

Department of Computer Science
 Proceedings of 2nd International Conference on Recent Innovations in Computer Science &
 Technology (ICRICT-2024)
 29th to 31st January 2024
 ISBN: 978-81-968265-0-5
 URL: [https:// pbsiddhartha.ac.in/ICRICT24/](https://pbsiddhartha.ac.in/ICRICT24/)

INDEX, VOLUME III

S.No	Title of the Article	Page. No
1	Blockchain Influence on the Digital Economy: Navigating Environmental Considerations Siri Mahitha Challa, Bhuvanewari Kosana, Lasya Priya Challa	1-7
2	Social Scopes in Virtual Reality Vasudha Jonnala, Shaik Ayesha, Shaik Nousheen	8-11
3	Enhancing Fake News Detection through Sentimental Analysis Meghana Durga Patnala, Gayathri Marrivada, Sri Lekha Ejnavarjala	12-18
4	Tabular-to-Image Transformations for the Classification of Anonymous Network Traffic Using Deep Residual Networks P. Bhargavi, S. Bhanu Sri, G.Keerthi	19-26
5	Predicting Churn Using (Statistical Tools) Mounika Emani, Jyothika Sankar Narayanan, Jaddu Harshini	27-29
6	Safeguarding Digital Landscapes: Cyber Threat Management Imadabattni.Sai Balaji, Dr.T.Srinivasa Ravi Kiran, Dr.V.Rama Chandran	30-34
7	Steganography Encryption Standard for Android Application Jaddu.Harshini, Sankaranarayanan.Jyothika, Emani. Mounika	35-39
8	Guardians of Data: A Comprehensive Study on Database Security Measures Marrivada Gayathri, Muddamsetty Sriya, Patnala Meghana Durga	40-45
9	Navigating The Digital Frontier: Technologies Shaping Digital Twins Ravada Vanitha, Dr.T.Srinivasa Ravi Kiran, Kadali Anjani	46-52
10	Navigating the Future of Bank Loan Status Predictions-Using(Python) Jyothika Sankar Narayanan, Mounika Emani, Devi Sri Ventrapragada	53-57
11	Robotic Process Automation (RPA) Shaik Nousheen, Shaik Ayesha Begum, Vasudha Jonnala	58-63
12	Intrusion Detection Using Machine Learning Shaik Obaid, Muddamsetty Sriya, S. Tulasi Prasad	64-68
13	Automatic Pavement Crack Detection and Classification Using Multiscale Feature Attention Network Seepana Nandini, Bora Uma Reddy, Sandhya Naidu	69-74
14	Forecasting Credit Risk a Comprehensive Analysis Using Advanced Models Devi Sri, S.Jyothika, E.Mounika	75-80
15	Augmented Reality: Transformative Applications in Aeronautical Maintenance, Building Information Models, and Beyond Annavarapu. Sridevi, Shaik. Ayesha Begum, Shaik. Nousheen	81-84
16	Improved Ant Colony Optimization Algorithm and its Application Bora Uma Reddy, Sandhya Naidu, Rasani Sunandini	85-88
17	SMOTE: Fusing Deep Learning and SMOTE for ImbaImbalanced Data Buraga. Bhavishya Sagarika, Gajjalakonda.Keerthi, Saggurthi.Pavitra	89-93

18	Navigating Cyber Threats and Securing Solutions Dirisala Pooja Sri, Dr.T.Srinivasa Ravi Kiran, Dr.N.Lakshmi Prasanna	94-98
19	The Relation of Big Data Analytics with CRM Sri Lekha Ejnavarjala, Gayathri Marrivada, Patnala Meghana Durga	99-104
20	An Optimal Algorithm for Finding Champions in Tournament Graph Gajjalakonda Keerthi, Penumudi Bhargavi, Singavarapu Bhanu Sri	105-110
21	Survey on Data Pre-Processing Methods for Improved ML Model Performance Neelima Kolli, Sri Latha Kolli, Mr.P.R.Krishna Prasad	111-115
22	YOLOv8-Based Helmet Detection with Integrated E-mail Notifications Muddamsetty Sriya, Shaik Obaid, Mr.Ch. Hari Prasad	116-121
23	Integrating Notifications and Manual Approval Workflows in Aws CDK Pipelines for Enhanced Devops Automation Sandhya Naidu, Rasani sunandini, Bora Uma Reddy	122-126
24	Swarm Intelligence Unleashed: Unraveling Nature's Algorithms for Computational Efficiency Saggurthi Pavitra, Bhavishya Sagarika, Muddamsetty Sriya	127-134
25	Containerized Security: Safeguarding Cloud Environments with Advanced Measures Shaik Ayesha Begum, Jonnala. Vasudha, Shaik. Nousheen	135-141
26	Bird Identification Using Deep Learning Siddhardha katipamula, Shaik Obaid, Mr. Sudha Kishore	142-145
27	Unraveling Deep Learning: Navigating the Fundamentals of Artificial Intelligence Jayamma Rodda, Dr.R. Vijaya Kumari	146-149
28	An Overview on Applications of Generative AI Dr.R.Vijaya Kumari, Jayamma Rodda	150-153

Blockchain Influence on the Digital Economy: Navigating Environmental Considerations

Siri Mahitha Challa
23DSC01, M.S.c(Computational Data Science)
Dept. of Computer Science
PB Siddhartha College of Arts and Science
Vijayawada, A.P, India
sirichalla2003@gmail.com

Bhuvaneswari Kosana
Dept of Computer Science
Teaching Assistant
P.B Siddhartha College of Arts and Science.
Vijayawada, A.P, India
bhuvikosana437@gmail.com

Lasya Priya Challa
B.Sc. (Computer Science)
Dept of Computer Science
Siddhartha Mahila Kalasala
Vijayawada, A.P, India
Lasyapriya.ch22@gmail.com

Abstract—Blockchain stands out as a cutting-edge and promising technology in today's economy. Its potential extends to addressing a range of issues in the industrial sector, including, challenges related to trust, transparency, security, and the reliability of data processing. Every transaction that gets incorporated is first verified by all the participants of that transaction. Blockchain is much more than a foundation for crypto currency. This paper summarizes the conditions of Blockchain research in terms of technology and its implementation. The practical implications and real-world impact remain essential considerations. Blockchain technology has a substantial influence on both the economic landscape and the environment. This paper presents a comprehensive overview on Blockchain Technology Applications (In Cyber security, IOT) Benefits and Challenges and their Solutions as per my view, and majorly focuses on the Environmental impact and Economy impact.

Keywords--Blockchain, IOT, Cybersecurity, Environmental Impact, Economic Impact.

I UNVEILING THE ESSENCE OF BLOCKCHAIN TECHNOLOGY

Blockchain technology holds the potential to usher in a promising future, contributing to enhanced reliability, trustworthiness, and security in business, government, and logistic systems. The robust benefits of this technology make it a valuable asset in achieving these objectives across diverse systems. At its fundamental level, a blockchain is a distributed database of records, whether they be historical or contemporary events, which are securely shared among participating parties. Blockchain technologies serve as a secure platform for the transfer of data integral to various transactions, encompassing financial transactions and contracts. At the core of blockchain lies cryptography, ensuring the integrity and authenticity of the transferred data by preventing tampering. Transactions within blockchains involve the transfer of assets, where these

assets primarily consist of data representing diverse information such as financial data, healthcare records, or corporate information. A prominent example of blockchain application is Bitcoin, a financial cryptocurrency, where the transfer of bitcoins among individuals is facilitated through blockchain technology. The National Institute of Standards and Technology (NIST) has defined blockchains to be "tamper evident and tamper resistant digital ledgers implemented in a distributed fashion (i.e., without a central repository) and usually without a central authority (i.e., a bank, company, or government)". Apart from Bitcoin, Ethereum stands out as another widely embraced application of blockchain. Ethereum is designed not only for transactions but also for implementing contracts that incorporate conditions and rules. These contracts, known as smart contracts, find relevance in domains like the Internet of Things (IoT), where countless devices collaborate by creating smart contracts for data exchange and process execution. Functioning as a peer-to-peer technology, Ethereum operates without centralized control. Blockchains primarily handle transactions, and transactions inherently involve data. The processing, analysis, and sharing of massive data sets are integral to various transactions. Data science techniques play a crucial role at the core of these transactions. Blockchains provide a secure means to execute these transactions.

II "DEMISTIFYING THE MECHANICS: UNDERSTANDING HOW BLOCKCHAIN TECHNOLOGY OPERATES:"

Blockchain technology operates as a decentralized and distributed ledger system designed to record and verify transactions securely. The fundamental principles of how blockchain works can be explained as follows
BLOCKS AND TRANSACTIONS: Blockchain essentially consists of a collection of blocks that are linked together via chains. A block is essentially a file that contains data pertaining to a transaction. The data from one block may be transferred to multiple blocks. Furthermore, a block may receive data from multiple blocks. The data in each block is permanent and immutable. Blocks can be added to the blockchain as the transaction progresses. Furthermore, each

transition has to be verified. However, unlike in non-blockchain applications where transactions are usually verified by a central authority, in a blockchain based transaction, it is verified by a distributed collection of processes.

DECENTRALIZATION: Traditional databases are often centralized, meaning a single entity or authority manages and controls the data. In contrast, blockchain operates on a decentralized network of computers (nodes). Each node on the network has a copy of the entire blockchain.

CRYPTOGRAPHY: An important component of blockchain is cryptographic hash functions. This is a form of a message digest where checksums are computed based on the contents. Cryptographic techniques, such as hash functions and digital signatures, are employed to secure the data within each block.

Hash functions create a unique identifier for each block, and digital signatures ensure the authenticity of transactions. Blockchains use asymmetric key technology which is essentially public key cryptography. Blockchains may also use network addresses which are derived from the public key cryptography.

CONSENSUS MECHANISM: Blockchain relies on a consensus mechanism to validate and agree on the state of the ledger. Common consensus algorithms include Proof of Work (used in Bitcoin) and Proof of Stake. These mechanisms ensure that all nodes in the network reach an agreement on the validity of transactions.

IMMUTABLE AND TRANSPARENT LEDGER: Once a block is added to the blockchain, it is almost impossible to alter the information within it. This immutability is achieved through the cryptographic links between blocks (hashes). The transparency of the ledger means that all participants in the network can view the entire transaction history.

SMART CONTRACTS: Some blockchains, like Ethereum, support the implementation of smart contracts. Smart contracts are self-executing contracts with the terms of the agreement directly written into code. They automatically execute and enforce the terms when predefined conditions are met.

IMMUTABILITY: Once a block is added to the blockchain, it is extremely difficult to alter or delete information within it. The decentralized and distributed nature of the network, along with cryptographic techniques, ensures the security and immutability of the data.

CRYPTOGRAPHIC HASHING:

Each block contains a cryptographic hash (a unique identifier) based on the contents of the block. If someone attempts to alter the data in a block, it will change the hash, and this change will be evident to all nodes on the network.

ADDING BLOCKS TO THE CHAIN:

Once a block is validated and approved, it is added to the existing blockchain in a chronological order, forming a chain of blocks. Each block contains a reference to the previous block, creating a secure and tamper-resistant chain.

MINING: In Proof of Work blockchains like Bitcoin, the process of mining involves solving complex mathematical problems to validate transactions and create new blocks. Miners compete to solve these problems, and the first one to solve it gets the right to add a new block to the blockchain.

NODES and PEER TO PEER WORK: Nodes are individual computers connected to the blockchain network. Each node has its copy of the entire blockchain, and they communicate with each other to maintain a synchronized and updated ledger.

III THE RIPPLE EFFECT

How Blockchain is Shaping Everyday Life of a Common Man: Blockchain technology is gradually becoming more integrated into the daily lives of common individuals. While its impact may not always be directly visible, several applications and use cases are influencing various aspects of daily life:

Blockchain's emphasis on cryptographic **security and decentralized architecture** contributes to improved data security and privacy standards. As businesses adopt blockchain for various purposes, there is a potential indirect benefit for individuals by reducing the risk of data breaches and unauthorized access to personal information. Blockchain has the potential to improve financial services and increase **financial inclusion**. As traditional banking systems integrate blockchain for faster and more secure transactions, individuals in underserved or remote areas may gain improved access to financial services. In industries such as pharmaceuticals, luxury goods, and food, blockchain can help combat **fraud and counterfeiting**. By ensuring the transparency and traceability of supply chains, individuals can have increased confidence in the authenticity of the products they purchase. Businesses adopting blockchain for **supply chain management**, logistics, and administrative processes can indirectly benefit individuals by contributing to more efficient and streamlined operations. This can potentially lead to improved services and reduced costs in various industries. **Decentralized applications** built on blockchain platforms may offer individuals alternatives to traditional centralized services. This could lead to increased autonomy, reduced reliance on intermediaries, and a greater level of control over personal data. Blockchain has the potential to enhance **digital identity protection**. As individuals increasingly conduct transactions and share personal information online, the implementation of blockchain-based identity systems can provide more secure and decentralized solutions, reducing the risk of identity theft. The growth of the blockchain industry can indirectly create **job opportunities and drive innovation**. As the technology

evolves, new businesses and startups may emerge, contributing to economic growth and providing individuals with opportunities in the rapidly expanding blockchain sector. Some blockchain projects focus on **sustainability and energy efficiency**. The adoption of these initiatives can indirectly benefit individuals by contributing to a more sustainable environment and promoting eco-friendly practices. Faster and more cost-effective transactions can lead to improved financial accessibility and efficiency for individuals engaged in global economic activities. Individuals can benefit from **educational opportunities** to acquire skills related to blockchain, enhancing their career prospects in a technology-driven job market.

3.1 BLOCKCHAIN REVOLUTION: TRANSFORMING BUSINESS ACROSS INDUSTRIES

Although blockchain technology initially gained prominence through cryptocurrency offerings like Bitcoin, its practical and transformative capabilities go well beyond the domain of digital currencies. In numerous sectors, blockchain technology has the potential to streamline operations, bolster security, and elevate transparency. Blockchain has the potential to redefine both business and social processes, driving innovation across various domains.

MEDICAL RECORDS MANAGEMENT:

Blockchain could revolutionize medical records management by providing a secure and immutable ledger for patient data. Patients would have complete control over their records, and healthcare providers could access accurate, up-to-date information, leading to better patient care and streamlined administrative processes.

SUPPLY CHAIN MANAGEMENT:

Blockchain technology can be used to create transparent and traceable supply chains. Blockchain enables the tracking of a product from its source to the end consumer; each step in the product's creation and distribution is recorded in a tamper-resistant, immutable ledger. Blockchain technology is particularly beneficial in sectors such as the food industry, where traceability can help producers and retailers swiftly identify and recall contaminated or unsafe products.

ASSET TRADING:

An obvious blockchain use case is tokenized stock trading, which would enable retail investors to buy parts of stocks that are too expensive for them to purchase in full. Blockchain could also enable the tokenization of other asset classes, including real estate, bonds or gold. This would make it possible for an investor to have a more diversified investment portfolio with various asset classes and provide low transaction costs and 24/7 trading.

VERIFYING THE PROVENANCE OF LUXURY GOODS:

Luxury goods brands can harness blockchain technology to showcase their products' provenance. These brands can

foster deeper trust with consumers by presenting a transparent, blockchain-validated history of a product. Such an approach would be welcomed by retail buyers and offer significant comfort to secondary market buyers. With high-quality fakes on the market, the time for this is now.

VOTING:

The traditional voting system can significantly benefit from blockchain technology. By utilizing blockchain's decentralized and immutable nature, democratic countries can ensure transparent, tamper-proof elections, making the voting process more trustworthy and verifiable for all citizens.

TOKENIZING AND TRACKING CREATIVE WORK:

Blockchain can combat piracy in the media and entertainment sectors by enabling artists to tokenize and track their content, ensuring rightful ownership and revenue streams. Furthermore, artists could facilitate direct sales to their audiences, eliminating middlemen and fostering a closer connection with their fans.

SATELLITE DATA MANAGEMENT:

The space industry can benefit from blockchain technology by implementing it for satellite data management. This approach enhances data security, verification, distribution and transparency and opens up new revenue streams. With blockchain's potential to revolutionize the way we handle satellite data; the space industry can better serve the needs of our evolving world and improve the quality of life on Earth.

DOCUMENTING CLINICAL TRIALS:

Clinical **trials** are **critical** for the evolution of healthcare but often suffer from data manipulation, incomplete reporting or even fraud. With blockchain, every step of the clinical trial process can be securely recorded and timestamped. This guarantees transparency, ensures data integrity and fosters trust among all stakeholders.

INSURANCE CLAIMS PROCESSING:

The insurance sector stands out as a prime potential beneficiary of blockchain technology. Specifically, blockchain could revolutionize claims processing. Through smart contracts, claims can be auto-verified and processed when predefined conditions are met, significantly reducing fraud, expediting settlements and bolstering customer trust in the insurance value chain.

SIMPLYFYING REAL ESTATE TRANSACTIONS:

The real estate industry can particularly benefit from blockchain technology by leveraging smart contracts and tokenization. Blockchain simplifies property transactions through digital titles and asset tokenization, allowing fractional ownership and boosting market liquidity, ultimately streamlining property transactions for investors and homeowners.

PROTECTING INTELLECTUAL PROPERTY:

In the realm of patents, blockchain technology can be used to protect intellectual property. Innovators and patent holders can utilize blockchain technology to register their

inventions, establishing ownership and guaranteeing equitable compensation. Smart contracts can be employed to automate royalty disbursements, establishing an impartial and transparent framework for patent holders.

IV "ECONOMIC EVOLUTION: THE AUGEMENTED INFLUENCE OF BLOCKCHAIN ON BUSINESS".

Blockchain for a new digital economy: Blockchain is revolutionizing the way we do business, making transactions faster, safer, and more efficient. It has the power to catalyze the shift to a digital economy, and here's how. The beauty of blockchain lies in its decentralized nature. Unlike traditional systems that rely on a central authority, blockchain operates on a network of computers, each with its own copy of the blockchain. This means that there's no single point of failure, making the system more secure and less susceptible to fraud. This is similar to the way a spider's web works, with each string connected to the other and no single point of failure. This network of connections makes the web strong and resilient, allowing it to withstand any external forces or damage. Understanding the effect of cryptocurrencies on the financial sector is a yardstick to measure the extent of the impact it may have on the general economy. **The acceleration** of the global digital economy is in borderless adoption and diffusion of blockchain technology. Hyperscale's in the corporate world have realized it too. Goldman Sachs, Deloitte, Cboe Global Markets, Microsoft, and Digital Asset have all joined forces to create a groundbreaking blockchain network that will revolutionize the world of financial product smart contracts for institutional crypto assets.

FINANCIAL TECHNOLOGY MODIFICATION:

When you talk about how cryptocurrency can change the world, you are invariably talking about the change that comes with the blockchain era. The blockchain industry has grown into a multi-billion-dollar industry as many professionals had anticipated. One can boldly assume that the blockchain era in cryptocurrencies will influence the financial system given how international transactions amongst financial institutions have been enabled through blockchain. The impact of cryptocurrency on the general economy also extends to adopting blockchain in keeping accurate and credible financial transactions. With blockchain technology, many methods like executing clever contracts through cloud computing, and vehicle leasing are simplified. Also, this blockchain era ushered in the possibility of making cryptocurrency payments to employees by businesses like Kodak.

IMPROVED FINANCIAL STABILITY:

People are slowly losing confidence and trust in traditional economic institutions like banks over the years. With the advent of blockchain and cryptocurrency, people have discovered a new system which offers them total control over their finance and also paves way for financial inclusion. You can freely enter into any financial services without any problems or any banks interfering in

between. Blockchain technology and cryptocurrencies have offered people the power and opportunity to transact in a more stable financial market without the involvement of any third party, hence, bringing the whole economy to equilibrium. With blockchain and cryptocurrency applications, some economic nations whose domestic currencies are constantly underperforming can be easily stimulated.

JOB OPPURTUNITIES:

It is too obvious to deny the brand-new labour market created solely by the blockchain and cryptocurrency industry. This is a major answer to the question of "how blockchain and cryptocurrency can impact the world," The general acceptability and recognition of cryptocurrencies have masterminded a huge demand for crypto professionals and experts around the world. There are large solutions, participants and exchanges in many international crypto industries which demands the management of humans. Therefore, with the enormousness of this industry and the increasing adoption of cryptocurrencies, the technology creates room for many more new job opportunities in the world which can strengthen the global economy. Many companies are regularly creating more room for the demand in crypto-related task roles given the heightened blast in crypto activity listings. **According to LinkedIn, blockchain and crypto experts are one of the highly demanded labors sought by employers.**

IV FUTURE SCOPE:

In a digital economy, where transactions happen in the blink of an eye, security is paramount. And blockchain delivers just that. For example, blockchain technology is used to facilitate secure payments, verify data and authenticate digital identities, all of which are necessary for efficient and secure digital transactions. Blockchain is slated to generate an annual business value of more than \$3 billion by 2030

Tapping the Power of Tokenization:

Imagine a fascinating world of blockchain and cryptocurrency, where crypto-tokens reign supreme. These tokens are a powerful tool that can transform the way we conduct transactions, cutting down on costs associated with verifying transaction attributes and networking. By utilizing blockchain technology, we can wave goodbye to intermediaries like banks and brokers and embrace a new era of peer-to-peer transactions. With crypto-tokens, you can trade, exchange, and utilize your assets with ease, without ever having to worry about unnecessary fees or complications.

Impact on the Financial Markets and Economic implications:

Tokenization underpinned by blockchain technology has the power to revolutionize the world of finance, heralding a new era of accessible and easily tradable financial assets. This seismic shift will fundamentally alter the financial landscape as we know it, ushering in a myriad of benefits for investors and market participants alike.

4.1 Unveiling the Environmental Dilemma: Blockchain's Ecological Footprint and Potential Environmental Hazards:

Blockchain technology has a significant **carbon footprint** due to its energy-intensive process of verifying transactions and creating new blocks on the blockchain. The energy consumption of blockchain technology results in significant greenhouse gas emissions, which contribute to climate change. The energy consumption of blockchain technology can be attributed to verifying transactions and creating new blocks on the blockchain. This verification process is done through a process called **"mining."** The energy consumption of mining is primarily due to using high-powered computing equipment, such as **ASIC (Application-Specific Integrated Circuit)** miners and **GPUs (Graphics Processing Units)**. The escalating popularity of blockchain technology has raised a significant apprehension regarding its energy consumption. The process of validating transactions and generating new blocks on the blockchain necessitates a substantial amount of computational resources, which consequently demands a considerable quantum of energy. The present study examines the carbon footprint of **Blockchain Technology and its potential impact on climate change**. The carbon footprint of blockchain technology can be attributed to the energy consumption of mining. Mining involves solving complex mathematical algorithms to verify transactions and create new blocks on the blockchain. Bitcoin presently uses around 110 Terawatt Hours each year, or about the yearly energy consumption of small nations like Malaysia or Sweden, according to the Cambridge Centre for Alternative Finance (CCAF).

Ecological Consequences of Crypto Currency Mining:

The amount of energy consumed by cryptocurrency mining will likely vary over time, assuming that prices and user adoption continue to change. Cryptocurrency mining is a competitive process: as the value of the block reward increases, the incentives to start mining also increase. Higher cryptocurrency prices mean more energy consumed by crypto networks because more people join the mining networks trying to profit from the increases. Calculating the carbon footprint of cryptocurrency is more complicated. Although fossil fuels are the predominant energy source in most countries where cryptocurrency is mined, miners must seek out the most inexpensive energy sources to remain profitable. Digiconomist estimates that the Bitcoin network is responsible for about 73 million tons of carbon dioxide per year—equal to the amounts generated by Oman. Based on data through December 2022, Ethereum produced an estimated 35.4 million tons of carbon dioxide emissions before dropping to 0.01 million tons following its transition to proof of work.

Environmental Repercussions of Bitcoin Mining:

Bitcoin mining is the process of validating the information in a blockchain block by generating a cryptographic solution that matches specific criteria. When a correct

solution is reached, a reward in the form of bitcoin and fees for the work done is given to the miner(s) who reached the solution first. Bitcoin mining requires the mining program to generate a random hash and append another number to it called the nonce, or "number used once." When a miner begins, it always starts this number at zero. Every miner on the network does this until a **i** combination is created that is less than or equal to the target hash. The first to reach that target receives the reward and fees, and a new block is opened. Once that block fills up with information (about one megabyte), it is closed, encrypted, and mined. Every miner on the network does this until a hash and nonce combination is created that is less than or equal to the target hash. The first to reach that target receives the reward and fees, and a new block is opened. Once that block fills up with information (about one megabyte), it is closed, encrypted, and mined.

4.2 Strategies to mitigate the environmental and economic impact of Bitcoin and crypto mining.

First thing to do is to reduce the carbon footprint in blockchain space. Reducing carbon emissions in the blockchain space requires a multifaceted approach that addresses various aspects of the technology and its operations. One effective strategy is to integrate renewable energy sources into blockchain mining operations. This involves powering the mining facilities with solar, wind, geothermal, or hydropower. As a result, blockchain operations can significantly reduce their carbon footprint by shifting away from fossil fuel-based energy.

- Transitioning from the traditional Proof of Work (Pow) consensus algorithm to Proof of Stake (PoS) can also significantly reduce carbon emissions in the blockchain space.
- Pow algorithms, like the one used by Bitcoin, are highly energy-intensive as they require miners to solve complex mathematical puzzles.
- Also, taking transactions off-chain can help. Implementing off-chain transactions or layer-two scaling solutions can alleviate the burden on the leading blockchain network, reducing energy consumption and carbon emissions.

Promoting decentralization and scalability is essential to reduce blockchain networks' energy consumption and carbon emissions. By encouraging a distributed network infrastructure, blockchain platforms can avoid the concentration of mining power in a few energy-intensive locations.

- Some cryptocurrencies have intense energy requirements and special equipment needs, generating lots of waste. In that sense, some are not environmentally friendly.

V RESEARCH RESULT:

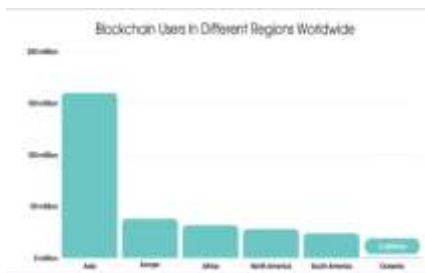
The blockchain market has steadily grown over the years. Its market value increased from \$4.19 billion in 2020 to

\$19.36 billion in 2023. It is fascinating to see that the blockchain industry is worth billions Even though just 3.9% of people worldwide use it.

Blockchain users: Nearly 1 in 20 people worldwide use Blockchain. That is quite low compared to the number of people who are aware of the blockchain and cryptocurrency

market. Besides, the highest number of Blockchain users are from the Asian continent, followed by Europeans.

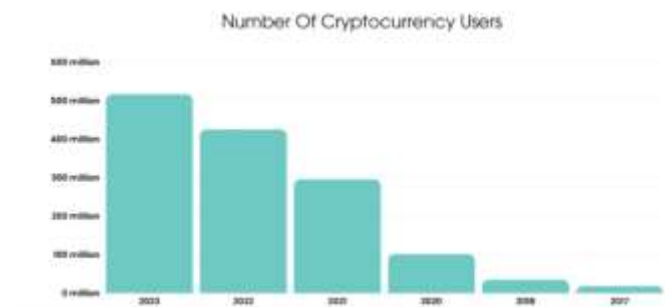
- More than 300 million people worldwide use Blockchain.
- Over 85 million people worldwide use Blockchain Wallet.
- The blockchain market is predicted to reach \$32.69 billion in 2024.
- The global spending on blockchain solutions is forecasted to reach \$19 billion.
- There are nearly 20,000 cryptocurrencies in existence. Of these, 8,866 cryptocurrencies are considered active.
- Nearly 90% of the businesses surveyed reported deploying blockchain technology in some capacity. 86% of people believe Blockchain technology can enhance their business.
- Crypto users lost \$1.8 billion in 2023 hacks and scams.



5.1 RELATED RESEARCH AND DEVELOPMENT:

The related work on blockchain encompasses a wide range of research and development efforts. Scholars and practitioners have explored various aspects, including scalability solutions, interoperability standards, energy efficiency improvements, regulatory frameworks, privacy-enhancing features, user-friendly interfaces, smart contract auditing, token standards, and community collaboration. There are many considerations that revolutionized in these developments, there are upcoming challenges and opportunities regarding this work. This body of work aims to address challenges and leverage opportunities for the responsible integration of blockchain technology into the digital economy. Ongoing efforts emphasize education, awareness, and the establishment of best practices to ensure a secure, efficient, and inclusive blockchain ecosystem. Various concerns have been raised regarding Bitcoin, cryptocurrencies, and their environmental impacts. These issues are not standardized, highlighting the need for well-defined methodologies to implement effective solutions.

Addressing the environmental effects of these technologies requires a structured approach, incorporating specific methodologies to normalize and mitigate the identified impacts. It is natural for people to want as much money as possible, but with that desire comes the drive to get it in any way possible—many people resort to illegal or unethical means to fulfill their dreams of wealth. Blockchain can solve corruption issues within financial systems by eliminating the risk of duplicating or double-spending assets. Transactions are initiated, executed, recorded, encrypted, verified, and stored by nodes who manage "an irreversible ledger of transactions" and must reach a consensus to approve a new block—making it the perfect solution. Over 85 million people worldwide use blockchain wallet, the blockchain users have significant growth over the years. There were just 10 million blockchain wallet users in 2016. This number has reached to 80 million in 2021, there as a increase of 70 million users in just 5 years. The blockchain market was valued at \$19.36 billion in 2023. It is predicted to reach \$162.84 billion by the end of 2027. The Blockchain Technology Market Size Is Forecasted to Reach \$32.69 Billion in 2024 further, the forecast states that the market will be valued at \$162.84 billion by the year 2027. Increased investments in blockchain technology and the rising adoption of distributed ledger technology (DLT) systems are major factors contributing to the growth of the blockchain market. Additionally, we have **cryptocurrency statistics**. This rapid growth in the number of users is due to the increasing awareness about crypto and blockchain among people worldwide. Here are further details about the cryptocurrency users. There Are Over 516 million Cryptocurrency Users Worldwide This is an increase of 90 million cryptocurrency users compared to 2022.



- Men’s interest has always been more significant than that of females in the blockchain and crypto markets.
- This is because females consider the market very risky and avoid its usage or investing in it. The same goes for Asians, as they prefer not to invest in the crypto markets.
- 22% of the males reported that they had used cryptocurrencies, while just 10% of females said

so. As per a survey by CNBC and Acorn, 16% of men and 7% of women invest in blockchain technology.

- Considering the fact that 69% of the population in the United States is White, it comes as no surprise that the highest percentage of Crypto owners in the United States are White.
- 57% Of the Total Number of Crypto Owners in The United States Are Millennials.
- Nearly 90% Of the Businesses Surveyed Reported Deploying Blockchain Technology in Some Capacity.
- Furthermore, 87% of businesses said they plan to invest in blockchain in 2024.42% of the Businesses that are already utilizing the technology are benefiting from its security capabilities.
- Another 42% of the businesses said they benefitted from its copy protection capabilities. 86% Of People Believe Blockchain Technology Can Enhance Our Integration Toward More Touchless Business Processes.

VI CONCLUSION:

Blockchain technology is a decentralized and distributed ledger system that securely records and verifies transactions across a network of computers. It has gained significant attention and adoption across various industries due to its unique features and potential benefits. Here are some key points to consider in conclusion about blockchain technology. There are many efficient measures like Decentralization, as block chain work on decentralized nodes, eliminating the need for a central authority or intermediary. The use of cryptographic measures ensures security of transactions on the blockchain. Also, it provides a transparency and immutable record of transactions. Smart contracts automate and enforce contract execution, reducing the need for intermediaries and potentially streamlining process. These also help in increasing efficiency and faster transaction. Blockchain has the potential to lower the transaction costs for business. In conclusion, blockchain technology has the potential to revolutionize various industries by providing secure, transparent, and efficient solutions. While it is not without challenges, ongoing

innovations and improvements are likely to shape its future and contribute to its widespread adoption.

VII REFERNCES:

[1] Siddharth Rajput, Archana Singh, Miti Khurana, Tushar Bansal, Sanyukta Shrehtha “Blockchain technologies and Cryptocurrencies.

[2] Zibin Zheng, Shaon Xie, Hongning Dai, Xiangping Chen, Huaimin Wang.” Overview of Blockchain Technology Architecture, Consensus, Future Trends.

[3] Henry Rossi Andrian, Novianto Budi Kurniawan, Suhardi “Blockchain Technology and implementation: A systematic Literature Review”.

[4] Tareq Ahram, Arman Sargolazer, Saman Sargolzaei, Jeff Daniels, Ben amaba “Block chain Technology Innovations”.

[5] Bhavani Thuraisingham “Blockchain Technologies and their Applications in Data Science and Cyber security.

[6] Julija Golosova, Andrejs Romsnovs, “The Advantages and Disadvantages of the Blockchain Technology.

[7] Tapscott, Don and Alex Tapscott “Blockchain Revolution, How the Technology behind the Bitcoin is Changing Money, Business, and the World.

[8] Walport, Mark “Distributed Ledger Technology: Beyond Block Chain”.

[9] Narayanan, A. Bonneau, J., Felten, E., Miller, A., & Goldfeder “Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.

[10] Mougayar “The Business Blockchain: Promise, Practice and Application of the Next Internet Technology.

[11] Nakamoto “Bitcoin: A peer-to –peer Electronic Cash System.

[12] Swan “Blockchain Blueprint for a New Economy”.

Social Scopes in Virtual Reality

Vasudha Jonnala
 23DSC03, M. Sc (Computational Data
 Science)
 Dep. of Computer Science
 P.B.Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 vasudhareddyjonnala@gmail.com

Shaik Ayesha
 23DSC31, M. Sc (Computational Data
 Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts &
 Science
 Vijayawada, India
 shaikayasha212126@gmail.com

Shaik Nousheen
 23DSC16, M. Sc (Computational Data
 Science)
 Dept. of Computer science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, India
 nshaik0311@gmail.com

Abstract—Virtual Reality (VR) is a computer-generated, immersive simulation that replicates a three-dimensional environment. Users typically interact with this environment using specialized devices like VR headsets. VR aims to create a sense of presence, enabling users to feel as though they are physically present in the simulated space. The technology is employed across diverse fields, such as gaming, education, healthcare, and training, offering novel ways to experience and engage with digital content.

Keywords— *Virtual Reality; presence; future, framework; VR Applications*

I. INTRODUCTION

Virtual reality (VR) immerses users in a computer-generated environment, often through headsets or other devices. It simulates a sensory-rich experience, allowing users to interact with and navigate digital spaces, creating a sense of presence and realism. VR finds applications in gaming, education, healthcare, and various industries, transforming how we perceive and engage with digital content.

Virtual Reality Types:

There are primarily two types of virtual reality:

1. Non-immersive VR: Involves less immersive experiences, often accessed through regular screens. Examples include 360-degree videos or basic simulations on a computer monitor.

2. Immersive VR: Provides a more immersive experience using specialized hardware like VR headsets. Users feel a sense of presence in a computer-generated environment, interacting with the virtual world in a more convincing way. Additionally, VR experiences can be classified based on the level of immersion they offer, ranging from basic 3D environments to fully immersive, interactive simulations.

II. VIRTUAL REALITY BACKGROUND

Virtual reality (VR) has a fascinating background that spans several decades:

- **Origins (1960s):** The concept of virtual reality emerged in the 1960s, with Morton Heilig's Sensorama (1962) and Ivan Sutherland's development of the first head-mounted display, known as the "Sword of Damocles" (1968).

- **Early Developments (1970s-1980s):** Research in the 1970s and 1980s laid the foundation for VR. Notably, Jaron Lanier coined the term "virtual reality" in the 1980s. Early attempts at VR systems, such as VPL Research's Data Glove and EyePhone, contributed to the field.
- **Commercialization Attempts (1990s):** The 1990s witnessed attempts to bring VR to the consumer market, including products like Nintendo's Virtual Boy. However, due to technological limitations and high costs, these efforts were not widely successful.
- **Dormancy and Research (Late 1990s-2000s):** VR experienced a period of relative dormancy in the late 1990s and early 2000s. Despite this, research and development continued, laying the groundwork for future advancements.
- **Resurgence (2010s Onward):** The 2010s marked a resurgence in interest and development for VR. The Oculus Rift, launched in 2012 through a Kickstarter campaign and later acquired by Facebook, played a crucial role in revitalizing the industry.
- **Diverse Applications (2010s-2020s):** VR has found applications beyond gaming, including education, healthcare, architecture, training simulations, and virtual meetings. Major companies like HTC, Sony, and Valve have entered the VR market with their respective headsets.
- **Technological Advancements:** Advances in display technology, graphics processing, motion tracking, and haptic feedback have significantly improved the immersive quality of VR experiences.
- **Challenges and Future Prospects:** Challenges such as motion sickness, the need for more compelling content, and affordability persist. Nonetheless, ongoing research and development continue to expand the possibilities of VR, with the potential to revolutionize various industries. In summary, virtual reality has a rich history of innovation, setbacks, and resurgence. Its evolution from conceptualization to widespread applications showcases the enduring interest and potential of this immersive technology. There are several virtual reality (VR) frameworks available that

developers can use to create VR experiences across different platforms.

III. FRAMEWORKS OF VIRTUAL REALITY

Some of the popular VR frameworks include:

- **Unity3D:** A widely used game engine that supports VR development for various platforms like Oculus Rift, HTC Vive, and others. It provides a robust set of tools and assets for creating VR experiences.
- **Unreal Engine:** Similar to Unity, Unreal Engine offers powerful tools for VR development. It's known for its high-quality graphics and is used for creating immersive VR experiences.
- **WebVR:** A JavaScript API that allows developers to create VR experiences that work directly in web browsers. It aims to make VR content accessible across different devices.
- **SteamVR/OpenVR:** Developed by Valve, SteamVR is an open VR software platform compatible with various VR headsets. OpenVR provides APIs and tools for creating VR experiences that can run on different VR hardware.
- **Google VR SDK/Cardboard:** Google provides SDKs for developing VR apps for Android and iOS using Cardboard and Daydream. These SDKs offer tools for creating VR experiences specifically for mobile devices.
- **Oculus SDK:** Oculus provides its own SDK for developing VR applications for Oculus Rift and Oculus Quest devices. It includes tools, APIs, and resources for building immersive VR experiences.

Each framework has its own set of features, compatibility, and learning curves. The choice often depends on the developer's familiarity, the target platform, and the specific requirements of the VR project.



Fig. 1. VR Frameworks

IV. LITERATURE REVIEW

Virtual reality (VR) has been extensively applied across various domains, offering innovative solutions and transformative experiences. Here's a literature review highlighting VR applications in different areas:

Healthcare:

- **Pain Management:** Studies (Bantin et al., 2020) have shown VR's effectiveness in reducing pain perception by

immersing patients in relaxing or engaging virtual environments, distracting them from discomfort.

- **Therapeutic Interventions:** VR has been used in exposure therapy for treating phobias (e.g., acrophobia, arachnophobia) and PTSD (Post-Traumatic Stress Disorder) by creating controlled, immersive scenarios (Rothbaum et al., 2014).

- **Rehabilitation:** VR-based rehabilitation programs (Laver et al., 2017) aid in motor skills recovery after stroke or injuries, providing engaging exercises and real-time feedback.

Education and Training:

- **Enhanced Learning:** VR-based educational content (Merchant et al., 2014) improves learning outcomes by offering immersive experiences in subjects like science, history, and complex concepts.

