition (ASR), enables seamless communication between humans and machines. This technology empowers organizations to transform human speech into written text. Speech recognition technology can revolutionize many business applications, including customer service, healthcare, finance and sales.

What are the features of speech recognition systems?

Speech recognition systems have several components that work together to understand and process human speech. Key features of effective speech recognition are:

- **Audio preprocessing:** After you have obtained the raw audio signal from an input device, you need to preprocess it to improve the quality of the speech input. The main goal of audio preprocessing is to capture relevant speech data by removing any unwanted artifacts and reducing noise.
- **Feature extraction:** This stage converts the pre-processed audio signal into a more informative representation. This makes raw audio data more manageable for machine learning models in speech recognition systems.
- **Language model weighting:** Language weighting gives more weight to certain words and phrases, such as product references, in audio and voice signals. This makes those keywords more likely to be recognized in a subsequent speech by speech recognition systems.
- **Acoustic modeling:** It enables speech recognizers to capture and distinguish phonetic units within a speech signal. Acoustic models are trained on large datasets containing speech samples from a diverse set of speakers with different accents, speaking styles, and backgrounds.
- **Speaker labelling:** It enables speech recognition applications to determine the identities of multiple speakers in an audio recording. It assigns unique labels to each speaker in an audio recording, allowing the identification of which speaker was speaking at any given time.
- **Profanity filtering:** The process of removing offensive, inappropriate, or explicit words or phrases from audio data.

REFERENCES

- [1] **Dong Yu and Li Deng**, "Automatic Speech Recognition: A Deep Learning Approach"
This book provides a comprehensive overview of ASR techniques, including deeplearning models and their applications.
- [2] **Daniel Jurafsky and James H. Martin**- "Speech and Language Processing" Although not exclusively focused on ASR, this book covers various aspects of naturallanguage processing, including speech recognition.

A SURVEY ON DATA ANALYTICS USING BIG DATA TOOLS

J.S. JAYA SURYA

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM Email: jsjayasurya9@gmail.com

ABSTRACT:

The age of big data is now coming but the traditional data analytics may not be able to handle such large quantities of data. The question that arises now is, how to develop a high-performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from big data. Data analytics has emerged as a crucial discipline in extracting insights, making informed decisions, and driving innovation across various domains. This paper provides a comprehensive overview of data analytics, covering its fundamental concepts, methodologies, and applications. We delve into the various techniques used in data collection, processing, analysis, and visualization, including descriptive, diagnostic, predictive, and prescriptive analytics. Furthermore, we explore the diverse applications of data analytics in business, healthcare, finance, marketing, and beyond, showcasing its transformative impact on decision-making processes and organizational performance. Additionally, we discuss the challenges associated with data analytics, such as data quality, privacy concerns, scalability, and interpretability, and propose strategies to address them. By understanding the principles and applications of data analytics, organizations can harness the power of data to gain a competitive advantage and drive innovation in today's data-driven world.

KEYWORDS:

Big data, data analytics, business, healthcare, finance, marketing, social networks, data mining, visualization, decision making.

INTRODUCTION

BIG DATA ANALYTICS

The larger the datasets, the more difficult they become to manage. Big Data can be coined as a term for datasets that become so huge that they become difficult to manage with traditional database management systems. These data sets enter a size which is beyond the handling capacity for common functional technologies used for capture, storage, analysis, management, visualization and presentations. Generally, data warehouses have been used to manage large datasets. Big data analytics can be broadly coined as the use of advanced analytic techniques against very diverse, large data sets. These data sets may be unstructured, semi-structured and structured data from different sources, in sizes ranging from terabytes (TB) to zettabytes.

VOLUME

The name Big Data itself is related to an enormous size. Big Data is a vast 'volume' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more. Facebook can generate approximately a billion messages, 4.5 billion times that the "like" button is recorded, and more than 350 million new posts are uploaded each day. Big data technologies can handle large amounts of data.

VARIETY

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources. Data will only be collected from databases and sheets in the past, but these days the data will come in array forms, that are PDFs, Emails, audios, SM posts, photos, videos, etc.

VERACITY

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development. For example, Facebook posts with hashtags.

VALUE

Value is an essential characteristic of big data. It is not the data that we process or store. It is valuable and reliable data that we store, process, and also analyze.

VELOCITY

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in real-time. It contains the linking of incoming data sets speeds, rate of change, and activity bursts. The primary aspect of Big Data is to provide demanding data rapidly. Big data velocity deals with the speed at the data flows from sources like application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.

REVIEW OF LITERATURE

Big data is transforming supply chain management by providing insights that enable better decision-making and optimization of operations. With the increasing availability of data and technological advancements, supply chain managers can leverage Big Data to gain a competitive advantage in the marketplace. Big data analytics can help supply chain managers to "optimize inventory levels, improve forecasting accuracy, reduce lead times, improve supplier performance, and enhance overall supply chain visibility" Supply Chain Digital.

In Modern world Big Data plays a vital role in Health Care Environment as data is growing exponentially, by this it very difficult to analyze data with traditional approaches. As the treatment cost is getting increased day by day, it has created so much burden on middle class people as they are not affordable for getting treatment by paying huge amount of money. So, the doors are opened for optimization in health care in curtailing cost involved in treatment for any type diseases by analyzing symptoms for various diseases.

Such that it would be easier for patients in getting awareness about the diseases before going for any treatment. For analyzing symptoms of various diseases there are so many challenges as data is in the form of both structured and unstructured. Once the disease is identified at early stage, it would be easier for doctors in providing better treatment for patients, so the life span of patients will increase to certain extent.

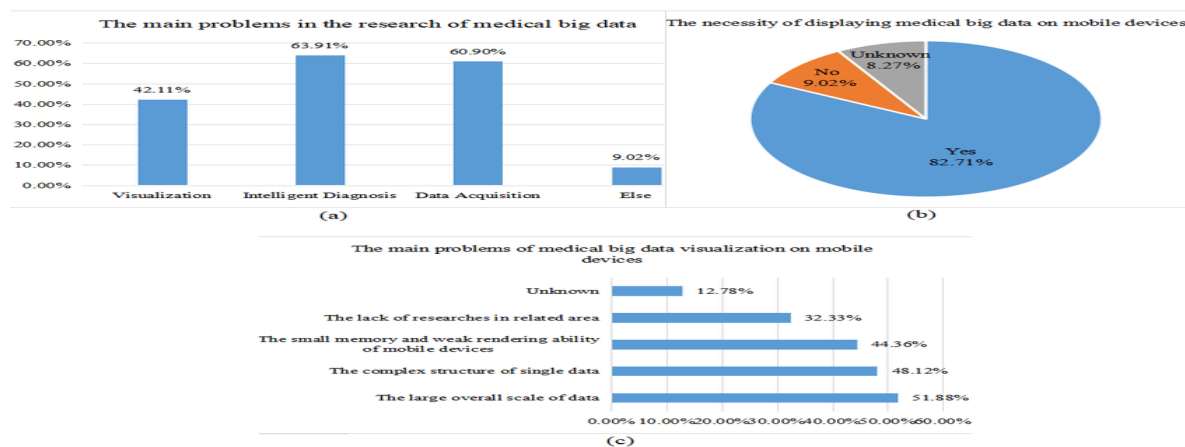
Social media are internet interaction platforms that allow people to share and consume information with one another. The Programmed Logic for Automatic Teaching Operations (PLATO) system, led by the University of Illinois and marketed by Control Data Corporation, was the first social network framework for multi-communication, which was developed in 1970. By permitting users to post contents that can be instantaneously edited and amended, social media began to revolutionize in 2004 with the launch of Facebook, even today, Facebook is one of the most prominent social media networks.

Given the above background, this study aims to address the second research question as in developing countries, health systems face significant quality, accessibility, affordability, and equity issues. On the one hand, developing countries such as India and South Africa has some of the world's most outstanding hospitals, adding to the expanding medical tourism industry. On the other hand, there is a severe scarcity of skilled medical personnel. Improved access to quality healthcare, particularly in rural and low-income settings; addressing the uneven ratio of qualified doctors to patients; improving the training and efficiency of doctors and nurses, particularly in complex procedures; and enabling the delivery of personalized healthcare at scale could all be addressed by new AI technologies.

METHODOLOGY

- ❖ SUPPLY CHAIN MANAGEMENT
- ❖ HEALTH CARE
- ❖ SOCIAL NETWORKS

RESULT AND DISCUSSION



CONCLUSION

Big data analytics, empowered by advanced big data tools, has revolutionized various industries including supply chain management, healthcare, and social networks. In supply chain management, big data tools enable real-time tracking of goods, optimize inventory management, and enhance demand forecasting, leading to improved efficiency and cost savings. In healthcare, big data analytics helps in personalized medicine, disease prediction, and treatment optimization by analyzing vast amounts of patient data. Social networks leverage big data tools to analyze user behavior, personalize content delivery, and enhance user engagement through targeted advertising and recommendation systems. Overall, big data analytics has become indispensable in these industries, driving innovation, improving decision-making, and ultimately delivering better services to customers and users.

REFERENCES

- [1] Chun-wei Tsai, Bigdata analytics – a survey, Department of Computer Science and Engineering, National Ilan University, Taiwan, 2015.
- [2] Md. Saifur Rahman, A Systematic Review of Bigdata in Social Media, School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, ND, USA. 2015
- [3] T. Ramesh, Exploring Big Data Analytics in Healthcare, Department of Computer Science and Engineering, Presidency University, Bangalore, India, 2020

A SURVEY ON ARTIFICIAL SUPER INTELLIGENCE (ASI)

JAYASRI G

GOVERNEMENT ARTS COLLEGE (Autonomous), KUMBAKONAM.

Email:jayasriraja1235@gmail.com

ABSTRACT

Artificial Intelligence is the science of making machines that can think like humans. So, AI is the backbone of smart assistants. Artificial Super Intelligence (ASI) is a software-based system with intellectual powers beyond those of humans across a comprehensive range of categories and fields of endeavor.

ASI is also known as 'Super AI'. It was different from regular AI, which involves as learning simulation of human intellectual capabilities, such as learning through the acquisition of information, reasoning and self-correction. With ASI, machines can think of the possible interpretations which are simply impossible for humans to think. This is because the human brain has a limit to the thinking ability which is constrained to some billion neurons.

It can also automate complex processes and improved accuracy and decision-making. AI is increasingly a part of our everyday lives in systems. Nevertheless, AI technology is in its early days of development. The intention behind ASI is to surpass human cognitive capacity, which is held back by chemical and biological limits of the human brain. Theoretically, ASI's superior capabilities would apply across many disciplines and industries and include cognition, general intelligence, problem-solving abilities, social skills and creativity. Advancements in AI for applications like Natural Language Processing (NLP) and Computer Vision (CV) are helping industries like financial services, healthcare, and automotive accelerate innovation, improve customer experience, and reduce costs. The goal for Super AI is to be able to do things such as recognize patterns, make decisions, and judge like humans.

Keywords: Artificial Intelligence (AI), Artificial Super Intelligence (ASI), Overview, Super AI, Learn from human thinks, Human brain limits, Accuracy, Early stage of development, Problem solving skills, Creativity, Advancements, NLP and CV, New invention on many sectors, Goal of Super AI.

INTRODUCTION

Artificial Super Intelligence (ASI)

Artificial intelligence (AI) refers to computer systems capable of performing complex tasks that historically only a human could do, such as reasoning, making decisions, or solving problems. AI technology can process large amounts of data in ways, unlike humans. And also AI is an umbrella term that encompasses a wide variety of technologies, including machine learning, deep learning, and Natural Language Processing (NLP). Artificial super intelligence is considered the most advanced, powerful, and intelligent type of AI that transcends the intelligence of some of the brightest minds, such as Albert Einstein. It can also automate

complex processes and minimize downtime by predicting maintenance needs. The first step in developing super intelligent technology is to establish artificial general intelligence (AGI). AGI is an AI system that can perform any task a human can with the same capabilities. AI is the backbone of smart assistants, which can be accessed through most phones on the market these days and are also being integrated into cars and smart home devices. Such as virtual assistants, expert systems and self-driving cars. It was different from regular AI, which involves as learning simulation of human intellectual capabilities, such as learning through the acquisition of information, reasoning and self-correction. Engineers, AI researchers, and practitioners are developing technology and machines with AGI, which is expected to pave the way for ASI development. ASI systems can quickly understand, analyze, and process circumstances to stimulate actions.



REVIEW OF LITERATURE

Artificial Intelligence (AI) is machine intelligence that mimics a human mind’s problem-solving and decision-making capabilities to perform various tasks.

Narrow AI focuses on a single task and is constrained by constraints to not go beyond that, leaving it unable to solve unfamiliar problems. But general AI can manage to perform a broad range of tasks by using humanlike cognitive capabilities. Super AI can be able to outperform human intelligence.

A big step toward developing an ASI would be to realize an artificial general intelligence (AGI) or Strong AI. AGI would be capable of cross-domain learning and reasoning with the ability to make connections across different fields. Just like ASI, true AGI has yet to be developed.

AI, while exceeding human beings in certain areas, still struggles to match the human ability to learn and adapt to new situations. Learning algorithms, inspired by how the human brain learns, enable AI to improve its performance over time. This continuous learning is crucial for achieving human-level intelligence, allowing AI to acquire knowledge and adapt to new situations without explicit programming.



METHODOLOGY

This requires the use of algorithms, statistical models, and linguistic rules to teach computers to recognize patterns in human language, and to derive meaning from text, speech, and other forms of natural language data.

Generative AI: Open AI's ChatGPT uses a large language model trained on a massive dataset of text and code, allowing it to process and generate human language with remarkable fluency and accuracy. Its ability to understand the complexity of written sentences, engage in conversation and generate creative output like poems, scripts and music is crucial to achieving human-level intelligence. Uses large language models (LLMs) to generate text in response to questions or comments posed to it. **Google translate** uses deep learning algorithms to translate text from one language to another.

Self-driving cars: Tesla has shown the potential of self-driving cars. Self-driving cars utilize a combination of sensors, cameras and powerful AI algorithms to navigate roads autonomously. The advanced perception and decision-making capabilities developed for self-driving cars are directly relevant to ASI. The ability to process complex sensory data and make real-time decisions in dynamic environments is a crucial aspect of general intelligence, a key goal of ASI research.

Healthcare: AI is also making significant strides in healthcare, with machine intelligence now analyzing medical images and data to assist doctors in diagnosing diseases. Companies like IBM Watson health and deep mind health are developing AI-powered systems that can detect cancer, heart disease and other conditions with high accuracy. These advancements in medical AI are paving the way for developing even more sophisticated systems that could one day diagnose and treat diseases autonomously.