- **Skill Training:** Simulations in VR (Sutherland et al., 2016) have proven effective for training surgeons, pilots, and other professionals, offering a safe and controlled environment for practice.

Entertainment and Media:

- **Gaming:** VR gaming (Bowman et al., 2019) provides highly immersive experiences, enhancing player engagement and interaction through realistic environments and interactions.

- **Cinematic Experiences:** VR has expanded storytelling by enabling immersive narrative experiences (Sanchez-Vives & Slater, 2016), allowing users to become part of the story.

Architecture and Design:

- **Visualization:** VR aids architects and designers in visualizing projects (Wu et al., 2017), allowing stakeholders to experience and interact with proposed designs in a realistic virtual environment.

- **Collaboration:** Teams in architecture and design (Forte et al., 2020) use VR to collaborate remotely, reviewing designs and making real-time changes in a shared virtual space.

Social Sciences:

- **Social Interactions:** VR facilitates remote social interactions (Pan et al., 2020), allowing people to meet, communicate, and collaborate in virtual spaces, providing a sense of presence.

- **Empathy and Perspective-taking:** Studies (Riva et al., 2007) suggest VR's potential to foster empathy by placing individuals in simulated situations, leading to better understanding and perspective-taking.

V. EVALUATION

Evaluating virtual reality (VR) involves assessing various aspects to ensure its effectiveness, usability, and impact across different applications. Here are key evaluation criteria:

User Experience:

- **Presence:** Measure the sense of "being there" in the virtual environment. This includes evaluating sensory immersion, interaction fidelity, and emotional engagement.

- **Usability:** Assess the ease of use, navigation, and interaction within the VR environment. Conduct usability testing to identify user interface issues and improve user experience.

Technical Performance:

- **Graphics and Rendering:** Evaluate the quality of visuals, frame rates, resolution, and overall graphical fidelity to ensure a smooth and immersive experience.
- **Latency and Responsiveness:** Measure the delay between user actions and system responses. Low latency is crucial to prevent motion sickness and maintain realism.

Impact and Effectiveness:

- **Learning and Training Outcomes:** Assess the effectiveness of VR-based training or educational programs in improving knowledge retention, skill acquisition, and performance.
- **Therapeutic Efficacy:** Evaluate the impact of VR in therapeutic settings, such as pain management, phobia treatment, or rehabilitation, by measuring patient outcomes and improvements.

User Comfort and Safety:

- **Motion Sickness:** Assess and mitigate factors contributing to discomfort or motion sickness in users. Monitor user reactions and adjust VR experiences accordingly.
- **Physical Safety:** Ensure that VR experiences do not pose physical risks to users. Design environments and interactions that prioritize user safety.

Cost and Practicality:

- **Cost-Benefit Analysis:** Evaluate the cost-effectiveness of VR solutions concerning their intended outcomes and benefits, considering both initial setup costs and long-term maintenance.
- **Scalability and Accessibility:** Assess the scalability of VR solutions across different platforms and devices. Ensure accessibility for diverse user groups.

Ethical and Social Considerations:

- **Privacy and Data Security:** Ensure user data protection and privacy within VR environments, especially in shared or collaborative experiences.
- **Inclusivity:** Evaluate the inclusivity of VR experiences to accommodate diverse users, considering factors such as disabilities, cultural differences, and age groups. Evaluating VR involves a mix of qualitative and quantitative methods, including user studies, surveys, usability testing, and performance metrics [4]. Continuous evaluation and iterative improvement are essential to enhance VR experiences and ensure their effectiveness in various applications.



VI.RESULTS AND DISCUSSION

A. VR Advantages

One of the key advantages of virtual reality in education is its ability to provide immersive learning experiences. By using VR technology, students can be transported to different places and environments that they might not otherwise have access to, such as a historical site, a foreign country, or even outer space. For example, Google Expeditions is a VR platform that allows students to take virtual field trips to various destinations around the world.

B. Key Challenges and Concerns

Absolutely, while virtual reality (VR) has seen significant advancements, it still faces several challenges and concerns:

Technological Hurdles:

- **Hardware Limitations:** High-quality VR experiences often require expensive, high-performance hardware, which can be a barrier to entry for some users.
- **Motion Sickness:** Some users experience motion sickness or discomfort while using VR, particularly with certain movements or prolonged sessions.
- **Tethered Systems:** Wired VR headsets limit movement and can be restrictive, hindering the immersive experience.

Content and Development:

- **Content Quality and Quantity:** Despite growth, there's still a relatively limited library of high-quality VR content, particularly in certain genres or applications.
- **Development Costs:** Creating VR content can be costly and time-consuming due to the need for specialized skills and tools.

- **Accessibility and Inclusivity:** Ensuring that VR experiences are accessible to all users, including those with disabilities, remains a challenge.

Ethical and Social Concerns:

- **Isolation and Social Disconnect:** Long-term use of VR might lead to social isolation or detachment from the real world, raising concerns about mental health impacts.
- **Privacy and Data Security:** VR collects user data, which can raise concerns about privacy breaches and data security, especially with biometric information.
- **Ethical Use of VR:** Addressing ethical considerations around VR content, especially in areas like gaming, education, and healthcare, remains important.

Adoption and Acceptance:

- **Public Perception and Acceptance:** Some people still perceive VR as a niche or gimmicky technology, hindering its widespread adoption.
- **Education and Training:** There's a need for better training and education about VR technology to maximize its potential across various industries.
- **Integration with Existing Systems:** Incorporating VR into existing workflows and systems in fields like healthcare, education, and businesses can be challenging.

Health and Safety:

- **Physical Health Concerns:** Prolonged use of VR might lead to physical discomfort or health issues, such as eye strain or musculoskeletal problems.
- **Cybersickness and VR Fatigue:** Users can experience fatigue or discomfort due to the sensory disconnect between physical movements and virtual experiences.
- **Regulation and Standards:** Lack of standardized guidelines for VR usage and safety measures presents challenges in ensuring user safety and wellbeing. Addressing these challenges involves technological advancements, user education, ethical considerations, and the collaboration of various stakeholders to ensure the responsible and beneficial integration of VR technology into different aspects of our lives.[3]

C. VR in Future

Social and Communication:

- **Social VR Platforms:** Growth in social VR platforms that facilitate virtual gatherings, events, and interactions, bridging geographical distances.
- **Augmented Reality (AR) Integration:** Integration of VR with [1] AR technologies for mixed reality experiences that blend the virtual and physical worlds seamlessly.
- **Telepresence and Remote Travel:** Advancements in telepresence technologies allowing people to virtually visit distant locations or attend events without physical travel.

Ethical Considerations and Regulation:

- **Privacy and Security Measures:** Enhanced measures to address privacy concerns and ensure data security, especially with the collection of sensitive user information in VR.

- **Ethical Content Guidelines:** Development of clearer ethical guidelines and standards for creating and distributing VR content across different domains.
- **Health and Safety Standards:** Implementation of standardized health and safety protocols for VR usage to mitigate potential health risks and ensure user wellbeing.

VII. CONCLUSION

Although there are numerous positive impacts of VR technology in many fields, it also has some drawbacks as described below, which should be considered technology is still experimental and it is not entirely acceptable for everyone due to its high-cost types of equipment. The immersive feature of VR can be harmful in many cases, such as users getting more focused on virtual things rather than the real ones that are important in real-life [7]. It has adverse impacts on VR users who are trained in VR environments whereby they perform better in the virtual environment than in real-life situations.

VIII. REFERENCES

- [1] VRS (2016). History of Virtual Reality Available from: <http://www.vrs.org.uk/virtualreality/history.html>
- [2] Gaudiosi, J. (2016). Mercedes-Benz Drivers Two Virtual Reality Experiences. Available from: <http://fortune.com/2016/04/30/mercedes-benz-drives-two-virtual-reality-experiences/>
- [3] Seth A, Vance JM, Oliver JH (2011) Virtual Reality for assembly methods prototyping: a review. *Virtual Reality* 15(1):5-20
- [4] De Gauquier L, Brengman M, Willems K, van Kerrebroeck, H (2018). Leveraging advertising to a higher dimension: Experimental research on the impact of virtual reality on brand personality impressions. *Virtual Real* 1-19
- [5] Lau KW, Lee PY (2018). Shopping in Virtual Reality: a study on consumers' shopping experience in a stereoscopic virtual reality. *Virtual Real* 1-14.
- [6] Cardos, RA, David OA, David DO (2017) Virtual Reality exposure therapy in flight anxiety: a quantitative meta-analysis. *Comput Hum Behav* 72:37-380
- [7] ZhangJing. Application of Virtual Reality Technology in Advertising form and[J]. *Modern Combiner*, 2012(30): 273-273.

Enhancing Fake News Detection through Sentimental Analysis

Meghana Durga Patnala
23DSC04, M.Sc. (Computational Data Science)
Dept. Of Computer Science
P.B.Siddhartha College of Arts & Science
Vijayawada, A.P, India
meghanapatnala786@gmail.com

Gayathri Marrivada
23DSC10, M.Sc. (Computational Data Science)
Dept. Of Computer Science, P.B.
Siddhartha College of Arts & Science
Vijayawada, A.P, India
marrivadagayathri@gmail.com

Sri Lekha Ejnavarjala
23DSC10, M.Sc. (Computational Data Science)
Dept. Of Computer Science.
Siddhartha College of Arts & Science
Vijayawada, A.P, India
srilekha.ejnavarjala03@gmail.com

Abstract: In the contemporary landscape, social media has emerged as the primary conduit for global news dissemination. The proliferation of misinformation on these platforms has evolved into a critical worldwide concern, exerting detrimental effects on political, economic, and societal realms, thereby adversely impacting the lives of individuals. Fake news, often infused with negative sentiments, elicits responses from the public characterized by emotions such as surprise, fear, and disgust. This article focuses on mitigating the impact of fake news by employing a sophisticated approach. By conducting sentiment analysis on news articles and emotion analysis on user comments related to the news, we extracted pertinent features. These features, in conjunction with the content feature of the news itself, were inputted into a proposed bidirectional long short-term memory (LSTM) model for fake news detection. Leveraging the Faked it dataset, which comprises news titles and associated comments, we trained and tested our model. The outcomes were promising, showcasing the efficacy of our proposed model. With a high detection accuracy of 96.77% measured by the Area under the ROC Curve, our model surpassed the capabilities of other state-of-the-art studies. This underscores the significance of features derived from sentiment analysis of news, reflecting the publisher's standpoint, and emotion analysis of comments, representing the collective sentiment of the audience. These features play a pivotal role in enhancing the efficiency of the fake news detection model.

Keywords—Deep Learning, Fake news, social media, Sentimental analysis, Emotional analysis

I.INTRODUCTION

Fake news exerts a profound impact on individuals' daily lives, manipulating their thoughts and emotions, influencing beliefs, and potentially guiding them towards erroneous decisions. The pervasive spread of misinformation on social media carries detrimental

consequences for society across various domains, including politics, economics, social issues, health, technology, and sports.

A 2016 study revealed that 23% of individuals in the US shared fake news, whether intentionally or inadvertently. Additionally, fake news has a 70% higher likelihood of dissemination compared to genuine news, as per survey findings. Despite gender, age, or educational background, many people struggle to discern between fake and real news, highlighting a pervasive challenge.

Social media platforms serve as virtual spaces for posting, discussions, exchanging views, and global interactions without constraints of location, time, or content volume. In 2017, a survey indicated that 67% of US citizens primarily obtained their news from social media. Notably, in 2021, Facebook reported the closure of approximately 1.3 billion fake accounts and the removal of over 12 million posts containing false information about COVID-19 and vaccines.

The urgency to address and curb the global problem of fake news, causing social panic and economic turmoil, is evident in recent research. Despite the challenges in detecting fake news, the proliferation of misinformation necessitates continuous development and exploration of innovative research directions to enhance identification techniques.

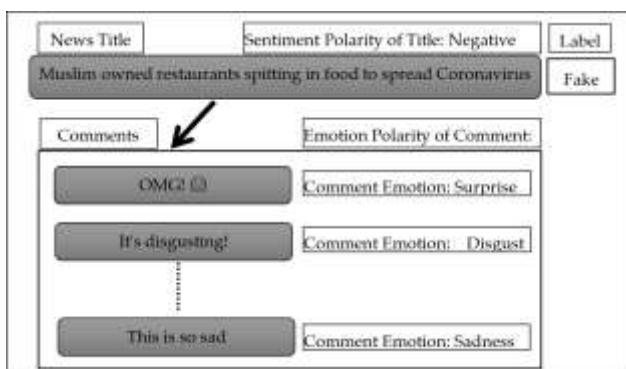
User responses to fake news often manifest emotions of fear, disgust, and surprise, contrasting with responses to real news characterized by anticipation, sadness, joy, and trust. The novelty factor significantly contributes to the spread and acceptance of fake news, capturing attention and stimulating idea exchange.

While sentiment analysis in existing studies often focuses on content sentiment signals from publishers, this research emphasizes the importance of emotion signals from user comments. Emotions, especially those expressed through emojis, play a crucial role in disseminating fake news on social media.

Deep learning techniques, particularly bidirectional long short-term memory (Bi-LSTM) models, offer substantial advancements in textual content classification, prediction,

and analysis. Bi-LSTM models, known for effectively capturing long-term dependencies and storing historical and future information, form the basis of the proposed model.

The research introduces a novel approach to detect fake news on social networks, leveraging features extracted from sentiment and emotion analysis, with special attention to emojis in comments. This approach aims to address contemporary challenges in identifying and combating the dissemination of misinformation, as exemplified by a real-life instance involving false accusations against the Muslim community in India during the early weeks of April 2020. This particular incident led to increased hostilities and



economic boycotts against the Muslim community.

Fig1. This sentiment analysis of fake news and emotion analysis of the public's comments about fake news.

Our contributions can be succinctly summarized as follows:

- Conducting a comprehensive analysis of sentiment in news titles and emotions in user comments within the Fakedit dataset, with a focus on understanding their correlation with the prevalence of fake news.
- Evaluating the efficacy of examining textual content in news headlines as a method for detecting fake news.
- Exploring the utility of sentiment-based features from news headlines and emotions derived from user comments in the identification of fake news.
- Introducing a novel model based on Bidirectional Long Short-Term Memory (Bi-LSTM) that leverages both textual features from news headlines and features extracted from sentiment analysis of news and emotion analysis of comments. The proposed model aims to achieve state-of-the-art results in the field of fake news detection.
- Presenting the practical application of our approach by demonstrating results on a real dataset, specifically the Fakeddit dataset.

Background of Study:

The term "Fake News" was officially declared the Collins Dictionary's Word of the Year for 2017 [14]. Fake news refers to the dissemination of false information in published news articles, deliberately crafted to mislead readers and serve malicious purposes [24] [20,21]. Given its global impact as a significant challenge and a threat to democracy, the economy, and societal harmony [17], various entities such as nongovernmental organizations, civil society organizations, journalists, politicians, and researchers have joined forces to mitigate its risks [5]. Notably, major technology companies like Facebook, Twitter, and Google have prioritized efforts to combat the proliferation of fake news, conducting extensive research in this domain [4]. According to a statistic from the Statista website (<https://www.statista.com/statistics/649221/fake-news-expose-responsible-usa/>, accessed on 4 March 2021), derived from a 2018 survey in the United States, social networks were implicated in the spread of fake news. The survey found that 29% of participants attributed primary responsibility to social media for spreading fake news, while 60% acknowledged these platforms as partially responsible, as illustrated in Figure 2

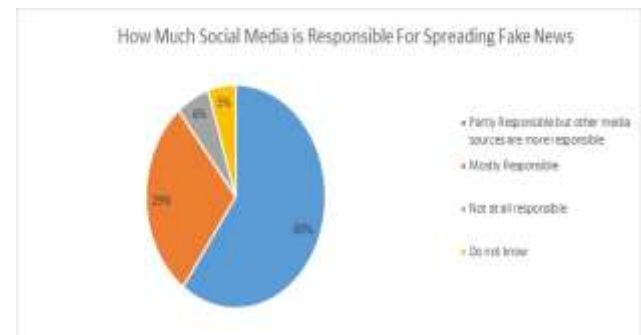


Fig. 2. How much social media responsible for spreading fake news

Sentiment and Emotional analysis in NLP

The Sentiment Analysis (SA) method finds widespread application across various domains, particularly within the realm of social media [7]. This includes tasks such as categorizing users' opinions on social media posts [18,19], discerning public sentiments during elections to predict outcomes, and influencing public opinion by gaining insights into people's attitudes through the analysis of opinions on specific situations [2].

Emotion analysis involves the categorization of data based on the emotions they convey, encompassing sentiments like joy, surprise, anger, fear, sadness, and disgust [13]. Typically, this analysis involves identifying specific words within a text, often present in emotional lexicons, that are indicative of particular emotions.

II RELATED WORK

Numerous studies have delved into the realm of fake news detection, employing sentiment analysis-based methodologies. These endeavors encompass various

strategies, including establishing connections between the sentiments expressed in posted news and the accuracy of the information, leveraging sentiment-based features to fortify their detection models, and introducing innovative approaches such as enriching datasets with sentiment as a pivotal feature.

These investigations have not only demonstrated the effectiveness of the proposed features but also emphasized their simplicity of integration into existing fake news detection models. Noteworthy methodologies include employing the emoCred strategy, utilizing a long-term memory (LSTM)-based model that accounts for emotional nuances to distinguish between genuine and fabricated claims. Additionally, an ensemble model, integrating convolutional neural networks (CNN) and Bidirectional Long Short-Term Memory (BI-LSTM) networks with an attention mechanism, has been proposed. Results from this ensemble model surpassed those of other employed models, with the observation that fake news tends to carry more stimulating emotions, whereas real news exhibits more subdued emotions. Despite the exploration of diverse algorithms in prior works related to fake news detection, their efficiency has been constrained for various reasons. In light of this, our paper proposes an improved approach by advocating for the application of sentiment analysis as a robust means to identify fake news

III PROPOSED WORK

The designed model for fake news detection comprises three distinct sub-units: an emotion analysis unit, a sentiment analysis unit, and a text classification unit. These units are seamlessly integrated through a concatenation layer, followed by a sigmoid layer for the final prediction, as illustrated in Figure 3. The implementation of our proposed model is based on Bidirectional Long Short-Term Memory (Bi-LSTM), and we conducted experiments using the python language on the google Collab platform (<https://colab.research.google.com/>, accessed on 11 March 2022). The model is organized into three subsections: 1) emotion analysis unit, 2) sentiment analysis unit, and 3) text classification unit.



Fig 3. The general design of the proposed model.

A. Emotion Analysis Unit:

The Emotion Analysis Unit focuses on leveraging emojis as a means of communication among users, considering them as a visual language that transcends cultural boundaries and has gained significant popularity [18]. To enhance the detection of fake news, we categorize emojis into three groups based on their relevance, as detailed in the Dataset Visualization section:

```

[ ] print('Found %s Emojis.' % len(Emoji_Dict))
    Found 2587 Emojis.

[ ] text="😱"
    convert_emojis_to_word(text)
    'face_screaming_in_fear'

[ ] text="😲"
    convert_emojis_to_word(text)
    'astonished_face'

[ ] text="I like to eat 🍕"
    convert_emojis_to_word(text)
    'I like to eat pizza'
  
```

Fig 4. The screenshot of code of replacing emojis with words.

Three groups will be created according to their relationship and contribution to the detection of fake news, as mentioned in the Dataset Visualization section, and the groups are:

- The novelty group contains the emotions of (fear, disgust, and surprise).
- The expectation group contains the emotions of (anticipation, sadness, joy, and trust).
- The neutral group, in which the emotions of the novelty group are equal to the emotions of the expectation group.

Since the model only deals with numeric data, we need to do data normalization for the emotion groups and convert the three resulting groups into numbers between 0 and 1, (0 = expectation, 0.5 = neutral, 1 = novelty). The generated column represents the emotion-based feature according to the groups proposed in this paper that will be combined with other features in the concatenate layer.

B. Sentiment Analysis Unit:

The Sentiment Analysis (SA) unit is dedicated to assessing the sentiment conveyed in news titles, employing a lexicon-based sentiment analyzer, Text Blob. Prior to sentiment analysis, the text undergoes several pre-processing techniques, including the removal of empty lines and tabs, stripping HTML tags, eliminating hyperlinks, removing accented characters, expanding contractions, discarding special characters (excluding! and?), eliminating stop words, and trimming whitespaces.

It's crucial to emphasize that these pre-processing steps are consistently applied to the text entering both the sentiment analysis unit and the news classification unit, as depicted in Figure 3. Text Blob, a Python-based Natural Language Processing (NLP) library leveraging the Natural Language Toolkit (NLTK), operates by taking a sentence as text input, typically represented as a collection of words. The sentiment analysis involves scoring each word individually, and the overall sentiment is determined through a pooling procedure, averaging all individual sentiments. The sentiment analysis outcomes of the news titles serve as a significant feature, contributing to the model's ability to detect fake news when combined with features from other units in the concatenate layer.

C. Text Classification Unit:

The Bi-LSTM classifier comprises two layers of long short-term memory (LSTM), as depicted in Figure 5: the forward LSTM layer and the backward LSTM layer. These layers operate collaboratively to aggregate extensive contextual information from both the front and back directions over a specific timeframe [4]. The architecture of the Bidirectional LSTM (BLSTM) is designed to capture a comprehensive array of noteworthy features from both directions, enhancing the model's ability to understand sequential patterns [8]. The forward layer focuses on learning the sequence of input data, processing information from left to right. The hidden state of this forward LSTM layer can be mathematically represented by the following formula:

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}) \quad (1)$$

The backward layer of the Bi-LSTM classifier is responsible for learning the reverse sequence of the input data. In other words, information is processed by the backward LSTM from right to left. The hidden state of this backward LSTM layer can be expressed using the following formula:

$$\overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \quad (2)$$

The two layers of LSTMs are linked to a single output layer, as depicted in Formula (3). They simultaneously traverse the input sequence from two distinct directions.

$$h = [\vec{h}_t, \overleftarrow{h}_t]$$

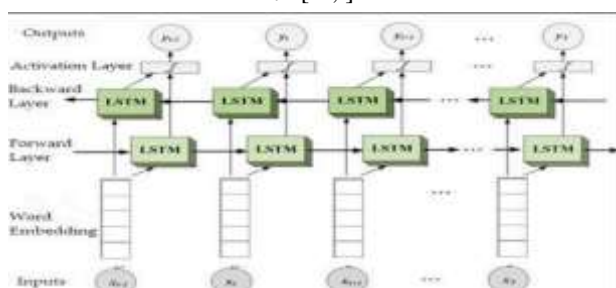


Fig 5. The basic architecture of Bi-LSTM uses word embedding.

To mitigate overfitting and enhance the generalization of our proposed model to new data [21,22], ultimately improving detection accuracy [19], we employed several techniques. These include streamlining the complexity of the model design [21,22], implementing weight regularization through L1 and L2 regularizers—where L1 represents the sum of absolute weights and L2 signifies the sum of squared weights [19,21]. Additionally, a dropout layer was introduced to prevent overfitting [19,21]. Each news article is treated uniquely based on its characteristics, extracted through sentiment and emotion analysis, and identified by its unique ID. Leveraging the Bi-LSTM model and the concatenation layer, various extracted features are amalgamated with text features derived from the Bi-LSTM model. The sigmoid layer is then employed to discern fake news. This section encompasses two subsections: (i) a concatenation layer and (ii) a dense layer.

Table 1. The architecture of the proposed fake news detection model

No.	Structure or Hyper parameter Name	Type or Value
1	Layers	Embedding layer Bidirectional LSTM layer Dense layer
2	Word embedding dimension	300
3	No. of hidden states	128
4	Dropout	0.2
5	Recurrent dropout	0.2
6	Regularize L1	0.0001
7	Regularizer L2	0.001
8	Activation function	Sigmoid
9	Loss function	Binary_crossentropy
10	Optimizer	Adam
11	Learning rate	0.1
12	Batch size	256
13	No. of epochs	5

I. Concatenation Layer:

In the Concatenation Layer, the features derived from emotions and sentiment, obtained from the preceding two units, are merged with the output of the Bi-LSTM layer within the text classification unit. This combined output is then forwarded to the subsequent layer, the sigmoid layer, within our proposed model, as depicted in Figure 6.

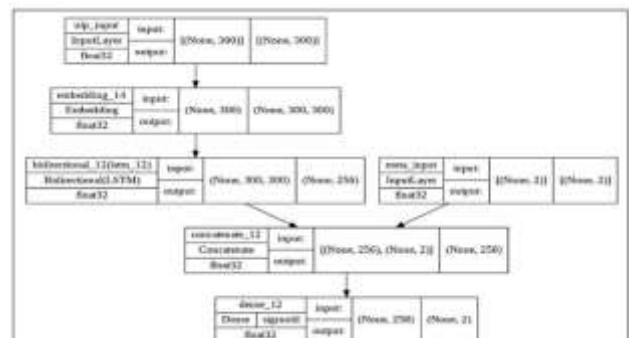


Fig 6. The concatenation layer in the proposed model.

ii. Dense Layer:

Subsequently, the output from the Concatenation Layer undergoes processing in the dense layer utilizing a sigmoid activation function. This step yields the ultimate output of our proposed model, determining the classification of news as either genuine or fake. The sigmoid function employed in this process generates output values within the range of 0 to 1, and its mathematical representation is defined by Equation (4).

$$\Phi(x) = 1 / (1 + e^{-x})$$

IV RESULTS AND ANALYSIS

Given the imbalanced nature of our dataset, we opted for the area under the curve (AUC) as a performance measure to evaluate the effectiveness of the proposed models. The AUC, representing the area under the ROC curve, is a crucial metric for comparing learning algorithms and constructing optimal learning models. Particularly in imbalanced classification tasks, such as fake news detection where the distribution of ground-truth fake news titles and real news titles is significantly skewed, AUC outperforms accuracy due to its statistical consistency and discriminatory power [13]. Thus, we employed the AUC measure, defined in Equation (5), alongside the F1-score metric presented in Equation (6), and accuracy to comprehensively assess the performance of our proposed models.

$$AUC = \frac{1 - FPR + TPR}{2}$$

The True Positive Rate (TPR) represents the percentage of positive examples accurately classified, whereas the False Positive Rate (FPR) signifies the ratio of instances incorrectly classified as negative to all other instances [9]. The F1-score, as depicted in Formula 6, provides a measure of performance by combining precision and recall, offering a comprehensive assessment of the model's ability to correctly classify positive instances while minimizing false positives.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

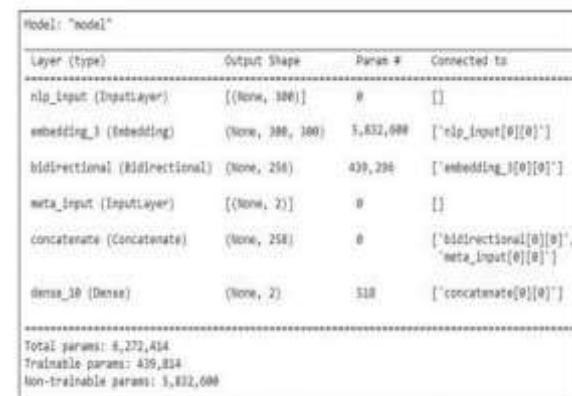
The F1-score, derived from the harmonic mean of precision and recall, gauges the balance between precision and recall. Precision for a class of predictions is the ratio of correctly classified positive cases to all instances predicted as positive. Recall, on the other hand, assesses the proportion of correctly classified positive cases out of all actual positive instances. Accuracy, in a broader sense, is the percentage of instances correctly classified, encompassing both positive and negative cases, among all instances.

In the pursuit of developing a high-accuracy model for detecting fake news, we delved into various factors influencing model performance. These factors encompassed the selection of features, choice of models, tuning hyperparameters, and refining the model structure. We conducted comprehensive tests on several powerful deep learning models, including LSTM, gated recurrent unit (GRU), Bi-LSTM, and CNN. Each model underwent two stages of evaluation. Initially, models were solely tested on the text features of news titles, while in the subsequent stage, they were applied to assess both the sentiments of titles and the emotions of comments derived from sentiment analysis and emotion analysis units. Additionally, the effectiveness of text-based features of news titles was evaluated. It's crucial to note that all proposed models adopted the most effective structures and hyperparameters derived from extensive experimentation to achieve optimal detection accuracy and model performance. Table 2 provides the results of the proposed detection models, categorized by the examined features, measured through AUC and F1-score metrics. These results were derived from experiments conducted on validation sets.

Table.2. The results of the proposed models according to the features.

Model	Textual content features(News titles)	Features Based on Titles Sentiment	Features Based on Comments Emotions	AUC	F1-Score
LSTM	✓			89.99%	90.98%
LSTM	✓	✓	✓	90.16%	91.78%
GRU	✓			91.65%	92.23%
GRU	✓	✓	✓	92.60%	94.09%
CNN	✓			94.14%	96.39%
CNN	✓	✓	✓	96.05%	97.76%
Bi-LSTM	✓			94.65%	95.54%
Bi-LSTM	✓	✓	✓	96.77%	97.81%

LSTM model, the AUC values during training and validation, as well as the corresponding training and validation loss.



```

Model: "model"
Layer (type)                Output Shape              Param #                    Connected to
-----
nlp_input (InputLayer)      [(None, 100)]            0                          []
embedding_1 (Embedding)     (None, 100, 100)         3,832,000                  ["nlp_input[0][0]"]
bidirectional (Bidirectional) (None, 256)              439,296                    ["embedding_1[0][0]"]
meta_input (InputLayer)     [(None, 2)]              0                          []
concatenate (Concatenate)   (None, 258)              0                          ["bidirectional[0][0]", "meta_input[0][0]"]
dense_10 (Dense)            (None, 2)                318                        ["concatenate[0][0]"]
-----
Total params: 4,272,414
Trainable params: 439,814
Non-trainable params: 3,832,600
    
```

Fig 7. The configuration of the Bi-LSTM.

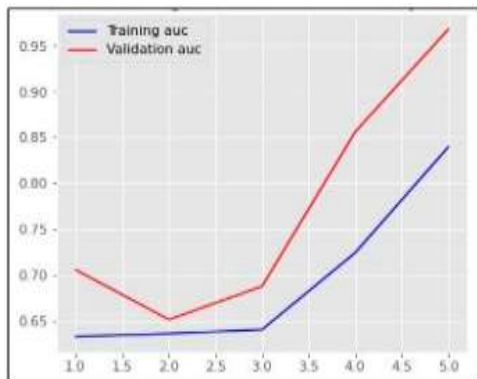


Fig 8. The training and validation AUC of the proposed Bi-LSTM model

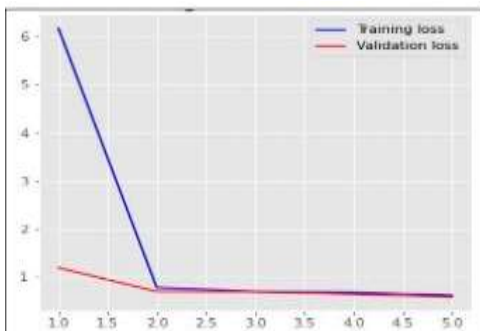


Fig 9. The training and validation loss of the proposed Bi-LSTM

Study	Model	Textual Content	Visual Content	Social-Based Features	Metadata-Based Features	Comments-Based Features	Emotions-Based Features	Sentiments-Based Features
Nakamura, Levy [74]	BERT + ResNet50	✓	✓	✓	✓	✓	✓	✓
Kalyan, Kumar [77]	DeepNet	✓	✓	✓	✓	✓	✓	✓
Kirchoepf, Slipek [78]	Multimodal architecture (BERT + CNN + MLP)	✓	✓	✓	✓	✓	✓	✓
Xie, Liu [79]	SERN model based on (BERT + ResNet + MLP)	✓	✓	✓	✓	✓	✓	✓
Raza and Ding [80]	FND-NS model based on BART	✓	✓	✓	✓	✓	✓	✓
Our Proposed Model	Bi-LSTM model	✓	✓	✓	✓	✓	✓	✓

Table 3 illustrates the results of our proposed model based on Bi-LSTM in comparison with models presented in relevant studies, using several measures such as AUC, accuracy, and F1-score.

Table 4. The results of the detection models for the benchmarking dataset.

Model	AUC	Accuracy	F1-Score
BERT+ResNet50	-	86.54%	-
DeepNet	-	86.4%	87.2%
Multimodal architecture(BERT+CNN+MLP+SERN Model)	-	94.4%	-
Based on (BERT+ResNet+MLP)	-	96.63%	96.63%
FND-NS model based on BART	70.4%	74.8%	74.9%
Our proposed model	96.77%	96.89%	97.81%

V CONCLUSION

The proliferation of fake news on social networks stands as a significant contemporary challenge, carrying adverse consequences for society and individual lives. This paper contributes to the ongoing efforts to combat the dissemination of fake news, with a predominant focus on textual content features. While many studies have explored features based on social networks or user behavior, others have adopted a multimodal approach incorporating both textual and image features for fake news detection. Notably, our research takes a unique approach by considering public attitudes towards news, examining the emotions expressed in user comments. Given that a majority of comments towards fake news convey emotions, and fake news often embodies negative sentiment and reflects the publisher's stance, we conducted sentiment and emotion analysis as features for fake news detection. Experimental results demonstrate that the inclusion of sentiment-based and emotion-based features significantly enhances the accuracy of fake news detection across all proposed deep learning models, surpassing the efficacy of text-based features alone. Our proposed Bi-LSTM model outperforms related works utilizing benchmark datasets. The findings suggest that features derived from sentiment analysis of news and emotion analysis of user comments can be valuable tools for social media platforms in mitigating the spread of fake news.

Addressing the challenge of an imbalanced dataset in future work could be accomplished through the application of Generative Adversarial Network (GAN) techniques. Additionally, we aim to further improve accuracy by exploring the integration of other state-of-the-art models, particularly those based on transformer architectures.

VI REFERENCES

- [1] Shrivastava, G.; Kumar, P.; Ojha, R.P.; Srivastava, P.K.; Mohan, S.; Srivastava, G. Defensive modeling of fake news through online social networks. *IEEE Trans. Comput. Soc. Syst.* 2020, 7, 1159–1167.
- [2] Li, J.; Kao, H.-Y. HAT4RD: Hierarchical Adversarial Training for Rumor Detection in Social Media. *Sensors* **2022**, 22, 6652.
- [3] Xu, K.; Wang, F.; Wang, H.; Yang, B. Detecting fake news over online social media via domain reputations and content understanding. *Tsinghua Sci. Technol.* 2019, 25, 20–27.
- [4] Kumar, S.; Asthana, R.; Upadhyay, S.; Upreti, N.; Akbar, M. Fake news detection using deep learning models: A novel approach. *Trans. Emerg. Telecommun. Technol.* 2020, 31, e3767.

- [5] Habib, A.; Asghar, M.Z.; Khan, A.; Habib, A.; Khan, A. False information detection in online content and its role in decision making: A systematic literature review. *Soc. Netw. Anal. Min.* 2019, 9, 50.
- [6] Rath, B.; Gao, W.; Ma, J.; Srivastava, J. Utilizing computational trust to identify rumor spreaders on Twitter. *Soc. Netw. Anal. Min.* 2018, 8, 64.
- [7] Xarhoulacos, C.-G.; Anagnostopoulou, A.; Stergiopoulos, G.; Gritzalis, D. Misinformation vs. Situational Awareness: The Art of Deception and the Need for Cross-Domain Detection. *Sensors* 2021, 21, 5496.
- [8] Ahmad, I.; Yousaf, M.; Yousaf, S.; Ahmad, M.O. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity* 2020, 2020, 8885861.
- [9] Umer, M.; Imtiaz, Z.; Ullah, S.; Mehmood, A.; Choi, G.S.; On, B.-W. Fake news stance detection using deep learning architecture (cnn-1stm). *IEEE Access* 2020, 8, 156695–156706.
- [10] Atodiresei, C.-S.; Tănăselea, A.; Iftene, A. Identifying fake news and fake users on Twitter. *ProcediaComput. Sci.* 2018, 126, 451–461.
- [11] Liang, X.; Straub, J. Deceptive Online Content Detection Using Only Message Characteristics and a Machine Learning Trained Expert System. *Sensors* 2021, 21, 7083.
- [12] Pathuri, S.K.; Anbazhagan, N.; Joshi, G.P.; You, J. Feature-Based Sentimental Analysis on Public Attention towards COVID-19 Using CUDA-SADBM Classification Model. *Sensors* 2021, 22, 80.
- [13] Eke, C.I.; Norman, A.A.; Shuib, L.; Nweke, H.F. Sarcasm identification in textual data: Systematic review, research challenges and open directions. *Artif. Intell. Rev.* 2020, 53, 4215–4258.
- [14] Liu, Y.; Wu, Y.-F.B. Fned: A deep network for fake news early detection on social media. *ACM Trans. Inf. Syst. (TOIS)* 2020, 38, 1–33.
- [15] Lin, L.; Chen, Z. Social rumor detection based on multilayer transformer encoding blocks. *Concurr. Comput. Pract. Exp.* 2021, 33, e6083.
- [16] Goksu, M.; Cavus, N. Fake news detection on social networks with artificial intelligence tools: Systematic literature review. In *Proceedings of the 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions-ICSCCW-2019*, Prague, Czech Republic, 27–28 August 2019; pp. 47–53.
- [17] Ali, A.M.; Ghaleb, F.A.; Al-Rimy, B.A.S.; Alsolami, F.J.; Khan, A.I. Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique. *Sensors* 2022, 22, 6970.
- [18] de Souza, J.V.; Gomes, J., Jr.; de Souza Filho, F.M.; de Oliveira Julio, A.M.; de Souza, J.F. A systematic mapping on automatic classification of fake news in social media. *Soc. Netw. Anal. Min.* 2020, 10, 48.
- [19] Guo, M.; Xu, Z.; Liu, L.; Guo, M.; Zhang, Y. An Adaptive Deep Transfer Learning Model for Rumor Detection without Sufficient Identified Rumors. *Math. Probl. Eng.* 2020, 2020, 7562567.
- [20] Varshney, D.; Vishwakarma, D.K. Vishwakarma, Hoax news-inspector: A real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles. *J. Ambient Intell. Humaniz. Comput.* 2020, 12, 8961–8974.
- [21] Kim, Y.; Kim, H.K.; Kim, H.; Hong, J.B. Do Many Models Make Light Work? Evaluating Ensemble Solutions for Improved Rumor Detection. *IEEE Access* 2020, 8, 150709–150724.
- [22] Yaakub, M.R.; Latiffi, M.I.A.; Zaabar, L.S. A review on sentiment analysis techniques and applications. *IOP Conf. Ser. Mater. Sci. Eng.* 2019, 551, 012070.
- [23] Santhoshkumar, S.; Babu, L.D. Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. *Soc. Netw. Anal. Min.* 2020, 10, 20.
- [24] Tian, L.; Zhang, X.; Wang, Y.; Liu, H. Early detection of rumours on twitter via stance transfer learning. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, 14–17 April 2020, Proceedings, Part I* 42; Springer: Cham, Switzerland, 2020; Volume 12035, p. 575.

Tabular-to-Image Transformations for the Classification of Anonymous Network Traffic Using Deep Residual Networks

P. Bhargavi
23DSC05, M.Sc. (Computational
Data Science)
Dept. of Computer Science P.B.
Siddhartha College of
Arts & Science
Vijayawada, A.P, India
bhargavibharu741@gmail.com

S. Bhanu Sri
23DSC32, M.Sc. (Computational
Data Science)
Dept. of Computer Science P.B.
Siddhartha College of Arts &
Science
Vijayawada, A.P, India
bhanusrisingavarapu@gmail.com

G. Keerthi
23DSC25, M.Sc. (Computational
Data Science)
Dept. of Computer Science P.B.
Siddhartha College of
Arts & Science
Vijayawada, A.P, India
gajjalakondakeerthi@gmail.com

Abstract: With the meteoric rise in anonymous network traffic data, there is a considerable need for effective automation in traffic identification tasks. Though many shallow and deep machine learning network traffic classification solutions have been proposed, they often rely on tabular data, making them unable to detect complex spatial relationships. However, recent advancements in computer processing power have increased the viability of transforming tabular data into images for training deep convolutional neural networks, transforming structured data problems into spatial ones. To identify the most effective methods for representing tabular anonymous network traffic data as images, we compared five deep learning classifiers trained on data from different tabular-to-image algorithms—Image Generator for Tabular Data (IGTD), Deep Insight, vector-of-feature wrapping (normalized and non-normalized), and our newly introduced Binary Image Encoding (BIE) technique in the classification of eight network application types. Furthermore, we examine whether deep residual models trained on tabular-to-image data can outperform the top-performing shallow learner, XGBoost, at classifying anonymous network traffic. We found that ResNet-50, a pre-trained instance of deep residual network, trained on image datasets using IGTD and the novel Binary Image Encoding outperformed XGBoost trained on tabular data. Our ResNet-50 models trained using IGTD and BIE achieved F1-scores of 96.0% and 98.49% respectively, improving on the baseline of 95.1% achieved by XGBoost.

INDEX TERMS: *Tabular-to-image techniques, binary image encoding, convolutional neural networks, network traffic, anonymous traffic, deep learning, XGBoost, image generator for tabular data.*

I INTRODUCTION

Network traffic classification is crucial for improving network management and security [1]. For instance,

real-time applications like video streaming may require lower latency for a better user experience than web browsing [2]. Classifying this traffic can enable better optimization by internet service providers (ISPs) to prioritize real-time applications with more suitable network nodes [1]. Moreover, classifying network traffic may aid in malicious network traffic interception, which local governments typically mandate [3], [4]. Automation of this task has garnered greater interest as the scale of network traffic increases and new threats to network security reveal themselves. Using the information in each packet that was transmitted or through a collection of packets and their metadata, called a flow, ISPs can classify traffic based on the application that produced it and optimize their infrastructure to scale to the evolving needs of their customers [6]. Due to the emergence of a suite of encryption and anonymization technologies such as Secure Shell Protocol (SSH), Hypertext Transfer Protocol (HTTPS), The Onion Router (TOR), and Virtual Private Networks (VPNs), it can be difficult to rely on conventional techniques to discover the origin of the anonymous and encrypted traffic [7]. To solve this problem, machine learning algorithms have been successfully employed to classify the applications producing network traffic [8], [9], [10]. storage by over 90% and minimizing computational overhead.

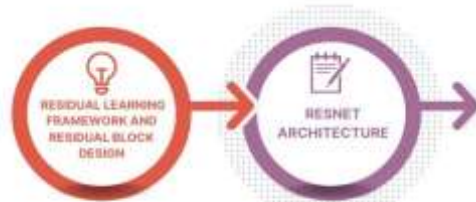
NETWORK TRAFFIC CLASSIFICATION WITH CNN

CNN-based deep learning techniques are becoming a more popular approach to network traffic classification problems as the rapid growth of computation power enables quicker model training. Lashkari et al. [9] introduced Deep Image, a tabular-to-image pipeline for detecting and classifying Darknet traffic using CIC-DN dataset. Deep Image synthesizes gray-scaled images composed of the most important features from the dataset. A custom CNN was trained on the images to detect and characterize Darknet traffic with an accuracy of 86% when classifying among eight application types.

TABULAR TO IMAGE ALGORITHMS In order to leverage the strengths of CNNs and improve classifier accuracy on tabular data, Tabular-to-Image (T2I) algorithms were introduced in previous works. SuperTML works by arranging feature values onto a 2D image. Features of greater importance are projected with larger font sizes. Moreover, SuperTML reduces the need for data preprocessing as missing values are projected as '?', and non-numeric values are placed on the image without the need for encoding. They tested SuperTML data with a pre-trained CNN and compared it with XGBoost on three separate datasets. SuperTML performed equally

Traditional Deep Learning Methods

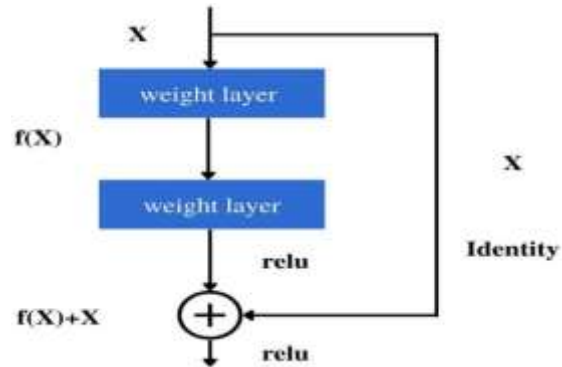
Image Recognition



well or out-performed XGBoost in each test. Buturović and Miljković developed a method for classifying tabular data with CNNs through an image-generating algorithm called Tabular Convolution (TAC). This method treats input vectors as kernels and converts the data into an image using convolutions of a fixed base image. Features are converted to kernels by creating a square matrix with an odd number of rows and columns. If the number of features is not the square root of an odd 220 number, the square is either padded or trimmed towards the nearest odd square. TAC was applied to gene expression data and trained using ResNet, and results were compared to shallow classifiers (XGBoost, Light GBM, and Support Vector Machines). TAC outperformed all shallow learning methods obtaining an accuracy of 91.1% compared to the highest performing shallow learner's accuracy of 89.6%. This result was obtained using several thousand epochs of training; whereas a similar performance to non-CNN methods was seen when using 50 epochs. They conclude that the additional computation time required for TAC is negligible on modern computer architecture. Table 1 compares the research of prior literature that explored the CIC-DN dataset, T2I techniques, or similar classifiers.

II MOTIVATION AND PURPOSE

From the literature review, we find that insufficient research on T2I methods and the application of these methods for anonymous network traffic classification have left the following gaps in knowledge:



DATASET: In this work, we use the CIC-DN dataset balanced with synthetic SMOTE data from CMU-SynTraffic-2022 (CMU) dataset. This data was then used in conjunction with the five T2I algorithms to create our CMU-SynTraffic2023-Image Dataset (CMU-I). We explain these datasets further in the next subsections.

CIC-Darknet2020

Lashkari et al. [9] amalgamated the CIC-Darknet2020 (CIC-DN) dataset by combining their ISCX-Tor2016 and ISCX-VPN2016 datasets. The CIC-DN dataset is provided in both raw Packet-Capture (PCAP) files as well as tabular data files that were preprocessed over a fixed time interval using CIC-Flow Meter v4.0. The tabular data samples consist of time-based features such as flow duration as well as statistical features which makes them highly representative of traffic flows. The dataset consists of eight anonymous traffic application types and contains 117,620 samples encompassing both Tor and VPN traffic. This dataset was chosen for our experiments due to its comprehensive selection of application types, having an adequate number of samples for training, and being well researched in prior works [9].

The eight application types comprising CIC-DN are audio streaming (Vimeo and YouTube), browsing (Firefox and Chrome), chat (ICQ, AIM, Skype, Facebook, and Hangouts), email (SMTPS, POP3S, and IMAPS), file transfer (Skype, FTP over SSH (SFTP) and FTP over SSL (FTPS) using FileZilla and an external service), p2p (uTorrent and Transmission), video streaming (Vimeo and YouTube), and VoIP (Facebook, Skype, and Hangouts voice calls). Class imbalance was an apparent problem with the original dataset as 47% of the samples are p2p traffic while VoIP and email samples consist of less than 1% of the total data. We address this

limitation with the synthetic data generation scheme discussed in the following section.

DATA BALANCING AND CLEANING Since models trained on imbalanced datasets often perform poorly in real-world deployment, it was necessary to balance the CIC-DN data to improve model in this work, we utilize the real network traffic data from the CIC-DN dataset up sampled with synthetic SMOTE data from the CMU dataset as the baselinetabular dataset for our experiments.

From our tabular dataset, 14 zero-valued features and six additional features—Flow-id, Source/Destination IP, Times- tamp, and Source/Destination port—were removed as they either overfit the model or contained duplicate information. This process left 64 features in the resulting dataset. After removing samples containing Nan and Inf values and up-sampling minority classes, our final training data contained 240,000 samples with 30,000 samples in each class. We used an 80/20 train-test split for the training of our models. This means that 80% of our data was used for training while the remaining 20% was used for testing the models and calculating our performance metrics.

TABULAR-TO-IMAGE DATASETS

We employ each of the five T2I techniques to transform all 240,000 samples into corresponding images. The images are grouped into folders based on their application type. The dataset dubbed CMU-SynTraffic2023-ImageDataset (CMU-I) is published online for further scrutiny. Section VI-C provides detailed insight into the generated images while providing visualizations of selected samples.

CLASSIFIERS: This section briefly discusses the XGBoost and ResNet-50 classifiers evaluated in our experiments as well as our reasoning for selecting these specific classifiers

XGBoost

XGBoost is an optimized distributed gradient boosting classifier that has become the tool of choice in many machine-learning applications due to its high performance XGBoost is built upon gradient boosting, which is an ensemble technique that combines the output of multiple weaker machine learning models to produce a more accurate prediction XGBoost provides L1 and L2 regularization to tune and further reduce overfitting and reduce loss; mainly it enables users to tune various hyperparameters to constrain trees, makes adjustments in the learning rate during the learning process, and provides random sampling techniques XGBoost was chosen as the baseline classifier as it outperformed all other classifiers in previous similar experiments and is a top-performing classifier over a wide range of experiments which were $\alpha = 0$ and $\lambda = 1$ respectively.

ResNet

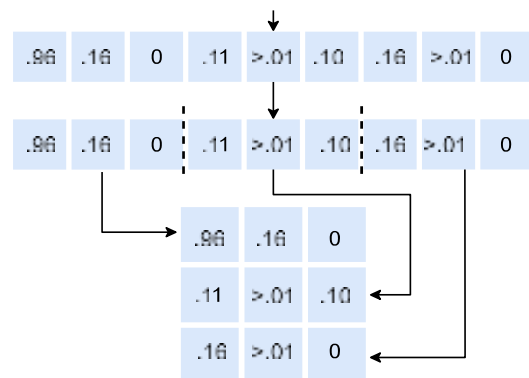
CNNs have produced high accuracy in image classification and increasing the depth of these networks results in improved classification accuracy. However, deeper neural networks are more difficult to train because simply stacking more layers onto a network introduces the problem of vanishing/exploding gradients as layers are stacked, the partial derivative of the loss function will either approach zero or ResNet is a deep residual neural network introduced to mitigate this drawback through residual learning. The residual aspect of ResNet allows for its enhancement over other CNNs because it can create a network with more depth. In a simple deep network, the output from each convolutional layer is passed directly as the input to the next layer which causes vanishing/exploding gradients. In comparison, ResNet introduces residual connections that enable the network to skip one or more layers. These connections allow information to directly propagate to all layers of the networks. As a result, ResNet models have fewer filters and lower complexity than other neural networks, such as VGG

We apply transfer learning to our ResNet models in order to improve performance. Transfer learning involves taking a pre-trained model, removing its output layer, and adding additional layers to be trained for a more specific task (in this case anonymous traffic classification). The advantage of transfer learning is that it can make use of the information learned from the previous, more general training to enhance performance on the new

Our goal is to compare the top performing shallow model trained on tabular data (XGBoost) with ResNet trained on T2I data. We also provide empirical evidence on whether T2I techniques are a viable approach to multi-class classification of various application types in anonymous traffic detection and categorization problems

Min-Max Normalization

FIGURE 1. Feature wrapping visualization.



873	151	0	101	7.2	90.9	147	.18	0
-----	-----	---	-----	-----	------	-----	-----	---

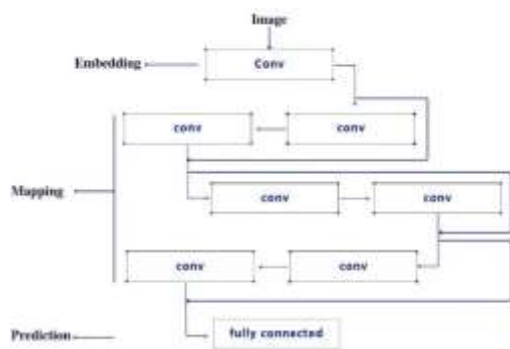
FEATURE WRAPPING

Ko et al introduced a method of converting raw network traffic data into images for classification by wrapping the binary data of a traffic flow into square images. A similar technique was then adapted to transform tabular data into images as a vector of feature wrapping the vector of feature wrapping technique takes a one-dimensional tabular data sample and normalizes its values before creating a square 2D image. We employ min-max normalization (1) for the feature wrapping technique used in our experiments The basic building block of a ResNet.

TABULAR-TO-IMAGE ARCHITECTURES

CNNs are effective at analyzing data with spatial differences between features. This makes them ideal for application on image and audio datasets where the important information about the data is based on the order of the features in these cases, the data is homogeneous which allows for

Since categorical features were discarded in our dataset, we did not employ the one-hot encoding technique. After normalization, each sample is then split into equal-length sub- vectors which are stacked on top of each other to form a square image. If a sample doesn't have enough values to form a square



image, it can be padded with additional zero This feature-wrapping method is illustrated in Figure 1 CNNs to distinguish spatial differences. However, when it comes to heterogeneous data, such as tabular data, CNNs cannot be directly applied. This limitation inspires the process of transforming tabular data into images to apply CNNs

Applying CNNs to the transformed data can result in superior prediction performance compared to other shallow models trained directly on tabular data. The potential for performance improvement motivates research on the most effective method of converting tabular data to images by evaluating new and pre-existing T2I algorithms. The development of Deep Insight pioneered this transformation of data to images, followed by more effective algorithms like

Super TML, TAC, and IGTD that improve the process.

Sections V-A through V-D explain the T2I algorithms we use in our experiments. The performance of the CNNs was then compared to traditional machine learning methods. The Deep Insight system achieved the highest accuracy metrics across each dataset, with an accuracy of 95% on average.

IGTD

Zhu et al. introduced Image Generator for Tabular Data (IGTD) to improve existing image generation techniques. The optimization algorithm converts tabular data to images by assigning each feature to a pixel. The assignment is determined by ranking the pairwise distances between features and the pairwise distances between the assigned pixels. The algorithm then minimizes the difference between these two measurements. Pairwise distances are calculated through a distance measure such as the Euclidean distance or the Pearson Correlation Coefficient. This assigns similar features to pixels close to one another and dissimilar features to ones farther apart. The efficiency of this method stems from a greedy iterative process of swapping the pixel assignments of features to best reduce the distance between them. Unlike Deep Insight, IGTD produces dense image representations where each pixel represents a unique feature. This results in smaller images that take less time when training CNNs. IGTD also does not require domain knowledge and has excellent feature preservation as closer features are more similar. The size and shape of the image generated can be adjusted, which makes it more applicable to a variety of domains. They compared IGTD to CNNs trained with Deep Insight and REFINED images on datasets for gene expression profiles of cancer cell lines and molecular descriptors of drugs. CNNs trained on IGTD provided similar or better prediction performance when compared to the other T2I methods and models trained on the original tabular data. Despite its origination in a different domain, we wanted to examine IGTD's applicability in the network traffic domain.

BINARY IMAGE ENCODING

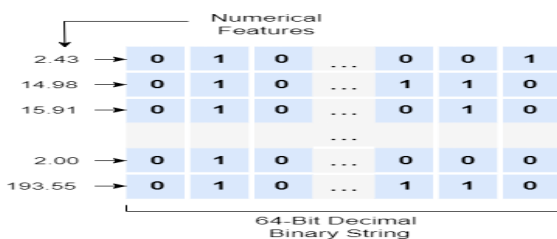
Inspired by the one-hot-encoding technique, we introduce Binary Image Encoding (BIE), a novel T2I scheme. The one-hot encoding method was originally introduced by Wang et al. and involved converting binary network flow data into a 2D image by applying one-hot encoding on each byte of the sample. The reasoning for this process is that raw network data often does not have an ordering and its values are better represented as categorical features. This is because the information in raw network data are features like protocol types or flags rather than

meaningful numerical values. Instead of treating features as unordered categorical values, BIE makes use of the structure of binary representations of floating point values. Fig. 2 outlines a binary encoded double as well as the conversion to a decimal representation. Double precision binary numbers consist of a sign bit, the exponent which dictates the magnitude of the number, and the mantissa which represents the significant digits of the value. The technique converts each numerical sample value into a double precision binary string as discussed above. Double precision floating point value.

The binary values are then stacked on top of each other to create a two-dimensional matrix to be interpreted as an image where zero values become black pixels and one values become white pixels. This process is illustrated in Fig. 3, which depicts numerical feature values being converted into 64-bit binary strings and then being situated on top of each other to form the full BIE image. The pseudocode for converting an input sample into an image is provided in Fig. 4.

FIGURE 5. Binary image encoding process

We believe that representing feature values as vectors of binary encoded floating point numbers could have many benefits for network traffic classification. First, this method does not rely on normalization as many of the previously discussed T2I techniques do. This is



advantageous because normalization reduces the range of potential feature values and can also be heavily affected by outliers. Secondly, a binary decimal string isolates the magnitude of a value (exponent) from the precise value of each digit (mantissa). When differentiating network traffic flows, one important factor to consider is the magnitude of the packets exchanged during the flow. For example, video streaming applications will have thousands of packets exchanged in a short time, whereas an email may typically have a lot fewer. To this end, isolating the mag

FIGURE 4. BIE pseudocode.

of the value in the image representation may make a classify based on packet quantity in a flow easier. Finally, the method expands the information of each value by partitioning it into meaningful parts as opposed to IGTD, where each pixel corresponds to a given feature, limiting image's information by the number of available features.

III RESEARCH METHODOLOGY

Fig. 5 outlines our research methodology. Subsequent sections present the experimental outline, processes for collecting metrics, tabular-to-image algorithmic conversion processes, and model training.

EXPERIMENTAL OUTLINE

First, we establish baseline results by training the shallow learning XGBoost classifier on CMU dataset. The XGBoost classifier was trained to distinguish among eight application types and its performance metrics were recorded. Next, we generate five (5) new image datasets using each of the T2I algorithms which are then used to train five ResNet-50 models. After collecting performance metrics on these models, we compared their performance to XGBoost with an eye toward providing empirical evidence of the importance of T2I techniques.

PERFORMANCE METRICS

F1-score is the harmonic mean of precision and recall where precision is the proportion of correctly classified positive classifications and recall is the percent of TPs that a model predicted accurately. ROC curves are a visual representation of the TP rate in relation to the FP rate. The area under the curve (AUC) is calculated by finding the total area under a ROC curve. MAE is simply the average of the absolute differences between the model's predicted values and actual values, whereas MSE is the squared differences between the predicted and actual values. MAE reflects the overall error giving equal consideration to all data samples. In contrast, MSE is more affected by outliers so a larger MSE can indicate that there are large outliers potentially from class confusion. Our loss function is categorical cross entropy, which is a standard loss function for measuring the general fit for multi-class models. F1-score and AUC are less susceptible to imbalanced data and can help determine whether a model is overfitting to training data. We primarily use F1-score to compare model performance. Tabular-to-image encoding time (in seconds) is also reported as it can be an important metric to consider when performing real-time anonymous traffic detection and continuous model learning.

had minimal effect on model performance; however, conversion time reduces considerably when

generating smaller images. All T2I algorithms were given the same input dataset and generated 240,000 corresponding images. The image samples shown in Figures 6, 7, 8, 9, and 10 were generated from the same samples for each of the T2I techniques.

Our novel BIE scheme generates images (Fig. 6) containing 64 rows and 64 columns. Each row represents a feature and each column is the corresponding 64-bit binary representation for that feature where ones (1) are represented as lightpixels feature where a darker value indicates a higher feature valueand vice-versa.

Unlike the other algorithms, Deep Insight does not create a grid-based image. The image (Fig. 8) is

Algorithm 1 Convert Samples To Binary Image

```

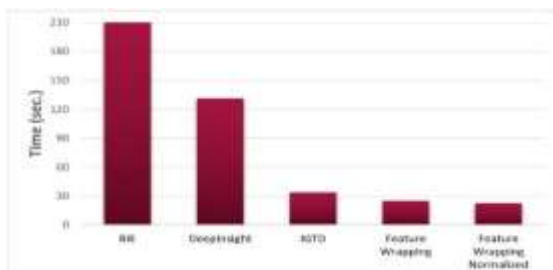
1: function CONVERT_SAMPLE_TO_BINARY(sample, feature.Types)
2:   Initialize an empty list sample_out
3:   for each feature in sample do
4:     Convert the feature to the 64-bit binary floating point representation
5:     Set the value of each bit to be equal to 255 if it is 1 and
6:     zero if it is a 0
7:     Append the binary representation of the feature to sample_out
8:   end for
9:   Create a new image with dimensions equal to the size of sample_out
10:  Set the pixel values of the image using the values in sample_out
11:  Save the image to the specified directory with the specified label
  
```

constructed as a bounding box that encompasses all the features using the convex hull algorithm. Dark pixels indicate no value and the lighter the color, the higher the feature value.

Similar to IGTD, each box in the Feature Wrapping images (Fig. 9) contains the value of a single feature. The Feature Wrapping Normalized follows the same process, but the features are normalized before encoding (Fig. 10).

ResNet MODEL TRAINING

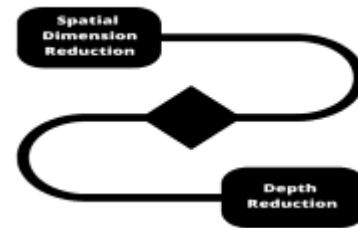
The system used to train the models ran on Ubuntu 20.04.3 LTS with an Intel i7-7700k CPU, GTX 1080 GPU,



IV RESULTS AND DISCUSSION

In this section, we present the results of the five ResNet classifiers trained on the CMU-I image datasets compared to the XGBoost classifier trained

Reduction in Depth & Width on



tabular CMU dataset. Then, we discuss the tradeoffs of the structured data and T2I approaches providing insights into the viability of each method in a potential real-world deployment.

CLASSIFIER RESULTS

Table 1 compares the performance metrics among the five ResNet-50 classifiers trained on each T2I method and XGBoost in the classification of eight application types. Values highlighted in green are the top metric across all classifiers whereas blue-highlighted values are metrics that exceeded those of XGBoost. It can be seen that the proposed Binary Image Encoding is the top-performing method across all measured metrics (excluding image generation time), improving over XG Boost's F1-score by 2.4 percentage points. IGTD was the only other method that saw higher metrics over the baseline, improving upon F1-Score by approximately 1 percentage point. Figure 11 provides better visual comparisons among the evaluated methods. Figure 12 depicts the differences in image generation time among the T2I methods. Notably, Binary Encoding took significantly longer (210 seconds) to convert the 240,000 tabular samples to their corresponding image representation which amounts to 0.9 ms per sample on average. This could be attributed to the fact that this technique is novel to this work and there may be room for further optimization. Deep Insight also took considerably longer, potentially due to its reliance on the computationally expensive Convex Hull and t-SNE algorithms. The other T2I methods had relatively shorter generation times, taking only 20-30 seconds to produce all 240,000 samples (0.1 ms per sample).

OCCUSION SENSITIVITY ANALYSIS

To better understand our BIE ResNet model's predictions, we employed occlusion sensitivity analysis. Occlusion sensitivity analysis is a popular way to visualize CNNs by blocking out a portion of a predicted image and seeing how the model's confidence is affected or important areas for the classification of that image should yield lower predicted confidence when covered. This process allows us to create occlusion sensitivity maps which visualize the parts of

FIGURE 12. Image generation times for T2I methods (240,000 samples).

BIE images that are important for classification for different classes.

We generated our occlusion sensitivity maps by replacing

part of the target image with a gray patch, classifying the image, and then mapping the model's confidence value to that region. We repeated this process for 1000 images for each class and averaged the values to find the most salient regions. Figure 13 shows the average occlusion sensitivity maps for all classes.

Observing these figures, we can see that critical regions are distinct among the classes. For instance, we observe that audio streaming is affected most by the group of features

detailed look at the critical regions. Once again, we see that audio, browsing, and chat have critical regions on the leftmost side of the image. The lower middle part of the image seems to be the most salient region for the File, P2P, and Video samples.

It should be noted that occlusion sensitivity analysis is highly dependent on patch size, so salient regions may be different when analysis is conducted with different patch sizes. Additionally, our averaged maps were conducted on a relatively small number of images, so they may not represent the entire distribution of the data. Drawing concrete inferences about how BIE images are classified is currently not possible, but these visualizations give a better idea of how the ResNet classifier differentiates different traffic types we have demonstrated that T2I methods can noticeably increase classification accuracy over shallow classifiers. Furthermore, IGTD offers an increase in accuracy while also keeping the image generation time comparatively lower.

Online learning, the process of continually updating a model based on new data, can be negatively impacted by the slow training time of deep learning models. Training each of our ResNet models took 2 hours (30 times longer than the same by XGBoost) on our hardware. If the proposed ResNet-50 models are deployed for anonymous traffic classification, online learning may not be viable depending on available computational resources.

Our experiments also showed that the choice of T2I method is important to the overall performance of the classification system as most of the T2I methods failed to improve upon or match the performance of XGBoost. IGTD and BIE also out-perform previous similar works [9] and [13] which achieved accuracies of 86% and 92% on the same eight application types. BIE may have performed well for the reasons stated in section V-D. IGTD has similarities to feature wrapping in the sense that each feature corresponds to a pixel value; however, IGTD is unique in that it correlates

features by importance which may have contributed to its superior performance. Finally, with >2%

Method	T2I Time	Loss	Acc	F1	Pre	Recall	AUC	MAE	MSE
BIE	211.4	0.4666	0.873	0.9749	0.8797	0.8703	0.9986	0.0097	0.0045
IGTD	13.7	0.1287	0.9666	0.9808	0.9818	0.8936	0.9976	0.0084	0.0083
DeepInsight	131.3	0.119	0.926	0.9225	0.905	0.900	0.998	0.0173	0.0134
Feature Wrapping Normalized	22.5	0.2717	0.8011	0.9039	0.9411	0.8700	0.9947	0.0168	0.0196
Feature Wrapping	34.9	0.4470	0.833	0.8375	0.8824	0.8234	0.9837	0.0402	0.0279
XGBoost	-	-	0.958	0.906	0.919	0.908	0.985	0.0151	0.0097

improvement on base-line classifier (XGBoost) and >1% improvement on the state-of-art T2I technique (IGTD) especially in a multi-class classification problem of determining various application types in anonymous network traffic data, we argue that our novel BIE scheme is a viable T2I technique in this domain

V LIMITATIONS AND FUTURE WORK

Due to limitations in computational resources, minimal hyper-parameter tuning was performed. Future works may benefit from experimenting with additional hyper-parameter tuning on the ResNet models and T2I parameter tuning (such as the dimensionality reduction technique used in Deep Insight and the image generation size). Moreover, other pre-trained CNN classifiers (such as ResNet-N with variation in depth, N [35]) should be evaluated and compared to non-pre-trained CNN-based models in addition to other computer vision techniques such as transformers. Finally, more visualization methods and sensitivity analysis should be applied to BIE trained models to better understand feature-class relationships.

We trained models to classify eight application types from the Tor and VPN protocols, but there are many anonymous protocols unexplored in this work. For instance, SSL/TLS, SSH, and HTTPS may not be detected or falsely classified by

the current models as they were not provided in the training data. Additionally, deep learning models benefit from larger datasets, so model performance may have been impacted by the relatively small dataset. Though synthetic data was generated using SMOTE to alleviate this concern, future work could look into gathering more anonymous network traffic to address this limitation. Nonetheless, two of the CNN models still outperformed the shallow counterpart, possibly mitigating the concern.

The CIC-DN dataset did not provide the flow interval used to generate the tabular dataset from the raw pcap file. A variety of flow intervals should be tested to find the interval that optimizes model performance.

The CMU-I dataset may be used to train/optimize additional classifiers or be used as a baseline for other

T2I techniques not considered in this work. Furthermore, the experimental workflow utilized to generate the CMU-I dataset should be applied to other domains to determine the general applicability of the approach. Since BIE is a novel technique, future work should test this method on other datasets and classifiers.

VI CONCLUSION

This work explored the viability of five T2I methods (IGTD, Deep Insight, vector-of-feature wrapping (normalized and non-normalized), and the novel BIE) and their efficacy in classification of eight anonymous network application types. These techniques were used to generate five image traffic datasets (CMU-I) for training ResNet-50. To establish baseline results for comparison, the XGBoost classifier was trained on the balanced tabular dataset (CMU). From these experiments, we found that IGTD and BIE introduced in this paper improved classification metrics when compared to XGBoost, with a tradeoff of greater computation time while encoding structured samples into images. As a novel method, the results from BIE are promising; however, further evaluation is warranted for general applicability of the technique across other problem domains.

VII ACKNOWLEDGMENT

Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding sources. The authors would like to thank Jackson Warren and Karson Fye for their early help with getting this research project started. To better identify variations in genomic and biological data, Sharma et al. introduced a T2I scheme called Deep insight. The algorithm clusters similar or related features into a two-dimensional feature space using different dimension reduction techniques such as t-SNE or PCA. Then the convex hull algorithm is used to find the smallest bounding rectangle of the feature data points. Once the rectangle has been calculated, it is rotated to be horizontal or vertical and then converted to a pixelated image.

VIII REFERENCES

- [1] A. Azab, M. Khasawneh, S. Alrabae, K.-K.-R. Choo, and M. Sarsour, "Network traffic classification: Techniques, datasets, and challenges," *Digit. Commun. Netw.*, Sep. 2022, Doi: 10.1016/j.dcan.2022.09.009.
- [2] H. Mohajeri Moghaddam, B. Li, M. Derakhshani, and I. Goldberg, "Skype Morph: Protocol obfuscation for tor bridges," in *Proc. ACM Conf. Comput. Commun. Secur.*, Oct. 2012, pp. 97–108, Doi: 10.1145/2382196.2382210.
- [3] A. Azab, M. Khasawneh, S. Alrabae, K.-K. R. Choo, and M. Sarsour, "Network traffic classification: Techniques, datasets, and challenges," in *Digital Communications and Networks*. Elsevier, Sep. 2022, Doi: 10.1016/j.dcan.2022.09.009.
- [4] S. AlDaajeh, H. Saleous, S. Alrabae, E. Barka, F. Breiting, and K.-K. R. Choo, "The role of national cybersecurity strategies on the improvement of cybersecurity education," *Comput. Secur.*, vol. 119, Aug. 2022, Art. no. 102754, Doi: 10.1016/j.cose.2022.102754.
- [5] S. Alrabae, M. Al-Kfairy, and E. Barka, "Efforts and suggestions for improving cybersecurity education," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Mar. 2022, pp. 1161–1168, Doi: 10.1109/EDUCON52537.2022.9766653.
- [6] M. Zhang, W. Sun, J. Tian, X. Zheng, and S. Guan, "An internet traffic classification method based on echo state network and improved slap swarm algorithm," *PeerJ Comput. Sci.*, vol. 8, p. e860, Feb. 2022, Doi: 10.7717/peerj-cs.860.
- [7] R. M. AlZoman and M. J. F. Alenazi, "A comparative study of traffic classification techniques for smart city networks," *Sensors*, vol. 21, no. 14, p. 4677, Jul. 2021, Doi: 10.3390/s21144677.
- [8] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2019, pp. 1–8, Doi: 10.1109/CCST.2019.8888419.
- [9] A. H. Lashkari, G. Kaur, and A. Rahali, "DI Darknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning," in *Proc. 10th Int. Conf. Common. Netw. Secur.*, Nov. 2020, pp. 1–13, Doi: 10.1145/3442520.3442521.
- [10] J. Halladay, D. Cullen, N. Briner, J. Warren, K. Fye, R. Basnet, J. Bergen, and T. Dolce, "Detection and characterization of DDoS attacks using time-base

Predicting Churn Using (Statistical Tools)

Mounika Emani
 23DSC06,
 M.Sc. (Computational Data
 Science)
 Dept. of Computer Science
 P.B. Siddhartha College of
 Arts & Science
 Vijayawada, A.P, India
 mounikaemani2016@gmail.com

Jyothika Sankar Narayanan
 23DSC15
 M.Sc. (Computational Data
 Science)
 Dept. of Computer Science
 P.B. Siddhartha College of
 Arts & Science
 Vijayawada, A.P, India
 jyothika2122@gmail.com

Jaddu Harshini
 23DSC06,
 M.Sc. (Computational
 Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 harshiniharsha525@mail.com

Abstract—Customer churn analysis is the process of predicting customers who tend to cancel the service (subscription) they receive for various reasons, especially in sectors such as telecommunications, finance and insurance, Determining the necessary operational steps to prevent this cancellation. The study used two separate datasets from kaggle.com to identify customers who tend to unsubscribe in the telecommunications industry. The analysis process was carried out by applying machine learning methods such as Logistic Regression, K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machines, AdaBoost, Multi-Layer Sensors and Naive Bayes methods on the relevant datasets. It was seen that the most successful method in the customer loss analysis performed on both datasets was the Random Forest method.

I. INTRODUCTION

Banking- Banking is the business of protecting money for others. Banking lends this money, generating interest that creates profits for the bank and also the customers. A bank is a financial institution licensed to accept deposits and make loans.

Marketing—Marketing refers to activities a company undertakes to promote the buying or selling of a product or service. Marketing includes advertising, selling and delivering products to consumers or other businesses. Some marketing is done by affiliates on behalf of a company.

Churn prediction—Churn prediction means detecting which Customers are likely to leave a service or to cancel a subscription to a service. It is a critical prediction for many businesses because acquiring new clients often costs more than retaining existing ones. Once you can identify those customers that are at risk of cancelling, you should know exactly what marketing action to take for each individual customer to maximize the chances that the customer will remain

Variables Taken:

By our given data we can see some variables. • Customer -Id: It contains random values and has no effect on

customers leaving the bank.it

- sur name: The surname of a customer has no impact on their decision to leave the bank.
- Credit score: We can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
- Geography: A customer's location can affect their decision to leave the bank.
- Gender: It's interesting to explore whether gender plays a role in a customer leaving the bank. we will include this column.
- Age: This is certainly relevant, since older customers are less likely to bank than younger ones.
- Tenure: Refers to the number of years that the customer has been a client of the bank. Normally, older clients are loyal and less likely to leave a bank.
- Balance: Also, a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balance.
- Num of products: Refers to the number of products that a customer has purchased through the bank.
- Has Card: Denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank(0=no,1=yes)
 - Is Active Member: Active customer is less likely to leave the bank, so we will keep this (0=no,1=yes)
 - Estimated salary: As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- Excited: Whether or not the customer left the bank. This is what we have to predict (0=no,1=yes)

II. MARKETING CONCEPT IN BANKING:

Marketing concept signifies a dramatic change in the approach of organizations towards their products and customers. In marketing, attention is focused on producing such goods which are wanted by customers rather than selling whatever goods have been needlessly produced. The starting point for the discipline of marketing lies in human needs and wants and their fulfilment by providing

them such a product which would satisfy the customer needs. Before nationalization of banks, banks were even more conservative and inward looking, concerned with their profits. After nationalization of 14 major commercial banks in 1969, banks woke up from their splendid isolation and found themselves placed in a highly competitive and rapidly changing environment. As a result, banks' approach towards customers and market underwent a change and focus was gradually shifted to marketing their products.

APPLICATIONS: -The data of the customer has 10000k in our data .in that we can see how the customer turns to us bank. We should show some offers. • By our we say what is the credit score of the customer. The credit of the customers is Average of credit is 650.53

- Active member percentage is – 51.51%
- Male is –38.61%
- female–39.2

THE NEED FOR MARKETING APPROACH:

In the present environment, bankers operate in a "buyer market" and here they are at the mercy of the customers, and so, there is a need for application of planned marketing approaches to find proper outlets for banking services. The concept of marketing in banks relates to the design and delivery of customer needed services in the way which satisfies them. Development of marketing on an organizational philosophy would enable banks to understand the customers and their needs better and to provide means to respond to these needs.



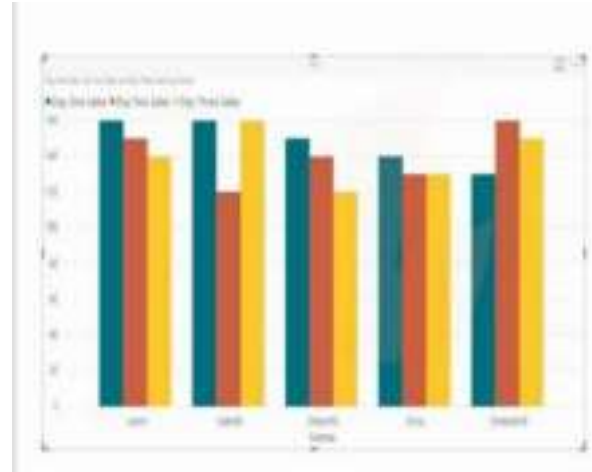
III STATISTICAL TOOLS

Clustered column chart:

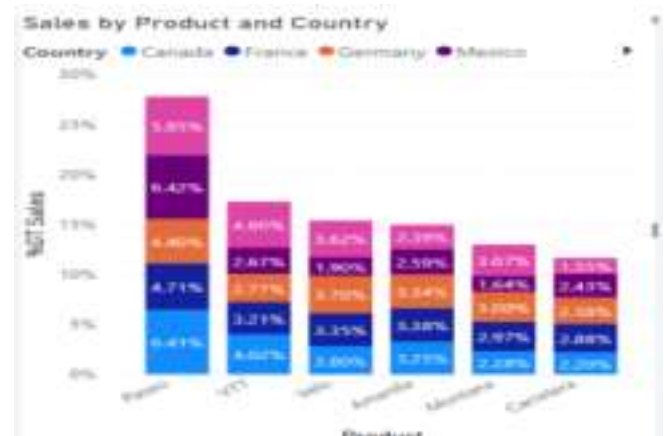
Clustered Column Chart is the default column chart behaviour where values from all series are displayed next to each other at the same category axis value. In this chart we see the count of customer id by ages. It will say how high the age is in the analysis.

Advantages:

The spacing between clusters makes comparisons clearer. Clustered charts emphasize the data within categories more than the data between them. However, you can make comparisons between categories more clearly by using consistent colour schemes; for example, in a quarterly sales chart, each quarter should have the same colour in each category.



Stacked column chart: A stacked chart is a form of bar chart that shows the composition and comparison of a few variables, either relative or absolute, over time. Also called a stacked bar or column chart, they look like a series of columns or bars that are stacked on top of each other. We should see these three variables that are Garde,



active rate and excited rate by geography. How many people in Germany, France, Spain?

Pie chart:

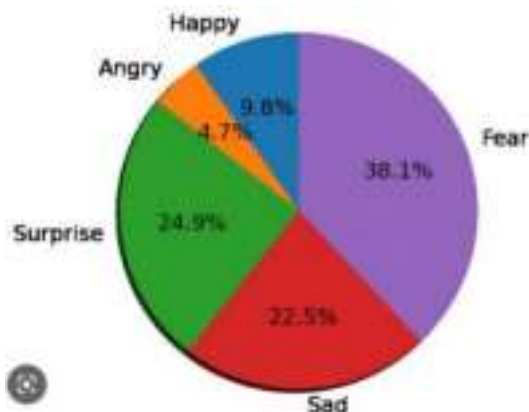
A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents. This chart shows the active rate by geography of our data.