RESULT AND DISCUSSION

ASI systems not only understand human sentiments and experiences but can also evoke emotions, beliefs, and desires of their own, similar to humans.

- ✓ Capability to predict and adapt
- ✓ Make decisions on their own
- ✓ Imitate human cognition
- ✓ Continually learn and evolve

CONCLUSION

With super intelligence, machines can think of the possible abstractions/interpretations which are simply impossible for humans to think. This is because the human brain has a limit to the thinking ability which is constrained to some billion neurons.

Upon mastering the general AI stage where it surpasses human intellect across

every field, we can begin to envision a future that marks the beginning of a super AI era. The future would then mean that we are surrounded by more intelligent, conscious, and self-aware entities.



REFERENCES

- [1] Stuart Russell and Peter Norvig - “Artificial Intelligence: a modern approach”
- [2] Nils J. Nilsson - “Artificial Intelligence: A New Synthesis”
- [3] Nick Bostrom - “Super intelligence: Paths, Dangers, Strategies”

A SURVEY OF POWERFUL VIRTUAL AGENTS

KALAIYARASI B

GOVERNMENT ARTS COLLAGE (AUTONOMOUS) KUMBAKONAM

Email: kalaibalakrishnan0203@gmail.com

ABSTRACT

Nowadays, customers expect to get support outside store hours, even 24/7, when live agents would be cost prohibitive. Therefore, companies deploy chatbots to provide a first-line response to the most popular inquiries submitted through various information channels. Powerful chatbots can initiate a dialogue with a human over the internet or by another channel, such as text messaging. They can also download information from various systems and present it to the user clearly and coherently.

Microsoft's technology makes it possible to build conversation flows with an intuitive graphical interface instead of advanced code. Power Virtual Agents has built-in AI language processing mechanisms, which it uses to create the most natural and human-like conversation possible. The chatbots are available for use in many channels, not just within the Microsoft ecosystem—on your website/online store without the help of any advanced developers. Naturally, PVA is perfectly integrated with Power Automate services, so users can create advanced custom workflows and retrieve information from external sources and other systems using pre built connectors.

Keywords: Virtual agents, PVA, Chat bots, micro soft.

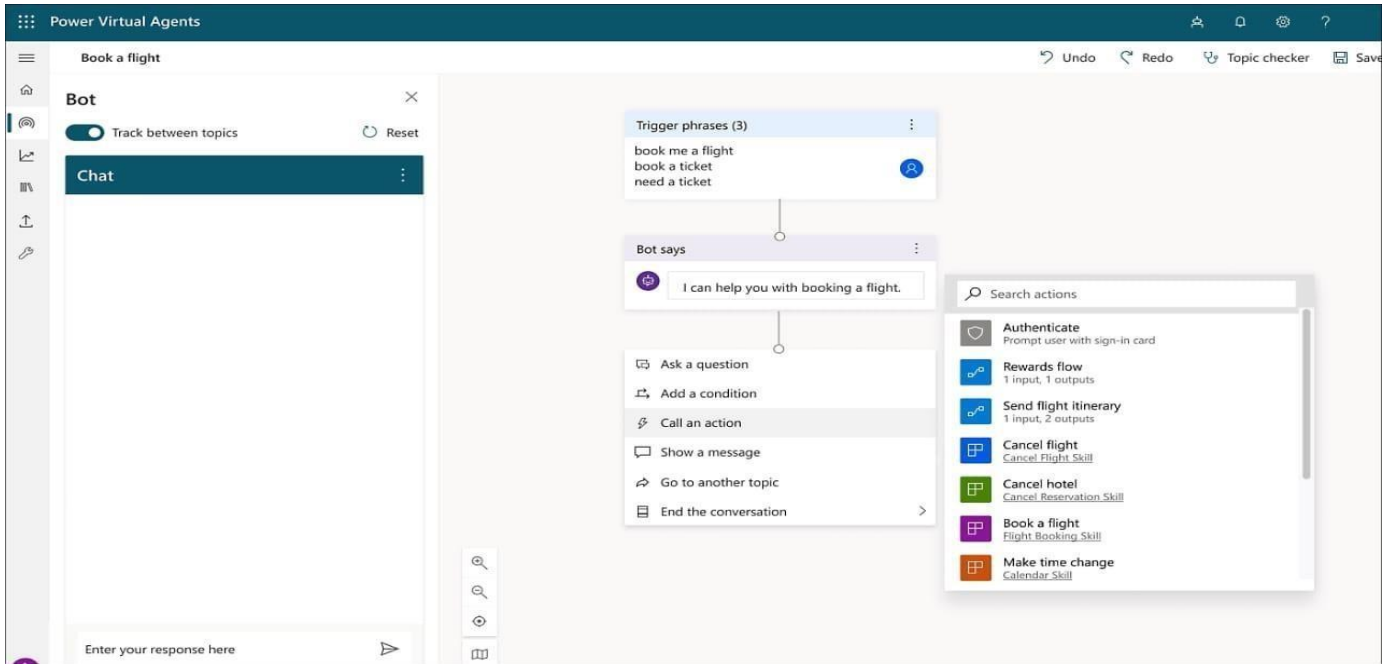
INTRODUCTION

Virtual Agents (VA) introduce conversational artificial intelligence (AI) and Natural Language Processing (NLP) to CCAI Platform. You can use the AI-powered virtual agents to act as the first line of customer support or configure them to handle support requests with limited to no human agent intervention.

You can use virtual agents to do the following:

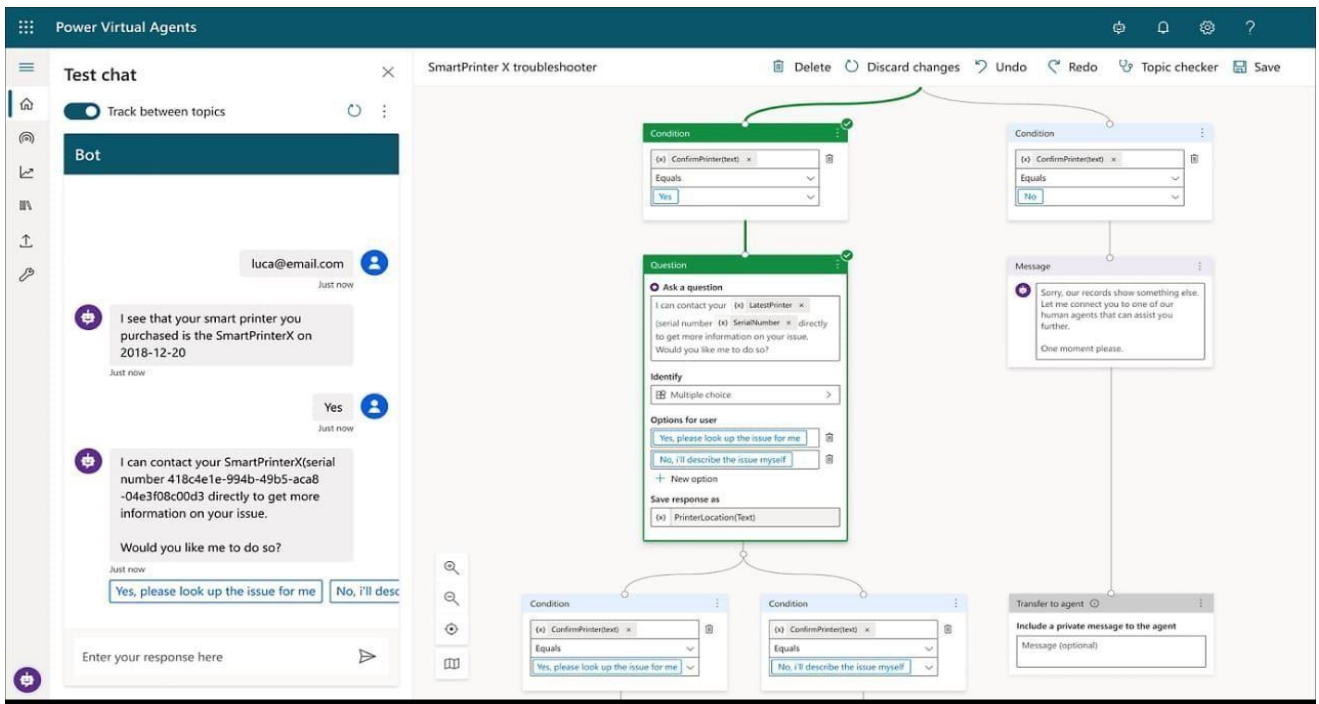
- Configure multiple, distinct VAs that focus on specific issues and assign to a specific queue.
- Assign a virtual agent to answer incoming calls, chats, or leverage existing routing options at the queue level.

Methodology

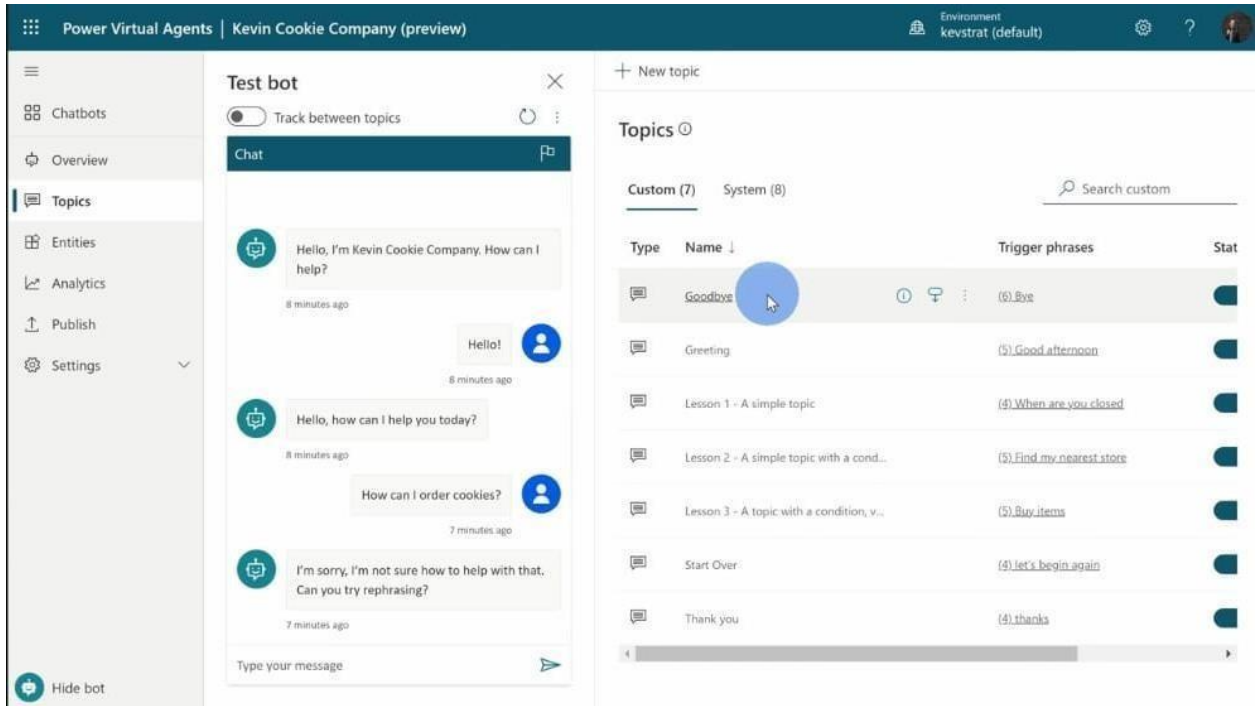


Enable any developer to build bots:

Keep it simple or extend your bot:



Adding topics for creating chatbots:



CONCLUSION

It provides lots of benefits. It helps to create a Power Virtual Agents bot and create a system fallback topic. Add Question Answering as an action to a topic as a Power Automate flow. Create a Power Automate solution. Add a Power Automate flow to your solution.

REFERENCE

- [1] Williams, D. (2021, January 25). How to Build Your First Power Virtual Agent Bot in Microsoft Teams. Retrieved from
- [2] Microsoft Power Virtual Agents Community. (n.d.).

A SURVEY ON NATURAL LANGUAGE GENERATION (NLG)

S.KAVIYA

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM

Email:kaviyas2512@gmail.com

ABSTRACT:

Personalization—adaptively to the individual—is becoming an essential component of any computer-based system. In e-health systems, personalization of health information is emerging as a key factor in the trend to patient-centric care. Patient-centric healthcare aims to engage patients. In their treatment to promote greater compliance and satisfaction with their therapeutic regimens. Resulting in both better patient outcomes and reduced healthcare costs. We have developed a Prototype Web-based Natural Language Generation system for the authoring and subsequent personalization of patient education materials. Our Initial domain of application is reconstructive Breast surgery, but our Natural Language software Tools and authoring methodologies are generally Applicable to all medical interventions.

Keywords: Education, Personalized e-Health, Natural Language Generation, Artificial Intelligence

INTRODUCTION

Natural language generation

Natural Language Generation, otherwise known as NLG, is a software process driven by artificial intelligence that produces natural written or spoken language from Structured and unstructured data. It helps computers to feed back to users in human language that they can comprehend, rather than in a way a computer might.

For example, NLG can be used after analyzing customer input (such as commands to voice assistants, queries to chatbots, calls to help centers or feedback on survey forms) to respond in a personalized, easily-understood way. This makes human-seeming responses from voice assistants and chatbots possible.

It can also be used for transforming numerical data input and other complex data into reports that we can easily understand. For example, NLG might be used to generate financial reports or weather updates automatically.

Data input: Structured data is the first input used by NLG systems. This information may originate from a number of sources, including spreadsheets, databases, and other organized formats.

Content Planning: Based on an analysis of the input data, the system decides what details to include in the text that is generated. Making choices regarding the selection of content, arrangement, and general structure is required.

Text Planning: The NLG system arranges the content's natural language expression after it has been decided upon. It chooses the right wording, tone, and style for the text that is generated.

Sentence Generation: Using the planned content as a guide, the system generates individual sentences. Choosing the right words, phrases, and syntactic structures is necessary for this. While some NLG systems generate text using pre-defined templates, others might use more advanced techniques like machine learning.

Advantages of NLG

► Once implemented, using NLG is less expensive and more time efficient than employing a person.

► NLG can also help businesses offer faster customer service response times. No matter the time of day or day of the week, customers receive immediate answers to their questions.

► Pertained machine learning models are widely available for developers to facilitate different applications of NLG, making them easy to implement.