IV ADVANTAGES:

- A simple and easy-to-understand picture.
- The need for readers to examine or measure underlying numbers themselves can be removed by using this chart.
- To emphasize points you want to make, you can manipulate pieces of data in the pie chart.

Disadvantages:

- To analyze and assimilate information quickly, this may make it more difficult for readers
- As the reader has to factor in angles and compare non adjacent slices, it has its problems in comparing the data slices.
- To make decisions based on visual impact rather than data analysis leads readers to draw



These are the statistical tools we have used
Python: -

The data Model of churn Prediction by using python: The dataset comes from the machine learning repository, and it is related to direct marketing campaigns (phone calls) of a

Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y).

Python code for predicting Accuracy of Churn Prediction

```
<class 'pandas. core. Frame. Data Frame'>
```

```

Range Index: 10000 entries, 0 to 9999
Data columns (total 14 columns):
# Column Non-Null Count D type
0 Row Number 10000 non-null int64
1 Customer Id 10000 non-null int64
2 Surname 10000 non-null object
3 Credit score 10000 non-null int64

```

```

4 Geography 10000 non-null int64
7 Tenure 10000 non-null int64
8 Balance 10000 non-null float64
9 Num of Products 10000 non-null int64
10 Has Cr card 10000 non-null int64
11 Is Active Member 10000 non-null int64
12 Estimated Salary 10000 non-null float64
13 Exited 10000 non-null int64
d types: float64(2), int64(11), object (1)
memory usage: 1.1+ MB

```

Output: -

```

Intercept [-4.39462413]
Coefficients [[ 0.44385017 -0.51711139 -1.07467782
0.07241397]]
train Data Shape (8000, 14)
test Data Shape (2000, 14)
accuracy on train data 0.8125
accuracy on test data 0.805

```

V CONCLUSION:

In conclusion, customer churn prediction analysis is helpful for companies trying to move toward a data driven approach and boost their financial metrics by lowering customer churn. Businesses can proactively address customer concerns and ensure long-term success by conducting regular analyses as a part of AI based customer service.

VI REFERENCES: -

- <https://www.cyberclick.net/marketing>
- <https://en.wikipedia.org/wiki/Bank>
- <https://www.safalta.com/blog/banks-in-india-types-functions>
- https://en.wikipedia.org/wiki/Churn_rate#:~:text=It%20is%20a%20possible%20indicator,of%20average%20customer%20life%20time

Navigating Cyber Threats and Securing Solutions

Imadabattni.Sai Balaji
23DSC07, M.Sc.(Computational Data
Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts &
Science
Vijayawada, A.P, India
balunaidu6424@gmail.com

Dr.T.Srinivasa Ravi Kiran
HoD & Associate Professor
Dept. of Computer Science
P.B.Siddhartha College of Arts &
Science
Vijayawada, A.P, India
tsravikiran@pbsiddhartha.ac.in

Dr. V. Rama Chandran
Professor & HoD,
Dept of CSE,
Vasireddy Venkatadri Institute of
Technology, Nambur
vrc.bhatt@vvit.net

Abstract- An Earthquake Early Warning System (EWS) gives us a heads-up about when strong seismic waves might hit after an earthquake's initial shaking. With today's advanced technology and data analysis tools, it's becoming more challenging to quickly process and understand this seismic data. Spotting earthquakes early is crucial for fast communication of warnings. In a recent study, deep learning was used to spot and categorize earthquake P waves and background noise using historical data from a specific monitoring station in West Sumatra. The study focused on selecting certain earthquake signals near the station and found that the deep learning approach performed well in telling apart earthquake signals from other noise. This study is an initial step towards using deep learning to classify earthquake signals and predict where earthquakes might happen, based on data from multiple recording channels.

Keywords- Seismic waves, Seismic data processing, Earthquake, Deep learning, P waves, Noise signals.

I INTRODUCTION

Past ways of sorting earthquakes used things like where they happened, how deep they were, why they occurred, how powerful they were, and how far the center was (Chandwani, 2013). No matter how we classify them, earthquakes are a big worry because they can harm people, buildings, and nature (Ramli and Razak, 2015). After an earthquake, lots of things are needed to handle the situation properly. We need stuff like tools for finding and rescuing people, things to help like food and shelters, and important medical supplies. Getting all these things right is a big job and needs a good amount of money to make sure we have what's needed. It's not just stuff, though—people are just as important. We need trained folks who can manage emergencies, volunteers to help out, and doctors and nurses too. People do things like getting others to safety, giving first aid, and searching for and rescuing those who need help.

When we face a huge disaster like an earthquake, it's smart to plan how to use our resources wisely. Some smart studies have used past cases to figure out how much help might be needed in future emergencies (Shao et al., 2021) and how to make good decisions during those tough times

(Wang et al., 2021). The plan made by Shao and others looks at past disasters and compares them to the current one to guess how much help will be needed. Similarly, Wang and the team made a plan that matches current emergencies to past ones to find the best way to help. These plans really help guide how we react during emergencies.

We need more studies to figure out better ways to sort out disasters, so we can give the right amount of help to manage them better. This study's main aim was to use a smart computer method (unsupervised machine learning) just for earthquakes. They wanted to group earthquakes based on how big they were and how much damage they caused. This grouping helps decide how to react to future earthquakes by giving the right kind of help based on each earthquake's unique qualities.

II RELATED WORK

2.1. Earthquakes Early Warning Systems

An earthquake early warning (EEW) system is made to find an earthquake, figure out how big it is, where it is, and when it started. It then sends warnings to places that might be affected by strong shaking. These systems use science and technology to alert devices and people in those areas so they can get ready before the strong shaking arrives. This early warning works by detecting fast, not-harmful waves called P waves before the more damaging S waves. The time gap between these waves decides how much warning time there is. If the earthquake is far away, this gap is longer—about 60 to 90 seconds for big, far, and deep earthquakes.

Countries that use early warning systems usually look at two things: how long the shaking lasts and how big the shaking is right after the first P wave hits a nearby monitoring station. Figuring out when the P wave arrives is super important because it helps predict how big the earthquake is and how far away it is. Even a tiny mistake in figuring out when the P wave comes can mess up the prediction of the earthquake's size and where it happened.

To get this timing right and calculate the earthquake details, a method called Integ High Order Statistical technique is used. It uses the signal's speed change to find when the P wave arrives. Then, using this info, they can figure out how big the earthquake is and where it happened.

Recently, people have been interested in using deep learning and data mining tricks to detect earthquakes. They've trained computers to spot and locate earthquakes using fancy networks like deep convolutional neural networks (CNN) and deep learning Long Short-Term Memory (LSTM) networks. Some even taught a computer system called a generative adversarial network (GAN) to recognize the first waves of an earthquake to avoid false alarms caused by other noises. They plan for the use of several evasive techniques to elude detection by their target's intrusion detection systems. They follow "low and slow" approach to increase the rate of their success.

2.2. Signal-to-noise ratio (SNR)

Intense noise usually disturbs seismic waves that are recorded by near-surface sensors. Therefore, the seismic data collected is often of poor quality; this phenomenon can be explained as a low signal-to noise ratio (SNR) [7]. The low seismic data SNR can decrease the rate of several subsequent seismological studies, such as inversion and imaging, for example. The reduction of unnecessary seismic noise is also of great significance.

It is possible to use Window D, part of the seismic record, for SNR:

$$D = [xi, j] M \times N \quad (0 < M \leq Nx, 0 < N \leq Nt) \quad (1)$$

The following assumption: waveform, amplitude, and seismic wavelet phase in window D hold constant in terms of distance "i" noise is randomly distributed "zero mean" along with the position of the survey line being independent (decorrelated) of the signal, such that the signal is independent (decorrelated)

$$xi, j = sj + ni, j \quad (2)$$

$$\sum ni, j = 0 \quad M \quad i=1 \quad (3)$$

In general, these assumptions suggest a restriction to this approach, but they can be fulfilled if the local window is selected in the seismic section's stable signal area. Where sj is amplitude of signal, and ni, j is amplitude of noise. So, if the energy of the signal in a window is:

$$Es = M \sum_{j=1}^N sj^2 = \frac{1}{M} \sum_{j=1}^N (\sum_{i=1}^M Xi, j)^2$$

2.3. Skewness

Skewness is a way to tell how data are spread out. When data follow a normal pattern, the skewness is zero, meaning the data are balanced around the average. But if the data lean more to one side or the other, the skewness shows that.

For earthquake wave arrival times, the size of the signals tends to be a lot bigger than usual. That's why skewness can help figure out when the first wave of an earthquake, called the P wave, arrives. The trick is, the best point on the skewness curve doesn't directly show the wave's arrival time. Instead, the biggest value on the skewness curve matches when the wave actually shows up. Using a method called differentiation skewness helps reduce the time delay and pinpoints when the wave arrives most accurately.

2.4. Generating location labels

Cluster analysis using K-means is a way to group a bunch of things based on similarities in their characteristics. This method divides all the data into clusters or groups, making sure each thing belongs entirely to just one cluster. It works by initially splitting the data and finding the average for each group. Then, it repeats this process, adjusting the groups based on how close they are to these averages.

The idea is to group similar things together. These groups might not be exactly the same as each other but aim to have similar things within each group. The "similarity" here means how close or far things are from each other in the data. When things are close, it suggests they are quite similar, and when they are far apart, they are less alike.

In this study, they used the K-means method to divide a list of earthquakes in an area into nine clusters based on how far apart they were from each other. This helped them identify distinct areas on a map where earthquakes tend to happen. They found that these nine clusters helped them separate the main earthquake patterns. They then labeled these clusters into two groups: Class 0 for places without earthquakes (just noise) and Class 1 for areas where earthquakes occur in that specific region.

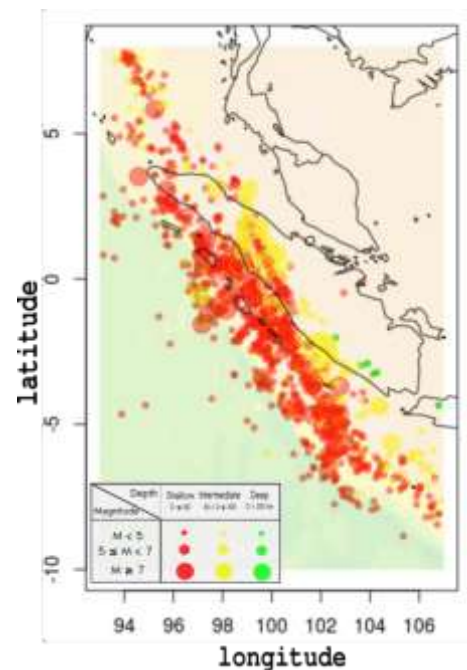


Figure 2. From 1 January 2014 to 30 September 2020, earthquakes and seismic stations in the Area of Interest (West Sumatra).

III PROPOSED WORK

Deep earthquake learning works by taking three sets of earthquake data and figuring out the chance of each data set belonging to different groups or categories. The method uses a specific structure made up of different parts—like

input, hidden layers, and an output—arranged in a certain way to process this data (shown in Figure 3). It's trained using a dataset and tries to minimize errors in its predictions. The core of this method is built using dense layers, which are like basic building blocks in this type of learning. In these layers, every piece of data connects to all the neurons, creating a network. Adjusting the batch size and number of times the network learns from the data (epochs) is important to make sure the computer can handle it without any memory issues or taking too long to learn. In this case, each learning session consists of 32 pieces of data at a time and goes through 100 sessions.

They used a specific way to activate and connect these neurons, like using ReLU, a type of activation function. This function helps in speeding up computations and makes the system more efficient. The study was done using a program called TensorFlow CPU 2.2.0, which helps create and run these kinds of learning systems.

Algorithm:



Figure.3. The deep neural network corresponding to the proposed model.

Dropout is a process to prevent overfitting and also speed up the learning process. Dropout refers to removing neurons that are visible hidden or visible layers on the network. By eliminating a neuron, means removing it temporarily from the existing network. The neurons to be removed will be randomly selected. Each neuron will be assigned a probability value between 0 and 1.

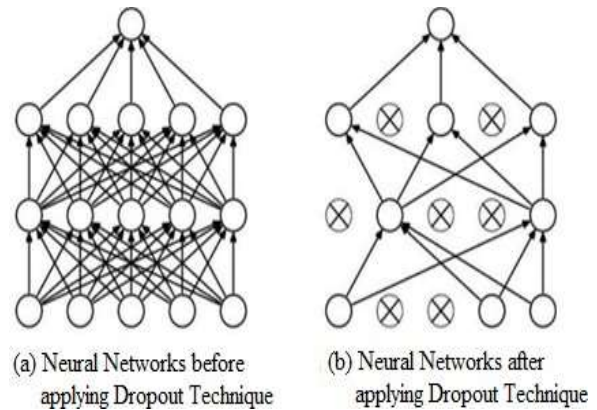
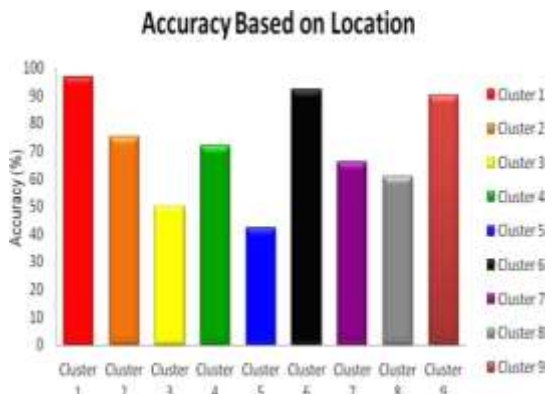


Figure 4. Dropout Technique

The picture above the neural networks (a) is an ordinary neural network with two hidden layers. Whereas in part (b), the neural network has applied dropout regularization technique where several activation neurons are no longer used. This technique is straightforward to implement in a deep learning model and will impact the model's performance in training and reduce overfitting.

For classification strategies with a vast number of classes, such as multinomial logistic regression, multiclass linear discrimination analysis, Naive Bayes Classifier, and Artificial Neural Network, the SoftMax function is used. SoftMax is a function that converts the K-dimensional vector X, which is an actual value, into a vector with the same shape but with values in the range 0-1, which is 1. The SoftMax function is used in layers in neural networks and is usually found in the last layer to get the output. SoftMax neurons receive input and then do weighting and adding bias. But after that, the neurons in the SoftMax layer do not apply the activation function but instead use the SoftMax function. It can be concluded that the SoftMax layer determines the most significant probability for its class result.

A confusion matrix is used to calculate performance metrics (accuracy) which aims to measure the performance of the model that has been made. The confusion matrix provides information on comparing classification results carried out by the system (model) with the actual classification results. The confusion matrix is in the form of a matrix table that describes the performance of the classification model on a series of test data whose actual value is known. There are four terms as a representation of the result of the classification process on the confusion matrix. The four terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) is positive data that is predicted to be correct. True Negative (TN) Is negative data that is predicted to be correct. False Positive (FP) Is negative data but predicted as positive data. False Negative (FN) Is positive data but predicted as negative data.



RESULTS AND ANALYSIS

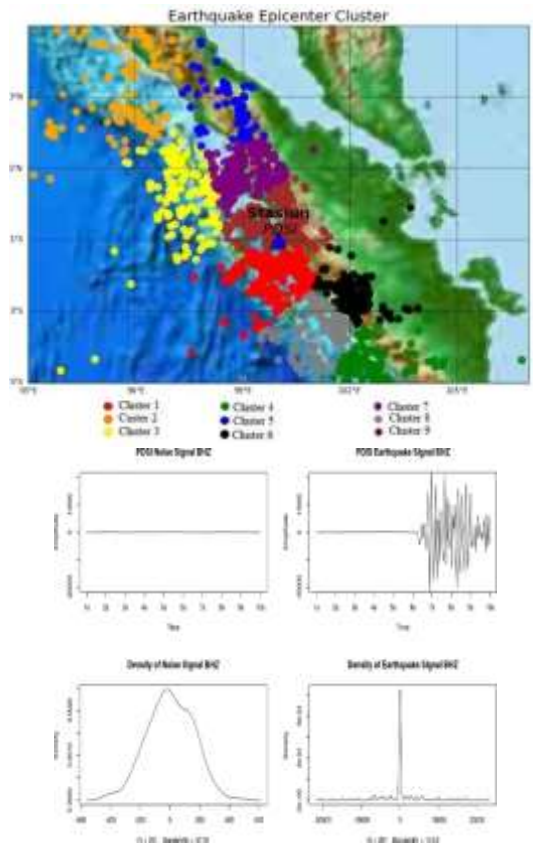


Figure 5. PDSI seismic waveform series
 Differentiating between earthquake waves (P waves) and background noise is important. Figure 5 shows how the patterns of these two types of signals are different, making it possible to create models to tell them apart. Sometimes, noise mixes in with the earthquake signals in recordings, which can make it harder to understand earthquakes. So, choosing a quiet place to record and preparing the data well

is crucial in earthquake studies. This helps improve the Signal-to-Noise Ratio (SNR), which helps in understanding seismic signals better. To understand these signals, math is used to break them down into simpler parts using a method called Fourier analysis. This helps spot irregularities in the earthquake signals, like when P waves aren't clear or when the noise signals differ a lot. They used a dataset of 1455 signals for training and 624 for testing their model. Their model was designed in a specific way using different layers and mathematical functions. It was trained using 100 rounds, with each round having 15 smaller parts. The model showed good accuracy, especially with a bigger dataset where both noise and earthquake signals were balanced. However, in classifying these signals, there's a balance between making mistakes like false alerts or missing real ones. Adjusting the criteria for detecting signals can change this balance. It depends on how well the system can pick out real signals from the noise and how much tolerance users have for false alerts. Also, they checked how accurately the model could pinpoint the location of the earthquake, and it showed good accuracy in some areas but needed improvement in others. They plan to focus on improving accuracy in areas where the model didn't perform as well, particularly by looking at the Signal-to-Noise Ratio in those spots.

From 624 tested signals, cluster 9 has the highest accuracy, see fig 6. From that point, and if we connect it to the location of the earthquake source, it turns out that the location of the earthquake source in cluster 9 is in an area close to the PDSI station. The far, the accuracy drops to 42 %. Cluster 3 and 5 have low average SNR value with the specification of SNR waveform see Table 1.

Table 1. SNR Value at low accuracy based on cluster

Cluster	Average SNR BHZ	Average SNR BHN	Average SNR BHE
Cluster 3	3,53973221	3,43782069	3,50723973
Cluster 5	3,56993703	3,42748344	3,51663228

The low seismic data SNR can decrease the quality of several subsequent seismological studies, such as inversion and imaging, for example [3]. The reduction of unnecessary seismic noise is also of great significance. So, for future work, low accuracy cluster needs specific processing techniques to improve the SNR value. One of the key challenges of applied seismology is maintaining high SNR or enhancing it by appropriate methods of data collection and analysis when conditions are poor. The effectiveness of SNR enhancement primarily depends on our experience of how seismic signals and noise varies. The performance of this method is better for near-source earthquake than the far source. These results are in accordance with previous studies using the different methods [4], but our accuracy was higher for the testing result.

Suppose the suggested technique is applied in EEWs to recognize earthquakes and noise with the near-source location stations detected in one second. In that case, an automated warning may occur well before the epicenter shakes strongly, so people who are close to the earthquake epicenter to be warned in the event of an earthquake [5].

IV CONCLUSION

This study is an early attempt using deep learning to tell apart earthquake signals from background noise and figure out where earthquakes might happen. The accuracy is good, especially with a larger dataset that has an equal mix of noise and earthquake signals. However, there's still work to do to reduce false alarms. We're focusing on using a method called Kmeans to group similar earthquake signals and improve the model. Some clusters in our analysis didn't predict well, particularly below clusters 3 and 5. To make the model better, we plan to add more earthquake signals from these clusters or use other ways to get more data that will help the model learn better, especially focusing on areas with more earthquakes. In the future, we'll include more data for training. Also, as more data becomes available, deep learning methods are getting better and more popular for this kind of work.

V REFERENCES

- [1] Pustlitbang PUPR, 2017 *Buku Peta Gempa 2017*.
Li Z Meier M A Hauksson E Zhan Z and Andrews J, 2018 Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning *Geophys. Res. Lett.* **45**, 10 p. 4773–4779.
- [2] Perol T Gharbi M and Denolle M A, 2017 Convolutional neural network for earthquake detection and location *arXiv* **2016**, November 2016 p. 2–10.
- [3] Kuyuk H S and Susumu O, 2018 Real-time classification of earthquake using deep learning
Procedia Comput. Sci. **140**, October 2018 p. 298–305.
- [4] Tajima F and Hayashida T, 2018 Earthquake early warning: what does “seconds before a strong hit” mean? *Prog. Earth Planet. Sci.* **5**, 1.
- [5] Gunawan H Puspito N T Ibrahim G and Harjadi P J P, 2011 Determination of earthquake early warning based on the initial phase of P-wave (case study of West Java) *IUGG Gen. Assem.* July.
- [6] Chen Y Zhang M Bai M and Chen W, 2019 Improving the Signal-to-Noise Ratio of Seismological Datasets by Unsupervised Machine Learning *Seismol. Res. Lett.* **90**, 4 p. 1552–1564.
- [7] Weatherill G and Burton P W, 2009 Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region *Geophys. J. Int.* **176**, 2 p. 565–588.
- [8] Everitt B and Hothorn T, 2011 An Introduction to Applied Multivariate Analysis with R Springer Sci. Media p. 163–164.
- [9] Novianti P Setyorini D and Rafflesia U, 2017 K-Means cluster analysis in earthquake epicenter clustering *3*, 2 p. 81–89.
- [10] Maas A L Hannun a Y and Ng A Y, 2013 Rectifier nonlinearities improve neural network acoustic models ICML Work. Deep Learn. Audio, Speech Lang. Process. 28.
- [11] He K Zhang X Ren S and Sun J, 2015 Delving deep into rectifiers: Surpassing human-level performance on imagenet classification Proc. IEEE Int. Conf. Comput. Vis. 2015 Inter p. 1026–1034.
- [12] Xu B Wang N Chen T and Li M, 2015.
- [13] Unterthiner T and Hochreiter S, 2016 Fast and Accurate Deep Network Learning arXiv p. 1–14.
- [14] Bormann P and Wielandt E, 2013 Seismic Signals and Noise New Man. Seismol. Obs. Pract. **2** June 2013 p. 1–41.
- [15] Alom M Z et al., 2019 A state-of-the-art survey on deep learning theory and architectures
a. *Electron.* **8**, 3 p. 1–67.

Steganography Encryption Standard for Android Application

Jaddu.Harshini
23DSC08, M.Sc. (Computational
Data Science)
Dept. of Computer Science
P.B. Siddhartha College of
Arts & Science
Vijayawada, A.P, India
harshiniharsha525@gmail.com

Sankaranarayanan.Jyothika
23DSC15, M.Sc.
(Computational Data Science)
Dept. of Computer Science
P.B. Siddhartha College of
Arts & Science
Vijayawada, A.P, India
jyothika2122@gmail.com

Emani. Mounika
23DSC06, M.Sc. (Computational Data
Science)
Dept. of Computer Science
P.B. Siddhartha College of Arts &
Science, Vijayawada, A.P, India
mounikaemani2016@gmail.com

Abstract- This paper presents an image Steganography algorithm that can work for cover images of multiple formats. Having a single algorithm for multiple image types provides several advantages. Steganography security is the main concern in today's informative world. The fact is that communication takes place to hide information secretly. Steganography is the technique of hiding secret data within an ordinary, non-secret, file, text message and images. To the best of our knowledge, the proposed algorithm is the first Steganography algorithm that can work for multiple cover image formats. The hidden message will be seamlessly included into the regular picture using advanced steganographic methods like the least significant bits approach.

Keywords: least significant bit (lsb); password-based encryption; data security; cover image, information hiding

I INTRODUCTION

The area of Steganography generally covers techniques that try to hide information in some media file being transmitted. The goal of Steganography is to ensure that an adversary should not suspect the presence of a hidden message. Text, image, audio or video files can be used as media to conceal the secret message. In Image Steganography, after Steganography operations, nobody should be able to notice the visual difference between original image (called cover image) and the resultant image (called stego image). Text, image, audio or video files can all be treated as sequence of data bits and can be transmitted by hiding in cover images using image Steganography. Data bits are also called payload data, secret data, data object and hidden data. Steganography has changed to take use of digital technology. Since it permits the insertion of concealed information in otherwise innocuous photographs, image steganography has proved to be a very efficient means of secret communication. The purpose of this work is to develop an Android app that makes use of Image Steganography, therefore allowing its users a simple and intuitive means of safely hiding messages or images under a cover image. We want to provide a widely used tool that takes advantage of mobile technology and is thus accessible to a large number of people. With the best of

our knowledge, the reported results for image Steganography focus on cover images of some particular format. Out of the numerous image Steganography algorithms, there are works that use cover images of type JPEG [7- 9, 18, 21], PNG [11-14] and RGB [1-6, 19, 20]. In this paper, we present a Generic Steganography Algorithm (called GSA) that is described for abstract image components. The novel aspect of GSA is that, we can implement GSA for all these types of cover images (i.e. JPEG, Bitmap or PNG) with trivial adaptations. In addition, we have implemented GSA for JPEG, Bitmap and PNG formats with very promising results compared to other reported methods. At this point, the following query may come up in the mind of the reader: what is the benefit of having an image Steganography algorithm that can work for multiple cover image formats? We can answer this question by noting the following points, which also provide the motivations for the current work.

Having the option to use any types of cover images using one algorithm provides flexibility and simplicity for a user.

- Capacity of a cover image can vary based on the image format. Based on the data length, network bandwidth and allowable distortions, GSA can adaptively select the best cover image format (for the same image) to hide data.
- Since the same algorithm is used for various cover image formats, security levels can be enhanced by modify ingonly a few parameters (like, using data spreading technique in GSA).

Combining several cryptographic and steganographic techniques is the heart of Image Steganography. The first step in the process is encrypting the original picture using a symmetric technique so that no one except the intended recipient can see its contents. The next step is to use an asymmetric encryption technique to keep the secret encryption key safe. The secret message is

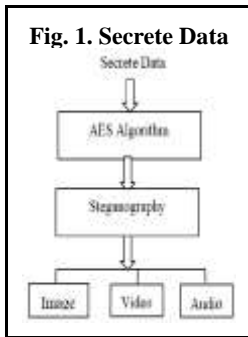


Fig. 1. Secrete Data

II RELATED WORK

There are a lot of works reported in image Steganography. In this section, we briefly review some of the works related to our approach.

Before we discuss different works, it is good to discuss some factors that we should consider while evaluating a Steganography algorithm. Fig. 1 illustrates some essential factors in Steganography. In confidentiality, the Stego image transmission must be confidential, where only the authorized person must be able to read the message, while others should not even suspect it. Integrity means that only an authorized person would be able to modify or change the message. Robustness is the ability of stego image to resist manipulations like filtering or compression. Hiding Capacity is the amount

of information hidden in the cover image media. It is measured as bpp (bits per pixel) or bpc (bits per coefficient). In Authentication, the origin of the message is recognized correctly with an assurance that identity is not false. Perceptibility implies that the Stego image must be same like cover image without visual difference and hidden message should not to be detectable by eyes.

PSNR (Peak Signal to Noise Ratio) is a visual quality estimator for stego images. It is used to measure the perceptible quality of the modified Steganography image in decibels (dB). A high value for PSNR indicates higher quality of an image. MSE (Mean Square Error) is an estimate for error between the original cover image and the output stego(reconstructed) images. There is an inverse relationship between PSNR and MSE, a lower MSE value indicates lower difference between input and output images. If MSE value is less than one, it implies that modified image has been properly restored. Stego key is a key that is embedded into the cover media along with the secret data to retrieve the embedded secret data back correctly (explained in detail in Section III). The various image Steganograsafe from prying eyes even if the encrypted picture is snatched because of the asymmetric encryption used. The encrypted secret key is then embedded in the cover picture via a steganographic approach, ensuring clandestine communication. The encrypted secret key will be hidden in the least significant bits (LSB) of the cover picture pixels using the least significant bits (LSB) steganographic method. Phy algorithms attempt to

optimize one or more of these factors.

Previous works present various Steganography algorithms for specific image formats [1 – 24, 26]. Algorithms for image formats like JPEG [7 – 9, 18, 21, 23, 24, 26], PNG [11 – 14], and RGB [1 – 6, 19, 20] are discussed here, as these types of

images are amongst the most common ones. Table I

Summarizes the various image Steganography algorithms.

In order to protect information in transit, the authors of this research suggest creating an Android app called Stego that uses cover graphics to conceal hidden text or images. The programme takes in images in JPEG, PNG, and BMP formats and keeps them in the same structure after data concealment using the Least Significant Bit (LSB) steganography method. The generated stego picture may be password-protected inside the programme, further bolstering security even if the algorithm's parameters are known [10]. This study dives into the technicalities of steganography, the practice of secretly exchanging information by embedding it in another file. The main emphasis is on picture steganography, including a look at various applications and methods.

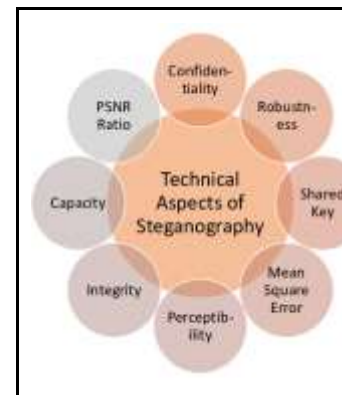


FIGURE 2. Major technical aspects of an image Steganography algorithm.

A. SPATIAL DOMAIN(RGB/BITMAP) STEGANOGRAPHY

Algorithms for bitmap images work in spatial domain, using direct modifications in the cover image pixels. RGB algorithms provide high capacity but less security because image pixels can be modified directly as per the scene's curve and edges. Examples of RGB algorithms are **LSB (least significant bit)** substitution method, pixel indicator technique optimal pixel adjustment procedure, secure key-based image realization Steganography, etc.

The techniques in [3, 4, 6, 32, 33] are all LSB substitution based methods, where the basic idea is to embed the message into the right most bits of the pixel array sequentially or randomly without disturbing the original pixel values much. Pixel indicator techniques are based on the concept of an indicator channel and an embedding

channel [1]. Amirtharajan et al. [2] used both LSB and pixel indicator techniques to enhance security.

B. FREQUENCY DOMAIN (JPEG) STEGANOGRAPHY

JPEG image format algorithms work in the frequency transform domain (they work on the rate at which the pixel values are changing in the spatial domain). Frequency transform domain is divided into two categories: high-In order to protect information in transit, the authors of this research suggest creating an Android app called Stego that uses cover graphics to conceal hidden text or images. The programme takes in images in JPEG, PNG, and BMP formats and keeps them in the same structure after data concealment using the Least Significant Bit (LSB) steganography method.

The generated stegopicture may be password-protected inside the programme, further bolstering security even if the algorithm's parameters are known [10]. This study dives into the technicalities of steganography, the practice of secretly exchanging information by embedding it in another file. The main emphasis is on picture steganography, including a look at various applications and methods. In order to protect information in transit, the authors of this research suggest creating an Android app called Stego that uses cover graphics to conceal hidden text or images. The programme takes in images in JPEG, PNG, and BMP formats and keeps them in the same structure after data concealment using the Least Significant Bit (LSB) steganography method.

The generated stego picture may be password-protected inside the programme, further bolstering security even if the algorithm's parameters are known [10]. This study dives into the technicalities of steganography, the practice of secretly exchanging information by embedding it in another file. The main emphasis is on picture steganography, including a look at various applications and methods.

III. PROPOSED WORK

The goal of the Android-based Image Steganography is to make the concealment of information even more secure and secret. In order to provide the highest level of protection for the information being stored, this system uses a two-pronged approach, one that encrypts data using a password and another that uses the Least Significant Bit (LSB) technique

A. The Least Significant Bit (LSB) Method: The LSB method is a prominent approach in the field of picture steganography. It is based on the premise of altering the least significant bits of the pixel data of an image with the bits of the hidden message. With this method, the visual quality of the picture may be protected while the hidden information continues to be undetectable to the human eye. The goal of the Android-based Image Steganography is to make the concealment of information even more secure and secret. In order to provide the highest level of protection for the information being stored, this system uses a two-pronged approach, one that encrypts data using a password and another that uses the Least Significant Bit (LSB) technique

The Least Significant Bit (LSB) Method: The LSB method is

a prominent approach in the field of picture steganography. It is based on the premise of altering the least significant bits of the pixel data of an image with the bits of the hidden message. With this method, the visual quality of the picture may be protected while the hidden information continues to be undetectable to the human eye. In modern digital steganography, data is first encrypted or obfuscated, and then inserted using a special algorithm into data that is part of a particular file format, such as a JPEG image, audio or video file. The secret message can be embedded into ordinary data files in many ways. In modern digital steganography, data is first encrypted or obfuscated, and then inserted using a special algorithm into data that is part of a particular file format, such as a JPEG image, audio or video file. The secret message can be embedded into ordinary data files in many ways

SYSTEM REQUIREMENTS

A. Hardware Requirements

1. Minimum i5 or above Processor
2. Minimum 4 GB or above RAM
3. Windows OS

B. Software Requirements

1. Java
2. XML
3. Android Studio

ARCHITECTURE DIAGRAM

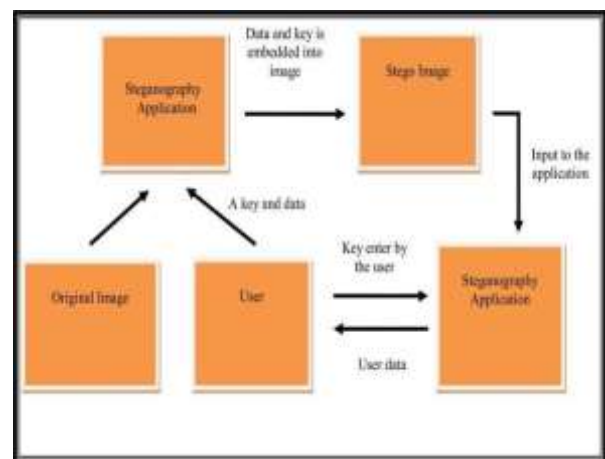


Fig. 3. The General Steganography System

The use of a symmetric encryption method comes with number of benefits, two of which being the method's efficiency and speed. Since symmetric encryption takes a lower amount of processing power than other encryption algorithms, it is an excellent choice for mobile apps, which often have constrained access to resources. In order for the

encryption procedure to be successful, it is necessary for the user to provide a password. It functions as a one-of-a-kind key that is only known to the sender and the receiver to whom it is meant to be sent. It is essential that this password be kept secret since anybody who has access to it would be able to decipher the communication that has been encrypted. In conclusion, the process of symmetric encryption offers an extra layer of security, making it possible to guarantee the message's secrecy even in the event that the steganography procedure is broken. This system is protected against intrusion by unauthorized users thanks to the password, which serves as the system's secret key. The data transfer is made much more secure by using both steganography and encryption in conjunction with one another.

C. Embedding of Cypher Text:

The last step in the Android-based Image Steganography project includes employing the Least Significant Bit (LSB)steganographic method to embed the encrypted secret message, also known as the cypher text, into the cover image. This is the procedure that completes the project. This step is essential because it not only conceals the existence of the message but also adds an extra degree of protection by making it very challenging for unauthorized persons to access the concealed data. As a result, the presence of the message is no longer detectable. The least significant bit (or bits) of the mage's pixel values are the ones that are replaced with the bits of the encrypted message when using the LSB steganography method, which is a method that is both straightforward and effective. This alteration is almost imperceptible to the human sight, and the resultant steganographic picture gives the impression to an unwary viewer that it is similar to the cover image that was used originally. As a result, the use of this method is extraordinarily advantageous in terms of maintaining the image's quality while simultaneously hiding sensitive information. For hiding a data behind cover image, first user has to login into the system. The embed message feature, embeds a message into a master file. The system asks for the master file & output file. After the user specifies the files, the system asks for the message to embed into the file. The system also asks to compress the output file password to be encrypted into output file. After the completion of above steps, the message is embedded into the output file. Then using LSB algorithm, secret message is hidden behind the cover image. LSB is a common and simple way to embed information in a cover image. The LSB of an image is replaced by bit of the secret message. Using image of 24 bit, a bit of each of the red, green and blue colour can be used to hide secret message. Now, once the message is hidden i.e. it is encrypted. It is called as Stego image and can be sent to the intended destination

LSB ALGORITHM

A. EMBEDDING DATA

1. Extract the pixels of the cover image
2. Extract the character
3. Insert characters of text file in each first component of next pixels by replacing it
4. Repeat till all the characters has been embedded.
5. Place some terminating symbol to indicate end of data.
6. Obtained stego image.

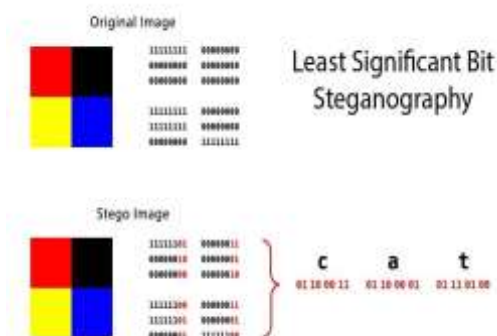


Fig. 4. Least Significant stenograph

B. DATA EXTRACTION STEPS

1. Extract the pixels of the stego image.
2. Now, start from first pixel and extract secret text characters from first component of the pixels. Follow up to terminating symbol
3. Then go to next pixels and extract secret message characters from first component of next pixels up to terminating symbol.
4. Extract secret message.

ADVANTAGES

- Messages do not attract attention to themselves i.e. difficult to detect. It can only be detected by desirous receiver. Provides better security for sharing data in LAN, MAN & WAN.
- The proposed technique uses LSB to hide data from a pre -defined position agreed between two parties. Same position is used only once to enhance security.
- Network surveillance and monitoring systems will not flag messages or files that contain Steganography data.
- Along with hiding secret information, Steganography also conceals the communicating parties.

IV APPLICATIONS

Fields of application

1. Defense and intelligence
2. Medical
3. Online banking

4. Online transaction

- Confidential communication and secret data storing.
- Protection of data alteration.
- The health care, and especially medical imaging systems, may very much benefit from information hiding techniques.
- In military applications. Transport highly private documents between international Governments.

V RESULT & ANALYSIS

The performance of the Android-based Image Steganography application was assessed based on the integrity of the concealed message, the amount of security given by the password encryption, and the perceptual transparency of the cover image. The programme was subjected to extensive testing, and the findings showed that it did a good job of protecting the confidentiality of the secret information. The original message was fully maintained and was able to be retrieved at the conclusion of the procedure, despite the complexity of the operation, which included encryption, embedding, extraction, and decryption. This demonstrated that the LSB embedding method as well as the password-based symmetric encryption that was used were both resilient and dependable for ensuring the confidentiality of the connection.

VI FUTURE SCOPE

- The compression ratio of images can be improved.
- It can be extended to a level such that it can be used for the different types of image formats like bmp, jpeg, .tif etc.
- So other image formats also will come in use for Steganography.
- The security using least significant bit algorithm is good but in future it can be improved to a certain level by varying the carriers as well as using the different keys for encryption and decryption.

both user-friendly and easy to use, and it was designed to provide users the ability to choose a cover picture and type in a secret message. The programme preserves the security of the cover picture by using strong symmetric encryption methods. This renders the image unavailable to unauthorized persons, therefore protecting its privacy.

capability to efficiently safeguard their sensitive information by integrating steganographic and encrypting approaches. The goal of the project was to develop an interface that is

VII CONCLUSION

Smart Steganography application software provided for the purpose, how to use embed message into image The master work of this application is in supporting the facility of compressing of output file, even encrypt the output file. Steganography can be the best security tool. Individuals who are interested in secure communication and the concealment of data may find the creation of an Android application based on image steganography to be a very helpful option. The programme provides users with the The compression of the images can be improved.