Disadvantages of NLG:

► Training can be time-consuming. If a new model needs to be developed without the use of a pre-trained model, it can take weeks before achieving a high level of performance

► another disadvantage of NLG is that ML is not percent reliable. There is always a possibility of errors in predictions and result.

REVIEW OF LITERATURE

The field of Natural Language Generation (NLG) has witnessed significant advancements in recent years, driven by the increasing demand for automated content generation in various applications. This review synthesizes key findings from seminal works and recent studies to provide insights into the current state and future directions of NLG research.

Early contributions by McKeown (1985) laid the foundation for NLG systems, introducing techniques such as template-based generation and surface realization. These pioneering efforts paved the way for subsequent developments in text planning, micro planning, and lexicalization, forming the basis for modern NLG architectures.

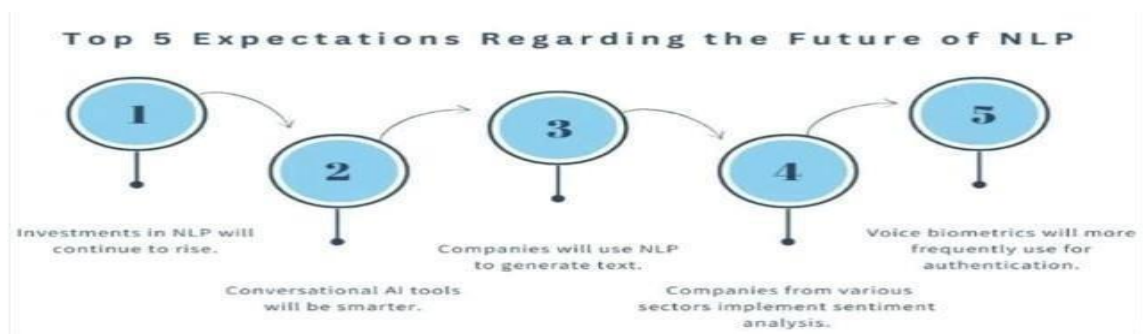
Reiter and Dale (2000) provide a comprehensive overview of NLG methodologies, covering rule-based, template-based, and statistical approaches. Their taxonomy of NLG systems remains influential in categorizing different generation techniques and guiding research directions. However, the emergence of deep learning models has expanded the scope of NLG beyond traditional rule-based paradigms, leading to new challenges and opportunities.

Recent advancements in deep learning, particularly with the advent of transformer-based models like GPT (Radford et al., 2019) and BERT (Devlin et al., 2019), have revolutionized NLG research. These models leverage large-scale pertaining on text corpora to generate human-like language with remarkable fluency and coherence. Fine-tuning these pertained models on specific NLG tasks has yielded state-of-the-art performance across various domains, including **text summarization, machine translation, and dialogue generation.**

NATURAL LANGUAGE GENERATION SIX STEPS



Future of NLG



CONCLUSION

► Natural Language Generation is the practice of teaching machines to understand and interpret conversational inputs from humans. NLG based on Machine Learning can be used to establish communication channels between humans and machines. Although continuously evolving, NLG has already proven useful in multiple fields. The different implementations of NLG can help businesses and individuals save time, improve efficiency and increase customer satisfaction.

REFERENCE:

- [1] Bental DS, Cawsey AJ, and Jones R. Patient information systems that tailor to the individual. *Patient Ed. and Counselling*, 36, 1999.
- [2] DiMarco C, Hirst G, and Hovy E. Generation by selection and repair as a method for adapting text for the individual reader. *Flexible Hypertext Wkshop, ACM Hypertext Conf*, 1997.

A SURVEY ON CYBER SECURITY

Kavya Arjunan

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM. Email: Kavyaarjunan2301@gmail.com

ABSTRACT:

As the digital landscape expands, so too does the threat landscape, making cybersecurity a critical concern for organizations and individuals alike. With the proliferation of sophisticated cyberattacks, there is an urgent need for innovative solutions to bolster defense mechanisms. Artificial Intelligence (AI) has emerged as a potent ally in this ongoing battle, offering capabilities that enhance threat detection, response, and prevention. This abstract explores the intersection of AI and cybersecurity, examining how AI technologies such as machine learning, natural language processing, and anomaly detection are being leveraged to strengthen cyber defenses. By analyzing vast amounts of data in real-time, AI systems can identify patterns, anomalies, and potential threats that traditional security measures might overlook. Moreover, AI-powered systems can adapt and evolve to combat emerging threats, providing a proactive defense strategy in an ever-evolving threat landscape. However, the deployment of AI in cybersecurity also raises ethical, privacy, and reliability concerns, underscoring the importance of responsible AI implementation and ongoing monitoring. Despite these challenges, the integration of AI into cybersecurity frameworks holds significant promise for fortifying digital infrastructure and safeguarding against malicious actors in the digital age.

INTRODUCTION

Cyber Security involves protecting information and systems from major cyber threats, such as cyber terrorism, cyber warfare, and cyber espionage. Cyber threats take aim at secret, political, military, or infrastructural assets of a nation, or its people. Cyber security is therefore a critical nation, or its people. Cyber security is therefore a critical part of any governments' security strategy.

Cyber terrorism is the disruptive use of information technology by terrorist groups to further their ideological or political agenda. This takes from the form of attacks on networks, computer systems, and telecommunication infrastructures.



What is Cyber Security?

Cyber security is the practice of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks. It's also known as information technology security or electronic information security. The term applies in a variety of contexts,

from **Network security** is the practice of securing a computer network from intruders, whether targeted attackers or opportunistic malware.

Application security focuses on keeping software and devices free of threats. A compromised application could provide access to the data its designed to protect. Successful security begins in the design stage, well before a program or device is deployed.

Information security protects the integrity and privacy of data, both in storage and in transit.

Operational security includes the processes and decisions for handling and protecting data assets. The permissions users have when accessing a network and the procedures that determine how and where data may be stored or shared all fall under this umbrella.

Disaster recovery and business continuity define how an organization responds to a cyber-security incident or any other event that causes the loss of operations or data. Disaster recovery policies dictate how the organization restores its operations and information to return to the same operating capacity as before the event. Business continuity is the plan the organization falls back on while trying to operate without certain resources.

End-user education addresses the most unpredictable cyber-security factor: people. Anyone can accidentally introduce a virus to an otherwise secure system by failing to follow good security practices. Teaching users to delete suspicious email attachments, not plug in unidentified USB drives, and various other important lessons is vital for the security of any organization.

business to mobile computing, and can be divided into a few common categories.

Challenges of cyber security

In the digital era, cyber security is a critical concern for people, corporations, and governments. With the increased use of technology and digital devices, it is more necessary than ever to secure electronic devices, networks, and data against unwanted access, theft, and damage. With the advancement of technology, the cybersecurity action of protecting an organization, employees, and critical assets from cyber threats faces several challenges. In this article, we will discuss the challenges faced by the cybersecurity industry and the future directions that can help address these challenges.

Sophisticated nature of cyber-attacks

The complex nature of cyber-attacks to gain unauthorized access to computer systems and networks poses a significant challenge in the area of cyber security. Intruders (cyber attackers) develop and employ very sophisticated and advanced techniques to gain unauthorized access or exploit vulnerabilities. Some of the complex cyber-attacks that use more advanced techniques to breach defenses and exploit a vulnerability are multi-vector attacks, Polymorphic and Fileless Malware, zero-day exploits, and advanced persistent attacks.

Advanced persistent threats (APT) are sophisticated, targeted, and organized attack that specifically targets the diplomatic, information technology industry, Military services, chemical industries, and other sensitive areas to filter confidential information and damage the targeted industry or area through maintaining persistence using different ways. The APT is a state-sponsored attack by governments and organized criminal groups to get economic, political, and strategic advantages by stealing information and resources from critical infrastructure and industries such as diplomats, national defense, government institutes,

manufacturing industries, military plans, and other sensitive areas. One of the APT attacks was the "Aurora operation" a sophisticated cyber-attack that happened in 2009 that targeted technology companies and IT industries such as Google, Adobe, Juniper Networks, and others.

Zero-day exploits are attacks that target newly discovered vulnerabilities that software developers have not yet addressed. Organizations become susceptible when attackers sell or use these vulnerabilities before they are found or fixed. Advanced threat intelligence, behavior-based analysis, and prompt software patching are necessary for identifying and thwarting zero-day exploits. The vulnerability attacks of the zero-day exploit have divided their cycle into five stages.

Zero-Day Attacks: These are the initial attacks that exploit vulnerabilities that are not known to the public or the software/hardware vendor. These vulnerabilities are typically discovered by malicious actors (black hat hackers), and they do not disclose them to the public or the affected party. Example: - Stuxnet was a highly sophisticated computer worm that targeted supervisory control and data acquisition (SCADA) systems, particularly those used in Iran's nuclear facilities. It exploited multiple zero-day vulnerabilities to gain access and manipulate industrial systems, causing physical damage to centrifuges used in uranium enrichment.

Pseudo-Zero-Day Attacks: Pseudo-zero-day attacks occur when the exploit is still relatively unknown, but it may have been discovered by a limited number of attackers. The vulnerability may not be widely publicized, but it's not as closely guarded a secret as in the first stage.

Potential for Pseudo-Zero-Day Attack: At this stage, there may be indications or clues that a vulnerability exists, but it hasn't been widely exploited yet. Security researchers or organizations might suspect the presence of a vulnerability based on unusual or suspicious activities.

Potential for Zero-Day Attacks: In this stage, details about the vulnerability have been made public. This might happen when a security researcher or a responsible disclosure process informs the affected party about the vulnerability, or it might become known through other means. Automated attack code or programs for exploiting the vulnerability may start to appear, making it more accessible to attackers.

Passive: This stage refers to the period after a vulnerability has been discovered, disclosed, and patched by the vendor or mitigated in some way. During this time, the vulnerability is no longer a zero-day, and organizations are expected to apply patches or counter measures to protect their systems.

Types of cyber threats:

The threats countered by cyber-security are three-fold:

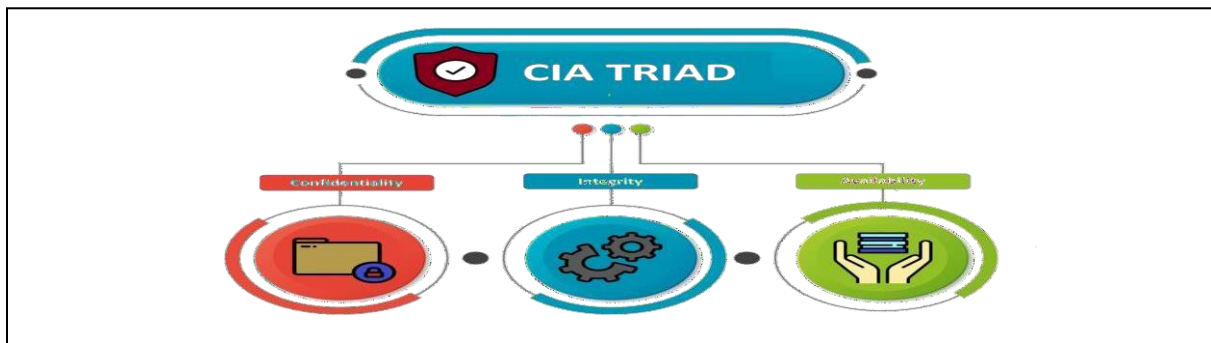
1. **Cybercrime** includes single actors or groups targeting systems for financial gain or to cause disruption.
2. **Cyber-attack** often involves politically motivated information gathering.
3. **Cyberterrorism** is intended to undermine electronic systems to cause panic or fear.



Types of Cyber Attacks

The CIA Triad:

The CIA triad which stands for Confidentiality, Integrity, and Availability is a design model to guide companies and organizations to form their security policies. It is also known as the AIC triad to avoid confusion with Central Intelligence Agency (CIA). The components of the triad are considered to be the most important and fundamental components of security. So let me brief you all about the three components



CONCLUSION

Cyber security is the process of protecting and safeguarding computer systems, networks, and data from cyber threats that attack the confidentiality, integrity, and accessibility of information systems. Cyber security is essential for protecting the safety of individuals and organizations since they highly depend on digital technologies. Cyber security is applicable in different application areas such as health centers, financial institutions, smart cities, grid systems, government organizations, education, and the military.

REFERENCES

[1] Tanishk Jharwal, cyber security tools and cyber-attacks, swami keshvanand institute of Technology Management and Gramothan SKIT, 2020

A SURVEY ON ARTIFICIAL INTELLIGENCE

KEERTHANA R

GOVERNMENT ARTS COLLAGE (AUTONOMOUS), KUMBAKONAM Email:krthnaraj1510@gmail.com

ABSTRACT

This project explores the applications and advancements of artificial intelligence (AI) in various domains, including healthcare, finance, transportation, and entertainment. Through a comprehensive review of literature and case studies, the project examines how AI technologies such as machine learning, natural language processing, and computer vision are revolutionizing industries, enhancing efficiency, and driving innovation. Additionally, it delves into the ethical implications and societal impacts of AI adoption, addressing concerns related to privacy, bias, and job displacement. The findings highlight the transformative potential of AI while emphasizing the need for responsible development and deployment to ensure its benefits are maximized while mitigating potential risks.

KEYWORDS: Artificial Intelligence, AI, Machine Learning, Deep Learning, Natural Language Processing, NLP, Computer Vision, Healthcare, Finance, Transportation, Entertainment, Ethics, Privacy, Bias, Job Displacement, Innovation, Society, Responsible AI.

INTRODUCTION

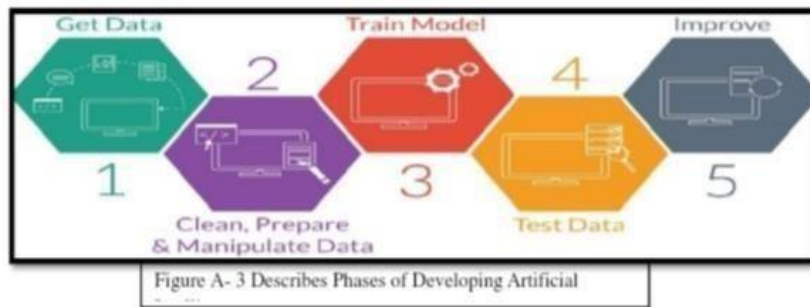
Since the invention of computers or machines, their capability to perform various tasks went on growing exponentially. Humans have developed the power of computer systems in terms of their drivers' working domains, their increasing speed and reducing size with respect to time. A branch of Computer Science named Artificial Intelligence pursues creating the computers or machines as intelligent as human beings.

HISTORY OF ARTIFICIAL INTELLIGENCE

Artificial Intelligence was first proposed by John McCarthy in 1956 in his first academic conference on the subject. The idea of machines operating like human beings began to be the center of scientist's mind and whether if it is possible to make machines have the same ability to think and learn by itself was introduced by the mathematician Alan Turing.

AI ALGORITHMS AND MODELS

AI is mainly based on algorithms and models as a technique which is designed based on scientific findings such as math, statistics, and biology (Li & Jiang, (n.d.)). AI works based on several models such as: Ant Colony Algorithm, Immune Algorithm, Fuzzy Algorithm, Decision Tree, Genetic Algorithm, Particle Swarm Algorithm, Neural Network, Deep Learning and in this report, I will discuss some of the most known models which are: Support Vector Machine, and the Artificial Neural Network.



APPLICATIONS ON ARTIFICIAL INTELLIGENCE

AI can be designed using lots of algorithms. Here are some of the greatest AI applications that we are probably using in our daily life without knowing

- Voice recognition Virtual agents
- Machine learning platform
- AI optimized hardware
- Decision management
- Deep learning platform
- Biomatters
- Robotic process automation
- Text analytics and NLP
- Adaptive Manufacturing

DOMAINS OF AI:

The major domains of AI (Fig. 1) are neural network, robotics, expert systems, fuzzy logic systems, natural language processing (NLP). Neural networks: these can be described as the representation of human neural system, i.e., neurons and dendrites in the form of layers and nodes representing data..

Robotics: It is the domain of AI which is mostly associated with the development of intelligent machines in the form of robot which obeys human instructions. Robots used in industry, medical surgery, restaurants, etc., are classified under this category.

Expert system: These are systems which make decisions with the help of data present in the knowledge base and getting guidance by an expert.

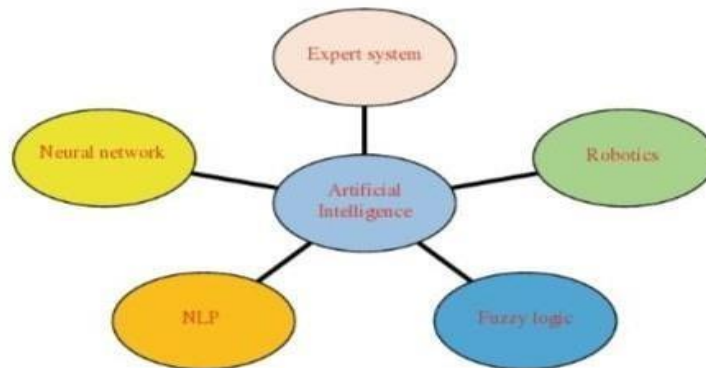


Fig.1 Various domains of Artificial Intelligence

Fuzzy logic system: This domain is considered as resembling the human thinking method and decision-making. It is quite similar to the way humans decide between 0 and 1, but it also deals with all the possibilities between 0 and 1. Examples of fuzzy logic systems used are in consumer electronics, automobiles, comparison of data etc.,

Natural language processing (NLP): This domain deals with bridging the gap of communication between the computer and human languages. It is basically the interaction between computer and human in a smart way.

Artificial Intelligence for Advanced Medical Diagnosis:

Medical field is the area where the AI has put its most important impact which has changed the way of medical diagnosis. It has made the diagnosis process much more effective, efficient, faster, and much more reliable. It never aims to replace the doctors but to help them in making their service easier. It basically acts as an assistant to the physician. The role of AI in medical diagnosis is classified into two types, i.e., virtual and physical.

CONCLUSION:

Artificial intelligence and its related technology like machine learning, deep learning, natural language processing has invaded the world undergoing fourth phase of industrial revolution in a very strong manner. The use of machine learning in health care has really modified the scenario in the healthcare system and still there is more scope for research and development

REFERENCE:

- [1] Spector, L.: Evolution of artificial intelligence. Artif. Intell. 170(18), 1251–1253 (2006)
- [2] Abbod, M.F., Linkens, D.A., Mahfouf, M., Dounias, G.: Survey on the use of smart and

adaptiveengineering systems in medicine. *Artif. Intell. Med.* 26(3), 179–209 (2002)

A SURVEY OF ATTENDANCE MANAGEMENT SYSTEM USING FACE RECOGNITION

S. KIRANISHA

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM. Email: kiranishakiranisha6@gmail.com

ABSTRACT:

In colleges, universities, organizations, schools, and offices, taking attendance is one of the most important tasks that must be done on a daily basis. The majority of the time, it is done manually, such as by calling by name or by roll number. The main goal of this project is to create a Face Recognition-based attendance system that will turn this manual process into an automated one. This project meets the requirements for bringing modernization to the way attendance is handled, as well as the criteria for time management. This device is installed in the classroom, where and student's information, such as name, roll number, class, sec, and photographs, is trained. The images are extracted using Open CV. Before the start of the corresponding class, the student can approach the machine, which will begin taking pictures and comparing them to the qualified data set. The face represents a unique part of the human body as well as a biometric identifier. Using facial features, we can achieve create and use applications or systems based on facial recognition. This project involves the construction of such a system which will use facial features to mark the presence, entry time, and finally the creation of an extracted document in the Excel application. The project was developed by researching and using libraries, models as well as Machine Learning based algorithms that are necessary for face detection, training, and recognition during system operation. Python programming language is used to create the system, and CSS is used to design interfaces. The system is built in the form that professors have access to the registration of new students, as well as the creation of relevant courses and obtaining the final attendance report after receiving notes from the system.

KEYWORDS: Open CV, Biometric, Machine Learning, Python

INTRODUCTION

Welcome to the future of attendance management! Our Face Recognition Attendance Management System revolutionizes traditional methods, ensuring accuracy and efficiency. Say goodbye to manual tracking – our system harnesses advanced facial recognition technology to seamlessly record attendance. With the power of bio metrics, it not only enhances security but also streamlines the entire process, saving time and reducing errors. Join us on a journey where technology meets convenience, making attendance management smarter and more accessible than ever before

Face recognition technology approaches are generally classified as feature-based approaches and holistic (comprehensive) approaches. In holistic approaches, recognition is based on global facial features, while in feature-based approaches, recognition is performed using local facial features. The face recognition system is the perfect way to solve and avoid the problems that the traditional system displays, the presence in this system will be obtained in real-time and fast, the data is safe at any time so there is no risk of loss by the lecturer, the data is always accurate as it is not accessible to students and is automatically generated by the system. It is a biometric technique that works in cases

where the image of a certain person matches the images of the face which are stored in a database.

REVIEW OF LITERATURE

Face recognition is a rapidly evolving field with extensive literature covering various aspects. Key areas of focus include face detection, feature extraction, and classification. Some notable approaches include Eigen faces, Fisher faces and deep learning-based methods.

Recent literature explores convolutional neural networks (CNNs) for improved accuracy. Notable models include VGGFace, FaceNet, and Deep Face. Challenges persist, such as handling variations in pose, lighting, and occlusion.

Review papers, like "Face Recognition: A Literature Review" by Zhao et al. (2018), provide comprehensive overviews. Stay updated on conferences like CVPR and ICCV for the latest research. Consider exploring datasets like LFW, CelebA, and MegaFace for benchmarking. Ethical concerns related to privacy and bias are also critical considerations in face recognition research.

Automated smart attendance management using face recognition: KoliPaka Preethi, swathy, vodithala (2021) The proposed method consists of different stages to mark the attendance live A. Face Detection B. DataSet Creation and Training C. Face Recognition and Updating attendance

METHODOLOGY

Register new ID students

Whereas by clicking the "Register new student" button, a new interface has been designed which will accept inputs that have previously been defined as necessary for the specification of new students, where each student in addition to the name will also contain an Id which is unique to all. After filling in the required data in the above fields then it is necessary to click the Take Image button in which case the system opens the camera and then captures images and finally gives the message if the images were taken successfully.

Take attendance UI

After capturing the images, the system training is required, a function which is defined in the file "trainImage.py" and which is called in the Train Image function in the interface that was presented above.

"Take attendance" is the function that enables the lecturer to start directly with the process of taking attendance but first by specifying the subject in question and then has the opportunity to start the process, but also to see the reports from past lectures, the interface of this process as well as the coding part and function are displayed below.

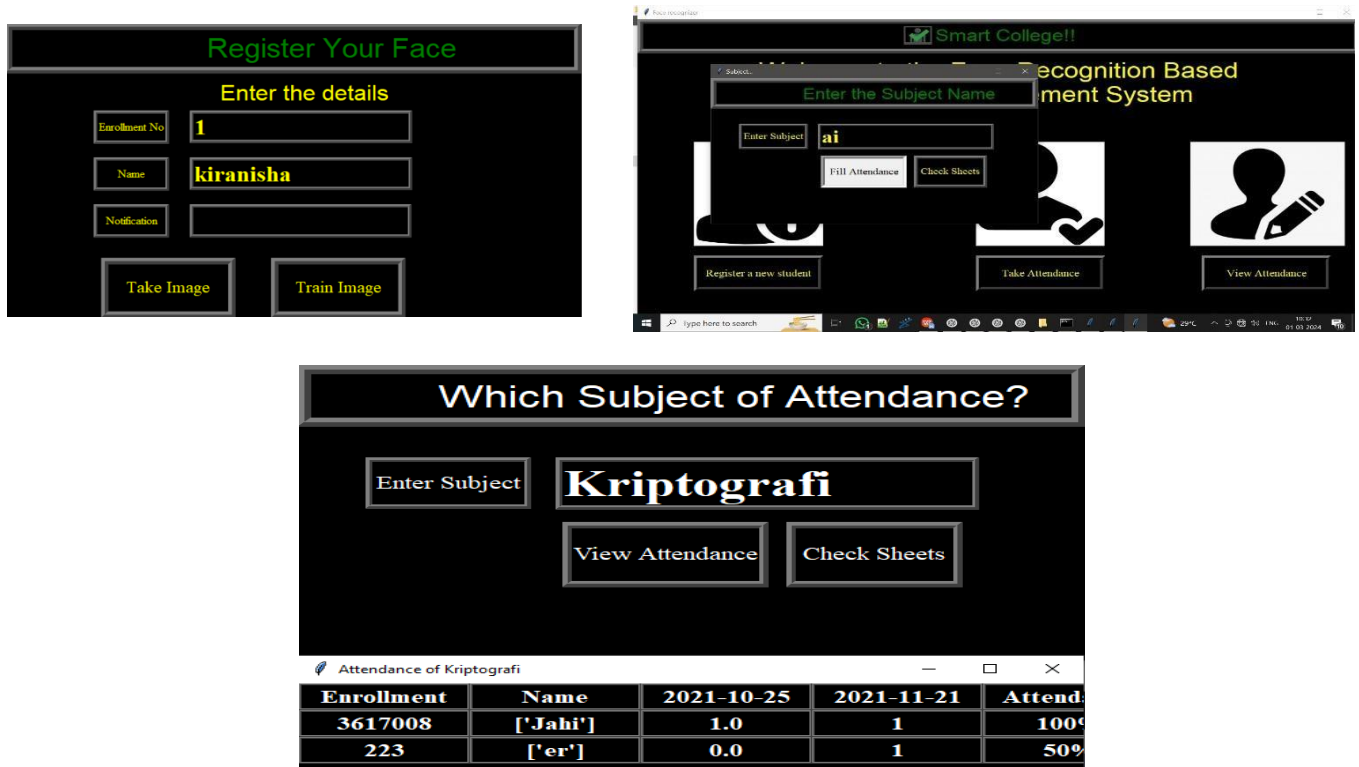
View attendance coding part

The "View Attendance" function enables lecturers to view reports from previously received presences, in tables that are generated within the system, but also in excel files which are created automatically.

The coding part includes two functions that are related to the buttons shown in the figure, by

clicking the "view attendance" button displays the table which is shown in the figure, while by clicking the "check sheets" button then the system directs us to excel documents in which the registered presence is also stored.

RESULT



CONCLUSION

The automated presence registration system is built to reduce the shortcomings and problems that appear in the traditional (manual) system. After extensive research into Python programming language and libraries needed to build such system.

The completion of the project resulted in a software product that can automate the registration process of participation in the university facility. The system requires minimal human intervention, even only during the start-up and initial enrollment of students and then the process is completely automatic, saving the time consumed by the traditional system.

REFERENCES

- [1] V. a. R. Tokas, Fast Face Recognition Using Eigen Faces, IJRITCC, vol.2, no.11,pp.3615-3618, November 2014.
- [2] N. J. M. M. K. a. H.A. Mayank Agarwal, Face Recognition Using Eigen faceapproach, IRCSE, vol.2, no. 4, pp.1793-8201, August 2010.
- [3] Book Face Recognition Based Smart Attendance System (ISBN-973-3-659-68627- January 2018.

EDGE COMPUTING

K.KIRUTHIGA

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM

kerithimadhu578@gmail.com

ABSTRACT:

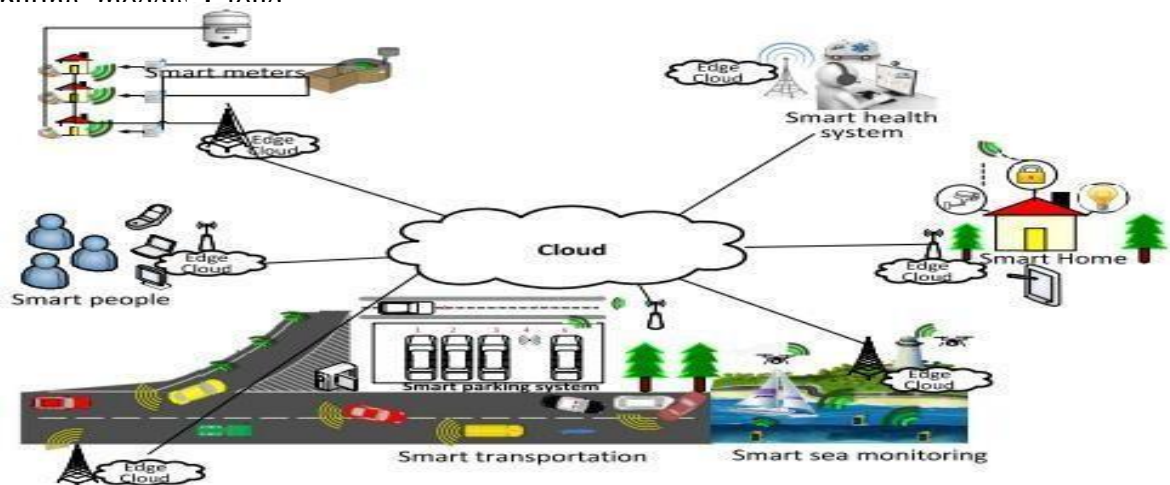
Edge computing has emerged as a paradigm to address the limitations of traditional cloud-centric approaches by bringing computation and data storage closer to the location where it is needed, thereby reducing latency, bandwidth usage, and reliance on centralized data centers. This thesis explores various aspects of edge computing, including its architecture, challenges, applications, and potential benefits. Through a comprehensive review of existing literature and case studies, it highlights the importance of edge computing in enabling real-time processing, low-latency applications, and improved scalability for emerging technologies such as Internet of Things (IoT), autonomous vehicles, and augmented reality. Additionally, this thesis investigates the impact of edge computing on network infrastructure, security, privacy, and energy efficiency, providing insights into the trade-offs and considerations involved in deploying edge computing solutions. Overall, this research contributes to a deeper understanding of edge computing and its implications for future distributed computing paradigms.