- It can be extended to a level such that it can be used for the different types of image formats like bmp, jpeg, .tif etc.
- So other image formats can also be used in steganography.

VIII REFERENCES

- [1] Comparative study of image steganography techniques, Himanshu Arora, Cheshta Bansal, Sunny Dagar, International Conference on Advances in Computing, Communication Control and Networking, 2018, 982-985
- [2] Implementation and analysis of image steganography using Least Significant Bit and Discrete Wavelet Transform techniques, Srushti S Yadahalli, Shambhavi Rege, Reena Sonkusare, IEEE, 2020 ,1325-1330
- [3] Review paper on image steganography, Deepali V. Patil, Mr. Shatendra Dubey, international journal of research in computer applications and robotic, 2014, Vol-02, 35-40
- [4] Image Steganography: A Review of the Recent Advances, Nandhini Subramanian, omar elharrouss, somaya al-maadeed, ahmed bouridane, IEEE, 2021, Vol-9, 23409-23423

Guardians of Data: A Comprehensive Study on Database Security Measures

Marrivada Gayathri
 23DSC10, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 marrivadagayathri@gmail.com

Muddamsetty Sriya
 23DSC28, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 muddamsetty.sriya@gmail.com

Patnala Meghana Durga
 23DSC04, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 meghanapatnala786@gmail.com

Abstract— Database security alludes to keeping unauthorized users from getting into the data set and to its core whether it is incidental or purposeful. Accordingly, every one of the organizations is giving uncommon consideration to potential dangers as stepping into database systems. CIA security triangle that notices the Confidentiality, Integrity, and Availability is normally holding the fundamental idea behind database security. Confidentiality intends to stay discreet. Integrity disappointment implies the information is adjusted and degenerate. Availability issues implies the information, or framework, or both cannot be accessed. Corporate companies should contribute time and exertion to distinguish and recognize the most genuine dangers. This research paper assesses existing explorations and research challenges on this specific area.

Keywords- Database, Database Technology, Security Technology, IT Management, Information Networking, Privacy and Security Management, Trust Management, Cloud Computing. Database Security, Threats, Breach, Access Control, Security Techniques.

I INTRODUCTION

Database security is a kind of collective measures which is required to protect as well securing the database from dishonest use. Moreover, Database Security protects malicious threats and attacks. Database Security is wide area and includes a multitude of processes, tools as well as methodologies keeps security within a database and allied environment [2], [7]. There should be proper tools and way to secure the database and complete storage systems. Today almost all types of organizations and institutions are using IT and there all the data are saved into the database; so proper and adequate database security are very important and required. There are different methods for securing database and among these important are include access control, access authorization etc. In a database there may be valuable and sensitive data. Therefore, it is important to secure these data. Both the user and owner

may demand the security of their data. Data Encryption before use of it is a one way of securing data. Searchable encryption has a growing interest and different ways of its implementations are developed recently. This paper describes challenges of database encryption. Lack of space for storage is becoming a major concern in today's technological world. Therefore, many organizations tend to outsource data storages in cloud. Because of that they tend to encrypt in a secure way before transmitting data (Song Dawn et al., 2015). Database Security can be mainly divided into three categories: physical database security, OS security and DBMS security (Prama nick N. & Ali S. T., 2017) (Mousa A., et al., 2020). Encryption is used to make data unavailable to unauthorized users. Symmetric key cryptography is mainly used for encryption. It encrypts and decrypts the same data with only one key. There are mainly two types of symmetric cyphers. They are stream cypher and block cypher. Stream cyphers are faster than Block cyphers, but it needs unique keys.



Fig 1. Database security threats

Common threats and challenges: Many software misconfigurations, vulnerabilities, or patterns of carelessness or misuse can result in breaches. The following are among the most common types or causes of database security attacks and their causes.

Insider threats: An insider threat is a security threat from any one of three sources with privileged access to the database:

- A malicious insider who intends to do harm
- A negligent insider who makes errors that make the database vulnerable to attack
- An infiltrator—an outsider who somehow obtains credentials via a scheme such as phishing or by gaining access to the credential database itself.

Insider threats are among the most common causes of database security breaches and are often the result of allowing too many employees to hold privileged user access credentials.

Human error:

Accidents, weak passwords, password sharing, and other unwise or uninformed user behaviors continue to be the cause of nearly half 49% data breaches.

Exploitation of database software vulnerabilities:

Hackers make their living by finding and targeting vulnerabilities in all kinds of software, including database management software. All major commercial database software vendors and open-source database management platforms issue regular security patches to address these vulnerabilities, but failure to apply these patches in a timely fashion can increase your exposure.

SQL/NoSQL injection attacks: A database-specific threat, these involve the insertion of arbitrary SQL or non-SQL attack strings into database queries served by web applications or HTTP headers. Organizations that don't follow secure web application coding practices and perform regular vulnerability testing are open to these attacks.

Buffer overflow exploitations: Buffer overflow occurs when a process attempts to write more data to a fixed-length block of memory than it is allowed to hold. Attackers may use the excess data, stored in adjacent memory addresses, as a foundation from which to launch attacks.

Malware: Malware is software written specifically to exploit vulnerabilities or otherwise cause damage to the database. Malware may arrive via any endpoint device connecting to the database's network.

Attacks on backups: Organizations that fail to protect backup data with the same stringent controls used to protect the database itself can be vulnerable to attacks on backups.

II RELATED WORK

Database security systems are made to guard against misuse, damage, and intrusion not just for the data in the database but also for the data management system and all applications that utilize it. Database security refers to the methods, procedures, and instruments used to create security inside a database environment.

Security threats can have various sources of origination such as Internal, External and Partner. Internal: These are the security threats sources that exist within the organisation like some company executives who have high access and privileges of the database. Internal sources enjoy certain levels of trust and privileges [19].

External: Sources outside the organisation pose as the external threats to database. Hackers, cybercrime groups and other government entities are some examples of external sources of threats. No trust or privileges are invested in external sources [7] [19].

Partners: These are the people outside the organisation that share business relationship with them. Customers, vendors, suppliers and contractors are a few examples of partner groups of organizations that can be a source of threat for the database. Since the communication. between both the parties are necessary for the functionality of business, moderate level of trust and privileges are associated with them [7].

On the basis of Data Breach Investigation Report (DBIR) 2016 [16], 2017[17], 2018[18] and 2019[19], where different origins of security threats such as Internal, External, Parties and Multiple parties [4] [5] were considered following chart has been derived. The conclusions made from the following graph [Figure 2] are:

There was an increase in the percentage of security threats originating from within the organisation from 11% in 2016 to 34% in 2019.

Threats originating from external sources decreased over the period of 2015 – 2019 from 86% to 69%.

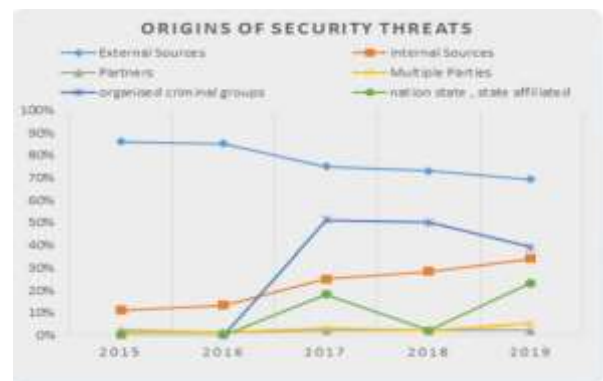


Fig. 2. Origins of Security Threats Chart

There are several security threats that can lead to data breach incidents. Top 10 threats over the past decade are:

A. Excessive and Inappropriate Privilege abuse: Database management systems and their corresponding data structures are complicated which makes administrators granting excessive rights to the users so as to prevent any application failure due to lack of rights [2]. When users are given privileges more than what is required for their job functionality, these privileges can be

used maliciously [12]. For example, course coordinator for any university is given the right to upload marks of every student. This privilege can be misused to change the marks of any student or any subject. This misuse is the result of granting generic access rights to a certain group of users even when it exceeds their specific job requirements [12].

B. Legitimate Privilege Abuse: This happens when a user is given only those privileges which are required by their job functionalities and these legitimate privileges are used for unauthorized purposes. User groups like Database System Administrator (DBA) and Developers have access to entire database due to their job requirements [2]. If a DBA tries to access the database data directly instead of application interface, all the application permissions and security mechanisms would be surpassed making the way for privilege abuse clear [2][7][12].

C. Privilege Elevation: Users with low-level privileges may use the vulnerabilities in the database to convert their access rights to high-level privilege. This can lead to the availability of critical information to unauthorized users [3] [12].

D. Platform Vulnerabilities: Any vulnerabilities in the underlying Operating System like Windows 2000, Windows XP, and Linux etc. can lead to privilege escalation, denial of service, data corruption and unauthorized access security threats [12]. For example, A potential security vulnerability in Intel WIFI Drivers and Intel PROSet/Wireless Wi-Fi Software extension DLL with severity rating as High was patched in November 2019 platform update. Memory corruption issues in Intel(R) WIFI Drivers before version 21.40 may allow a privileged user to enable escalation of privilege, denial of service, and information disclosure via local access [8].

E. Weak Audit Trails: Automated recording of any database transactions involving sensitive data should be a part of every database deployment. Failure to monitor transactions and collect audit details of database activities poses risk to the organization on many levels [2]. Many organizations rely on native audit tools provided by the database but the native audit tools do not record sufficient contextual information necessary to ensure security, detect attacks and provide incident forensics. Another reason native audit tools are not reliable is that users with administrative rights either legitimate or escalated can turn off database auditing to hide malicious activities [12] [14]. Therefore, database responsibilities and audit capabilities should be separate from both database server platform and database administrator to ensure strong separation of duties policy.

F. Denial of Service (DoS): This is a general category attack in which the legitimate users like employees, members or account holders are deprived of database services or resources which they require. This is done by shutting down the machine or the network making it inaccessible for its intended users. This can be done in two

ways either by flooding the destination with excess traffic or by sending them information that results in a crash [12]. Even though Dos doesn't directly result in data theft, loss or corruption they can cost a significant amount of time and money to handle. G. Unsecured Storage Media: Backup storage media is often less secured compared to the other database assets. This consequence to several high-profile data breaches involving theft or incidental exposure of database backup tapes and hard disks. Many regulations have made it mandatory to protect backup copies of sensitive data. One of the possible solutions to this is encryption of all the backup data [2] [7]. H. SQL Injection Attack (SQLIA) [1]: This is an attack which gives a potential attacker complete control over your database through the insertion of unauthorized or malicious SQL code in the database query.

III PROPOSED WORK

Authentication, or the system's way of confirming a user's identity, is one of the most fundamental ideas in database security. An authentication token, or proof of identification, can be given by a user in response to a request for authentication. Authorization is the second level of security that an authenticated user passes through. The authorization process is how the system learns details about the verified user, such as the database actions and data objects that the user is permitted to access. A safe system guarantees the privacy of information. This indicates that people can only view the data that they are intended to see. There are various facets of confidentiality, including communication privacy, sensitive data storage that is safe, user authentication, and authorization.

Database is the backbone of any organization. Therefore, it is important for the organization to implement any security solution. The security technique must ensure the safety of not only the data inside the system but also the database hardware, software and human resources. Database security techniques can be broadly classified into four categories, namely: Access Control, Techniques against SQLIA, Data Encryption and Data Scrambling.

A. Access Control (Mechanism): Data confidentiality can be ensured by using Access Control Mechanism. Most users are assigned or have authorized privileges to specific database resources and every time a user tries to access any data from the database, the access control mechanism will compare the required privileges to assigned privileges. Through this technique users can only access that data object for which they are adequately authorized. For example, for a university database teachers and students can be two categories of users with different access privileges. A student can only read grades and course offered and the teacher can update grades of students. A student can't make changes in the grades obtained whereas a teacher can't make changes in the courses offered.

- **Discretionary Access Control (DAC):** DAC grants or restricts the access to a data object based on an access policy created by the owner of data object. It is discretionary because the owner can transfer the authenticated objects and information access to other users.

- **Mandatory Access Control (MAC):** MAC allows a user to access a data object only when the authority level of the user matches the security level of needed data item.

- **Content Based Access Control:** In this model, the access control decisions are based on the contents of data objects [12]. For example, Employee table has salary details of all the employees of the organization. So only those employees of accounts department who are working on employee salary part should be able to access that data. This approach is implemented using views. Users are presented with the temporary view of the table with only those data they are authorized to access and not the complete table itself.

- **Fine Grained Access Control (FGAC):** General access control for database is coarse grained, i.e. it grants access to all the rows of the table or none at all. In contrast to this is fine grained access control that implements access control at the tuple level of the database. It enforces access control at the granular level. In this scheme each data object is given its own access control policy. This is implemented using specialization of views. Oracle Virtual Private Database (VPD) is one such database implementing FGAC.

B. Preventing SQLIA - Fighting Techniques are SQL injection attack gives complete control of our database to the attacker and thus it is one of the most dangerous security threats. The detection approaches for SQLIA can be categorized as

- **Pre-Generated:** Implemented during the testing phase of web application of database.

- **Post-Generated:** Used when the dynamic SQL generated by web application is analyzed.

- **Positive tainting and Syntax Aware Evaluation:** In this technique valid input strings are provided to the system initially to detect SQLIA. Positive tainting here means identifying, marking and tracking of trusted SQL queries and differentiating malicious queries from the legitimate ones using taint marking. Syntax aware evaluation allows us to actually use taint marking to identify trusted queries for the database. It allows the use of untrusted input data in a SQL query as long as it does not lead to SQLIA. Syntax evaluation of a query string is performed before the string is sent to the database for execution.

- **Context Sensitive String Evaluation:** It works on simple classification of data, User based data is considered as unreliable and data given by application is considered as reliable. Un-reliable data is then sent for syntax evaluation where string and numeric constants are differentiated from

each other and all unsafe characters are removed from the strings identified.

C. Data Encryption- The technique used to secure any kind of data or information can also be used to protect the data stored in database. Data encryption is a technique of transforming a plain text to intelligible form. This resulting information is known as encrypted data which can be converted back to its original form using encryption key. This technique can be used to secure the database by saving encrypted data in the database instead of plain text and converting the encrypted data to its original form when it is required for processing purposes. There are two different approaches to data encryption technique:

Symmetric Encryption: One common key is used for both encryption of data and decryption of data as well.

Asymmetric Encryption: Two keys are used, one key for encryption and the other for decryption.

D. Data Scrambling: It is a process of deliberately changing or removing the data saved in the database so as to make sensitive data safer for wider visibility. It is also known as Data masking, Data sanitization and Data obfuscation. It is used in the scenario where a user has access to a certain data but still the data needs to be protected from the user. For example, testers and third-party developers involved in working on the data in the database [1]. Even though they require working on the data but the actual values of data can be changed to hide the sensitive information. Basically, data is changed but the changed data resembles the actual data. Relationships between the columns in the original data would exist in the scrambled data as well. This way the actual sensitive information would be hidden from third party developers and they can still work with the data.

IV ANALYSIS

Hence, Database Security is about the protecting the databases such as data, such as the database applications including the stored functions; and complete database and similar systems. As far as Security is concerned few things are very common viz. Confidentiality, integrity and availability. Database security is affairs should be confirmed and measured by the following:

- Technical
- Procedural/administrative and
- Physical

There are different styles of protecting data can vary and increasing rapidly day by day. Among the possible, few important are include (but not limited to):

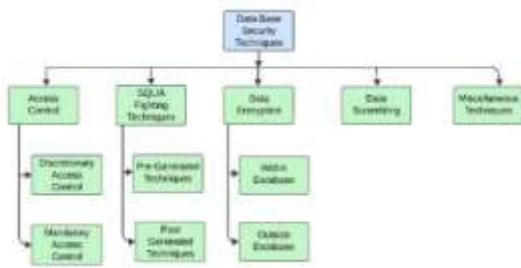


Fig. 3. Database Security Techniques

- Unauthorized activity by the authorized or unauthorized database users. And the users even may include the database administrators, systems managers, network administrators etc. Access of the hackers to the valuable, confidential and sensitive data.
- Unauthorized and similar activities within the databases or various changes in the repository or database data or the database programs including the structures as well as security configurations.
- Malware infections can be done with an unauthorized access as well as leakage of data proprietary data. Delete data or delete the data or programs. Moreover, sometimes the denial of authorized access may act as attacks on database systems and can lead the database failure.
- Overloads including the performance constraints are major database security issue. Moreover, the capacity issues resulting dad governance in database systems.
- Physical damage is another concern which may cause by the fires or floods. Even sometimes it may be based on overheating, lightning. Some expert expressed this as the electronic equipment failures as well.
- Design flaws (including the bugs in programming and simply in the databases programs can result the data loss as well as corruption even it may cause the performance degradation etc.
- Data corruption due to the entering invalid data and or commands, also mistake in the database can result weakness in database or repository systems. According to the NCC Group Study, UK following are the common database security issues (refer Fig: 2).

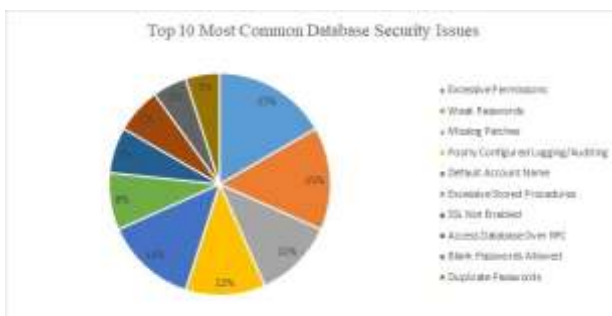


Fig. 4. Top 10 Most Common Database Security Issues

Database Security can be managed by different way and tools and among these few important are include as follows—

- Recovery of the Systems/ Database: Database Management System (DBMS) should offer backup facilities having the recovery of the concerned database after, if failed. Backup copies are need to maintain regularly in a perfect location as well.
- Minimizing unauthorized use of the database to the systems by ensuring strong as well as multifactor access including the data management controls.
- Load balancing and testing including the capacity testing id important in the database systems to ensure or user overload; it should ensure it is not crashing a DDoS attack.
- Physical security of the systems (i.e. Database) including the server; all types of backup tools, devices and equipment should be provided from the safety of natural disasters.

V FUTURE SCOPE AND CONCLUSION

The different organizations who create their own security guidelines and fundamental security controls for their database systems will find this review paper useful. They will be aware of the several threats that could jeopardize the database system and compromise its dependability. Future database security applications will make use of this review paper's cutting-edge technologies to support the design, implementation, and operation of data management systems that include security and privacy functions and provide reassurance that those systems are implemented in compliance with security and privacy regulations. All organizations and institutions this should take proper initiative and policies for the healthy, advanced and secure database system administration. Database administration plays a leading role in securing complete database systems and administration. A DBA is deals critical role in managing the databases of different kind of various organizations. Among the tools used by the database administrators to keep data safe and secure; few important are include—

- Controlling access to the database
- Providing the support services the client or users
- Managing proper strategies backup as well as recovery of data/ contents
- Ensuring data integrity
- Controlling data security
- Setting data privacy
- Formulation of IT Norms and Guidelines.

Moreover, the careful controlling access to the data helps in better data security as well as data privacy.

Databases and overall systems are normally facing a number of security threats. Many of these threats are common in small organizations but in case of big organizations and institutions vulnerability is very important as they hold sensitive information and different people and section used such. In Data loss, any parts of a

database can no longer be retrieved; so, for secure data security it is very important and valuable. Physical damage is an important concern in database security from the fire or water damage, human error or hardware failures etc. The norms and guidelines should be also an important matter for the healthy and sophisticated security practice as far as database systems is concerned. Database management systems provide an easy and efficient way to manage and manipulate data. Protecting data and the DBMS from any attacks is the goal with the highest priority for any organization. In this paper, we highlighted about the origins of security threats for an organization and how the pattern has shifted from external sources to internal sources over the period of last 5 years. Top 10 security threats to database were also highlighted in the paper along with the strategies which are being used to prevent the data attacks. Because the data kept in databases is often extremely sensitive and valuable, security is a crucial concern in database management. Thus, it is necessary to safeguard the data in a database management system from misuse as well as against illegal access and updates. The topic of potential vulnerabilities to database systems has been attempted to be explored in the database security paper. These consist of both integrity and confidentiality loss the article has also covered topics related to methods for dealing with threats by utilizing views and authentication. Using backup techniques, which guarantee that the data is kept somewhere else and can be restored in the event of an attack or failure, is a further strategy. Additionally, this paper has covered the numerous prerequisites that the database needs.

VI REFERENCES

- [1] Borgesius, F. Z., Gray, J., & Van Eechoud, M. (2015). Open data, privacy, and fair information principles: Towards a balancing framework. *Berkeley Technology Law Journal*, 30(3), 2073-2131.
- [2] Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS quarterly*, 34(3), 523-548.
- [3] George, B., & Valeva, A. (2006). A database security course on a shoestring. In *ACM SIGCSE Bulletin* (Vol. 38, No. 1, pp. 7-11). ACM.
- [4] Krishnan, V., McCalley, J. D., Henry, S., & Issad, S. (2011). Efficient database generation for decision tree-based power system security assessment. *IEEE Transactions on Power systems*, 26(4), 2319-2327.
- [5] Li, Y., Stewart, W., Zhu, J., & Ni, A. (2012). Online privacy policy of the thirty Dow Jones corporations: Compliance with FTC Fair Information Practice Principles and readability assessment. *Communications of the IIMA*, 12(3),5.
- [6] Mathieu, R. G., & Khalil, O. (1998). Data quality in the database systems course. *Data Quality Journal*, 4(1), 1-12.
- [7] Murray, M., & Guimaraes, M. (2008). Expanding the database curriculum. *Journal of Computing Sciences in Colleges*, 23(3), 69-75.
- [8] Murray, M. C. (2010). Database security: What students need to know. *Journal of information technology education: Innovations in practice*, 9, IIP-61.
- [9] Neto, A. A., Vieira, M., & Madeira, H. (2009). An appraisal to assess the security of database configurations. In *2009 Second International Conference on Dependability* (pp. 73-80). IEEE.
- [10] Said, H. E., Guimaraes, M. A., Maamar, Z., & Jololian, L. (2009). Database and database application security. *ACM SIGCSE Bulletin*, 41(3), 90-93. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, ISSN: 2581-6012, Vol. 4, No. 2, October 2019. SRINIVA
- [11] Sandhu, R. S., & Jajodia, S. (1993). Data and database security and controls. *Handbook of information security management*, Auerbach Publishers, 1-37.
- [12] Smith, G. W. (1991). Modeling security relevant data semantics. *IEEE Transactions on Software Engineering*, (11), 1195-1203.
- [13] Srinivasan, S., and Anup Kumar. (2005). Database security curriculum in InfoSec program. In *Proceedings of the 2nd annual conference on Information security curriculum development*, pp. 79-83. ACM.
- [14]. A.W Akanji, A.A. Elusoji and A.V. Haastrup, "A Comparative Study of Attacks on Databases and Database Security Techniques", *IEEE African Journal of Computing & ICT*, Vol 7. No. 5, ISSN: 2006-1781, December 2014, pp. 1-8,
- [15]. Emil Burtescu, "Database Security - Attacks and Control Methods", *Journal of Applied Quantitative Methods*, Vol. 4 No. 4 Winter 2009, pp. 449-454
- [16]. Verizon, 2016 Data Breach Investigations Report,
- [17]. Verizon, 2017 Data Breach Investigations Report 10th Edition,
- [18]. Verizon, 2018 Data Breach Investigations Report 11th edition,
- [19]. Verizon, 2019 Data Breach Investigations Report.
- [20]. Chris Brook, "What's the Cost of a Data Breach in 2019?", 30 July 2019,

Navigating The Digital Frontier: Technologies Shaping Digital Twins

Ravada Vanitha
 23DSC14, M.Sc. (Computational Data Science)
 Dept. of Computer Science P.B. Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 ravadavarsha@gmail.com

Dr.T. Srinivasa Ravi Kiran
 HoD & Associate Professor
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 tsravikiran@pbsiddhartha.ac.in

Kadali Anjani
 23DSC09, M.Sc.
 (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 anjanikadali296@gmail.com

Abstract-The digital transformation that is ongoing world-wide, and triggered by the industry 4.0 initiative, has brought to the surface new concepts and emergent technologies. One of these new concepts is the Digital Twin, which recently started gaining momentum, and is related to creating a virtual copy of the physical system, providing a connection between the real and virtual systems to collect and analyze and simulate data in the virtual model to improve the performance of the real system. The benefits of using the digital twin approach is attracting significant attention and interest from research and industry communities in the last few years, and its importance will increase in the upcoming years. Having this in mind, this paper surveys and discusses the digital twin concept in the context of the 4th industrial revolution, particularly focusing the concept and functionalities, the associated technologies, the industrial applications and the research challenges. The applicability of the digital concept is illustrated by the virtualization of a collaborative robot which used the V-REP simulation environment and the Modbus communication protocol.

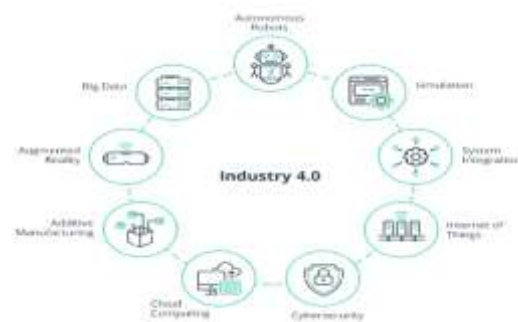
Keywords -Digital Twins (DT), Industry 4.0, IOT, Artificial Intelligence.

I INTRODUCTION

Industry 4.0 is largely concerned with digitalization and convergence of the real world with the virtual world and manufacturing pedagogy is facing a challenge now more than ever to produce workforce that can cope with this paradigm shift. Educators will need to adapt curriculum and teaching methodologies to help instill concrete understanding of the new trends and principles.



The Digital Twin can tackle the challenge of seamless integration between IoT and data analytics through the creation of a connected physical and virtual twin (Digital Twin). A Digital Twin environment allows for rapid analysis and real-time decisions made through accurate analytics. This paper provides a comprehensive review of Digital Twin use, its enabling technologies, challenges and open research for healthcare, manufacturing and smart city environments. Since the center of gravity of the literature relates to manufacturing application, the review has tried to capture relevant publication from 2015 onwards across three areas: manufacturing, healthcare and smart cities. The paper, uses a range of academic sources found through keywords related to IoT and data analytics, but with an overall aim of identifying papers relating to Digital Twin.



II WHAT IS A DIGITAL TWIN?

The origins of the Digital Twin are set out in this section. The review sets out clear definitions while also looking at some of the misconceptions found with wrongly identified Digital Twins. Formal ideas around Digital Twins have been around since the early 2000s.

That said, it may have been possible to define Digital Twins earlier owing to the ever-changing definitions.

Definitions: The first terminology was given by Grieves in a 2003 presentation and later documented in a white paper setting a foundation for the developments of Digital Twins. The National Aeronautical Space Administration (NASA) released a paper in 2012 entitled “The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles”, setting a key milestone for defining Digital Twins.

Since 1956, AI researches have succeeded in developing intelligent systems allowing machines doing not only all of the physical work, but also the reasoning, the predicting and the subsequent decision-making. Rather than trying to achieve a perfect replica of the human mind, AI systems exploit processes emulating human reasoning as a guide to provide both aiding tools and better services. For this reason, and thanks to the continuous advances in the computational power, in Big Data processing, and in the machine learning (ML) and pattern recognition (PR) fields, AI applications are becoming a fundamental part of our everyday life, providing surprising benefits in several fields. Examples are researches in the medical fields, where AI algorithms are developed with the aim of discovering novel biological relations and treatments. Similarly, AI algorithms modeling biological structures and human reasoning are integrated either to develop Computer Aided Diagnosis Systems, aiding clinicians during their everyday diagnostics procedures, or to study organs’ functioning and reaction to pharmacological treatments, eventually uncovering the hidden patterns and information encoded by the data, by reducing the data dimensionality to remove redundant information (closed-loop optimization) between the DT, its physical twin and the external, surrounding environment.

By using smart computer programs and analyzing a lot of data (Big Data), the digital twin learns and changes along with its real-world counterpart. It has a flexible structure that can quickly adapt. The digital twin is always kept in sync with its physical counterpart, and any changes in the real object are reflected in the digital version. With the help of artificial intelligence (AI), the digital twin can find information, discover patterns, and understand connections in the system it represents. It’s also capable of recording, controlling, and monitoring the conditions of the real system. This allows AI to predict and suggest solutions for potential issues, test the results of different solutions, and even activate self-repair mechanisms.

a) Nasa 2012:

“A Digital Twin is an integrated Multiphysics, multiscale, probabilistic simulation of an as-built

vehicle or system that uses the best available physical models, sensor updates, fleet history, etc., to mirror the life of its corresponding flying twin.”

b) Chen 2017

“A digital twin is a computerized model of a physical device or system that represents all functional features and links with the working elements.”

c) Liu et al. 2018

“The digital twin is actually a living model of the physical asset or system, which continually adapts to operational changes based on the collected online data and information, and can forecast the future of the corresponding physical counterpart.”

d) Zheng et al. 2018

“A Digital Twin is a set of virtual information that fully describes a potential or actual physical production from the micro atomic level to the macro geometrical level.”

e) Vrabic et al. 2018

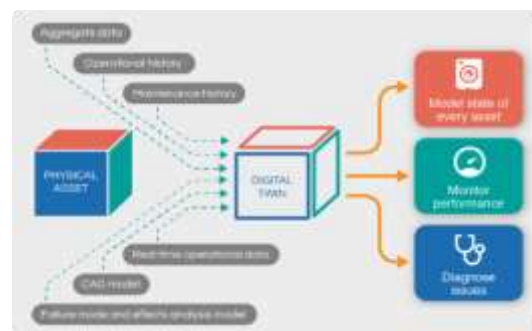
“A digital twin is a digital representation of a physical item or assembly using integrated simulations and service data. The digital representation holds information from multiple sources across the product life cycle. This information is continuously updated and is visualized in a variety of ways to predict current and future conditions, in both design and operational environments, to enhance decision making.”

f) Mandi 2019

“A Digital Twin is a virtual instance of a physical system (twin) that is continually updated with the latter’s performance, maintenance, and health status data throughout the physical system’s life cycle.”

Definition:

a) is an ambiguous definition specific for NASA’s interplanetary vehicle development and is one of the early papers that defines Digital Twins. Despite there being over six years between publications a) and f), the consensus remains that there is not a fundamental or meaningful change. Academia and industry alike have not helped in distinguishing DTs from general computing models and simulations. Future work requires a more definitive definition for a Digital Twin. This research aims to aid in the development of an updated definition, while also helping in analyzing related work and pointing out wrongly identified Digital Twins



B. DIGITAL TWIN MISCONCEPTIONS

1) Digital Model A digital model is described as a digital version of a preexisting or planned physical object, to correctly define a digital model there is to be no automatic data exchange between the physical model and digital model. Examples of a digital model could be but not limited to plans for buildings, product designs and development. The important defining feature is there is no form of automatic data exchange between the physical system and digital model. This means once the digital model is created a change made to the physical object has no impact on the digital model either way.

2) Digital Shadow A digital shadow is a digital representation of an object that has a one-way flow between the physical and digital object. A change in the state of the physical object leads to a change in the digital object and not vice versus. Figure 1. illustrates a Digital Shadow

3) Digital Twin If the data flows between an existing physical object and a digital object, and they are fully integrated in both directions, this constituted the reference “Digital Twin”. A change made to the physical object automatically leads to a change in the digital object and vice versa. illustrates a Digital Twin. These three definitions help to identify the common misconceptions seen in the literature. However, there are several misconceptions seen but they are not limited to just these specific examples. Amongst the misconceptions is the misconception Digital Twins have to be an exact 3D model of a physical thing. On the other hand, some individuals that think a Digital Twin is just a 3D model. this review presents a range of publications, highlighting the claimed level of integration against the actual integration based on the above definition. The definitions and figures should help in the development and identification of future Digital Twins.

III DIGITAL TWIN APPLICATIONS

The next part of this review focusses on the applications of Digital Twins. It will first start by looking at the potential applications for Digital Twins, discussing the domain, sectors, and specific problems for Digital Twin technology. For the moment the term and concept of a Digital Twin are growing across academia, and the advancements in IoT and artificial intelligence (AI) are enabling this growth to increase. At this stage, the primary areas of interest are smart cities and manufacturing with some healthcare-related applications of Digital Twin technology found.

1) Smart cities

The use and the potential for Digital Twins to be dramatically effective within a smart city is increasing year on year due to rapid developments in connectivity through IoT. With an increasing number of smart cities developed, the more connected communities are, with this comes more Digital Twins use. Not only this, the

more data we gather from IoT sensors embedded into our core services within a city, but it will also pave the way for research aimed at the creation of advanced AI algorithms.

2) Manufacturing

The next identified application for Digital Twin is within a manufacturing setting. The biggest reason for this is that manufacturers are always looking for a way in which products can be tracked and monitored in an attempt to save time and money, a key driver and motivation for any manufacturer. Thus, why Digital Twins look to be making the most significant impact within this setting. Likewise, with the development of a smart city, connectivity is one of the biggest drivers for manufacturing to utilize Digital Twins. The current growth is in line with the industry 4.0 concept, coined the 4th industrial revolution, this harnesses the connectivity of devices to make the concept of Digital Twin a reality for manufacturing processes

3) Healthcare

The healthcare sector is another area for the application of Digital Twin technology. The growth and developments enabling technology are having on healthcare is unprecedented as the once impossible is becoming possible. In terms of IoT the devices are cheaper and easier to implement, hence the rise in connectivity. The increased connectivity is only growing the potential application of Digital Twin use within the healthcare sector. One future application is a Digital Twin of a human, giving a real-time analysis of the body. A more realistic current application is a Digital Twin used for simulating the effects of certain drugs. Another application sees the use of a Digital Twin for planning and performing surgical procedures.

IV DIGITAL TWIN IN INDUSTRY

General Electric (GE) first documented its use of a Digital Twin in a patent application in 2016. From the concept set out in the patent, they developed an application called the “Predix” platform which is a tool for creating Digital Twins. Predix is used to run data analytics and monitoring. In recent years, GE has scaled back their plans for a Digital Twin, planning to focus on their heritage as an industrial multinational rather than a software company. Siemens, however, has developed a platform called “Mind Sphere” which has embraced the Industrial 4.0 concept with a cloud-based system that connects machines and physical infrastructure to a Digital Twin. It uses all the connected devices and billions of data streams with the hope of transforming businesses and providing Digital Twin solutions. An alternative platform for developing Digital Twin and AI technology is “Thing Worx”. This platform created by PTC is an Industrial Innovation Platform with the main focus of harvesting IIoT/IoT data and presenting via an intuitive, role-based user interface that delivers valuable insight to users. The

platform facilitates the smooth development of data analytics while also developing an environment for a Digital Twin solution. IBM developed a platform called “Watson IoT Platform” marketed as an all-round IoT data tool that can be used to manage large scale systems, in real-time, through data collected from millions of IoT devices. The platform has several add on features: cloud-based services, data analytics, edge capabilities and blockchain services. All of which makes this a possible platform for a Digital Twin system.

V DIGITAL TWIN CHARACTERISTICS

This section focuses on RQ2. By analyzing the selected papers, we have been able to identify the main characteristics that DTs are supposed to possess. Both the physical and the digital twins must be equipped with networking devices to guarantee a seamless connection and a continuous data exchange either through direct physical communications or through indirect cloud-based connections. Thanks to the seamless connection, the DT continuously receives dynamic (eventually sensed) physical twin data, which describe the physical twin status and change with time along its lifecycle, and dynamic environment data describing the surrounding environment status. Moreover, it continuously sends back to its physical twin, to the domain experts, and to other DTs in the environment, predictions and prescriptions for system maintenance and for function optimizations. There are mainly three types of communication processes that need to be designed:

- 1) Between the physical and the virtual twin.
- 2) Between the DT and different DTs in the surrounding environment.
- 3) Between the DT and domain experts, which interact and operate on the DT, through usable and accessible interfaces.

All the exchanged data must be stored in a data storage system, accessible by the digital twin. Together with dynamic data, the data storage contains historical static data, which reflect the physical twin memory and record historical information provided by human expertise or by past actions, and descriptive static data, which describe important characteristics of the physical twin that must not change over time (e.g. its requirements and constraints, in the case of a product or device).

VI PROPOSED WORK

DESIGN IMPLICATIONS

In this section, we provide an overview of design implications we derived from this study. In particular, we first illustrate the need of a sociotechnical and collaborative approach to the design process, and then we outline two different lifecycles that describe a DT’s life, from its design to its dismissal.

SOCIOTECHNICAL AND COLLABORATIVE DESIGN

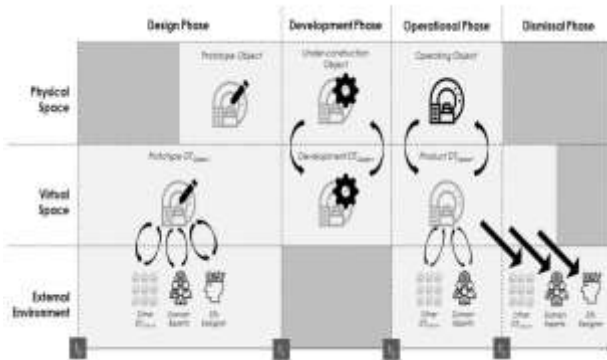
The variety, complexity and the increasing scale of DT design projects require all end users (also called “domain experts” because they are specialized in specific application domains different from Computer Science) to act in concert and collaborate in teams, by applying the respective knowledge to extend or modify the system. This allows satisfying needs and requirements that cannot be anticipated at design time. The need of finding new strategies to support such collaboration therefore becomes an open issue. The challenge is to bridge the communication gaps among stakeholders with diverse cultural and professional backgrounds. It is necessary to develop open-ended software environments that can be evolved and tailored in opportunistic ways to tackle the co-evolution of users and systems.

A sociotechnical design approach is needed to bridge the communication gaps raised during collaborative design activities. Such approach can be framed into Human Work Interaction Design (HWID), a lightweight version of Cognitive Work Analysis, addressing the concept of Work in Human-Computer Interaction.

LIFECYCLES

This study led us to describe two possible lifecycles for DTs, from their design to their dismissal. The former refers to a case where the object that has to be twinned still does not exist and, in this case, the design process simultaneously conceives both the object and its DT. The latter is about an object that already exists but has no DT in place; in this case, the design process focuses on the extension of the objects to make it connected.

Both lifecycles share the same timeline: a first Design phase, followed by a Development phase, an Operational phase, and finally a Dismissal phase. For describing these two lifecycles, we use a running example of a medical device (the object) – i.e. a computer tomography scanner. The first lifecycle is shown in Fig. 3. In this first case, the DT starts living before the physical object as a Prototype (Prototype DTObject), which is then used by designers during the Design phase of the Prototype Object. During the initial part of the Design phase, the Prototype DTObject is used, as if it was the real prototype, to simulate, test, change, and eventually validate design choices, until the best solution is found. During this part of the design phase, designers exploit: 1) Historical data the Prototype DTObject acquires from any other already existing DTs linked to similar devices. 2) Static data (e.g., data describing the product requirements, customer preferences, bill of materials). 3) The results of simulations performed by the Prototype DTObject, the result of predictions computed by the Prototype DTObject, and its suggestions and optimization schema.