KEYWORDS:

Architecture, Challenges, Applications, Benefits, Real-time Processing, Low-latency, IoT, Autonomous Vehicles, Augmented Reality, Network Infrastructure, Security, Privacy, Energy Efficiency.

INTRODUCTION

Edge computing constitutes a new concept in the computing landscape. It brings the service and utilities of cloud computing closer to the end user and is characterized by fast processing and quick application response time. The currently developed internet-enabled applications such as surveillance, virtual reality, and real-time traffic monitoring require fast processing and quick response time. End users normally run these applications on their resource-constrained mobile devices while the core service and processing are performed on cloud servers. Leveraging services of cloud by mobile devices result in high latency and mobility-related issues. Edge computing fulfills the above-mentioned application requirements by bringing the processing to the edge of the network. The cloud computing issues can be resolved through the three Edge computing models: Cloud



LATENCY

More industries are implementing applications that require rapid analysis and response. Cloud computing alone can't keep up with these demands because of the latency introduced by network distance from the data source, resulting in inefficiency, lag time, and poor customer experiences.

BANDWIDTH

Adding transmission bandwidth or more processing power could overcome latency issues. However, as companies continue to increase the number of edge devices on their network and the amount of data they generate, the cost to send data to the cloud may reach impractical levels that could be alleviated if data can be processed, stored, and analyzed at the edge.

SECURITY AND PRIVACY

Securing sensitive data, such as private medical records, at the edge and transmitting less data across the internet could help increase security by reducing the risk of interception. In addition, some governments or customers may require that data remain in the jurisdiction where it was created. In healthcare, for example, there may even be local or regional requirements to limit the storage or transmission of personal data.

AI

With the need for actionable intelligence in near real-time, companies need AI at the data source to allow faster processing and to take advantage of the potential in previously untapped data.

REVIEW OF LITERATURE

In today's interconnected digital landscape, the exponential growth of data generation has sparked a transformative shift in computing paradigms. Among the most significant developments is the emergence of edge computing, a groundbreaking approach that revolutionizes how data is processed, analyzed, and utilized. This essay delves into the evolution, principles, applications, and impact of edge computing on various domains of technology and society.

Edge computing represents a departure from the traditional centralized model of data processing, where information is transmitted to remote data centers for analysis and storage. Instead, it advocates for the distribution of computational resources closer to the data source, or "edge," thereby minimizing latency, enhancing efficiency, and enabling real-time decision-making. This distributed architecture leverages a network of decentralized nodes, such as routers, gateways, and IoT devices, to execute computing tasks at the periphery of the network.

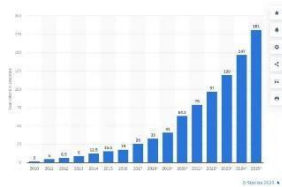
The principles underpinning edge computing are rooted in the imperative to address the limitations of centralized processing, particularly in contexts where latency and bandwidth constraints are critical factors. By bringing computation closer to the data source, edge computing enables faster response times, reduces network congestion, and enhances data privacy and security. Furthermore, it optimizes resource utilization by offloading computational tasks from centralized servers, thereby improving overall system scalability and resilience.

The applications of edge computing span a diverse range of domains, each benefiting from its unique advantages. In the realm of Internet of Things (IoT), edge computing facilitates real-time data analysis and decision-making at the device level, enabling smarter and more responsive IoT deployments. In autonomous vehicles, edge computing powers onboard processing capabilities, enabling rapid interpretation of sensor data and facilitating split-second

decision-making for navigation and safety. Additionally, in the realm of augmented reality (AR) and virtual reality (VR), edge computing reduces latency and enhances user experiences by delivering immersive content with minimal delay.

However, the adoption of edge computing also poses challenges and considerations, including interoperability, security, and resource management. As organizations embrace edge computing architectures, they must navigate these complexities and invest in robust frameworks and standards to ensure interoperability, data integrity, and privacy protection.

RESULT FOR EDGE COMPUTING



CONCLUSION

In conclusion, edge computing represents a paradigm shift in how data is processed and managed in modern computing environments. By bringing computational resources closer to the data source, edge computing minimizes latency, reduces bandwidth usage, and enables real-time decision-making. This decentralized approach is particularly valuable in scenarios where data needs to be processed quickly, such as in IoT applications, industrial automation, autonomous vehicles, and augmented reality. However, edge computing also presents challenges, including security concerns, scalability issues, and the need for efficient resource management. Despite these challenges, the potential benefits of edge computing make it a critical component of the evolving computing landscape, driving innovation and enabling new applications across various industries.

REFERENCES

- [1] N. Hassan, S. Gillani, E. Ahmed, I. Yaqoob, M. Imran, The role of edge computing in internet of things, *IEEE Communications Magazine* (99) (2018) 1–6.
- [2] M. Liu, F. R. Yu, Y. Teng, V. C. Leung, M. Song, Distributed resource allocation in blockchain-based video streaming systems with mobile edge computing, *IEEE Transactions on Wireless Communications* 18 (1) (2019) 695–708.
- [3] E. Ahmed, A. Akhunzada, M. Whaiduzzaman, A. Gani, S. H. Ab Hamid, R. Buyya, Networkcentric performance analysis of runtime application migration in mobile cloud computing, *Simulation Modelling Practice and Theory* 50 (2015) 42–56

A SURVEY ON NATURAL LANGUAGE PROCESSING

MADHUMITHA G

GOVERNMENT ARTS COLLEGE (Autonomous), KUMBAKONAM

Email madhumithsgunasekar0510@gmail.com

ABSTRACT

Natural language processing is a branch of computer science and artificial intelligence which is concerned with interaction between computers and human languages. Natural language processing is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. This includes the spoken language systems that integrate speech and natural language. Natural language processing has a role in computer science because many aspects of the field deal with linguistic features of computation. Natural language processing is an area of research and application that explores how computers can be used to understand and manipulates natural language text or speech to do useful things. The applications of Natural language processing include fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence (AI) and expert systems.

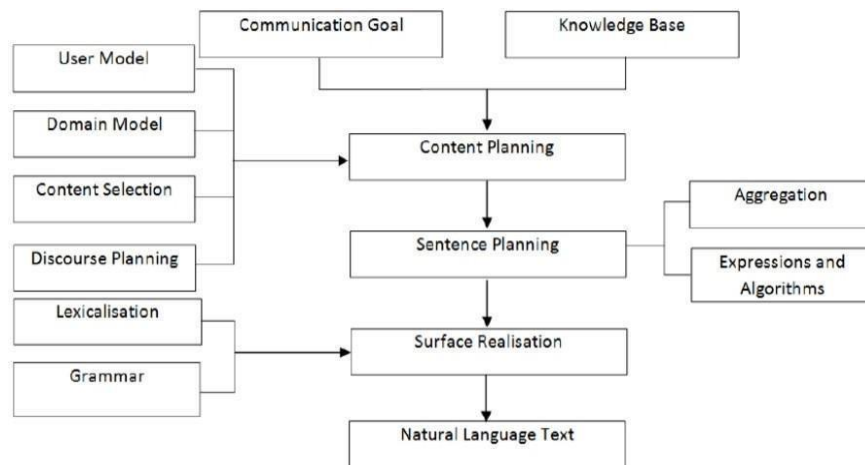
Keywords: Natural language Processing (NLP), machine translation, Cross Language Information Retrieval (CLIR), Artificial intelligence (AI), expert systems.

INTRODUCTION

Natural language processing (NLP) is the intersection of computer science, linguistics and machine learning. The field focuses on communication between computers and humans in natural language and NLP is all about making computers understand and generate human language. Applications of NLP techniques include voice assistants like Amazon's Alexa and Apple's Siri, but also things like machine translation and text-filtering. Language can be defined as a set of rules or set of symbols. Symbol are combined and used for conveying information or broadcasting the information. Symbols are tyrannized by the Rules. Natural Language Processing basically can be classified into two parts i.e. Natural Language Understanding and Natural Language Generation which evolves the task to understand and generate the text.

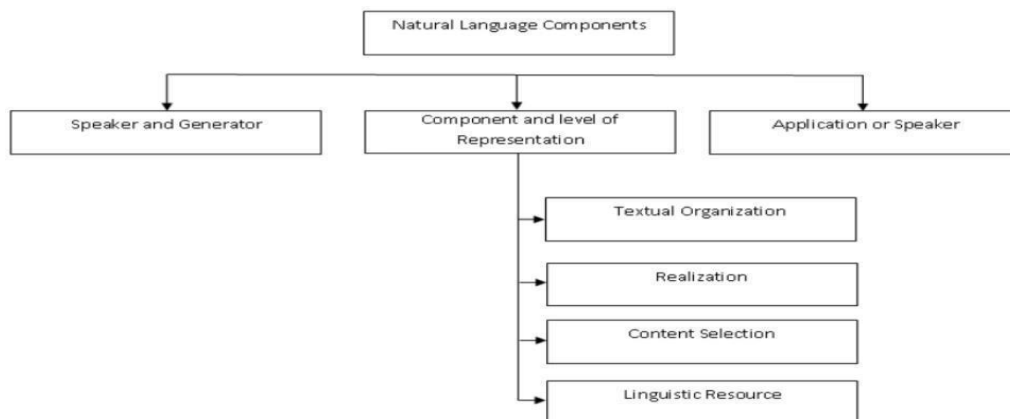
LEVELS OF NLP:

The 'levels of language' are one of the most explanatory method for representing the Natural Language processing which helps to generate the NLP text by realizing Content Planning, Sentence Planning and Surface Realization phases.



NATURAL LANGUAGE GENERATION:

Natural Language Generation (NLG) is the process of producing phrases, sentences and paragraphs that are meaningful from an internal representation. It is a part of Natural Language Processing and happens in four phases



HISTORY OF NLP:

In late 1940s the term wasn't even in existence, but the work regarding machine translation (MT) had started. Research in this period was not completely localized. Russian and English were the dominant languages for MT, but others, like Chinese were used for MT (Booth ,1967). MT/NLP research was almost died in 1966 according to ALPAC report, which concluded that MT is going nowhere. But later on some MT production systems were providing output to their customers (Hutchins, 1986). By this time, work on the use of computers for literary and linguistic studies had also started.

CROSS-LANGUAGE INFORMATION RETRIEVAL:

Cross-Lingual Information Retrieval (CLIR) refers to the retrieval of documents that are in a language different from the one in which the query is expressed.

NLP MACHINE TRANSLATION:

Machine translation is the process of automatically translating text from one language to another. This task has gained significant attention with the advent of neural machine translation models. These models, often based on deep learning architectures, learn to understand the contextual and semantic aspects of language, improving the accuracy of translations. Machine translation systems are widely used in online translation services, facilitating communication and breaking down language barriers in various domains.

EXPERT SYSTEM:

An expert system is a computer program that uses artificial intelligence (AI) technologies to simulate the judgment and behavior of a human or an organization that has expertise and experience in a particular field.

CONCLUSION:

In conclusion, NLP bridges the gap between human language and computers, enabling communication and interaction through natural language. Its applications are diverse and continue to evolve, with ongoing research and development driving advancements in language understanding, generation, and human-computer interaction.

REFERENCES:

- [1] Chomsky, Noam, 1965, Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press.
- [2] Hutchins, W. J. (1986). Machine translation: past, present, future (p. 66). Chichester: Ellis Horwood.

A SURVEY ON EXPERT SYSTEMS FOR CONSTRUCTION PROJECT

MALATHI.K

GOVERNMENT ARTS COLLAGE (Autonomous), KUMBAKONAM

Email:kk6685860@gmail.com

ABSTRACT

Potential applications of knowledge based expert systems in the area of construction project monitoring and control are described. Originally developed from research in artificial intelligence, these systems are computer programs that can undertake intelligent tasks currently performed by highly skilled People. While some project monitoring can be accomplished by algorithmic Procedures, the capability of knowledge based expert systems to deal with ill structured problems and to be extensively modified over time make them desirable for application in this area. Sample applications and heuristic rules in Scheduling and inventory control are provided.

KEYWORDS:

Cost and time control, Purchasing and inventory control, Expert system, artificial intelligence.

INTRODUCTION

Project Monitoring is one aspect of project management involving checking, regulating and controlling the performance and execution of a construction project. This includes cost control, scheduling and time control, purchasing and inventory control as well as quality control through the entire planning, design and construction process. Project Monitoring is conducted independently or in concert by owner represent, construction managers, engineers, project managers and others.

REVIEW OF LITERATURE

The literature on knowledge-based expert systems in construction project monitoring offers valuable insights into their potential applications and benefits. This review synthesizes key findings and identifies gaps for further exploration.

Duda et al. (1979) present a pioneering work on computer-based consultants in mineral exploration. While not directly related to construction projects, their approach highlights the feasibility and efficacy of expert systems in complex decision-making domains. This study underscores the importance of domain-specific knowledge and heuristic algorithms, which are fundamental to the design of knowledge-based expert systems.

Fenves (1983) provides class notes on expert systems in civil engineering, laying the groundwork for applying artificial intelligence techniques to construction management. These notes emphasize the adaptability of expert systems to address ill-structured problems, a critical

aspect in project monitoring where uncertainties and changing conditions are common. However, the practical implementation of these systems in real-world construction projects remains a topic requiring further investigation.

Fox (1982) presents a comprehensive overview of computer applications in engineering, including expert systems. This review discusses the potential of expert systems to enhance project management processes, particularly in areas such as cost control, scheduling, and inventory management. Despite the promising outlook, challenges such as system rigidity and limited customization options are noted, suggesting the need for continuous refinement and adaptation to diverse project requirements.

KNOWLEDGE-BASED EXPERT SYSTEMS

Knowledge-based expert systems are programs that can undertake intelligent tasks currently performed by highly skilled people. Expert systems use domain specific knowledge and heuristics to perform many of the functions of a human expert.