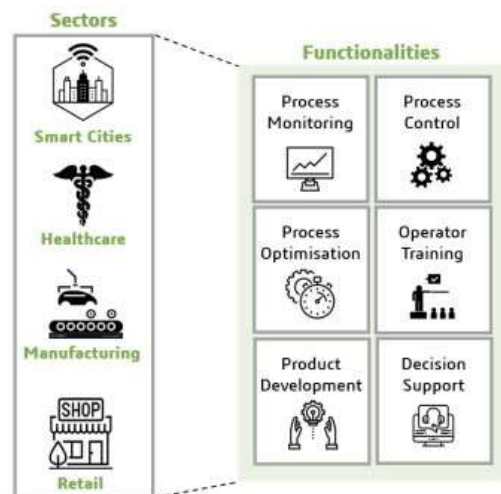


When the design of Prototype DTOBJECT is completed, the process moves to the Design of the prototype Object, during which the Prototype DTOBJECT is eventually modified to address technical constraints that may arise during the prototyping of the physical Object. During the Development Phase, the Prototype DTOBJECT evolves becoming a Development DTOBJECT, which must interact with the production machines to follow and optimize the assembly/construction of the Object, i.e. its physical twin. When the Object is finally built, the Development DTOBJECT starts being a Product DTOBJECT, and this moves the process into the Operational phase. The Product DTOBJECT fully resembles the Object: it has the AI acquired by the preceding stages of its life, and is therefore ready to follow and mirror the medical device (Object) while it is being used. During its existence, the intelligence of the DTOBJECT grows and self-adapts to the Object (in the case of the medical equipment, for example, it might start learning the most requested examinations, and the days when more or less examinations are performed). When the Object stops being used (due to obsolescence or any other reason) it must be disassembled, and the Dismissal phase begins, first for the Object and then for the DTOBJECT. The stored historical data of the Product DTOBJECT are backed-up and made available to other DTOBJECT as well as to domain experts; in this way, designers, or any other domain expert, will be able to use the collected information to optimize the production of future devices.

V EXISTING SYSTEMS

The DT concept started to be applied in several areas, such as manufacturing, smart cities, healthcare and retail, providing different functionalities. DT has also been studied in the areas of shipping oil and gas, constructions and agriculture. The application of DT in the manufacturing sector impacts the way the products are designed, manufactured and maintained. On a high level, the DT can evaluate the production decisions, access the product performance, command and reconfigure machines remotely, handle the troubleshoot equipment remotely and connect systems/processes to improve monitoring and optimize their control. As example, the

CNH Industrial company implemented a DT for a production line that combines simulation with the analysis of collected data. This DT was responsible for establishing new and more efficient operations in the company. The DT can also be applied for process control, process monitoring, predictive maintenance, operator training, product development, decision support, real-time analytics and behaviour simulation. To perform the process control, the DT uses the real-time and historical data to feed the virtual decision support system that will help the user to take strategic or operational decisions or implement directly the adjustment in the system operation. The DT also performs the remote process monitoring in real-time, as illustrated in. One of the main concerns that has been raised with the digital transformation, and particularly with the industry 4.0 initiative, is the role that the human will play in this new digital world. The use of DT can contribute for the integration of the human in CPS, particularly focusing on the operator training for complex operations. In fact, all manual tasks can be replicated in an in-line DT, which will allow the operators to improve their skills and confidence before handling the real system/device, e.g. the real machine. In the case of design and development of new products, there is a hard path, but with the design and simulation capabilities of the DT it will be possible to speed up the process. According to DT can speed up the commissioning process, is also responsible for performing machine optimisation, reducing the risk of project and its able to perform early fault detection.



APPLICATIONS AND FUNCTIONALITIES OF DT

The DT in smart cities includes the setup of a virtual replica of a city, more specifically a digital representation of urban networks of the city, such as urban power systems. This would allow to monitor different devices scattered throughout the city e.g., traffic lights or water pumps. The DT of the city could be used to develop prevention

strategies or to analyze in advance the impact of the introduction of new infrastructures. One example of this city DT is being developed in Singapore, where the virtual model of the city is being created with support of 3D maps of the city, aiming to increase the resilience of the city in unpredictable scenarios. The DT will allow to e.g., monitor the environmental conditions of the city or to plan emergency paths under critical situations. In the case of the healthcare industry, the DT can be applied in two perspectives, one more directed to the medical devices and another more related to the patients. Related to the medical devices, the DT will allow to identify maintenance needs before they even arise, to remotely monitor operations and to perform simulations on the devices before building the first prototype. In a futuristic perspective, it would be ideal to have a DT of a patient, which will provide the physician with the support on the diagnosis and treatment actions. The application of DT in the retail sector is still very recent, but offers a great potential in terms of consumer experience and marketing. The DT can be implemented for creating virtual models of the customers and modelling fashion for them, for creating virtual copies of the products and for identifying the consumer patterns of interest. As referred, the DT can be applied in a variety of sectors, but in manufacturing this technology will be indispensable to reach the factories of the future, taking advantage not only of the reduction of costs, but also in reducing resources, increasing productivity and effectiveness in the design of new products, and integrating faster the human operators.

VI OPEN ISSUES AND CHALLENGES

There are currently some important issues and challenges that need to be further studied and addressed, and are related to different aspects, all important for the future of the research in this field.

A. ETHICAL ISSUES

Developers must address the ethical issues raised by the exchange of data describing/being produced by/being analyzed by multiple sources, such as the manufacturing company exploiting the developed DTs, its partners and customers, or by twinned hospitals, clinical experts and patients. This requires developers and users to treat data according to privacy statements and legal limits that must still be set. This is especially true and critical with personal medical records. Indeed, especially in healthcare and medicine, access to high-quality data with high cardinality and containing enough variation will be crucial to opportunely train effective AI models. Such datasets could be formed as a mixture of publicly available data, data from clinical trials or from collaborations with hospitals, as well as some data from customers. Proper regulations should guarantee that all of these records are made anonymous and are only used with patients' consent.

B. SECURITY AND PRIVACY

Due to the usage of the IoT and cloud computing, any DT environment must be developed with a particular attention to robustness with respect to hacking and viruses. Hacking of private, confidential or valuable information could damage all the sources involved in the physical environment being twinned. Especially for DT technology in medicine and healthcare sectors, security and privacy should be deeply taken care of.

C. COST OF DEVELOPMENT

Developing a DT environment requires to reconsider and reconfigure the underlying software platform, as well as the hardware of production machines and their cloud/physical interconnection. This implies huge costs and might open the way to the spread of DT technologies only for large companies with the necessary capital and human resources.

D. GOVERNMENT REGULATIONS FOR MEDICAL DTS

Government needs to set regulations and rules establishing how predictive physiological and biological computational models can be validated and approved before any physician is willing to trust a diagnosis generated by a machine, or any patient is trusting the diagnostic evaluation of any expert analyzing a simulation on a virtual model. In other words, proper validation methodologies must be set to assess the credibility of computational models in biology and medicine [130]. Moreover, regulations should be set to define the extent of virtualization of the human being.

VII CONCLUSION AND FUTURE WORK

The launching of the industry 4.0 program contributed to a momentum in the adoption of the DT, being one of the digital transformation technologies of the upcoming years. This technology has evolved from the simple virtualization into a complex digital copy of real assets, being characterized by the combination of several technologies, such as optimization, real-time data and machine learning. This paper presents an overview of the DT concept, namely its origins and definitions, the technologies that can be used to construct a DT and the possible applications and research fields in which can be applied the various operating scenarios and reduction of costs by saving resources. An important conclusion from the survey of the DT applications is that currently, very few applications are reported in the literature can bring significant benefits, such as real-time monitoring, decision-support based on real data, simulation/optimization

VIII REFERENCES

- [1] H. Kagermann, W. Wahlster, and J. Helbig, "Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0," German National Academy of Science and Engineering (ACATECH), Tech. Rep., 2013

- [2] M. Grieves, "Digital Twin: Manufacturing Excellence through Virtual Factory Replication," white Paper, NASA, 2014
- [3] M. Grieves, "Digital Twin: Manufacturing Excellence through Virtual Factory Replication," white Paper, NASA, 2014.
- [4] E. Glaessgen and D. Stargel, "The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles," in 53rd AIAA Structures, Structural Dynamics and Materials Conference, (Honolulu, Hawaii), American Institute of Aeronautics and Astronautics, Apr. 2012.
- [5] Y. Chen, "Integrated and Intelligent Manufacturing: Perspectives and Enablers," *Engineering*, vol. 3, pp. 588–595, Oct. 2017.
- [6] Z. Liu, N. Meyendorf, and N. Mrad, "The role of data fusion in predictive maintenance using digital twin," in *Annual Review of Progress in Quantitative Nondestructive Evaluation*, (Provo, Utah, USA), p. 020023, 2018.
- [7] Y. Zheng, S. Yang, and H. Cheng, "An application framework of digital twin and its case study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 1141–1153, June 2018.
- [8] R. Vrabic, J. A. Erkoyuncu, P. Butala, and R. Roy, "Digital twins: Understanding the added value of integrated models for through-life engineering services," *Procedia Manufacturing*, vol. 16, pp. 139–146, 2018.
- [9] A. Madni, C. Madni, and S. Lucero, "Leveraging Digital Twin Technology in Model-Based Systems Engineering," *Systems*, vol. 7, p. 7, Jan. 2019.
- [10] D. Howard, "The Digital Twin: Virtual Validation in Electronics Development and Design," in 2019 Pan Pacific Microelectronics Symposium (Pan Pacific), (Kauai, HI, USA), pp. 1–9, IEEE, Feb. 2019.
- [11] A. Coraddu, L. Oneto, F. Baldi, F. Cipollini, M. Atlar, and S. Savio, "Data driven ship digital twin for estimating the speed loss caused by the marine fouling," *Ocean Engineering*, vol. 186, p. 106063, Aug. 2019.
- [12] J. David, A. Lobov, and M. Lanz, "Leveraging Digital Twins for Assisted Learning of Flexible Manufacturing Systems," in 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), (Porto), pp. 529–535, IEEE, July 2018.
- [13] T. DebRoy, W. Zhang, J. Turner, and S. Babu, "Building digital twins of 3d printing machines," *Scripta Materialia*, vol. 135, pp. 119–124, July 2017.
- [14] S.-K. Jo, D.-H. Park, H. Park, and S.-H. Kim, "Smart Livestock Farms Using Digital Twin: Feasibility Study," in 2018 International Conference on Information and Communication Technology Convergence (ICTC), (Jeju), pp. 1461–1463, IEEE, Oct. 2018.
- [15] G. Knapp, T. Mukherjee, J. Zuback, H. Wei, T. Palmer, A. De, and T. DebRoy, "Building blocks for a digital twin of additive manufacturing," *Acta Materialia*, vol. 135, pp. 390–399, Aug. 2017.
- [16] K. Sivalingam, M. Sepulveda, M. Spring, and P. Davies, "A Review and Methodology Development for Remaining Useful Life Prediction of Offshore Fixed and Floating Wind turbine Power Converter with Digital Twin Technology Perspective," in 2018 2nd International Conference on Green Energy and Applications (ICGEA), (Singapore), pp. 197–204, IEEE, Mar. 2018.
- [17] H. Pargmann, D. Euhäusen, and R. Faber, "Intelligent big data processing for wind farm monitoring and analysis based on cloud-technologies and digital twins: A quantitative approach," in 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), (Chengdu), pp. 233–237, IEEE, Apr. 2018.

Navigating the Future of Bank Loan Status Predictions-Using (Python)

Jyothika Sankar Narayanan
 23DSC15, M.Sc.
 (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of
 Arts & Science
 Vijayawada, A.P, India
 jyothika2122@gmail.co

Mounika Emani
 23DSC06, M.Sc. (Computational
 Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of
 Arts & Science
 Vijayawada, A.P, India
 mounikaemani2016@gmail.com

Devi Sri Ventrapragada
 23DSC19, M.Sc. (Computational Data
 Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 ventrapragadadevisri@gmail.com

Abstract—In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So, they can earn from interest of those loans which they credit. Bank's profit or a loss depends to a large extent on loans i.e. Whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model.

The data is collected from the Kaggle for studying and prediction. Logistic Regression models have-been-performed-and-the-different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results. Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loans but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

Keywords—loan, outlier, Prediction, component, accuracy

I INTRODUCTION

The project deals with the concept of “Bank Loan Status “. The main function of the Project is the ‘Credit risk’. In general Credit is defined as an “agreement between lender and a borrower concept. Credit also refers to an individual's or business's creditworthiness or credit history. In accounting, a credit may either decrease assets or increase liabilities as well as decrease expenses or increase revenue”. Whereas the credit risk is risk of default on a debt that may arise from a borrower failing to make required payments. In the first resort, the risk is that of the lender and includes lost principal and interest, disruption to cash flows, and increased collection costs. The loss may be complete/partial.

Credit Spread Risk: - Credit Spread Risk is typically caused by the changeability between interest rates and the risk-free return rate.

Default Risk: - When borrowers are unable to make contractual payments, default risk can occur.

Downgrade risk: -Risk ratings of issues can be downgraded, thus resulting in downgrade risk.

Variables Taken

Variables according to our project is: -

- >Loan ID
- >Customer ID
- >Loan status
- >Current loan amount
- >Term
- >Credit score
- >Annual income
- >Years in current job
- >Home ownership
- >Purpose
- >Monthly Debt
- >Years of credit
- >Months since last delinquent

- >Number of open accounts
- >Number of credit problems
- >Current credit balance
- >Maximum open credit

Variable	example	Description
Loan id	820000	Should be unique value
Customer id	87246	It should be unique
Loan status	77%paid	Fully paid / charged off
Current loan amount	10,52,100	Shows the loan amount amount they have to pay
Term	72%427	Shows term long term
Credit score	385	Shows the credit score of the person
Annual income	5,00,000	Shows the annual income of the person
Years in current job	5 years	Shows how many years he is being in his current job
Home ownership	50%	Shows that they have own house or they are staying in rent house
Purpose	40%	Shows for what purpose they are looking for loan
Monthly debt	3212.74	Monthly debt payments are any payments you make to pay back a creditor or lender for money you borrowed.
Years of credit	17.2	Shows the length of your history
Months since last delinquency	8	Shows how many times the payer is behind in the payments.
Number of open accounts	6	Shows that how many open that the payer have
Number of credit problems	1	Shows the lack of credit history
Current credit balance	228190	Shows the current credit balance
Maximum open credit	416718	Shows the maximum open credit

II PURPOSE OF POWER BI

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

Power BI is a Data Visualization and Business Intelligence tool by Microsoft that converts data from different data sources to create various business intelligence reports. It provides interactive visualizations using which end users can create reports and interactive dashboards by themselves.

The different components of Power BI Architecture:

1. Data Sources
2. Power BI Desktop
3. Power BI Service
4. Power BI Report Server

1. Data Sources:

Microsoft Power BI can supply data and information from

a wide array of sources and extends support to various kinds of files. The information can either be directly imported into Power BI or through a live service link. To import big information sets, users get two options:

- Power BI Premium
- Azure Analytics Services

Some of the most common data sources are XML, txt/CSV, Excel, JSON, Azure SQL Data Warehouse, SQL Server Analysis Services Database, Power BI service, and many others.

2. Power BI Desktop:

Power BI Desktop is used to create reports and visualize the data on a given dataset. It is the development tool for Power Query, Power Pivot, and Power View.

3. Power BI Service:

Power BI Service is an on-cloud platform that lets you share the reports that you made on Power BI desktop. You can also use it to collaborate with other users and even create new dashboards. The Power BI Service platform comes in three different versions:

- Free version
- Pro version
- Premium version

It is also known as Power BI Workspace and comes packed with features such as natural language Q&A and alerts.

4. Power BI Report Server:

Power BI Report Server is somewhat identical to Power BI Service. The difference is that the Power BI Report Server is an on-premise platform. When an organization does not want to store reports on the cloud (owing to data security), they go for the Power BI Report Server. If you wish to use this tool, you need to have a Power BI Premium license.

Benefits of Power BI:

1. Rich, Personalized Dashboards
2. No memory shortage
3. Advanced data services
4. Easy implementation
5. Extract hidden information
6. Accuracy

Advantages of Power BI over software:

Power BI Dashboard offers numerous benefits and advantages that make visualizing complex data easy. Here are some of them:

1. **Easy to use:** Power BI offers you a pretty user-friendly interface that makes the entire dashboard creation process easy. So much so that you do not even need to write a single line of code to create them. Just drag and drop the features, and you're good to go.
2. **Low Learning Curve:** As stated earlier, the Power BI dashboard requires no coding; it's easy to use and master. Also, Power BI was developed on the foundation of Microsoft Excel, which further lowers the learning curve when it comes to creating Power BI dashboards.

3. **Customizable Dashboards:** Power BI offers amazing customization when it comes to creating and sharing dashboards. You can create Power BI HR Analytics Dashboard to simplify the HR process, a Power BI for Banking Dashboard for analyzing finances, or Power BI Marketing Dashboard to determine the success of your campaigns.

4. **Cost-Effective:** Power BI is quite an affordable Business Intelligence solution. The desktop version of Power BI is free, and you can create immersive dashboards and reports both simple and complex without paying a penny.

5. **Q&A Function:** The Q&A function is probably one of the best advantages of the Power BI dashboard. The Q&A feature Power BI comes with allows you to ask questions using a natural language and get the information you want.

Fig 1: -Count of Loan Status by Loan Id
Count of Loan Status by Loan ID

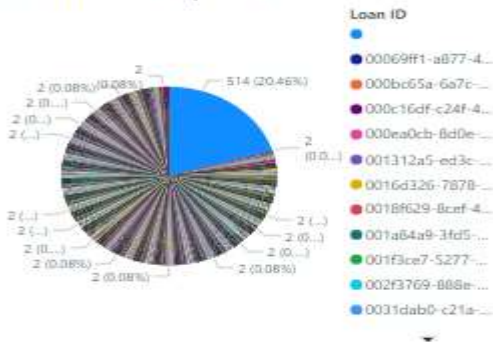


Fig 2: Representing Sum of Years of Credit History by Loan Status through Boxplot

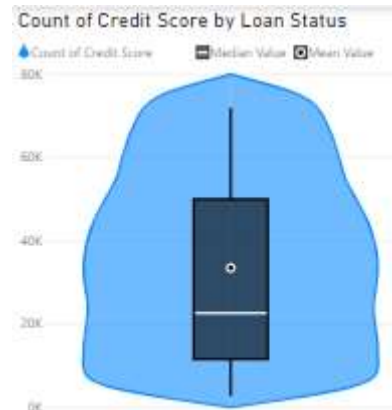
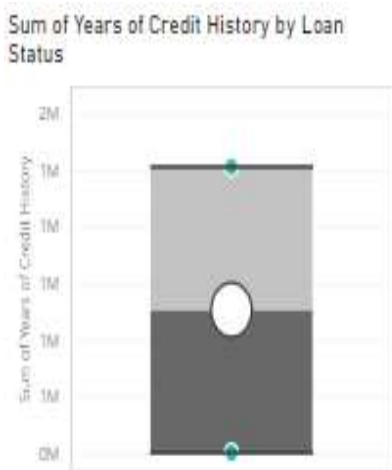


Fig 3: -Representing Count of Credit Score through violin plot

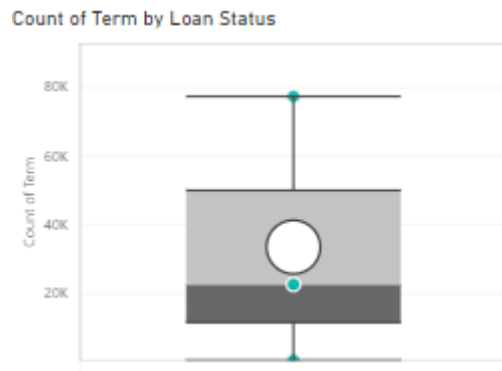


Fig 4:-Representing Count of term through box plot



Fig 5: -Dashboard

Python: -

Here I used the Python code to find the accuracy between the Train data and Test data.

Python Coding for Logistic regression Model

```

import pandas as pd
import numpy as np
import matplotlib as matlab
import stats models as sm
data1
  
```

```
pd.read_csv("https://raw.githubusercontent.com/drgvasu/
drgvasu/main/Data%20sets%20student%20projects/proje
ct9.csv")
```

```
data1.jinfo ()
```

```
O/p:
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 80846 entries, 0 to 80845
```

```
Data columns (total 19 columns):
```

#	Column	Non-Null Count	Dtype
0	Loan ID	80846	non-null object
1	Customer ID	80846	non-null object
2	Loan Status	80846	non-null int64
3	Current Loan Amount	80846	non-null int64
4	Term	80846	non-null object
5	Credit Score	80846	non-null int64
6	Annual Income	80846	non-null int64
7	Years in current job	77434	non-null object
8	Homeownership	80846	non-null object
9	Purpose	80846	non-null object
10	Monthly Debt	80846	non-null float64
11	Years of Credit History	80846	non-null float64
12	Months since last delinquent	38173	non-null float64
13	Number of Open Accounts	80846	non-null int64
14	Number of Credit Problems	80846	non-null int64
15	Current Credit Balance	80846	non-null int64
16	Maximum Open Credit	80845	non-null float64
17	Bankruptcies	80684	non-null float64
18	Tax Liens	80840	non-null float64

```
Dtypes: float64(6), int64(7), object (6)
```

```
memory usage: 11.7+ MB
```

```
Len (data1["Loan Status"])
```

```
80846
```

```
from sklearn.linear_model import Logistic Regression
```

```
logistic1= Logistic Regression (max_iter=200)
```

```
###fitting logistic regression for active customer on rest o
f the variables#####
```

```
logistic1.fit (data1[["Annual Income"] +
['Years of Credit History'] +
['Number of Credit Problems'] + ['Credit Score']],
data1[['Loan Status']])
```

```
print ("Intercept", logistic1.intercept_)
```

```
print ("Coefficients", logistic1.coef_)
```

```
Intercept [9.05142369e-08]
```

```
Coefficients [[ 1.40875261e-06 1.41883774e-06
```

```
1.84881084e-08 -6.06003634e-04]]
```

```
from sklearn import model selection
```

```
train_data, test_data = model_selection.train_test_split
```

```
(data1, test_size=0.2)
```

```
print ("train Data Shape", train_data.shape)
```

```
print ("test Data Shape", test_data.shape)
```

```
train Data Shape (64676, 19)
```

```
test Data Shape (16170, 19)
```

```
from sklearn.metrics import confusion matrix
```

```
logistic= Logistic Regression(max_iter=200)
```

```
###fitting logistic regression for active customer on rest o
f the variables#####
```

```
logistic2.fit (train_data [["Annual Income"] +
```

```
['Years of Credit History'] +
```

```
['Number of Credit Problems'] + ['Credit Score']], train
```

```
data [['Loan Status']])
```

```
predict=logistic2.predict(train_data [["Annual Income"] +
```

```
['Years of Credit History'] +
```

```
['Number of Credit Problems'] + ['Credit Score']])
```

```
cm_train = confusion matrix (train_data [['Loan Status']],
```

```
predict)
```

```
accuracy_train= (cm_train [0,0]
```

```
+cm_train[1,1])/sum(sum(cm_train))
```

```
print ("accuracy on train data", accuracy_train)
```

```
predict=logistic2.predict(test_data [["Annual Income"] +
```

```
['Years of Credit History'] +
```

```
['Number of Credit Problems'] + ['Credit Score']])
```

```
cm_test = confusion matrix (test_data [['Loan Status']],
```

```
predict)
```

```
accuracy_test= (cm_test [0,0]
```

```
+cm_test[1,1])/sum(sum(cm_test))
```

```
print ("accuracy on test data", accuracy_test)
```

```
Output: -
```

```
accuracy on train data 0.8377914527800111
```

```
accuracy on test data 0.8395794681508967
```

III OBJECTIVE: -

The objective of our project is " PREDICTING BANK LOAN STATUS", bank depend on the 4 independent factors to provide loan to the lender they are: -

- Annual income
- Credit score
- No of credit problems
- Years of credit history

Depending on these independent variables the dependent variable 'LOAN STATUS' can be changed. We can't provide a loan to the one who does not have the 'annual income' without the source of income. One can't repay the loan amount. Credit score gives us the idea about the person's financial stability about whether he can repay the loan amount or not. No credit problems gives us the idea about whether the lender has any credit problems such as that he won't repay the previous loan amounts.

IV CONCLUSION:

To analyze the data, we use Power BI. In this analysis we observed that, to predict the status of the bank loan we build a logistic regression model with influencing factors annual income, number of credit problems, Years of Credit History and credit score on dependent variable Loan status. Its mathematical model is

$$y = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}}$$

Using python, we estimate the parameter values as

Intercept is $\beta_0 = 9.05142369e-08$

$\beta_1 = 1.40875261e-06$

$\beta_2 = 1.41883774e-06$

$\beta_3 = 1.84881084e-08$

$\beta_4 = 6.06003634e-04$

To test the accuracy of the model we divide our data into Train data and Test data with 80% and 20% sizes. For the Train data we fit the same model and calculate the Accuracy. Accuracy on train data = 0.8377914527800111
 Now using Test data, we fit the model once again and get the Accuracy as

Accuracy on test data = 0.8395794681508967

Here accuracy of Test data and Train data are approximately equal. From this we can conclude that this

model is the best fit for our project data set. With this model, we predict whether the customer is likely to default payment or not with 83% of accuracy.

V REFERENCES: -

- [1] Mastering Microsoft Power BI by Devin knight and brain knight
- [2] Python programing language Fluent Python: Clear, Concise, and Effective Programming by Luciano Ramalho,
- [3] Regression Analysis with Python-Luca Massaron
- [4] <https://www.geeksforgeeks.org/python-programming-language/www>
- [5] <https://pythonbasics.org/>

Robotic Process Automation (RPA)

Shaik Nousheen
23DSC16

M.Sc. (Computational Data Science)
 Dept. of Computer Science P.B.
 Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 nshaik0311@gmail.com

Shaik Ayesha Begum
23DSC31

M.Sc. (computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 Shaikayasha21216@gmail.com

Vasudha Jonnala
23DSC03

M.Sc. (Computational Data Science)
 Dept. of Computer Science P.B.
 Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 vasudhareddyjonnala@gmail.com

Abstract—Robotic Process Automation (RPA) represents a transformative technological advancement in the field of business process automation. It involves the use of software robots or "bots" to mimic human interactions with digital systems and execute rule-based tasks across various applications. RPA offers organizations the ability to streamline repetitive, rule-driven processes, enhancing operational efficiency and reducing human intervention. This abstract explores the foundational concepts of Robotic Process Automation, including its key characteristics, benefits, and applications across diverse industries. It delves into the integration of RPA with artificial intelligence and machine learning, showcasing the evolution toward Intelligent Automation. Additionally, the abstract discusses challenges associated with RPA implementation and strategies for successful adoption.

Keywords—Robotic Process Automation, Human-Machine Collaboration, Digital Transformation, Task Automation, Rule-Based Tasks, Workflow Automation, Process Optimization, Cognitive Automation

I. INTRODUCTION

Robotic Process Automation (RPA) is a groundbreaking technology that revolutionizes the way organizations manage and execute their business processes. It involves the use of software robots or "bots" to automate repetitive, rule-based tasks traditionally performed by humans across various applications and systems. At its core, RPA aims to enhance operational efficiency by allowing these digital workers to mimic human interactions with software systems. [1] Unlike traditional automation solutions, RPA does not require extensive coding or integration efforts. [2] Instead, it operates at the user interface level, interacting with applications in the same way a human user would. Rule-Based Task RPA excels at automating tasks governed by clear rules and structured data, making it well-suited for repetitive processes. User Interface Interaction Bots interact with applications through the user interface,

enabling seamless integration with existing systems without the need for deep integration efforts. Scalability RPA facilitates scalable automation, allowing organizations to deploy multiple bots to handle increasing workloads and adapt to changing business requirements. [3] Efficiency and Accuracy By eliminating manual intervention, RPA enhances process efficiency, reduces errors, and ensures consistent and accurate task execution. RPA finds applications across various industries, including finance, healthcare, manufacturing, and customer service. [4] Common use cases include data entry, invoice processing, customer onboarding, and order fulfillment. As Organisations strive for greater agility, improved customer experiences, and operational excellence, Robotic Process Automation emerges as a transformative technology, reshaping the way businesses approach and execute their day-to-day operations.

II. LITERARY ANALYSIS OF ROBOTIC PROCESS AUTOMATION

Robotic Process Automation (RPA) has become a focal point in the discourse on digital transformation and business process optimization. The literature surrounding RPA encompasses a diverse range of perspectives, exploring its technological foundations, practical applications, and the broader implications for organizational efficiency.

- **Technological Foundations:** Scholars delve into the technical aspects of RPA, examining its underlying mechanisms, software architecture, and integration capabilities. This analysis provides a comprehensive understanding of how RPA operates at the intersection of automation and artificial intelligence.
- **Operational Efficiency and Cost Savings:** A significant portion of the literature focuses on the tangible benefits of RPA, emphasizing its role in enhancing operational efficiency and achieving cost savings. Studies and case analyses often highlight specific industries or business processes where RPA implementation has resulted in notable improvements.
- **Human-Machine Collaboration:** The concept of human-machine collaboration emerges as a recurring

theme in the literature.[5] Researchers explore how RPA can augment human capabilities, allowing employees to focus on higher-value tasks while digital workers handle routine, rule-based activities.

- Challenges and Considerations:** Literary analyses acknowledge the challenges associated with RPA adoption. Issues such as security concerns, ethical considerations, and the need for workforce upskilling are examined, providing a balanced view of both the advantages and potential pitfalls.
- Regulatory and Compliance Implications:** RPA's impact on regulatory compliance is scrutinized, particularly in industries with stringent regulatory requirements. The literature evaluates how organizations navigate compliance challenges and maintain a balance between automation and regulatory adherence.
- Future Trends and Implications:** Forward-looking analyses discuss potential future trends in RPA, considering advancements in technology, evolving business needs, and the role of RPA in shaping the future workplace. These insights contribute to a broader understanding of the long-term implications of RPA adoption.

In essence, the literary analysis of Robotic Process Automation offers a multidimensional perspective, ranging from technical intricacies to the broader societal and organizational implications.[6][7] It forms a knowledge base that informs practitioners, researchers, and decision-makers in navigating the evolving landscape of automation technologies.

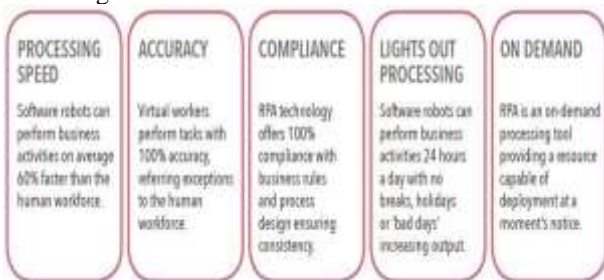


Fig. 1. Key Benefits of Robotic Process Automation

III. REAL-WORLD APPLICATIONS

Automation (RPA) span across various industries, providing tangible benefits in terms of efficiency, accuracy, and cost savings. Here are some notable real-world applications:

Finance and Accounting: RPA is extensively used in financial institutions for tasks like invoice processing, accounts payable and receivable, reconciliation, and fraud detection. Bots can handle repetitive financial processes with precision and speed.

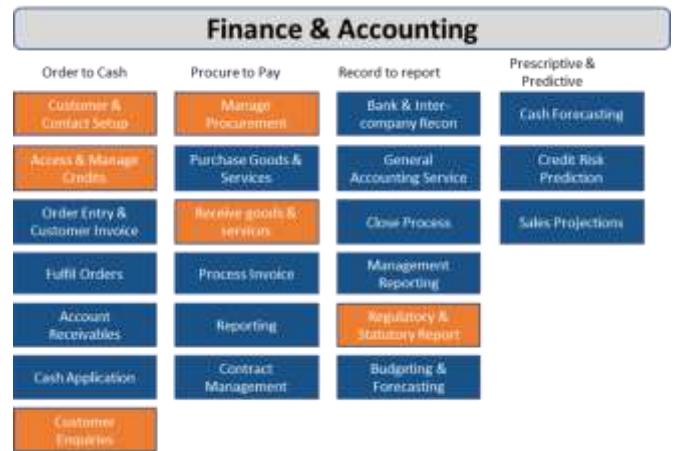


Fig. 2. Finance and Accounting in RPA

Human Resources: HR departments leverage RPA for activities such as resume screening, employee onboarding, payroll processing, and benefits administration.[4] This streamlines HR processes and allows human resources professionals to focus on strategic tasks.

Customer Service: In the realm of customer service, RPA is employed for tasks like data entry, order processing, and handling routine inquiries. [8] Bots can automate responses to common queries, improving response times and customer satisfaction.

HealthCare: RPA finds applications in healthcare for tasks like appointment scheduling, claims processing, and billing. Automation helps reduce administrative burdens, minimize errors, and enhance overall operational efficiency.

Supply Chain and Logistics: In logistics and supply chain management, RPA is used for inventory management, order processing, and tracking shipments.[9] Bots can automate updates across multiple systems, ensuring real-time visibility and accurate data.

Insurance: Insurance companies utilize RPA for claims processing, underwriting, and policy administration. Bots can extract and validate information from various documents, accelerating the claims settlement process.

Legal Services: In the legal sector, RPA assists with contract review, document management, and compliance tasks. Automation ensures that legal professionals can focus on.

IT Processes: RPA automates IT-related tasks, including software installation, system monitoring, and user account management. It helps in resolving routine IT issues, reducing downtime, and enhancing overall IT service delivery [10].

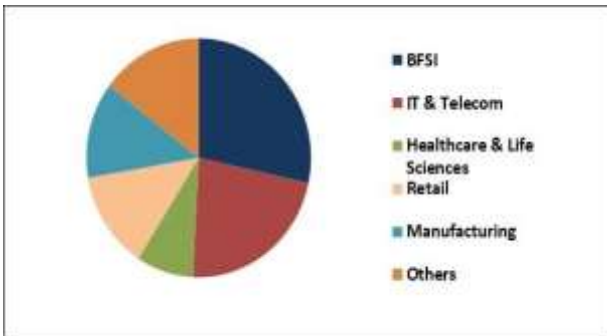


Fig .3. Robotic Process Automation Market Share

IV. WHAT FEATURES AND CAPABILITIES ARE IMPORTANT IN RPA TECHNOLOGY?

- **User-Friendly Interface:** An intuitive and user-friendly interface facilitates ease of use and allows users with varying technical backgrounds to design, configure, and manage automation workflows.
- **Bot Development and Design Tools:** Comprehensive bot development tools that support the creation of automation scripts and workflows, including features for drag-and-drop functionality, code generation, and process mapping.
- **Process Discovery and Documentation:** Capabilities for process discovery to analyze and identify automation opportunities within existing workflows. Documentation features aid in creating a clear understanding of processes before automation.
- **Integration Capabilities:** Robust integration capabilities to seamlessly connect with a variety of systems, applications, databases, and APIs. This includes integration with legacy systems as well as modern applications.
- **Scalability:** Advanced security measures, including encryption, secure access controls, and compliance with data protection standards, to ensure the protection of sensitive data processed by RPA bots [6].
- **Audit and Compliance Trails:** Logging and audit trail features to track and record actions performed by bots. This is essential for compliance, accountability, and troubleshooting purposes.
- **Exception Handling:** Capabilities for handling exceptions and error scenarios effectively, ensuring that automated processes can recover from unexpected situations without manual intervention.
- **Cognitive Automation and AI Integration:** Integration with artificial intelligence (AI) and cognitive capabilities for handling unstructured data, making decisions, and performing more complex tasks beyond rule-based automation [10].
- **Analytics and Reporting:** Robust analytics tools that provide insights into the performance of automated processes. Reporting features enable stakeholders to

track key metrics and assess the impact of automation on business outcomes.

- **Support for Multiple Environments:** Compatibility with various operating systems, cloud platforms, and environments to support a diverse IT landscape within organizations
- **Version Control:** Version control features that enable tracking, managing, and reverting to different versions of automation scripts, ensuring consistency and reliability in automated processes.
- **Monitoring and Management Dashboard:** A centralized dashboard for monitoring and managing the health, performance, and status of RPA bots and automated processes in real-time.
- **Comprehensive Documentation and Training Resources:** Well-documented resources, including training materials and documentation, to support users in learning and leveraging the capabilities of the RPA platform.
- **Community and Support:** Active community forums, support channels, and regular updates from the RPA provider to address issues, share best practices, and stay informed about the latest features and enhancements.

These features collectively contribute to the effectiveness, adaptability, and success of RPA technology in automating diverse business processes.

V. HOW AI AND RPA RELATE?

AI (Artificial Intelligence) and RPA (Robotic Process Automation) are related technologies that can complement each other, often combined to create more sophisticated automation solutions. Here's how they are related

- **RPA for Rule-Based Tasks:** RPA is effective for automating rule-based, repetitive tasks. It excels at mimicking human interactions with software applications and following predefined rules for processing data and performing routine activities.
- **Intelligent Process Automation (IPA):** The integration of AI and RPA is often referred to as Intelligent Process Automation. This combination leverages the strengths of both technologies to automate processes more intelligently and handle a broader range of tasks.
- **AI for Cognitive Abilities:** AI, on the other hand, provides cognitive capabilities. Machine learning, natural language processing, and computer vision are examples of AI technologies that enable systems to learn from data, make decisions, understand unstructured information, and adapt to changing circumstances.
- **Combining Rule-Based Automation with Intelligence:** By integrating RPA and AI, organizations can enhance automation solutions.[4] RPA handles structured tasks with clear rules, while

AI adds cognitive abilities to analyze unstructured data, make complex decisions, and adapt to dynamic scenarios.

- **Use Cases of Intelligent Process Automation (IPA):** Intelligent Process Automation can be applied to tasks such as invoice processing, document understanding, fraud detection, customer service, and other processes that involve a mix of structured and unstructured data.
- **Automation of End-to-End Processes:** While RPA is effective for automating specific tasks, AI extends the automation capability to handle end-to-end processes by understanding, learning, and adapting to variations and exceptions.
- **Continuous Improvement:** AI's ability to learn from data allows automation solutions to continuously improve over time. This is particularly valuable in dynamic environments where processes evolve, and decision-making requires adaptation.
- **Enhanced Decision-Making:** The combination of RPA and AI enables systems to make more intelligent decisions by processing and understanding complex information, allowing for a higher level of automation in decision-centric processes.

In summary, while RPA focuses on automating repetitive tasks based on predefined rules, AI brings cognitive capabilities to the table, allowing systems to learn, adapt, and handle more complex scenarios.[9] Together, they create a powerful synergy in Intelligent Process Automation, providing organizations with advanced automation solutions.

VI. RPA PROCESS MODEL

RPA process modeling involves creating a visual representation of the automated workflow. The process model serves as a blueprint for developing the automation scripts or bots.[10] Key elements of RPA process modeling include:

- **Process Mapping:** Create a step-by-step visual representation of the entire process to be automated. This includes activities, decision points, and interactions with various systems.
- **Data Flow:** Clearly define the flow of data within the process. Identify input sources, data transformations, and output destinations. Ensure that the data used by the bots is accurate and up-to-date.
- **Task Sequencing:** Arrange tasks in a logical sequence, considering dependencies and prerequisites. This ensures that the automation script follows a coherent and efficient path.
- **Decision Trees:** Represent decision points in the process using decision trees or flowcharts. [5] Define the conditions that determine the path the automation should take based on specific scenarios.
- **Input/Output Handling:** Specify how the automation process will handle input data and generate output.