1. A Change Monitor detects changes in the short memory that may require attention.
2. A Pattern Matcher compares the short term memory with the knowledge base.
3. A Scheduler decides which action is the most appropriate.

POTENTIAL ROLE OF EXPERT SYSTEMS IN PROJECT MONITORING

A major complaint concerning the available computer systems is that programs are often too limited and rigid for a company's individual needs.

1. Algorithmic methods are either not feasible, too cumbersome or too restrictive.
2. There are recognized experts in the field.
3. The task requires from ten minutes to a few days when performed by an expert.

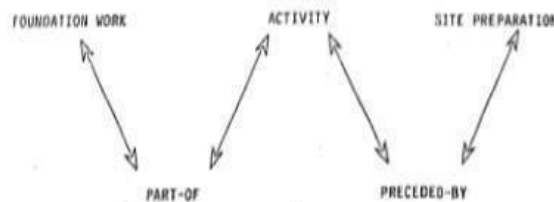


FIG. 1.—Illustration of Semantic Network

COST AND TIME CONTROL

Percent Complete
 0
 15
 35

50
 65
 85
 100
 Adjustment Vah
 1.0
 $(25/15) = 1.67$
 $(40/35) = 1.14$
 $(50/50) = 1.0$
 $(60/65) = 0.92$
 $(75/85) = 0.88$

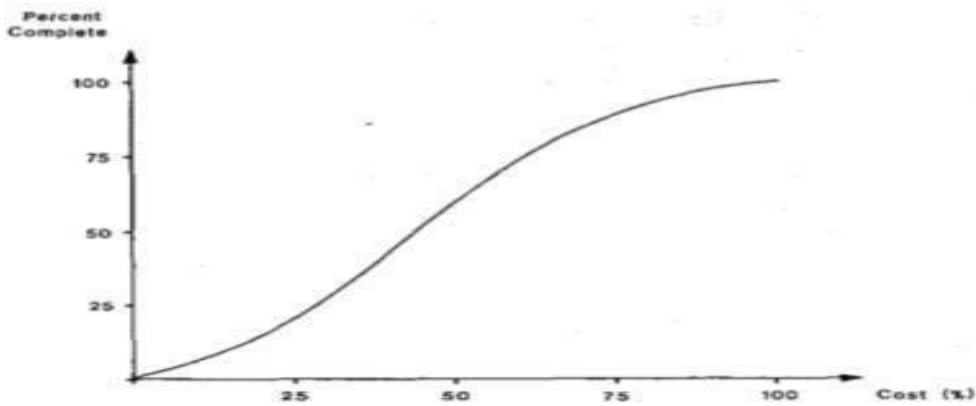


FIG. 2.—Illustrative Forecast Expenditure Curve

PURCHASING AND INVENTORY CONTROL

Purchasing and inventory control can have a direct impact on project performance. Improper material or late deliveries can cause cost over runs, time delays, and quality problems. One expert system application in the area of purchasing and inventory control would be an aid to a project manager for determining appropriate inventory levels.

CONCLUSIONS

After the development of large project management data bases and related software, the application of knowledge-based expert systems is a desirable extension. These systems can deal with the ill-structured problems common in this area. They also have the potential for incremental development as users' experience accumulates and is incorporated in additional rules.

REFERENCE

- [1] Duda, R. O., et al., "A Computer-Based Consultant for Mineral Exploration," Final Report SRI Project 6415, SRI International, 1979.
- [2] Fenves, S. J., "Class Notes for Expert Systems in Civil Engineering ,"Technical Report, Department of Civil Engineering, Carnegie-Mellon University ,Pittsburgh, Pa., 1983.
- [3] Fox, A. J., Ed., "Special Computer Issue—Getting On Line," Engineering News Record, Vol. 209, No. 23, Dec. 2, 1982, pp. 31-67

A SURVEY ON MACHINE LEARNING PROJECT

NANCY.D

GOVERNMENT ARTS COLLEGE Autonomous KUMBAKONAM Email: nancymercy585@gmail.com

ABSTRACT

In this project, we were asked to experiment with a real world dataset, and to explore how Machine learning algorithms can be used to find the patterns in data. We were expected to gain Experience using a common data-mining and machine learning library, Weka, and were expected to submit a report about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my final report.

Keywords

Machine Learning, Pattern Recognition, Classification, Supervised learning, Artificial Intelligence.

INTRODUCTION

MACHINE LEARNING

Machine learning is a sub-domain of computer science which evolved from the study of Pattern recognition in data, and also from the computational learning theory in artificial Intelligence. It is the first-class ticket to most interesting careers in data analytics today. As Data sources proliferate along with the computing power to process them, going straight to the Data is one of the most straightforward ways to quickly gain insights and make predictions. Machine Learning can be thought of as the study of a list of sub-problems, viz : decision Making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic Algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labeled data and they each have their own merits and demerits.

REVIEW OF LITERATURE

Machine learning has emerged as a powerful tool in data analysis, allowing researchers to extract patterns and insights from complex datasets. The literature on machine learning encompasses various algorithms, techniques, and applications, providing a rich foundation for conducting empirical studies and real-world projects.

Problems and Issues in Supervised learning

Before we get started, we must know about how to pick a good machine learning Algorithm for the given dataset. To intelligently pick an algorithm to use for a supervised learning

Task, we must consider the following factors

1. Heterogeneity of Data

Many algorithms like neural networks and support vector machines like their feature vectors to be homogeneous numeric and normalized. The algorithms that employ distance metrics are very sensitive to this, and hence if the data is

Heterogeneous, these methods should be the afterthought. Decision Trees can handle heterogeneous data very easily.

Redundancy of Data

If the data contains redundant information, i.e. contain highly correlated values, then it's useless to use distance based methods because of numerical instability. In this case, some sort of Regularization can be employed to the data to prevent this situation.

Dataset

The dataset used is a sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The dataset that was used for this project is a subset of a much larger dataset, as described in Rousseau et al, 1983, South African Medical Journal, and has the following Feature vectors:

1. Sbp. systolic blood pressure
2. Tobacco cumulative tobacco (kg)
3. Ldl low density lipoprotein cholesterol

Baseline Classifier:

As the baseline classifier, we chose a Naïve Bayesian Network because it is easy to compute, and because the features in the given dataset are all aspects of a person's physical habits or medical history, and hence can be assumed to be independent of each other, which is the primary assumption in Naïve Bayes Classifier. It is a conditional probability model: given a problem instance to be classified, represented by a vector representing some n features (independent variables), it assigns to this instance probabilities for each of K possible outcomes or classes.

Naïve Bayes Classifier

Class

Attribute. 0 1
 (0.65) (0.35)

=====

Sbp

Mean 135.6278. 143.8165
 Std. dev. 17.8582 23.5657
 Weight sum 302. 160
 Precision 1.918. 1.918

Tobacco

Mean 2.6347. 5.5268
 Std. dev. 3.6078 5.551
 Weight sum 302 160
 Precision 0.1465 0.1465

Ldl

Mean 4.3436. 5.489

Std. dev. 1.867 2.2194

Weight sum 302 160

Precision 0.0438. 0.0438

As we can see that the Naïve Bayes classifier works really well with the given dataset, with the True Positive classification rate being 71.6 percent on an average, i.e. this classifier can correctly classify 71.6 percent of all the examples it sees. However, there is still a vast majority of the dataset, i.e. 28.4% which can't be correctly classified. This means that our expert medical diagnosis system still misdiagnoses one third of its cases, and one third of the patients' symptoms who may have the disease are not being scrutinized by the doctor. We will now attempt to improve result by using other more sophisticated classifiers.

CONCLUSION

We conclude that the dataset is not a complete space, and there are still other feature vectors missing from it. What we were attempting to generalize is a subspace of the actual input space, where the other dimensions are not known, and hence none of the classifiers were able to do better than 71.6% (Naïve Bayes). In the future, if similar studies are conducted to generate the dataset used in this report, more feature vectors need to be calculated so that the classifiers can form a better idea of the problem at hand.

REFERENCES

- [1] Intro to Machine Learning | Udacity. Intro to Machine Learning | Udacity. Accessed April 27, 2016.
- [2] No Free Lunch Theorems. No Free Lunch Theorems. Accessed April 27, 2016

A SURVEY ON BIOMETRIC

R. NISHANTHINI

GOVERNMENT ARTS COLLEGE (AUTONOMOUS) KUMBAKONAM

Email:nishanthini4122002@gmail.com

The biometrics have been used as a solution for access control systems for many years, but the simple use of biometrics cannot be considered as final and perfect solution. There are many risks that should not be ignored. Most problems are related to the transmission path between the system where the users require access and the servers where the captured biometric data is stored. Various types of attacks can be made by impostors who want to use the system improperly. Besides the technical aspects, there is the social aspect. There is a growing concern of users about both data storage and the misuse of their biometrics, which is a Unique identifier and, being invariant in time, may be lost forever if compromised. The fact that several companies keep their biometric data in different servers is causing discomfort to users because it makes their biometric data more susceptible to attacks. In this thesis, the use of smart cards is adopted as a possible solution to the above problems. Smart cards prepared for multi-applications are used to perform biometric comparisons internally. Thus, it would not be necessary to use different servers because biometric features will always be on a single card in the possession of the owner. It was developed and implemented three different algorithms using different biometric identification characteristics: fingerprint, palmprint and iris. Considering the used memory, average execution time and accuracy, palm print biometrics obtained the best results, achieving minimum error rates and processing time lower than half a second.

KEYWORDS

Biometrics. Smart cards. Minutiae. PalmCode. IrisCode

INTRODUCTION

The automatic recognition of individuals based on their anatomical (e.g., face, fingerprint, iris, retina) and behavioral (e.g., signature, posture) individualities is called Biometrics. It is a form of information that helps in identifying one's physical characters such as psychosomatic, behavioral characters, etc.

The term Biometrics is made from two words - 'Bio (Greek work)' and 'Metrics' where bio means life and metrics indicate measurements. Biometrics is widely used for security purposes as it provides a high degree of accurateness in recognizing an individual.

Every human being is different from one another regarding physical and behavioral characteristics. We can analyze one's attribute with the help of these characteristics. The physical traits such as fingerprints, the color of the iris, the color of hair, hand geometry, and behavioral features such as voice and accent of tongue, signature, grid, or the way of striking keys of the computer keyboard, etc., makes a person stand out from the rest.

HISTORY OF BIOMETRICS

Biometrics is not a new concept; it is the oldest form of identification. Bertillon Systems (1882) took subject's photography, height, the length of one-foot, an arm and index finger. FBI setup a fingerprint identification division in the year 1924. By 1926, law enforcement officials in several U.S. cities had begun submitting fingerprint cards to the FBI in an effort to create a database of fingerprints from known criminals. In the early 1960's the FBI invested a large amount of time and effort into the development of automated fingerprint identification systems. This automation of biometric identification for law enforcement purposes coincided with the development of automated systems for non-forensic applications, such as high-security access control. AFIS installed in 1965 with a database of 810,000 fingerprints. During the 1970's a biometric product based on measuring the geometry of the hand was introduced in a number of access control applications. First face recognition paper was published in the year 1972.

BIOMETRICS

Biometrics encompass a variety of different technologies that use probabilistic matching to recognize a person based on their biometric characteristics. Biometric characteristics can be physiological features (for example, a person's fingerprint, iris, face or hand geometry), or behavioral attributes (such as a person's gait, signature, or keystroke pattern).

As biometric characteristics are generally unique to individuals, they can be more effective and reliable at uniquely verifying individuals' identities than other methods such as knowledge-based verification systems (for example, a password or PIN) or token-based systems (for example, an ID card or licence)

TYPES OF BIOMETRICS

- ❖ Fingerprints
- ❖ Vein recognition
- ❖ Iris recognition
- ❖ Retina scanning
- ❖ Facial recognition

Fingerprints

It's well-known that every person has unique fingerprints. Fingerprints have been used to identify people since long before the digital age, so it's a logical choice for biometrics. Fingerprint identification is cheap, affordable and typically extremely accurate.

Vein Recognition

Vein patterns, as it happens, are also unique to individuals. Technology currently exists to examine finger or palm vein patterns. Vein identification is more secure than fingerprint identification

because while it’s conceivable that fingerprints could be altered, it’s difficult to imagine how a vein pattern, being sub dermal, could be altered in a useful way.

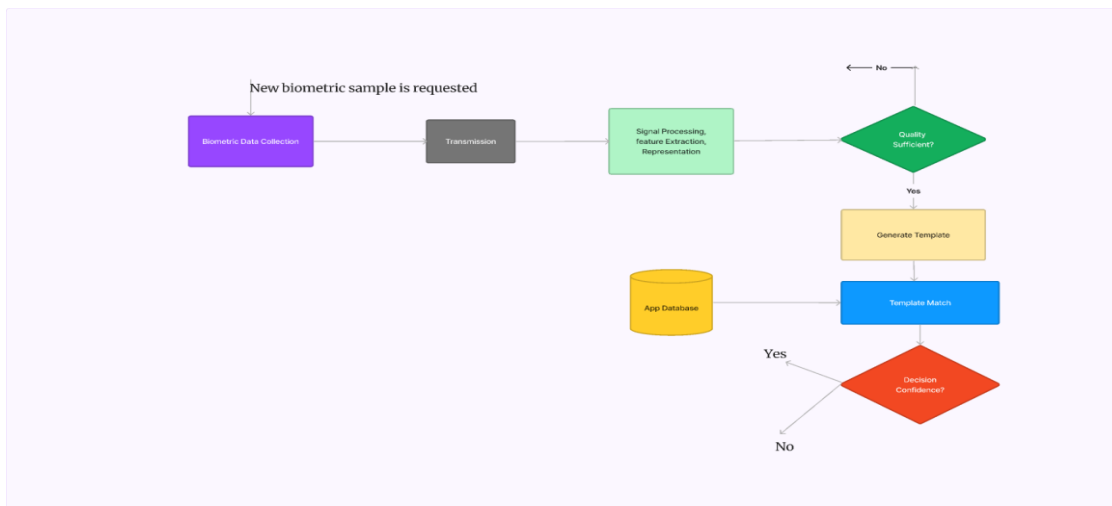
Retina Scanning and Iris Recognition

Another popular method of biometric identification is eye pattern recognition. The user simply looks into an eye reader, which may analyze either the iris pattern or the retinal structure in order to match it to an approved user profile.

Facial Recognition

A security measure that has been used for years in large-scale environments requiring tight security like casinos, facial recognition is another well-known option. Today, technology is so advanced that even many phones are able to map key points on a person’s face to match with that person’s identity.