Ensure that the bots can interact seamlessly with other systems and applications.

- **Exception Handling:** Incorporate mechanisms for handling exceptions or errors that may occur during the automation process. Clearly define the steps to be taken when the automation encounters unexpected situations [9][10].
- **Validation and Testing:** Develop a process for validating and testing the automation script. This involves running simulations, testing different scenarios, and ensuring that the bots perform as expected.
- **Continuous Improvement:** Implement mechanisms for continuous improvement by monitoring the performance of the automated process over time. This may involve refining the automation script based on feedback and evolving business requirements.

By conducting a thorough process analysis and creating a comprehensive process model, organizations can effectively plan, design, and implement successful RPA initiatives that align with business objectives and deliver tangible benefits.

VII. LIMITATIONS AND THREATS OF RPA

Despite its numerous advantages, Robotic Process Automation (RPA) also comes with limitations and potential threats. Understanding these aspects is crucial for organizations considering RPA implementation:

- **Lack of Cognitive Abilities:** RPA is rule-based and lacks true cognitive capabilities. It may struggle with tasks that involve complex decision-making, creativity, or understanding unstructured data.
- **Dependency on Stable Processes:** RPA works best with stable and well-defined processes. Changes in underlying systems or processes may require manual adjustments, limiting adaptability to dynamic environments.
- **Initial Implementation Costs:** While RPA can lead to long-term cost savings, the initial implementation can be costly. Organizations need to invest in software, training, and development before realizing the full benefits.
- **Need for Continuous Monitoring:** Automated processes require ongoing monitoring to ensure they function correctly. Any changes in the environment may necessitate adjustments to RPA workflows.
- **Inability to Handle Ambiguity:** RPA struggles with tasks that involve ambiguity or require subjective judgment. It may not be suitable for processes that require nuanced decision-making.

Threats:

- **Job Displacement Concerns:** The automation of routine tasks raises concerns about job displacement. While RPA is designed to complement

human work, organizations need to manage the impact on the workforce and address potential resistance.

- **Security Risks:**

RPA introduces new security considerations.[6] Bots interacting with various systems may pose a security risk if not properly configured and monitored. Ensuring secure access and data protection is essential.

- **Regulatory Compliance:**

Organizations using RPA must ensure compliance with data protection and privacy regulations. Automation processes need to adhere to industry standards and legal requirements.

- **Overreliance on Technology:**

Overreliance on RPA without strategic planning may lead to a "set and forget" mentality. Organizations must actively manage and update automated processes to avoid inefficiencies and errors.

- **Integration Challenges:**

Integration with legacy systems or complex IT environments can be challenging. Ensuring seamless interaction between RPA bots and existing systems requires careful planning and execution.

- **Resistance to Change:**

Employees may resist automation due to fear of job displacement or uncertainty about how their roles may evolve. Effective change management strategies are essential to address this resistance.

- **Limited Scalability in Some Scenarios:**

While RPA is generally scalable, certain processes may reach scalability limitations due to their complexity or the volume of exceptions they encounter.

Organizations should conduct thorough risk assessments and consider these limitations and threats when implementing RPA. Strategic planning, effective change management, and ongoing monitoring are crucial to maximizing the benefits of RPA while mitigating potential challenges.

VIII. CONCLUSION

Robotic Process Automation (RPA) stands as a transformative force, reshaping the landscape of business operations across various industries. As organizations strive for enhanced efficiency, accuracy, and agility, RPA emerges as a powerful tool, automating routine tasks and allowing human resources to focus on strategic, high-value activities.

The adoption of RPA brings forth a myriad of benefits. The reduction in operational costs, achieved through the automation of repetitive tasks, contributes to financial optimization. The newfound efficiency and accuracy in data processing result in streamlined workflows and improved decision-making. RPA's adaptability to diverse industries and its ability to seamlessly integrate with existing systems make it a versatile solution for businesses navigating complex and dynamic environments.

While the journey with RPA is marked by success stories, it's crucial to acknowledge the limitations and potential challenges. The lack of cognitive abilities in traditional RPA may limit its applicability to tasks requiring nuanced decision-making. The concerns related to job displacement underscore the importance of managing the impact on the workforce and fostering a collaborative relationship between humans and digital workers.

Looking ahead, the evolution of RPA extends beyond rule-based automation. The integration of Artificial Intelligence (AI) in Intelligent Process Automation (IPA) promises a future where systems not only mimic human actions but also possess cognitive capabilities. This synergy between RPA and AI represents the next frontier, opening new possibilities for handling complex tasks and adapting to evolving business landscapes.

In the grand symphony of organizational efficiency, RPA plays a key note, orchestrating a harmonious blend of automation, innovation, and human ingenuity. As businesses embark on the RPA journey, they are poised to unlock unprecedented potential, ushering in an era where processes are not only automated but intelligently optimized for the challenges of tomorrow.

IX. REFERENCES

- [1] Kampik, Timotheus, and Peter Hilton. "Towards Social Robotic Process Automation." SIAS Conference 2019
- [2] Aguirre, Santiago, and Alejandro Rodriguez. "Automation of a business process using robotic process automation (rpa): A case study." Workshop on Engineering Applications. Springer, Cham, 2017.
- [3] van der Aalst, Wil MP, Martin Bichler, and Armin Heinzl. "Robotic process automation." (2018): 269-272.
- [4] Zur Muehlen, M., & Ho, D. T. (2008, January). Service process innovation: a case study of BPMN in practice. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008) (pp. 372-372). IEEE.
- [5] Ratia, M., Myllärniemi, J., & Helander, N. (2018, October). Robotic Process Automation-Creating Value by Digitalizing Work in the Private Healthcare? In Proceedings of the 22nd International Academic Mindtrek Conference (pp. 222-227). ACM.
- [6] Cewe C., Koch D., & Mertens R. (2017, September). Minimal effort requirements engineering for robotic process automation with test driven development and screen recording. In International Conference on Business Process Management (pp. 642-648). Springer, Cham.
- [7] Tripathi, A. M. (2018). Learning Robotic Process Automation: Create Software robots and automate

business processes with the leading RPA tool–UiPath.
Packt Publishing Ltd.

- [8] Chappell D. (2017). *Introducing Blue Prism: Robotic Process Automation for the Enterprise* (2. utg.). San Francisco: David Chappell and Associates.
- [9] Smys, S., & Ranganathan, G. (2019). Robot assisted sensing, control and manufacture in automobile industry. *J ISMAC*, 1(03), 180-187.

Intrusion Detection Using Machine Learning

Shaik Obaid
 23DSC17, M.Sc.(Computational Data
 Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 obaidsk7865@gmail.com

Muddamsetty Sriya
 23DSC28, M.Sc.(Computational Data
 Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 muddamsetty.sriya@gmail.com

S.Tulasi Prasad
 Assitant Professor,
 Department of CSE,
 Potti Sriramulu chalavadi Mallikarjuna
 Rao College of Engineering &
 Technology
 stprasad@pscmr.ac.in

Abstract- An intrusion detection system is like a security guard for computer networks, keeping an eye out for any suspicious activity. With more and more computers connected, these systems are super important for network safety. People use fancy math and computer techniques to build these systems, and they need to be really good at spotting bad stuff without making mistakes. To make them better, researchers use different tricks. They look at loads of data from network traffic and use smart ways to organize it. They try out cool math tricks like Support Vector Machine and Naïve Bayes to see which one works best. They tested these tricks using a dataset called NSL-KDD. Turns out, the Support Vector Machine trick works better than Naïve Bayes. They checked how accurate these tricks are and how often they make mistakes to see which one is better at spotting problems in networks.

Keywords— *Intrusion Detection, Support Vector Machine Naive Bayes, Machine Learning.*

I INTRODUCTION

An intrusion detection system (IDS) helps spot weird stuff happening on a computer or network. There are two main types: one looks for known bad things, like a snort, and it's pretty good at catching those without making mistakes. But it can't find new kinds of bad stuff it doesn't already know about. The other type builds a normal behavior picture and then looks for anything that doesn't fit that picture, calling it a possible problem. This one can find both known and new bad stuff, but it often thinks something is wrong when it's not. People use fancy computer tricks to make it better at not making mistakes.

A. Intrusion Detection System

When someone sneaks into a computer or network and causes trouble by stealing or damaging information quickly, it's called an intrusion. This is a big problem for network security because it can harm the system's hardware too. People use different methods to try and spot these intrusions, but they often struggle with being accurate. They want to find a balance between catching the bad stuff and not raising too many false alarms. To help with this, they use special computer techniques like Support Vector

Machine and Naïve Bayes, which are good at sorting things out. They also use tricks like Normalization and Feature Reduction to compare different ways of spotting intrusions. The goal is to make sure they catch the bad stuff without causing too much unnecessary panic.

B. Machine Learning

Machine Learning is a way to teach computers to learn and make decisions on their own by looking at data. It's part of Artificial Intelligence and is all about training systems to spot patterns without too much help from people. There are two main types: one where the computer learns from examples that are labeled (like showing it what's right and wrong), and another where it figures things out from data that doesn't have labels. Sometimes, they use tricks like using a few labeled examples and lots of unlabeled ones. In other cases, it's like a trial-and-error game where the computer does things to get the best rewards. They use this for things like sorting stuff or predicting what might happen in the future. The main goal is to make the computer learn how to make good choices and reach its goals faster.

C. Support Vector Machine

Support Vector Machine (SVM) is special way computers learn from different types of information. It works by drawing lines or shapes to separate different groups of data in a very smart way. These lines or shapes help classify the information into different categories. There are special tricks called kernel functions that help SVM make these lines or shapes in a really effective way. The goal is to create these lines or shapes so that they are as far apart as possible, and there are different types of tricks to do this.

SVM is super useful in tasks like recognizing pictures and patterns. When using SVM for tasks like sorting things into categories, there are two main sets of data: one to teach the computer (training) and another to see how well it learned (testing). In these sets, the categories are called "target variables" and the information about those categories are called "features" or "observed variables."

D. Naive Bayes

Bayesian classifiers are like smart tools that guess the chance of something belonging to a certain group. They use a formula called Bayes' theorem to do

this. These classifiers work by assuming that when something belongs to a certain group, its different features don't really depend on each other. This idea is called class conditional independence. Essentially, they predict the likelihood of something fitting into a

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)}$$

particular category based on its features, assuming these features work independently from each other when making the guess.

II LITERATURE SURVEY

Protecting information on computers and networks is really important because if someone gets in and messes things up, it can lead to big losses. To stop this, we use something called an intrusion detection system (IDS). People are working on making these systems better by using different computer learning methods. The goal of a new IDS is to be really good at spotting both known and new kinds of attacks. It's made up of three main parts: one that organizes things (Clustering Manager), one that makes decisions (Decision Maker), and one that keeps everything updated (Update Manager). They tested this new system using a dataset called NSL-KDD. They used different ways of teaching [2].

A new model was created by using machine learning methods like SVM and Extreme Learning Machine (ELM) together. They used a smarter way to build a smaller but really good set of data from a big one, which helped the computer learn faster. They tested this model using a dataset called KDDCUP 1999 and found it was about 95.75% accurate. Different computer tricks like SVM, Random Forest, and ELM were tested to solve a problem. ELM did the best job in accuracy compared to the others. They tested these tricks using different amounts of data and found that SVM did well with smaller amounts, while ELM handled really large amounts of data the best. A new computer program that combines the ideas of Artificial Bee Colony and Artificial Fish Swarm was suggested to deal with information theft on computer systems. They used special techniques like Fuzzy C-Means Clustering and Correlation-based Feature Selection to pick out the most important parts of the data. By using these methods on different datasets, they achieved really good results in spotting unusual stuff happening in the computer systems. Another method used something called Correlation-based feature selection, [3] which is a simple way to find important things in the data. By applying this method to different datasets, they were able to detect almost all unusual things happening in the systems while hardly making any mistakes. They also combined different tricks to get really good at spotting problems without raising too many false alarms [6].

III EXISTING METHODOLOGIES

To check how well their new method works, they used a big dataset called KDD Cup. Instead of using the entire dataset, they took only 10% of it to train computers like SVM and ELM. This smaller set was used because using the whole thing would cause some problems, especially with certain kinds of information. They changed or removed some specific details like protocol, service, and flag. Then, they grouped all the examples into four categories: Normal, DoS, Probe, and R2L.

They trained SVM and ELM with this smaller dataset and tested their method using another corrected version of the KDD dataset. They found that their new method was pretty accurate, reaching about 95.75%, and had a low false alarm rate of 1.87%.

IV PROPOSED WORK

Dataset pre-processing, classification and result evaluation are the vital phases in the proposed model. In proposed system each phase is essential and enhances important influence on its performance. To examine the function of SVM and Naïve Bayes classifiers are the essential steps of this work.

A. Pre-Processing:

The data we use has symbols and stuff that computers can't easily understand, so we clean it up. We remove or change the non-number information to make it easier for the computer to learn. Things like protocol, service, and flags are changed or taken out. Then, we group everything into four categories: Normal, DoS, Probe, and R2L.

B. Methodology:

We want to see which method is better for finding bad things in our data: SVM or Naïve Bayes. First, we organize our data into different attack categories. Then we clean it up even more to make it easier for the computer to learn. We mix up the data randomly to avoid any patterns that might affect our results.

- Now, SVM is like a smart way for computers to learn from different kinds of data. It tries to draw lines or shapes to separate different groups of data. Naïve Bayes is a bit different; it's based on a formula that helps the computer guess which category something belongs to.
- After all this, we compare how well SVM and Naïve Bayes do in finding the bad stuff. We calculate how accurate they are and make a graph to see which one is better. From the graph, we see that SVM is better at this job than Naïve Bayes.
- The data we use has symbols and stuff that computers can't easily understand, so we clean it up. We remove or change the non-number information to make it easier for the computer to learn. Things like protocol, service, and flags are changed or taken out. Then, we group everything into four categories: Normal, DoS, Probe, and R2L.

The block diagram of this approach is given in “Fig. 1”. Accuracy has been calculated and a graph has been plotted based on the obtained results. From the graph, we have can analyze, SVM outperforms Naïve Bayes.

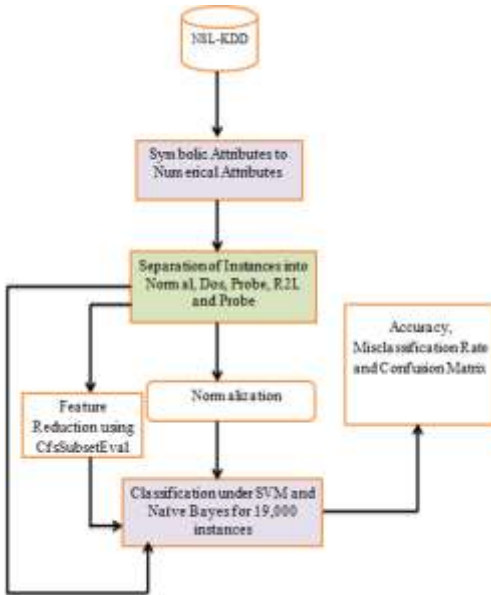


Fig.1. Block Diagram

NAIVE BAYES ALGORITHM

INPUT: Training Set T Predictor Variable P

OUTPUT: A group of datasets for testing.

STEPS:

1. Read the Training set T.
2. Calculate the conditional probability P for every class $H \leftrightarrow$ dependent class
 $X \leftrightarrow$ class variable
 $P(X|H) =$
 $P(H|X) * P(X)/p(H)$
3. Find the class with maximum probability.
4. Generate confusion matrix
5. Find Accuracy and Misclassification rate.

SUPPORT VECTOR MACHINE

INPUT: Preprocessed Data

OUTPUT: Output Classes

STEPS:

1. Calculate Objective Function T.
2. Objective function = $\min_w \lambda \|w\|^2 + \sum (1 - y_i(x_i, w))$ Where x_i is the input sample, y_i is the output label, W is weight vector, λ is regularization parameter
3. Apply gradient descent learning w.r.t weight
4. Update rule for weight for misclassified output

$$w = w + \eta (y_i x_i - 2\lambda w)$$

5. Update rule for weight for correctly classified output $w = w + \eta (i - 2\lambda w)$, where η is the learning vector
6. Return T
7. end function

V RESULTS & ANALYSIS

By analyzing accuracy rate and misclassification rate, the performance of SVM and Naïve Bayes algorithm has been evaluated for 19,000 instances. The performance metrics of these algorithms is evaluated by the information from confusion matrix

Methodology	Accuracy Rate	Misclassification Rate
SVM	97.29	2.705
Naïve Bayes	67.26	32.73
SVM-CfsSubsetEval	93.95	6.04
Naïve Bayes-CfsSubsetEval	56.54	43.45
SVM-Normalization	93.95	2.705
Naïve Bayes-Normalization	71.001	28.998

A. Evaluation

The model is evaluated based on NSL-KDD dataset, after applying methodologies like pre-processing and randomization. The dataset consists of 19,000 samples. Accuracy rate and Misclassification rate are taken as evaluation metrics.

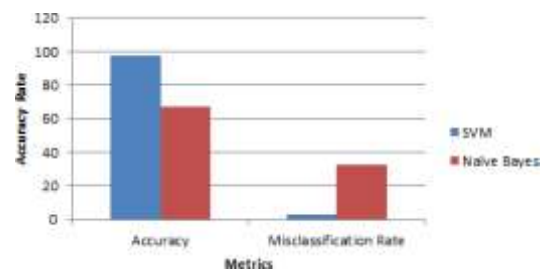


Fig 2. Accuracy and misclassification rate of SVM and Naive Bayes for 19,000 instances

The above graph describes the comparison of classification accuracy and Misclassification rate of the original dataset after preprocessing. From the graph it can be infer that SVM attains accuracy of 97.29 percentages and Naive Bayes attains accuracy rate of 67.26 percentages for 19000 instances. Naive Bayes has high Misclassification rate than SVM and for 19000 instances.

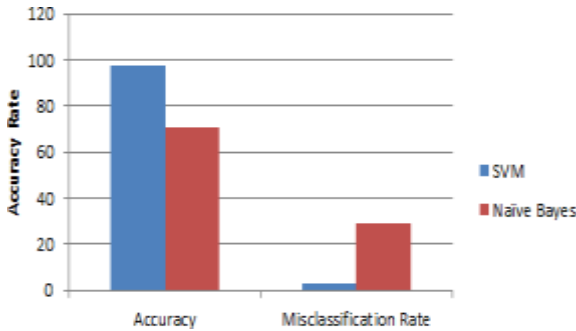
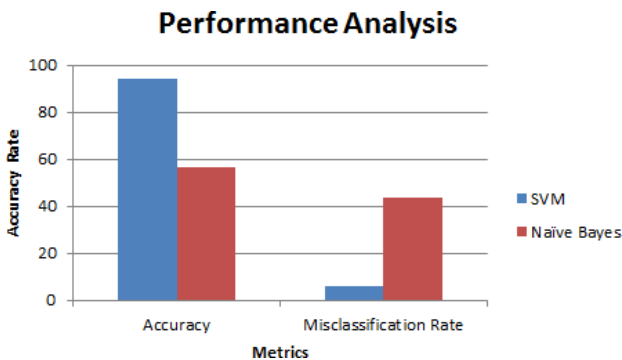


Fig 3. Accuracy and misclassification rate of SVM and Naive Bayes for 19,000 instances after Normalization

The above graph describes the comparison of classification accuracy and Misclassification rate of the dataset after Normalization. From the graph it can be infer that SVM attains accuracy of 93.85 percentages and Naive Bayes attains accuracy rate of 71.001 for 19000 instances. Naive Bayes has high Misclassification rate than SVM and for 19000 instances. the accuracy rate has been decreased for Naive Bayes for 19000 instances.



The above graph describes the comparison of classification accuracy and misclassification rate of the dataset after Feature reduction. From the graph it can be infer that SVM attains accuracy of 93.95 percentages and Naive Bayes attains accuracy rate of 56.54 for 19000 instances. Naive Bayes has high Misclassification rate than SVM for 19000 instances.

VI CONCLUSION

Detecting and stopping intrusions in our networks and systems is super important today. We've tried lots of methods to do this, and using machine learning like SVM and Naive Bayes is a big part of it. After checking with 19,000 examples, we found that SVM works better than Naive Bayes for this job.

VII FUTURE WORK

Future work deals with large volume of data, a hybrid multi-level model will be constructed to improve the accuracy. It deals with building a more effective model based on well-organized classifiers which are capable to categories new attacks with better performance

VIII REFERENCES

- [1] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," Knowl. -Based Syst., vol. 136, pp. 130–139, Nov. 2017.
- [2] Setareh Roshan, Yoan Miche, Anton Akusok, Amaury Lendasse; "Adaptive and Online Network Intrusion Detection System using Clustering and Extreme Learning Machines", ELSEVIER, Journal of the Franklin Institute, Volume.355, Issue 4, March 2018, pp.1752-1779.
- [3] Wathiq Laftah Al-Yaseen, Zulaiha Ali Othman, Mohd Zakree Ahmad Nazri; "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", ELSEVIER, Expert System with Applications, Volume.66, Jan 2017, pp.296-303.
- [4] Iftikhar Ahmad, Mohammad Basher, Muhammad Javed Iqbal, Aneel Raheem; "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", IEEE ACCESS, Survivability Strategies for Emerging Wireless Networks, Volume.6, May 2018, pp.33789-33795.
- [5] BuseGulAtli1, YoanMiche, AapoKalliola, Ian Oliver, SilkeHoltmanns, AmauryLendasse; "Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space" SPRINGER, Cognitive Computation, June 2018, pp. 1-16
- [6] Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu; "A Feature Reduced Intrusion Detection System Using ANN Classifier", ELSEVIER, Expert Systems with Applications, Vol.88, December 2017 pp.249-247
- [7] Vajiheh Hajisalem, Shahram Babaie; "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection", ELSEVIER, Department of Computer Engineering, Vol. 136, pp. 37-50, May 2018.
- [8] Karen A. Garcia, Raul Monroy, Luis A. Trejo, Carlos Mex-Perera and Eduardo Aguirre, "Analyzing Log Files for Postmortem Intrusion Detection", IEEE Transactions on Systems, Man, and Cybernetics, part C (Application and Reviews)42.6(2012), pp.1690-1704.

[9] R.M. Elbasiony, E.A. Sallam, T.E. Eltobely, and M.M. Fahmy, 'A hybrid network intrusion detection framework based on random forests and weighted k-means,' *Ain Shams Eng. J.*, vol. 4, no. 4, pp. 753–762, 2013.

[10] Hudan Studiawan, Christian Payne, Ferdous Sohel; "Graph Clustering and Anomaly Detection of Access Control log for Forensic Purposes", *ELSEVIER, Digital Investigation*, Vol. 21, pp.76-87, June 2017.

[11] Mazini, Mehrnaz, Babak Shirazi and Iraj Mahdavi; "Anomaly network-based intrusion detection

system using a reliable hybrid artificial bee colony and AdaBoost algorithms", *Journal of King Saud University-Computer and Information Sciences*, 2018.

[12] Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. "Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics*" 42(2), 513–529, 2012.

Automatic Pavement Crack Detection and Classification Using Multiscale Feature Attention Network

Seepana Nandini
 23DSC18, M.Sc. (Computational Data Science)
 Dept. of Computer Science P.B. Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 nandiniseepana11@gmail.com

Bora Uma Reddy
 23DSC21, M.Sc.
 (Computational Data Science)
 Dept. of Computer Science P.B. Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 umakrishna7620@gmail.com

Sandhya Naidu
 23DSC29, M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 sandhyanaidu879@gmail.com

Abstract-This study introduces an advanced approach for automatic pavement crack detection and classification employing a multiscale feature attention network. The challenges of non-uniformity, topological complexity, and noise in crack textures are addressed through a novel crack detection network with a multiscale dilated convolution module and an attention mechanism for refining high-level features. The proposed method further utilizes an up-sampling module to enhance detailed crack detection results. For crack classification, a characterization algorithm categorizes crack types (transversal, longitudinal, block, alligator), and severity levels are assessed by calculating average width and distance between branches. The proposed system achieves state-of-the-art accuracy, surpassing manual classification results, with transversal and longitudinal cracks exceeding 95% accuracy and block and alligator classifications remaining above 86%.

Keywords- Pavement crack detection, crack classification, convolutional neural network, multiscale feature extraction, attention mechanism.

I INTRODUCTION

The increasing demand for efficient and sustainable transportation infrastructure underscores the critical importance of effective pavement maintenance. Pavement deterioration, particularly in the form of cracks, poses a significant challenge to road integrity, safety, and overall functionality. Traditional manual inspection methods are not only time-consuming but also prone to human error, necessitating the development of automated systems to address these issues. The complexity of pavement crack detection and classification arises from the diverse nature of cracks, varying in size, orientation, and severity. Non-uniformity, topological intricacies, and noise within crack textures further compound the challenges faced by automated systems. In response to these complexities, this paper introduces a groundbreaking solution utilizing a Multiscale Feature Attention Network, a state-of-the-art

framework designed to enhance the accuracy and efficiency of pavement crack detection and classification.

II MATERIALS AND METHODS

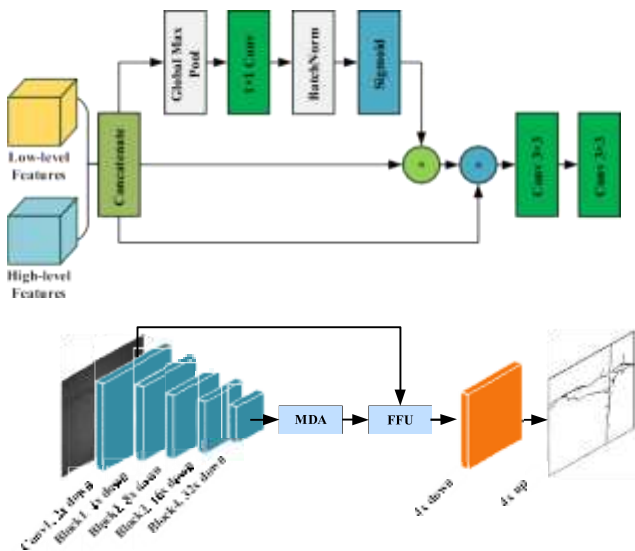
1)FEATURE FUSION UPSAMPLING MODULE:

Although the MDA module in the encoding stage could capture rich semantic features from the input image, these features have a coarse spatial resolution [29] and the purpose of this up sampling module is to restore these features to the input image resolution. Inspired by the decoder network in Deeplabv3 [39], the up-sampling module proposed in this study mainly contains two inputs: low-resolution features with discriminative semantic information generated by the MDA module, and high-resolution features in shallow layers. We therefore use different scales of extracted features to provide local and global context information. The up-sampling module first concatenates the low-level and high-level features, then uses batch normalization to balance the feature scales. Secondly, the weighted feature vector is calculated by using the attention mechanism similar to that in the MDA. This weight vector can re-select and combine the features, further try to refine the merged features, and improve the feature representation ability. Finally, the two 3x3 convolutions are continuously used to improve the feature representation and restore to that of the original input pavement image. The up-sampling module can use high-level and low-level hierarchical features to restore the positioning of crack pixels.

2)NETWORK ARCHITECTURE:

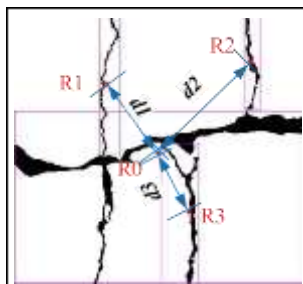
According to the proposed MDA and FFU modules, a crack detection network is developed as shown in Fig. 5. Given the input crack image, the ResNet [40] pre-training model is first used to extract the crack features. After the high-level features of deep neural networks are extracted, the MDA is employed to extract the crack features of multiple sizes under multi-scale. Then the semantic information at different levels is merged to obtain the global prior, which

is taken as the high-level feature of the network. Next, through fusing the low-level features generated in a shallow layer by the FFU module, the feature map size of the network output is consistent with the input image resolution, and finally, the probability that each pixel belongs to a crack or a non-crack is calculated.



CRACK CLASSIFICATION AND SEVERITY LEVELS

After detecting the pixel region containing cracks, the characterization of crack types based on the connected component labeling and the spatial distribution of each crack joint branch is investigated. The crack type is divided into transversal, longitudinal, block, and alligator types in this paper. We investigate the assignment of crack severity levels based on the average width of crack pixels and the distance between branch spaces. From this we propose a new algorithm for severity level assignment in this section.



1) CRACK CLASSIFICATION

Different from a simple transversal and longitudinal crack, the space distribution of netted type cracks is more complicated. To classify cracks as a whole, we merge the single extracted adjacent crack branches into a new target and then characterize the crack type as a whole. In our method, the crack connection analysis and classification are mainly divided into the following steps: firstly, the

connected component labeling is performed on the extracted crack binary images, and the cracks are divided into independent objects. To analyze the relationship between adjacent cracks, we generate a Minimum Enclosing Rectangle (MER) for each crack target. Each MER record consists of the target coordinates (x and y), width and height. Secondly, the centroid coordinate of each crack rectangle is calculated to obtain the distance between adjacent cracks. Then, these MERs are merged into a new one by determining whether the MERs are adjacent or intersecting, and then the distance between the two branches of centroids and the number of branches are calculated. An example of calculating the distance between the centroids is as shown in Fig., and R0 is adjacent to R1, R2, and R3, respectively. The average distance of the crack branches included in the generated new rectangle can be expressed as follows:

$$d = \frac{\sum_{n=1}^N d_n}{N} = \frac{d_1 + d_2 + d_3}{2}$$

where d_n refers to the distance between contiguous branches. Finally, if the number of branches is smaller than the branch threshold, the angle between the diagonal of the rectangle and the horizontal direction is determined as a transversal or a longitudinal crack. When the number of branches is greater than the branch threshold, the merged target is determined as a netted crack. Then, the block and alligator types are classified based on the average distance of recorded centroids. The characteristic threshold of the crack classification is shown in Table 1. The example of merged results of crack branches is shown in Fig.

TABLE 1. Crack classification feature threshold.

Crack Types	Angle	Numbers of Branches	Average Distance
Alligator	-	$n \geq 3$	$d > 0.5m$
Block	-	$n \geq 3$	$d \leq 0.5m$
Longitudinal	$\alpha < 45$	-	-
Transverse	$\alpha > 45$	-	-

TABLE 2. Type of pavement distresses and definitions.

Type	Severity level	Definitions	Weight
Alligator	Light	$W < 2mm, 0.2m < D < 0.5m$	0.6
	Medium	$2mm < W < 5mm, D < 0.2m$	0.8
	Heavy	$W > 5mm, D < 0.2m$	1.0
Block	Light	$1mm < W < 2mm, 0.5 < D < 1.0m$	0.6
	Heavy	$W > 2mm, 0.5 < D < 1.0m$	0.8
Longitudinal	Light	$W < 3mm$	0.6
	Heavy	$W > 3mm$	1.0
Transverse	Light	$W < 3mm$	0.6
	Heavy	$W > 3mm$	1.0

2) SEVERITY LEVEL ASSIGNMENT

According to different types of distresses, the damage severity of cracks can be divided into different levels. The severity of alligator cracks can be divided into light, medium, and heavy levels. The severity of block, transversal, and longitudinal cracks can be divided into light and heavy. Detailed classification criteria are shown in Table. The main feature for judging the damage degree is the average crack width W, and the distance D between the main crack branches, where D can be calculated by .

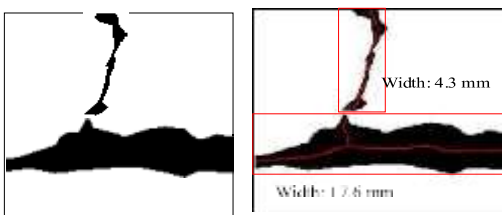
As the crack width can be measured at different locations, it is difficult to quantify its width. Similar to Oliveira and Correia, the average width of the crack is calculated at the pixel level. We can calculate the average width of the crack as follows:

$$Wcs = Wc \cdot Rc$$

where Wc is the total number of cracked pixels in the image and Ws is the total number of cracked pixels in the skeleton. Then, the average width Wcs of the crack is calculated according to the spatial resolution Rc of the pavement images. As shown in the two marked black areas are detection results, and the number of pixels is 473 and 4313, respectively; whereas the number of pixels after skeleton detection is 101 and 223, respectively. Based on the known image spatial resolution (for the data set herein, one pixel corresponds to an actual distance of 0.91 mm), from which the average width of the crack is calculated to be 4.3 mm and 17.6 mm, respectively.

Finally, the pavement crack damage rate DR can be calculated by the following formula:

$$DR = 100 \times \frac{\sum_{i=1}^n w_i A_i}{A}$$



where A_i is the area of the crack type i (m^2), A is area of the surveyed pavement surface (m^2), and w_i is the weight of i th crack type, which is valued as described in Table 2. i refers to the crack types, which contain severity levels (light, medium, and heavy), and i_0 is the total number of crack type

III EXPERIMENT AND ANALYSIS

EXPERIMENTAL SETUP:

1) Crack Dataset:

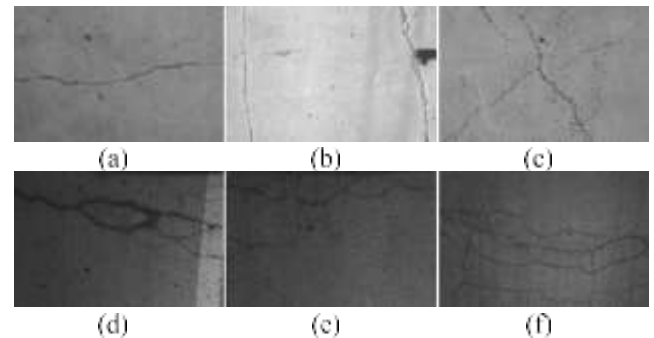
The dataset consists of pavement survey images from 14 cities in the Liaoning Province, China. The data employed in this study were primarily from plane array and charge-coupled device (CCD) cameras, which cover most of the road surface conditions and contains images from different roads and illuminations. The pavement image from the sensors has resolutions of 2330 1750 and 3120 2048. As the large size of pavement crack images, training our network with them would require a large amount of memory, resulting in overburdening of the training process. Additionally, the crack areas occupy only a small

proportion of the whole image, and the remaining background areas are useless for the training process. Therefore, we divided the original road crack images into several small blocks with a size of 256 256 pixels. Subsequently, a subset was manually labeled by human experts and taken as ground truth. The ground truth in the dataset provides two types of labels: cracks and non-cracks. We also divided the dataset into three parts, in which the training set and the validation set comprised 4736 and 1036 crack images, respectively, and the test set contained 2416 images. Furthermore, the dataset contained 300 hand-marked classification labeling results from human experts. Example of the crack images are shown in Fig. 8, where some of the images are accompanied by noise such as shadows, oil spots, and water stains; the cracks contained in the same image also have more complex topologies.

2) IMPLEMENTATION DETAILS:

We implemented our proposed crack detection network using TensorFlow, which is an open-source platform for deep learning. To improve the robustness of the model, several transformations were applied to the data, including

random flip, color enhancement, and enlargement. We also utilized the Adam optimizer to converge the network. The network was trained with an initial learning rate of 0.0001, and the momentum and weight decay were set to 0.9997 and 0.0005, respectively. All experiments in our work were performed using an NVIDIA GTX 1080 GPU and 8 GB of onboard memory.

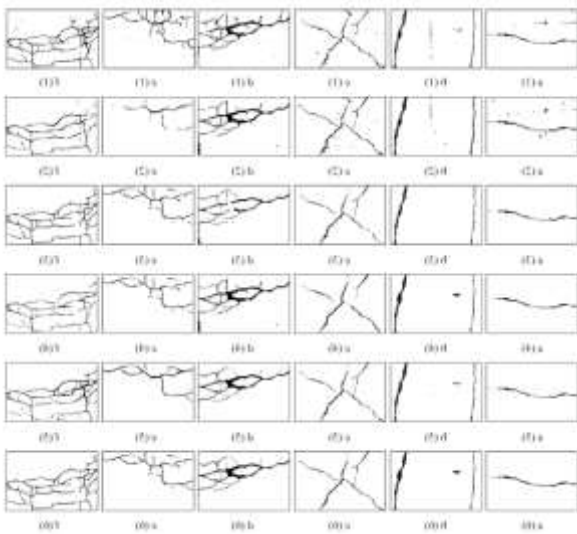


2) CRACK DETECTION RESULT:

The automatic crack detection network proposed in this study was compared with other state-of-the-art deep learning semantic segmentation models, including SegNet [17], U-Net [16], PSPNet [34], Deep Lab v3 (DL-v3) [39] and Discriminative Feature Network (DFN) [37]. Table 3 shows the quantitative comparison results of the test in the crack data set. Compared to the other deep learning-based segmentation methods, the crack detection network achieves the highest performance with Precision 97.70%, Recall 98.00%, F-score 97.34%, and mIoU 75.24%. In

detecting images with a resolution of 2330 1750, our network detected cracks at a rate of 0.71 s per image. PSPNet and DFN are faster, at 0.63 and 0.69 s per image, respectively. Conversely, SegNet, U-Net, and DL-v3 detect cracks at slower speeds of approximately 0.84 s, 1.16 s, and 1.30 s, respectively

Method	P (%)	R (%)	F (%)	mIoU (%)	Average Time
SegNet [17]	96.77	97.08	96.92	70.56	0.84s
U-Net [16]	96.99	97.09	97.04	71.49	1.16s
PSPNet [34]	96.90	96.88	96.89	69.63	0.63s
DL-v3- [39]	97.01	97.64	97.32	71.77	1.30s
DFN [37]	96.97	97.43	97.20	71.58	0.69s
Ours	98.74	98.05	98.39	74.81	0.71s



The crack detection visualization comparison results were performed on the six crack images (a-f) in Fig. 8, in which some images are affected by noise such as shadows, oil spots, and water stains; the cracks that are contained in the same image have complicated topologies and are shown in Fig. 9. From a (1) to f (1), in which the results of crack detection using the SegNet network are shown, a (1) and b (1) are incorrectly extracted with more shadows and stains, indicating that the method is sensitive to noise

3) CRACK CLASSIFICATION AND SEVERITY LEVELS:

This paper provides more detailed qualitative and global assessment results than other published crack classification methods. It is possible to divide multiple crack types simultaneously in one image and assign corresponding severity levels to them. In [31], although the crack type can be effectively divided, the classification criteria are relatively simple, and only the influence of the crack width on the severity is considered. In [41], according to the angle between the crack and the horizontal, the classification of longitudinal and transversal cracks can be classified. If there is a crack branch in the extracted image, no matter what angle it is, it will be considered a block crack. In addition, the method does not include assignment of

severity levels, and the corresponding category cannot be assigned a corresponding weight. In this study, according to the spatial distribution of cracks after the binary detection, the multiple crack targets in each image are discriminated, and the corresponding crack types are assigned: transversal, longitudinal, block, and alligator cracks. Crack classification and severity level assignment are mainly based on the number of crack branches, the average crack width, and the distance between branches. The classification accuracy is evaluated in the crack type test dataset. The crack classification results of the six images in Fig. 8 are shown in Fig. 10, and the MER is calculated for the combined cracks into independent crack types. The detailed classification results are shown in Table 4. In classifying crack types from 2330 1750 resolution binary images, our classification method can perform crack evaluation at an approximate rate of 0.5 s per image.