FLOWCHART OF BIOMETRIC



ADVANTAGE OF BIOMETRICS:

- ✓ Less Processing time
- ✓ Accuracy
- ✓ Increased Security
- ✓ Ease of work
- ✓ Screening

DISADVANTAGE OF BIOMETRICS:

Like all other security methods, biometrics also has limitations and threats which can impact its effectiveness and efficiency which are as follows:

- ✓ Intra-class variability and inter-class similarity
- ✓ Segmentation
- ✓ Noisy input & population coverage
- ✓ System performance (error rate, speed, cost)
- ✓ The individuality of biometric characteristics
- ✓ Fusion of multiple biometric attributes
- ✓ Scalability
- ✓ Attacks on the biometric system
- ✓ Privacy Issues

CONCLUSION:

Biometrics is a promising and exciting area, where different disciplines meet and provide an opportunity for a more secure and responsible world. There are a number of popular biometrics mechanisms currently deployed, some with strong histories, and some relatively new mechanisms. Each mechanism has its own strengths and weaknesses. When properly applied, biometrics can be used to combat fraud, and ensure that timekeeping systems are honest and accurate.

REFERENCE:

- [1] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology 14 (1) (2004) (4-20).
- [2] Andrea F. Abate, Michele Nappi, Daniel Riccio, Gabriele Sabatino, "2D and 3D facerecognition: A survey", www.elsevier.com, 2007.

A SURVEY OF PEER-TO- PEER

PRIYANKA.S

GOVERNMENT ARTS COLLEGE (AUTONOMOUS) KUMBAKONAM

Email:priyankasankar29@gmail.com

ABSTRACT

Created When two or more PCs are connected and share resources together without going through a separate server computer. The P2P network allows a couple of computers to communicate and connect via a Universal Serial Bus to transfer data. The P2P network creates severe impact on a company's communication process, and it is the most accessible type of architecture to build. On a magnificent scale, the P2P network helps set up direct relationships among users with specific protocols and applications added to the network.

The peer-to-peer network was first developed in the late 1970s. A computer acts as a node for file sharing within the network, and every node acts as a server. Thus, there is no central server to the network. Every node has an equal share of the workload.

Keywords:

Large networks, fault-tolerant, search efficiency, load balancing, peer to peer computing, search scheme, indexing, rooting, bloom filter, routing table, the frequency inverse document.

INTRODUCTION

Significance and Emergence:

Several important desktop computing applications have emerged in recent years that use an Internet-scale decentralized architecture to simultaneously connect millions of users to share content, form social groups and communicate with their contacts. These applications are classified as peer-to-peer because of the elimination of servers to mediate between end systems on which the applications run, and their network behavior is described as an overlay network because the peer protocols form a virtualized network over the physical network. While peer-to-peer (P2P) applications have had a rapid ascent and wide impact, in the future P2P overlays are likely to enable important new applications following from these technology trends.

Table 1 Specialized overlay networks for internet services

Type	Example	First use or definition
Email	SMTP	1970s
Internet news	NNTP	1986
Multicast	MBone	1992
Web caching	Internet cache protocol	1995
Content delivery network	Akamai	1999
Anonymous communication	FreeNet	1999
Application layer multicast	Narada	2000
Routing	RON	2001

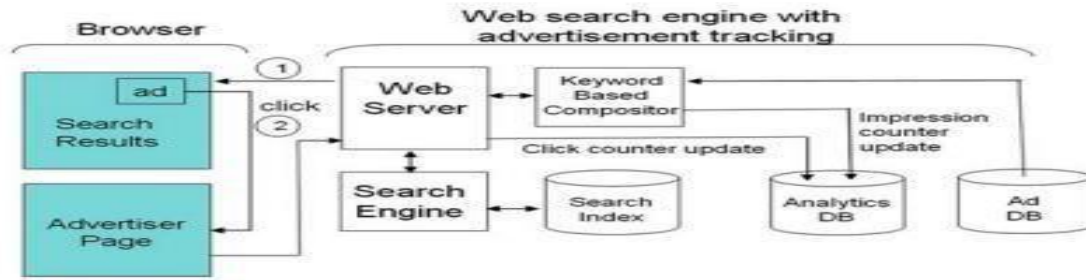
Applications

The first widely used file sharing system, Napster, featured a hybrid architecture in which the directory was stored on a server, but peers directly transferred files between them. Napster became the first legal test case for file sharing of licensed content, and was subsequently forced to change to protect such content. A number of peer-to-peer file sharing systems were developed (Table 2) to avoid the legal challenges faced by Napster. The majority of these second-generation file sharing systems were based on unstructured overlays. While these systems had no mechanisms for protecting the rights of content owners, in some cases the P2P application developers obtained revenue by either selling their applications or by embedding

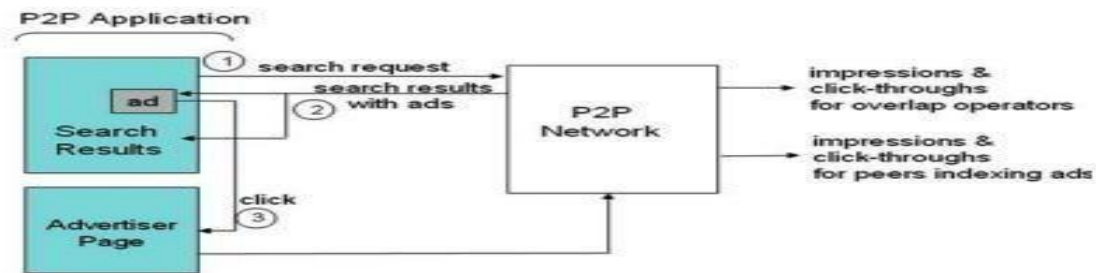
Technology Drivers:

Peer-to-Peer Networking and Applications: Synopsis and Research Directions

9



(a) Click-through and impression tracking in web search



(b) Click-through and impression tracking in P2P networks

CONCLUSION

With reference to our above discussion, it is clear that peer-to-peer network is one of the most sophisticated technologies that we have today. What makes it even more prominent is its link with Block chain, AI and the potential it holds in terms of its usability. However, it is equally important to know and understand that it may not always come to use for the right reasons due to such open and easy access to all types of content. While it can be a possible argument that this technology should not be promoted, we as individuals must understand our responsibilities as users while accessing online content.

REFERENCES

- [1] MacQuire, A. Brampton, I. Rai and L. Mathy. Performance Analysis of Stealth DHT with Mobile Nodes. In Proceedings of the 4th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW 2006), March 2006.
- [2] J. Buford. Mobile P2P after Five Years – Where are we and where are we headed? (Contributed Talk) Fifth IEEE Workshop on Mobile.
- [3] S. Banerjee, B. Bhattacharjee, and C. Kommareddy. Scalable application layer multicast, in Proceedings, ACM SIGCOMM2002, September 2002.
- [4] E. Adar and B. Huberman, Free riding on Gnutella, First Monday, 2000

A SURVEY OF DEEP LEARNING PROCESSING

RAJALAXMI M

GOVERNMENT ARTS COLLAGE (AUTONOMOUS), KUMBAKONAM

Email:rajalaxmimurugan3108@gmail.com

ABSTRACT:

Deep learning is an emerging area of machine learning (ML) research. It comprises multiple hidden layers of artificial neural networks. The deep learning methodology applies nonlinear transformations and model abstractions of high level in large databases. The recent advancements in deep learning architectures within numerous fields have already provided significant contributions in artificial intelligence. This article presents a state-of-the-art survey on the contributions and the novel applications of deep learning. The following review chronologically presents how and in what major applications deep learning algorithms have been utilized. Furthermore, the superior and beneficial of the deep learning methodology and its hierarchy in layers and nonlinear operations are presented and compared with the more conventional algorithms in the common applications. The state-of-the-art survey further provides a general overview on the novel concept and the ever-increasing advantages and popularity of deep learning. Through this platform, users can accelerate their research and development processes, leading to advancements in areas such as computer vision, natural language processing, and reinforcement learning.

Keywords:

Deep learning, machine learning, applied deep learning platform, neural networks artificial intelligence, training, deployment, management, computer vision, training, reinforcement learning.

INTRODUCTION

Artificial intelligence (AI) as an intelligence exhibited by machines has been an effective approach to human learning and reasoning. In 1950, "The Turing Test" was proposed as a satisfactory explanation of how a computer could perform a human cognitive reasoning. As a research field, AI is divided in more specific research sub-fields. For example: Natural Language Processing (NLP) can enhance the writing experience in various applications. The most classic subdivision within NLP is machine translation, which is understood as the translation between languages.

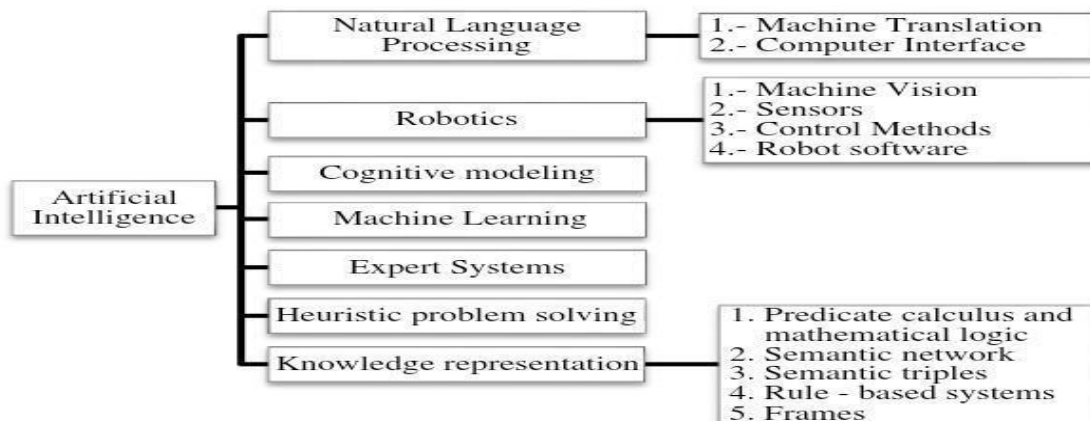


Fig. 1. Research in artificial intelligence (AI) Source: [1].

BACKGROUND

The Deep Learning (DL) concept appeared for the first time in 2006 as a new field of research within machine learning. It was first known as hierarchical learning at the and it usually involved many research fields related to pattern recognition. Deep learning mainly considers two key factors: nonlinear processing in multiple layers or stages and supervised or unsupervised learning.

APPLICATIONS

Deep learning implies an abstract layer analysis and hierarchical methods. However, it can be utilized in numerous real life applications. Applying a deep learning algorithm, coloring can be performed automatically by a computer. Similarly, sound can be added into a mute drumming video by using Recurrent Neural Networks (RNN) as part of the deep learning methods.

- ✓ Image Processing
- ✓ Biometrics
- ✓ Medicine

OVERVIEW

Table 1 summarizes several applications carried out throughout previous years regarding deep learning. Mostly, speech recognition and image processing are mentioned. This review only considers a few from the large list of applications.

Table 1. Deep Learning Applications, 2003 - 2017.

Author	Application	Method/algorithm	Year
Tai Sing Lee, David Mumford	Hierarchical Bayesian inference in the visual cortex	Particle filtering and Bayesian - belief propagation	2003
Hinton, Geoffrey E., Simon Osindero, Yee-Whye Teh.	Digit Classification	Complementary Priors on Belief networks	2006
Mohamed, Abdel-rahman, George Dahl, Geoffrey Hinton	Deep Belief Networks for phone recognition	Back propagation and associative memory architecture	2009
Abdel-Hamid Ossama, Mohamed Abdel-rahman, Jiang Hui, Penn Gerald	Multi-speaker speech recognition	Local filtering and max-pooling infrequency domain	2012
Kiran B. Raja, R. Raghavendra, Vinay Krishna Vemuri, Christoph Busch	Iris Recognition by using smartphones' cameras	Deep sparse filtering	2015
Silver David, et al	Mastering the Game of Go with Deep Neural Networks and Tree Search	Supervised learning and reinforcement learning	2016
Francesco Marra, Giovanni Poggi, Carlo Sansone.	Iris sensor model identification	Convolutional neural networks	2017

CONCLUSION

Deep learning is indeed a fast growing application of machine learning. The numerous applications described above prove its rapid development in just few years. The usage of these algorithms in different fields shows its versatility. The publication analysis performed in this study clearly demonstrates the relevance of this technology and gives a clearly illustrates the growth of deep learning and the tendency regarding for future research in this field.

REFERENCES

- [1] Abdel, O.: Applying convolutional neural networks concepts to hybrid NN-HMM model for). speech recognition. *Acoustics, Speech and Signal Processing* 7, 4277-4280 (2012).
- [2] Mosavi A., Varkonyi-Koczy A. R.: Integration of Machine Learning and Optimization for Robot Learning. *Advances in Intelligent Systems and Computing* 519, 349-355 (2017).
- [3] Bannister, A.: Biometrics and AI: how FaceSentinel evolves 13 times faster thanks to deep learning (2016).
- [4] Bengio, Y.: Learning deep architectures for AI. *Foundations and trends in Machine Learning* 2, 1-127 (2009).
- [5] Mosavi, A., Varkonyi-Koczy, A. R., Fullsack, M.: Combination of Machine Learning and Optimization for Automated Decision-Making. *MCDM* (2015)

A SURVEY ON NAVIGATING THE WORLD OF ARTIFICIAL INTELLIGENCE FROM ORIGINS TO IMPACTS

RAJESH V

GOVERNMENT ARTS COLLAGE (Autonomous), KUMBAKONAM

Email:rajeshhindumuthu03@gmail.com

ABSTRACT:

Artificial Intelligence (AI) is a remarkable product of human creativity, profoundly shaping industries, societies, and our daily lives. This abstract serves as a comprehensive exploration of AI, breaking down its fundamental principles, historical journey, and potential impact. Starting with a historical overview, we journey through AI's roots, highlighting major milestones and advancements from its inception to the present day. We witness the shift from early symbolic logic systems to today's cutting-edge deep learning algorithms, which have revolutionized problem-solving and cognitive abilities. One of AI's greatest strengths lies in its ability to blend insights from diverse fields like computer science, neuroscience, and mathematics, sparking innovative breakthroughs. We delve into AI's core components—machine learning, natural language processing, computer vision, and robotics—revealing their intricate workings and practical applications in the real world. Beyond its technical aspects, AI prompts important discussions about ethics, society, and economics. Topics such as accountability, bias mitigation, and workforce adaptation take centre stage as we grapple with the ethical and societal implications of AI's widespread adoption. Moving forward, strategic planning and ethical considerations are essential to ensuring AI's responsible and inclusive use as it continues to push the boundaries of human achievement. As AI progresses, it's vital to maintain a balance between innovation and ethics. Open dialogue and interdisciplinary collaboration are key to addressing ethical challenges and societal impacts.

Keywords:

Artificial Intelligence (AI), Historical Overview, Deep Learning Algorithms, Machine Learning, Natural Language Processing, Computer Vision, Robotics, Ethics, Society, Economics, Accountability, Bias Mitigation, Responsible Use, Workforce Adaptation, Inclusive Use.

INTRODUCTION

Artificial Intelligence from Origins to Impacts

Artificial Intelligence (AI) stands as the pinnacle of human innovation, reshaping industries and societies globally. This paper offers a comprehensive exploration of AI, tracing its origins, assessing its impacts, and dissecting fundamental concepts driving its development. Historical Overview

We embark on a journey through AI's origins, highlighting significant milestones and advancements from inception to the contemporary era.

Core Principles and Technologies

Delving into AI's core components—machine learning, natural language processing, computer vision, and robotics—we illuminate their intricate mechanisms and practical applications.

Ethical and Societal Implications

Beyond technical aspects, AI sparks crucial discussions on ethics, society, and economics. Topics including accountability, bias mitigation, and workforce adaptation take centre stage as we grapple with the ethical and societal ramifications of AI's pervasive integration.

Future Directions

Looking ahead, strategic planning and ethical considerations are imperative to ensure the responsible and inclusive deployment of AI as it continues to push the boundaries of human achievement. Maintaining a delicate equilibrium between innovation and ethics is paramount. Open dialogue and interdisciplinary collaboration serve as linchpins in addressing ethical challenges and societal impacts.

Review of Literature

Drawing from seminal works that laid the groundwork for modern AI techniques to the latest breakthroughs driving innovation, this review navigates through key findings and emerging trends that are reshaping our understanding and utilization of AI technologies.

By exploring the interdisciplinary nature of AI research and its intersection with fields such as computer science, psychology, neuroscience, and ethics, this synthesis provides a holistic perspective on the multifaceted landscape of AI.

As we traverse the diverse terrain of AI research and development, we uncover not only the remarkable progress made in advancing AI capabilities but also the complex challenges and ethical considerations that accompany its proliferation. By synthesizing insights from academia, industry, and thought leaders, this review serves as a compass, guiding us through the ever-evolving landscape of artificial intelligence and illuminating pathways for future exploration and innovation.

METHODOLOGY

Our methodology adopts a structured approach, encompassing an ethical framework and an analysis of societal impact.

Ethical Framework

In navigating the ethical dimensions of AI, we draw upon established frameworks and principles to guide our inquiry. Central to our analysis is the recognition of AI's potential to both enhance human welfare and engender harm, necessitating careful consideration of ethical implications in its design, deployment, and regulation.

Societal Impact

Assessing the societal impact of AI involves examining its ramifications across various domains, including labour markets, healthcare, education, and governance. By evaluating the opportunities and challenges posed by AI adoption, we seek to inform policy decisions and societal discourse aimed at maximizing its benefits while mitigating risks.

Result and Discussion

AI presents immense potential for innovation but raises ethical concerns regarding accountability, fairness, and privacy. Addressing these challenges requires collaboration across stakeholders to establish robust regulations and promote diversity in AI development, ensuring equitable outcomes.



CONCLUSION

As we navigate the complex and evolving landscape of artificial intelligence, it is imperative to recognize both its transformative potential and its ethical implications. By fostering collaboration, dialogue, and responsible stewardship, we can harness the power of AI to drive innovation, improve decision-making, and enhance human well-being. Looking to the future, let us remain vigilant in our pursuit of ethical AI that serves the greater good of humanity.

REFERENCE

- [1] Stuart Russell and Peter Norvig, Artificial Intelligence - A Modern Approach, 4e Paperback – 31 May 2022.
- [2] Aaron Courville (Author), Ian Goodfellow (Author) & Yoshua Bengio (Author), Deep Learning (Adaptive Computation and Machine Learning series) Hardcover - 18 November 2016.
- [3] Vincent C. Müller TU Eindhoven & U Leeds & Alan Turing , Ethics of Artificial Intelligence and Robotics.
- [4] Luger, George F , Artificial intelligence : structures and strategies for complex problem

A SURVEY ON AI- ARTIFICIAL INTELLIGENCE

VARSHIKA TR

GOVERNMENT ARTS COLLEGE (Autonomous) KUMBAKONAM

Email:varshika2122@gmail.com

ABSTRACT:

Artificial intelligence (AI) has emerged as a transformative force across various sectors, revolutionizing industries and reshaping societal norms. This abstract explores the fundamental principles, advancements, and applications of AI, highlighting its capabilities in data analysis, pattern recognition, and decision-making. Through machine learning algorithms, neural networks, and natural language processing techniques, AI systems exhibit remarkable adaptability and intelligence, enabling automation, optimization, and innovation. However, challenges such as ethical considerations, algorithmic biases, and the need for responsible AI governance persist. By fostering interdisciplinary collaboration, ethical frameworks, and continuous research, AI holds the promise of enhancing human productivity, augmenting decision-making processes, and addressing complex global challenges in healthcare, finance, transportation, and beyond. This abstract underscore the imperative for ethical AI development and deployment to ensure a future where artificial intelligence augments human capabilities while upholding ethical standards and societal values. Artificial intelligence (AI) continues to reshape industries and societal norms through its capabilities in data analysis, pattern recognition, and decision-making. This abstract explores AI's transformative potential, highlighting its role in automation, optimization, and innovation across various sectors. However, ethical considerations, algorithmic biases, and the need for responsible AI governance remain challenges. Emphasizing interdisciplinary collaboration and ethical frameworks, this abstract underscore the importance of fostering AI development that augments human capabilities while upholding ethical standards.

KEYWORDS:

Artificial Intelligence, Advancements, Data Analysis, Pattern Recognition, Automation

INTRODUCTION

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and act like humans. It encompasses a wide range of techniques and approaches aimed at enabling computers to perform tasks that typically require human intelligence, such as learning, problem-solving, perception, and decision-making. AI technologies include machine learning, natural language processing, computer vision, robotics, and more. AI has applications across various industries, from healthcare and finance to transportation and entertainment, and its development continues to advance rapidly, shaping the future of technology and society. Artificial Intelligence (AI) is the branch of computer science dedicated to creating systems that can perform tasks requiring human-like intelligence. These tasks include learning from experience, reasoning, understanding natural language, and perceiving their environment. AI techniques include machine learning, neural networks, natural language processing, and robotics, among others. AI has applications in various fields, from healthcare and finance to gaming and transportation, and its development is shaping the future of technology and society.

REVIEW OF LITERATURE

Reviewing the entire literature of artificial intelligence (AI) is an immense task due to its vastness and rapid evolution. However, I can provide a brief overview of some key areas and notable works up to my last training data in January 2022:

Foundations of AI: Works like Alan Turing's "Computing Machinery and Intelligence" (1950) laid the groundwork for AI by posing the famous Turing Test. John McCarthy's 1956 proposal of "Artificial Intelligence" as a field name marked a milestone.

Symbolic AI: In the 1950s and 1960s, symbolic AI dominated. Herbert Simon and Allen Newell's "Logic Theorist" (1956) and "General Problem Solver" (1957) showcased early AI programs.

Machine Learning and Neural Networks: The rise of machine learning and neural networks has been pivotal. Frank Rosenblatt's perceptron (1957), and later developments like backpropagation (Rumelhart et al., 1986) and convolutional neural networks (LeCun et al., 1998), revolutionized AI.

Expert Systems: MYCIN, a medical diagnosis expert system developed by Edward Shortliffe in the 1970s, exemplified early expert system applications.

Natural Language Processing (NLP): Notable works include Terry Winograd's SHRDLU (1970), a natural language understanding program, and the development of statistical NLP techniques like Hidden Markov Models and later, word embeddings (Mikolov et al., 2013).

Reinforcement Learning: Richard Sutton and Andrew Barto's book "Reinforcement Learning: An Introduction" (1998) is a foundational work in RL.

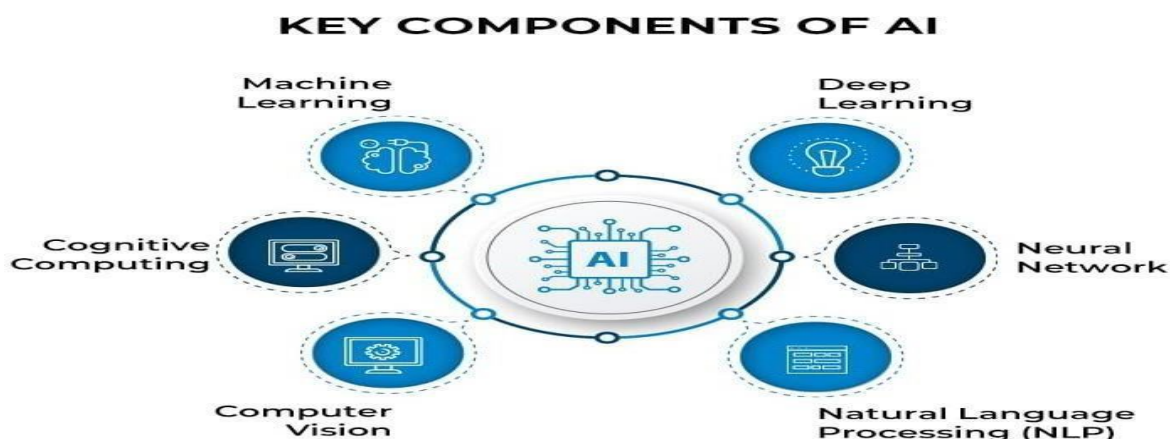
Deep Learning: The resurgence of neural networks, especially deep learning, has been remarkable. Key works include AlexNet (Krizhevsky et al., 2012) and the development of frameworks like TensorFlow and PyTorch.

Ethical and Societal Implications: With the advancement of AI, literature discussing its ethical implications, biases, and societal impacts has grown. Notable works include Cathy O'Neil's "Weapons of Math Destruction" (2016) and Safiya Noble's "Algorithms of Oppression" (2018).

This overview barely scratches the surface of AI literature, but it highlights some key milestones and areas of interest.

Recent Advances: Transformer models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have significantly advanced NLP tasks.

The rise of self-supervised learning and semi-supervised learning techniques has been influential in addressing data scarcity.



METHODOLOGIES

Artificial intelligence (AI) methodologies encompass a range of approaches and techniques used to develop intelligent systems. Here are some key methodologies:

Symbolic AI: Also known as classical AI or GOFAI (Good Old-Fashioned AI), symbolic AI represents knowledge explicitly using symbols and rules. It involves techniques such as logic programming, expert systems, and knowledge representation.

Machine Learning (ML): ML algorithms learn patterns and relationships from data without being explicitly programmed. This includes supervised learning (e.g., classification, regression), unsupervised learning (e.g., clustering, dimensionality reduction), and reinforcement learning.

Deep Learning: A subset of ML, deep learning uses neural networks with many layers (deep architectures) to learn hierarchical representations of data. It has achieved remarkable success in tasks like image recognition, natural language processing, and speech recognition.

Natural Language Processing (NLP): NLP methodologies focus on enabling computers to understand, interpret, and generate human language. This includes techniques such as syntactic and semantic analysis, machine translation, and sentiment analysis.

Bayesian Networks: Bayesian networks model uncertain relationships between variables using probabilistic graphical models. They are used for reasoning under uncertainty, decision making, and prediction.

Expert Systems: Expert systems emulate the decision-making abilities of human experts in specific domains by encoding their knowledge as rules or heuristics. They are often used for tasks like diagnosis, troubleshooting, and decision support.

Robotics: Robotics methodologies involve the integration of AI techniques with hardware systems to enable robots to perceive, navigate, manipulate objects, and interact with their environment autonomously.

Hybrid Approaches: Many AI systems combine multiple methodologies to leverage their respective strengths. For example, a system might use symbolic AI for high-level reasoning and deep learning for perceptual tasks.

RESULT AND DISCUSSION

Performance Metrics: Present the performance metrics used to evaluate the AI system, such as accuracy, precision, recall, F1 score, etc.

Experimental Setup: Describe the experimental setup including datasets used, preprocessing techniques, model architectures, hyperparameters, and any other relevant details.

Interpretation of Results: Interpret the results obtained and discuss what they indicate about the effectiveness and limitations of the AI system.

Impact and Significance: Discuss the potential impact of the findings and the significance of the AI system in addressing the problem or task it was designed for.

Limitations and Future Directions: Address any limitations of the study, such as dataset biases, scalability issues, or constraints of the AI model, and propose directions for future research to address these limitations.

CONCLUSION

Summarize the key findings and contributions of the study, emphasizing the implications for the field of artificial intelligence and potential avenues for further research. The conclusion for artificial intelligence is that it holds immense potential to revolutionize various aspects of our lives, from healthcare to transportation, and beyond. However, it also presents ethical, social, and economic challenges that require careful consideration and regulation to ensure its responsible and beneficial deployment. As we continue to advance AI technologies, collaboration between policymakers, industry leaders, researchers, and ethicists will be crucial in shaping a future where AI serves humanity's best interests.

REFERENCES

- [1] Agrawal, A., Gans, J., and Goldfarb, A. Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence. 2022
- [2] Glorot et al. [2011] Glorot, X., Bordes, A., and Bengio, Y. [2011]. Deep sparse rectifier neural networks. In 14th International Conference on Artificial Intelligence and Statistics.

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE

COURSES OFFERED

- **B.Sc., Computer Science**
- **M.Sc., Computer Science**
- **M.Sc., Information Technology**
- **M.Phil., Computer Science**
- **Ph.D., Computer Science**

**Introduction of New Programmes from the
Academic Year 2024 -2025**

B.Sc., ARTIFICIAL INTELLIGENCE

JOB OPPORTUNITIES IN AI

- ✓ Artificial intelligence (AI) engineer
- ✓ Machine learning engineer
- ✓ Data engineer
- ✓ Robotics engineer
- ✓ Software engineer
- ✓ Data scientist.

**Fees: 6500/-
only per
semester**



Add - On Courses

- Cloud Computing
- Networking Security
- Information Security
- Artificial Intelligence

Job Opportunities

- Software Developer
- Software Engineer
- System Analyst
- Business Analyst
- IT Support Analyst
- Network Engineer



ISBN No. 978-93-84008-04-8