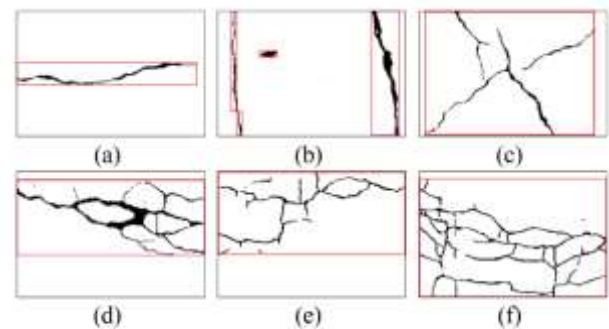
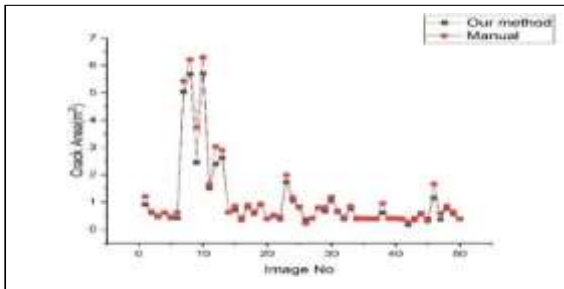


FIGURE 10. The results of automatic classification. Even in the case of fracture, the overall analysis of cracks can still be achieved.

To validate the influence of weight on the area measurement in the severity of crack characterization, 50 crack images were randomly selected in the crack classification test dataset and compared to the area of the hand-drawn crack area. The comparison results are shown in Fig. 11. These results illustrate that the area evaluated by the automatic detection method in this paper has a high degree of uniformity with the manually measured area, demonstrating that the crack classification method is highly reliable.

TABLE 4. Detailed crack classification results.

Image	Crack ID	Block	Transverse	Longitudinal	Alligator	Block	Crack Type	Severity Level	Time
001	1	1	0.77	-	0.8	0.8	Transverse	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
002	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
003	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
004	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
005	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
006	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
007	1	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s
	2	1	0.89	-	0.9	0.9	Longitudinal	Heavy	0.05s



IV DISCUSSION

In this section, the crack detection and classification results are discussed in detail for each module proposed. In this experiment, ResNet50 was used as the backbone network, and the effectiveness of the method was evaluated given the crack data set proposed in this paper.

CRACK DETECTION

MULTI-SCALE DILATED ATTENTION MODULE:

To compare the influence of the MDA module on the crack extraction more clearly, the feature map was up-sampled 16 times by simple bilinear interpolation in the up-sampling stage to obtain the final prediction result. In the experiment, the ResNet-50 network structure was used as the backbone network for verifying the multi-scale dilated convolution module, and several variant experiments were carried out on the multi-scale dilated convolution module. The experimental results are shown in Table 5, the mIoU of the baseline model using ResNet50 as the feature detection network is only 65.07%, the ASPP [29] module improves the performance of baseline by 1.25%, indicating that the dilated convolution improves the crack detection result. The dilation rate introduced here is an important hyperparameter that enables changing of the size of the receptive field of the MDA module in the network. To verify the ability of the dilated convolution mechanism to capture rich features, the dilation and convolution operations of three different dilated rates groups—{6,12}, {2,4} and {3,5}—were performed on the final high-level features, and the size of the receptive field increased. The experimental results show that the multi-scale dilated convolution module increased the mIoU by 1.40%, 1.23% and 2.05%, demonstrating that the advanced features with different dilation rates and convolution kernel sizes at multiple scales have stronger characterization capabilities and are better able to help locate cracked pixels during the encoding process. Although dilated convolution with a larger dilation rate, {6,12}, has a larger receptive field, it introduces other unrelated regions while capturing crack characteristics, which affects the final crack identification outcome. However, an overly small dilation rate, {2,4}, cannot effectively increase the receptive field. With the dilation rate {3,5}, better optimal convergence and better detection effect can be obtained in model training.

TABLE 5. Comparison results of different dilation rates.

Method	P (%)	R (%)	F-score (%)	mIoU (%)
Baseline	97.32	95.61	96.46	65.07
ASPP	97.41	96.29	96.85	66.32
MD {6,12}	97.64	96.37	97.00	66.47
MD {2,4}	97.53	96.31	96.92	66.30
MD {3,5}	97.85	97.41	97.63	67.12
With Attention	98.41	97.87	98.14	69.30

TABLE 6. Different upsampling features and structural comparison results.

Method	P (%)	R (%)	F-score (%)	mIoU (%)
DL-v3+ [39]	98.68	97.93	98.30	72.68
With attention	98.84	98.05	98.39	74.81

UPSAMPLE MODULE:

It can be observed from the results in Table 5 that the mIoU obtained from the simple bilinear interpolation up sampling method is only 69.30%. To further improve the crack detection performance, fusion of high-level semantic information and low-level features by the FFU module was introduced. The MDA features were used as the advanced input of the up-sampling module, as it has a stronger discrimination ability. The low-level features in the network have a higher spatial resolution in which crack edge detail information is preserved. After the low-level semantic information was merged with the high-level features with discriminative power, the convolution operation is then used in the up-sampling module to restore to the original size for obtaining finer segmentation results. The experimental results are shown in Table 6. The selection of different convolution times had a great influence on the final crack detection results of the model. The best effect was obtained by using two [3 3, 256] convolutions in Deeplabv3 [39]; the mIoU increased by 3.38% compared to the direct up sampling method, indicating that the combination of advanced features and low-level features can greatly improve the detection performance. To further refine the fused features, the attention mechanism is used to refine the feature representation in the feature fusion stage. Compared to the simple feature fusion method in Deeplabv3 [39], the mIoU value is increased by 2.13%. In summary, the FFU module is used to combine the shallow crack information with deep powerful semantic information, helping to fuse the multi-level features of the cracks and improve the overall crack detection accuracy.

TABLE 7. Crack classification results.

Crack type	Correct	Error	Accuracy	Severity level (No)
Transversal	132	7	95.0%	Light (20)
				Heavy (119)
Longitudinal	149	6	96.1%	Light (47)
				Heavy (108)
Block	51	7	87.9%	Light (17)
				Heavy (41)
Alligator	56	9	86.2%	Light (13)
				Medium (28)
				Heavy (24)

CRACK TYPE LABELING AND SEVERITY ASSIGNMENT:

For simple linear cracks, assigning the severity level of the detected crack segment depends on the measurement of the crack width, which is calculated as the ratio of the crack area to the number of cracked pixels in the skeleton. The severity level of damage is assigned as a light crack with a width of less than 3 mm, and a heavy crack with more than 3 mm width. The crack classification algorithm was verified in 300 crack classification samples; and Table 7 shows the crack classification system evaluation results.

The netted cracks include block and alligator cracks. In general, they have different crack densities and widths. However, due to the crisscrossing of the netted cracks, the width was difficult to calculate and measure, and the more intensive density of block cracks often exceeded the alligator cracks. Therefore, it was difficult to estimate the width and density of each block and alligator crack; thus, it is impossible to characterize the type and severity level accordingly. By further observing and comparing, it was found that the key difference between block and alligator cracks is that both of them break the road surface into pieces, but the former has fewer pieces and the distance between the branches is larger, whereas the latter has more blocks and the distance between the crack branches is smaller. From the perspective of intuitive perception and machine recognition, the block and alligator cracks can be effectively distinguished according to the branch distribution between the cracks. They also have practical physical meanings, such as the smaller the distance between the crack branches, the higher the severity level.

V CONCLUSION

In this paper, a novel trainable convolutional network was proposed for automatic detection of cracks in complex environments. In consideration of the different characteristics of different level features, we designed an MDA feature extraction module containing different dilated convolutions at multiple scales and a channel-wise attention module to capture the semantic high-level features. Then, crack pixel-level prediction is achieved by an FFU module that is combined with low-level features and continuous convolution. The experimental results show that both the MDA module and the FFU module contribute to the improvement of crack detection performance. Compared to other segmentation networks, our proposed crack detection network achieves state-of-the-art performance with Precision 98.74%, Recall 98.05%, F-score 98.39% and mIoU 74.81%. The experimental results indicate that our network is insensitive to noise crack marking and can effectively distinguish the low contrast caused by shadows, stains, and exposures during data acquisition.

Cracks were labeled according to the types defined in the Chinese distress category, with each different crack present in a given image receiving the appropriate label. Moreover, a novel methodology for the assignment of crack severity levels was introduced. The conclusion can be drawn from

the experiment that our method is a universal and robust automatic method to simultaneously determine the type of crack and severity levels—both of which are crucial for roadway agencies to assess pavement quality. For future developments in research, we will continue to investigate the influence of the attention mechanism on crack feature extraction. Like-wise, the crack classification algorithm will be optimized, especially for classification of the block and alligator cracks. Additionally, other types of distress, such as potholes and crack sealings will be taken into account to improve the procedure of automatic crack detection.

VI REFERENCES

- [1] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, and P. Fieguth, "A review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure," *Adv. Eng. Inform.*, vol. 29, no. 2, pp. 196–210, Apr. 2015.
- [2] H.-S. Yoo and Y.-S. Kim, "Development of a crack recognition algorithm from non-routed pavement images using artificial neural network and binary logistic regression," *KSCE J. Civil Eng.*, vol. 20, no. 4, pp. 1151–1162, May 2016.
- [3] A. Mohan and S. Poobal, "Crack detection using image processing: A critical review and analysis," *Alexandria Eng. J.*, vol. 57, no. 2, pp. 787–798, Jun. 2018.
- [4] A. Ahmadi, S. Khalesi, and M. Bagheri, "Automatic Road crack detection and classification using image processing techniques, machine learning and integrated models in urban areas: A novel image binarization technique," *J. Ind. Syst. Eng.*, vol. 11, pp. 85–97, Sep. 2018.
- [5] Q. Li, Q. Zou, D. Zhang, and Q. Mao, "FoSA: F* Seed-growing approach for crack-line detection from pavement images," *Imag. Vis. Comput.*, vol. 29, no. 12, pp. 861–872, Nov. 2011.
- [6] Q. Li and X. Liu, "Novel approach to pavement image segmentation based on neighboring difference histogram method," in *Proc. 1st Congr. Image Signal Process. (CISP)*, vol. 2, May 2008, pp. 792–796.
- [7] F. Liu, G. Xu, Y. Yang, X. Niu, and Y. Pan, "Novel approach to pavement cracking automatic detection based on segment extending," in *Proc. Int. Symp. Knowl. Acquisition Modeling (KAM)*, Wuhan, China, Dec. 2008, pp. 610–614.
- [8] R. Medina, J. Llamas, E. Zalama, and J. Gómez-García-Bermejo, "Enhanced automatic detection of road surface cracks by combining 2D/3D image processing techniques," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 778–782.

Forecasting Credit Risk a Comprehensive Analysis Using Advanced Models

V.Devi Sri
 23DSC19, M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts & Science, Vijayawada.
 ventrapragadadevisri@gmail.com

S. Jyothika
 23DSC15, M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts & Science, Vijayawada.
 jyothika2122@gmail.com

E. Mounika
 23DSC06, M.Sc. (Computational Data Science)
 Dept. of Computer Science P.B. Siddhartha College of Arts & Science, Vijayawada.
 Mounikaemani2016@gmail.com

Abstract—Credit risk is a significant focus in the banking and finance industry since evaluating the borrower's ability to repay a loan is crucial before extending credit. Also, in emerging nations, the underbanked population lacks access to the collateral and identification often necessary by banks before they will issue loans. This research study proposes a novel approach for predicting credit risk in financial institutions using ensemble machine learning models. The data is preprocessed, and relevant features are selected by evaluating the feature's importance using the information gain method. The first ten relevant features are selected for training the machine learning models. To predict credit risk, the suggested method used gradient boosting algorithms, including XG Boost, XG Boost RF, and CAT Boost. The proposed approach is compared with other state-of-the-art algorithms like Adaboost, Random Forest, and neural networks. Moreover, the findings prove that gradient-boosting algorithms like Xgboost and CAT Boost outpace other algorithms by achieving the highest training accuracy of 93.7% and 93.6%, respectively, and testing accuracy of 93.6% and 93.8%, respectively. While XG-Boost takes comparatively one-third of the time for training concerning the CAT Boost. Hence, the XG-Boost outperforms all the models regarding the accuracy and time trade-off. Hence, the proposed approach can be applied to financial institutions to provide credit to high-security borrowers.

Keywords—component, formatting, style, styling, insert

I INTRODUCTION

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make a break investment decision. The project deals with the concept of "Predicting credit risk". The main function of the project is credit risk. Credit risk is the possibility of a loss resulting from a borrower's failure to repay a loan. This project comes under the Banking industry.

Banks give credit, in return for using their services, banks pay clients a small amount of interest on their deposits. It is an agreement between banks and borrowers where banks make loans to borrowers. Bank credit is the total amount of funds a person or business can borrow from a financial institution. Bank credits may be secured or unsecured. These loans require the borrower to pledge collateral for the money being borrowed. In case if the borrower is unable to repay it then the bank reserves all the rights to utilize the pledged collateral to recover the pending payment. Unsecured loans are those that do not require any collateral for loan disbursement. The bank analyzes the past relationship with the borrower, the credit score, and other factors to determine whether the loan should be given or not. Education loans are financing instruments that aid the borrower pursue education. You must have the admission pass provided by the institution to get the financing. The financing is available both for domestic and international courses. The purpose of taking personal loan can be anything from repaying an old debt, going on vacation, medical emergency etc. Vehicle loans finance the purchase of two-wheeler and four-wheeler vehicles. Further the four-wheeler vehicle can be a new one or a use done. Home loans are dedicated to receiving funds in order purchase a house construct a house, renovate an existing house or purchase a plot for construction of house/flats. Gold loan is a secured loan taken by the borrower from a lender by pledging their gold articles as collateral. The loan must be repaid in monthly installments so the loan can be cleared by the end of the tenure and the gold can be taken back into custody by the borrower. The main object of the project "Predicting Credit Risk". With this predict for risky customer based their influencing factors age, monthly income, number of open credit lines and loans and number of dependents with dependent factor seriousDlqin2yrs.

II RELATED WORK

1.SeriousDlq in 2years-Person experienced 90 days past due delinquency or worse-Yes/No

2. Revolving utilization of unsecured lines-Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divides by the sum of credit limits-percentage
3. Age Age of borrowers in years-Integer
4. NumberOfTime30-59DaysPastDueNotWorse-Number of times borrower has been 30-59 days past due to no worse in last two years-Integer
5. Debt Ratio-Monthly debt payments-Percentage
6. Monthly Income Monthly Income Real
7. Number Of-Open CreditLines AndLoans Number of loans-Integer
8. NumberOfTimes90Dayslate Three months late frequency Integer
9. Number Real Estate Loans or Lines-House loans - Integer
10. Number of Time60-89DaysPastDueNotWorse-Two months late frequency-Integer
11. NumberOdDependents-Dependent excluding themselves-Integer

3. Bank Account Registration Form
4. Debit Card Application Form
5. Visa Credit Card Application Form
6. Credit Card Application Form
7. Account Closing Form
8. Bookkeeping Client Intake Form.

The most problems based on banking are increasing competition, Cultural Shift, Regulatory Compliance Changing Business Models, expectations, customer Retention, Outdated Mobile experiences, security breaches. The reality is bankers are fantastic problem solvers. It's in their blood. The problem is that, as an industry, we leave much to be desired about DIAGNOSING the right problem. Present a challenge to a banker, and they will quickly switch to solution mode without first analyzing what the issues are. They don't fully understand the "intent" of the customer and usually fail to understand the universe of solutions. This article presents the seven-step framework that we have found helpful in coaching banking teams in solving the right problem in the right way.

These are some strategies that help financial services managers meet the challenges of doing business in today's market:

- Attract and retain client Know your customer
- Promote confidence in the economy
- Use technology that customers expect
- Watch your reputation.

III PROPOSED METHOD

1) Statistical tools

Frequency tables: A frequency table lists a set of values and how often each one appears. Frequency is the number of times a specific data value occurs in your dataset. These tables help you understand which data values are common and which are rare. These tables organize your data and are an effective way to present the results to others. Frequency tables are also known as frequency distributions because they allow you to understand the distribution of values in your dataset. Frequency distribution tables are a great way to find the mode for datasets. For example, if 18 students have pet cats, cat's ownership has a frequency of 18. A frequency table of pet ownership will list various types of pets and their frequencies, including cats. You can make frequency tables for various types of data, including categorical, ordinal, and continuous. Categorical and ordinal data have natural groupings that you'll use in the frequency distribution. However, for continuous data, you need to create logical groups for the frequency distribution.

2) Pivot tables in excel:

3) Pivot table in excel is used to categorize, sort, filter, and summarize any length of data table which we want

S.no	Variable name	Description	Type
1	Serious Dign, yrs	Target variable (loan defaulter)	VN
2	Revolving Utilization of unsecured lines	Credit utilization	Percentage
3	Age	Age	Integer
4	Number of time 30-59 days past due not worse	One month late frequency	Integer
5	Debt ratio	Debt to income ratio	Percentage
6	Monthly income	Income	Real
7	Number of open credit lines and loans	Number of loans	Integer
8	Number of times 90 days late	Three months late frequency	Integer
9	Number of real estate loans or lines	House loans	Integer
10	Number of times 60-89 days past due not worse	Two months late frequency	Integer
11	Number of dependents	dependents	Integer

The Bank Management System (BMS) is a web-based application used for paying financial institutions for the services they provide to the Bureau of the Fiscal Service. BMS also provides analytical tools to review, and approve compensation, budgets, and outflows, the top 10 banking software tools rely on. NET, Python, Ruby, and Java. Also, there are specific technologies for core banking development: Oracle FLEXCUBE, Finastra, Temenos, etc.

Types of banking application:

1. Account Opening Form
2. W9 Form

to get count, sum, values either in tabular form or in the form of 2 column sets. To insert the pivot table, select

the Pivot table option from the Insert menu tab, which will automatically find the table or range. We can use the shortcut keys Alt+D +P simultaneously, which will detect the range of cells and take us to the final pivot option. We can also create a customized table by considering those columns which are actually required.

Cross tabs: Cross-tabulation is one of the most useful analytical tools and a mainstay of the market research industry. Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyze categorical (nominal measurement scale) data. Cross-tabulation (also cross-tabulation or crosstab) is one of the most useful analytical tools and a mainstay of the market research industry. Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyze categorical (nominal measurement scale) data. At their core, cross-tabulations are simply data tables that present the results of the entire group of respondents, as well as results from subgroups of survey respondents. With them, you can examine relationships within the data that might not be readily apparent when only looking at total survey responses.

4) Pictorial representation

5) **BAR GRAPHS:** Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars is proportional to the measure of data. They are charts. Bar graphs are one of the means of data handling in statistics. The collection, presentation, analysis, organization, and interpretation of observations of data are known as statistics. The statistical data can be represented by various methods such as tables, bar graphs, pie charts, histograms, frequency polygons, etc.

1. **Pie chart:** The "pie chart" is also known as a "circle chart", dividing the circular statistical graphic into sectors or sections to illustrate the numerical problems. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie-chart works the best at that time. In most cases, pie charts replace other graphs like the bar graph, line plots, histograms, etc.

The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total (percentage). The sum of all the data is equal to 360°. The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total (percentage). The sum of all the data is equal to 360°.

2. Histogram: A histogram is an approximate representation of the distribution of numerical data. The term was first introduced by Karl Pearson. To construct a histogram, the first step is to "bin" (or "bucket") the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be

adjacent and are often (but not required to be) of equal size. If the bins are of equal size, a bar is drawn over the bin with height proportional to the frequency—the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies showing the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1. However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency

POWER BI —

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

Power BI is a Data Visualization and Business Intelligence tool by Microsoft that converts data from different data sources to create various business intelligence reports. It provides interactive visualizations using which end users can create reports and interactive dashboards by themselves.

The different components of Power BI Architecture:

1. Data Sources
2. Power BI Desktop
3. Power BI Service
4. Power BI Report Server

DATA SOURCE

Microsoft Power BI can supply data and information from a wide array of sources and extends support to various kinds of files. The information can either be directly imported into Power BI or through a live service link. To import big information sets, users get two options:

Power BI Premium Azure Analytics Services

Some of the most common data sources are XML, txt/CSV, Excel, JSON, Azure SQL Data Warehouse, SQL Server Analysis Services Database, Power BI service, and many others.

POWER BI DESKTOP

Power BI Desktop is used to create reports and visualize the data on a given dataset. It is the development tool for Power Query, Power Pivot, and Power View.

POWER BI SERVICE

Power BI Service is an on-cloud platform that lets you share the reports that you made on Power BI desktop. You can also use it to collaborate with other users and even create new dashboards. The Power BI Service platform comes in three different versions:

- Free version
- Pro version
- Premium version

It is also known as Power BI Workspace and comes packed with features such as natural language Q&A and alerts

Benefits Of POWER BI

1. Rich, Personalized Dashboards
2. No Memory shortage
3. Advanced data services
4. Easy implementation
5. Extract hidden information
6. Accuracy

PYTHON

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant-indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming. Python is commonly used for developing websites and software, task automation, data analysis, and data visualization. Since it's relatively easy to learn, Python has been adopted by many non-programmers such as accountants and scientists, for a variety of everyday tasks, like organizing finances

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Concepts of Learning is the process of converting-experience-into-expertise-or-knowledge. Learning can be broadly classified into three categories, as mentioned

below, based on the nature of the learning data and interaction between the learner and the environment.

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning

Similarly, there are four categories of machine learning algorithms as shown below

- Supervised learning algorithm
- Unsupervised learning algorithm
- Semi-supervised learning algorithm
- Reinforcement learning algorithm

Here is the list of commonly used machine learning algorithms that can be applied to almost any data problem –

- Linear Regression
- Logical Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting like GBM, XGBoost, LightGBM and CatBoost

Linear Regression

Linear regression is used to estimate real world values like cost of houses, number of calls, total sales etc. based on continuous variable(s). Here, we establish a relationship between dependent and independent variables by fitting a best line. This line of best fit is known as regression line and is represented by the linear equation $y = \beta_0 + \beta_1 x$.

In this Equation

- Y – Dependent Variable
- β_1 – Regression Coefficient
- X – Independent variable
- β_0 – Intercept

Logistic Regression

Logistic regression is another technique borrowed by machine learning from statistics. It is the preferred method for binary classification problems, that is, problems with two class values

$$y = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

IV. SIMULATION RESULTS AND ANALYSIS

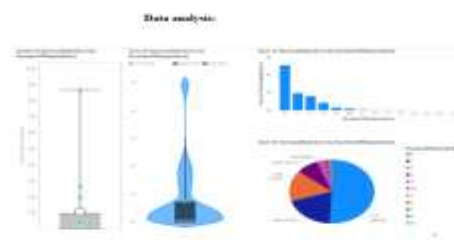


Fig 1. Number of depositors distribution

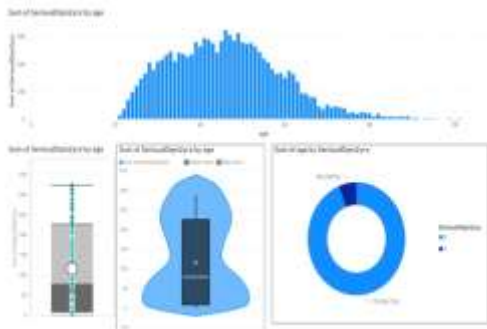


Fig. 2: Age distribution



Fig. 6: Average loan defaulter distribution by number of loans, dependents and monthly income

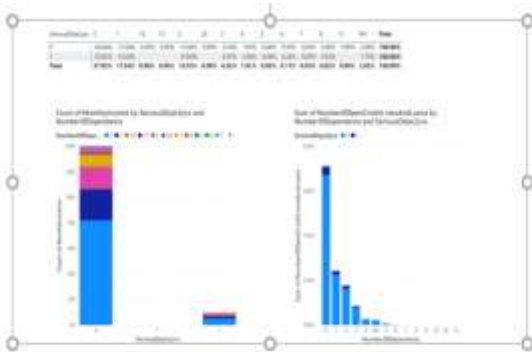


Fig. 3: Loan defaulter distribution by monthly income and loans

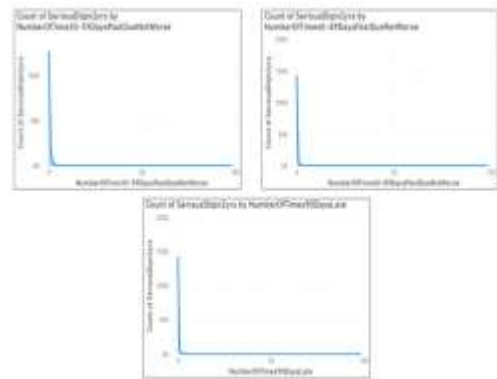


Fig. 7: Count of Loan defaulter distribution by number of days past due

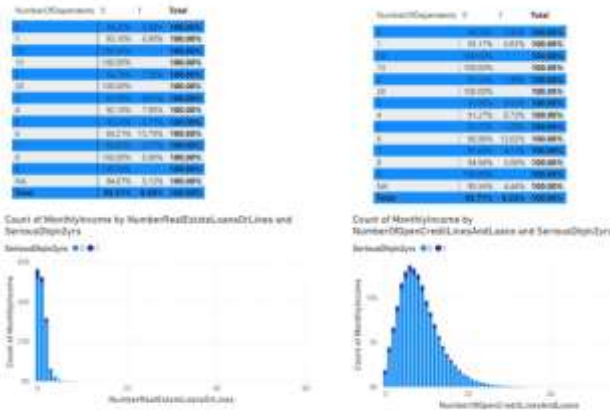


Fig. 4: Monthly income distribution of number of loans and dependents

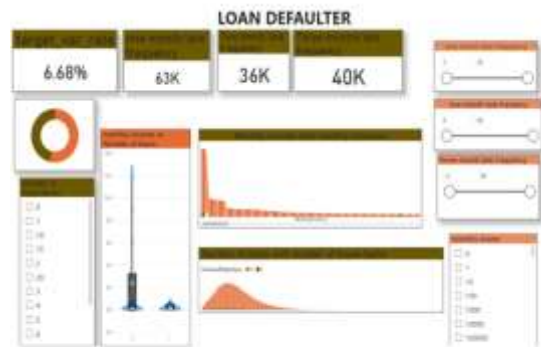


Fig. 8: Dashboards



Fig. 5: Average loan defaulter distribution by number of days past due

Model building for logistic regression using python

```

import pandas as pd
import numpy as np
import matplotlib as matplotlib
import statsmodels as sm

data = pd.read_csv('http://www.gutenberg.org/cache/epub/60000/60000-h/60000-h.txt')
data.head()

# class 'pandas.core.frame.DataFrame'
# Attributes: 10000 entries, 4 to 14999
# Data columns (total 11 columns):
#  #   column                non-null count  dtype
# ---  ---
# 0  age                    10000 non-null  int64
# 1  her_loan_in_2yrs       10000 non-null  int64
# 2  monthly_utilization   10000 non-null  float64
# 3  open_lines             10000 non-null  int64
# 4  number_of_dependents   10000 non-null  int64
# 5  debt_ratio             10000 non-null  float64
# 6  her_loan_in_1yr       10000 non-null  int64
# 7  number_of_credit_lines  10000 non-null  int64
# 8  number_of_loans        10000 non-null  int64
# 9  number_of_credit_lines  10000 non-null  int64
# 10 number_of_loan_in_1yr  10000 non-null  int64
# 11 number_of_dependents  10000 non-null  int64
# Dtypes: float64(2), int64(9)
# Memory usage: 11.7 MB

from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression(solver='lbfgs')
# Fitting logistic regression for active customer on rest of the variables
logistic.fit(data[['age', 'her_loan_in_1yr', 'number_of_credit_lines', 'number_of_loans', 'open_lines', 'monthly_utilization', 'debt_ratio', 'number_of_dependents']])
print('Intercept', logistic.intercept_)
print('Coefficients', logistic.coef_)

Intercept [-0.01069619]
Coefficients[[-4.82084567e-02-2.35509916e-051.79932234e021.23774288e-02]]
  
```

month late frequency is 63k, two months late frequency is 36k and three months late frequency is 40k. The monthly income vs number of loans is 13k, the average age of seriousDlqin2yrs 0 is 53.46% and the average age of 1 is 46.54%. To predict risk customers, we build a logistic regression model using python with influencing factors age, monthly income, number of open credit lines and loans and number of dependents on dependent factor seriousDlqin2yrs and estimate the coefficients of the model as follow Intercept is $\beta_0 = -0.01069619$

$$\beta = -4.82084567e-02$$

To test the accuracy of the model we divide our data into Train data and Test data with 80% and 20% sizes. For the Train data we fit the same model and calculate the Accuracy. Accuracy on train data = 0.9328333333333333. Now using Test data, we fit the model on ceagain and get the Accuracy. Accuracy on test data = 0.9344666666666667. Here accuracy of Test data and Train data are approximately equal. From this we can conclude that this model is the best fit for our project dataset. With This model we can predict risk customers with 93 % accuracy

VI REFERENCES

- [1]. <https://www.geeksforgeeks.org/python-programming-language/>
- [2]. <https://www.techtarget.com/searchcontentmanagement/definition/Microsoft-Power-BI>
- [3]. <https://pythonbasics.org/>
- [4] Mastering Microsoft PowerBI by Devink night and brain knight
- [5] Python programing language Fluent Python: Clear, Concise, and Effective Programming by Luciano Ramalho,
- [6]. Regression Analysis with Python- LucaMassaron
- [7] Handbook of-REGRESSION-ANALYSIS Samprit Chatterjee, Jeffrey S. Simonoff 2013.

```

from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression(solver='lbfgs')
# Fitting logistic regression for active customer on rest of the variables
logistic.fit(data[['age', 'her_loan_in_1yr', 'number_of_credit_lines', 'number_of_loans', 'open_lines', 'monthly_utilization', 'debt_ratio', 'number_of_dependents']])
print('Intercept', logistic.intercept_)
print('Coefficients', logistic.coef_)

Accuracy on train data 0.9328333333333333
Accuracy on test data 0.9344666666666667
  
```

V CONCLUSION AND FUTURE WORK

In our project we represent the data using PowerBI and we observed that the average age of customer is 37 with average-dependence 3 and average monthly income 140k per Annum, and the average number of real estate loans or lines is above 50k and the average number of open credit lines and loan is above 10k. The Average loan defaulter is 6.68% and the count of loan defaulter in number of time 30-59 days past due not worse, number of time 60-89 days past due not worse and number of 90 days late almost 140k and the one

Augmented Reality: Transformative Applications in Aeronautical Maintenance, Building Information Models, and Beyond

Annavarapu. Sridevi
 23DSC20
 M.Sc. [Computational Data Science]
 P.B Siddhartha College of Arts and
 Science
 Vijayawada, A.P, India
 srideviannavarapidevi@gmail.com

Shaik. Ayesha Begum
 23DSC31
 M.Sc. [Computational Data Science]
 P.B Siddhartha College of Arts and
 Science
 Vijayawada, A.P, India
 shaikayesha21216@gamil.com

Shaik. Nousheen
 23DSC16
 M.Sc. [Computational Data Science]
 P.B Siddhartha College of Arts and
 Science
 Vijayawada, A.P, India
 nshaik0311@gmail.com

ABSTRACT– The integration of Augmented Reality (AR) in complex operations, especially in maintenance tasks, represents a promising avenue for enhancing efficiency and knowledge transfer. AR, by merging virtual and real-world elements, provides users with innovative tools and perspectives that can significantly improve various processes across different environments. The research community has proposed numerous AR solutions, with a particular focus on maintenance operations, promising substantial advancements.

Keywords-Augmented Reality, Maintenance operations

I. INTRODUCTION

AR tools offer unique perspectives and have the potential to bring about dramatic improvement in maintenance tasks. By overlaying digital information onto the physical world, technicians can access real-time data, schematics, and instructions, thereby streamlining the execution of complex operations. The transfer of knowledge in diverse processes and environments is facilitated by the immersive nature of AR. However, it is essential to acknowledge that Augmented Reality is a demanding technology facing significant challenges, especially in its implementation within industrial contexts. Despite the promising potential, several serious flaws currently impede its widespread adoption and effectiveness. AR tools offer unique perspectives and have the potential to bring about dramatic improvements in maintenance tasks. By overlaying digital information onto the physical world, technicians can access real-time data, schematics, and instructions, thereby streamlining the execution of complex operations. The transfer of knowledge in diverse processes and environments is facilitated by the immersive nature of A.R

II RELATED WORK

AR is based on Computer Aided Design (CAD) modeling to create 3D objects. What separates this current research from previous approaches is that the properties and relationships in the BIM model are still retained and used for further application. Software such as Smart Reality may display the building, but does not provide any information on objects in

the scene without additional dedicated programming effort. A key part of AR is the localization and orientation of the user in the scene, to ensure that the real world and the augmented scene merge seamlessly together and outline localization techniques. Some methods require specialized equipment. Furthermore, they range in accuracy depending on placement and location of the building (indoor or outdoor). Less invasive methods rely on using the camera to detect the user's location, including image or object markers. These require work on the model designers end to ensure that markers are properly identified by the tracking software. Furthermore, detail, size, and illumination of the marker can have a substantial effect on tracking. AR enhances the view of the surrounding scene by overlaying diverse content, such as visual animations, sounds, written instructions, or static images. This augmentation enriches perception, offering a comprehensive and interactive experience. For maintenance tasks, this means technicians can access relevant information in real-time, contributing to improved decision-making and task execution.

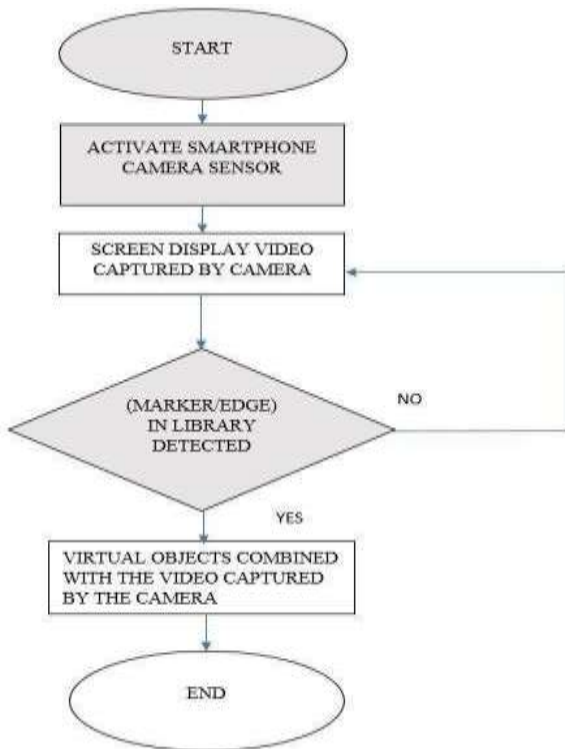
III CHALLENGES IN AR IMPLEMENTATION:

Cost-Effective Maintenance Improved maintenance facilitated by AR can result in cost-effective solutions for industries such as automotive and aviation. Reliability Enhancement Beyond cost reduction, the article suggests that better maintenance leads to increased reliability, potentially reducing unforeseen failures and associated costs. Unspecified Challenges The specific challenges are not detailed in the provided excerpt, leaving room for further exploration of issues related to hardware limitations, user experience, or integration complexities.

IV PROPOSED WORK

The evolution of Virtual Reality (VR) and Augmented Reality (AR) technologies has a rich history, with significant milestones shaping their development over the years. An Augmented Reality (AR) system is designed to enhance the perception of the real world by overlaying virtual information or objects onto their view. These systems, regardless of their specific design, share common hardware components that

enable them to function effectively. description of the key elements typically found in an AR



Cameras: Cameras are essential for capturing images of the real-world environment. Number of Cameras While some AR solutions use a single camera, others may employ multiple cameras for more comprehensive scene perception. **Purpose:** Cameras serve to provide visual input to the AR system, allowing it to understand and interpret the surroundings.

Hardware Function: Tracking hardware is crucial for monitoring the movement and changes in both the real and virtual worlds. Types of Trackers Various tracking methods are utilized, including infra-red trackers and mechanical trackers.

Purpose: The tracking system ensures that virtual objects align and interact seamlessly with the real-world environment. It allows the AR system to adapt to changes in the surroundings, maintaining an accurate overlay of virtual elements onto the real scene. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs.

Registration of virtual objects: In the depicted AR application for airplane fuel filter maintenance, the virtual filter seamlessly integrates with the real scene through the use of visible markers. These markers serve as reference points,

enabling precise registration of virtual objects within the field of view. This registration process ensures that the virtual elements align accurately with their corresponding real-world counterparts, creating a cohesive and immersive user experience.

- **Guided Tool usage:** A notable feature of the application is its ability to guide users through maintenance procedures.

- By leveraging augmented visuals, the application overlays instructions on which tools to use and how to operate them directly onto the view of the real-world scenario. This guidance enhances user

- proficiency by providing step-by-step instructions in real-time, reducing the likelihood of errors and facilitating a smoother execution of maintenance tasks.

- **User Warnings and Alerts:** In addition to tool guidance, the AR application incorporates a warning system.

- Users are alerted to potential issues or critical steps in the maintenance process through visual cues and alerts superimposed onto the real-world scene.

- This proactive approach enhances safety, minimizes the risk of errors, and ensures that users are well-informed about crucial aspects of the maintenance operation.

- These features collectively contribute to the effectiveness of Augmented Reality in maintenance tasks, offering a comprehensive solution that does not.

- **Processing Tasks** It processes real-time information, interprets the environment, and calculates.

- The appropriate augmentation to be displayed.

- Typically equipped with sufficient computational

- **Processing Unit:** power to handle the complex tasks associated with real-time image processing and tracking.

- **Processing Tasks** It processes real-time information, interprets the environment, and calculates the appropriate augmentation to be displayed.

- **Computational Requirements** The processing unit is equipped with sufficient computational

- power to handle the complex tasks associated with real-time image processing and tracking.

- However, it is crucial to acknowledge and address challenges such as marker visibility, accuracy, and system responsiveness to further refine and optimize AR applications for maintenance operations. Subsequent sections will delve into these challenges and propose potential avenues for improvement. maintaining a high level of functionality and effectiveness.

- High level of functionality and effectiveness. sections will delve into these challenges and propose

potential avenues for improvement. maintaining a high level of functionality and effectiveness.

- Responsiveness to further refine and optimize AR applications for maintenance operations. Subsequent sections will delve into these challenges and propose potential avenues for improvement. maintaining a high level of functionality and effectiveness.

Superimposition of Virtual Objects: Objective the AR system superimposes virtual objects congruently with the estimated real-world reference frame.

- **Alignment** Virtual objects are correctly aligned with the identified real-world objects, creating a seamless blend of the virtual and physical environments.
- This alignment facilitates the AR system ability to offer useful instructions to users.
- During tasks, enhancing user guidance and interaction.
- **Flexibility and Easy Implementation:** Adaptability AR technology is highly flexible and can be easily integrated into various processes.
- It can be applied across different industries, making it a versatile solution for a wide range of applications. **Access to Additional Knowledge** AR systems provide additional information that may not be easily accessible or could be demanding to retrieve. **Guided Assistance** The added knowledge helps reduce errors during maintenance tasks by offering guided assistance and real-time information, improving the accuracy and efficiency of operators.
- **Cost Savings and Economic Benefits:** Operational Cost Reduction AR systems contribute to lowering operational costs by streamlining processes and reducing errors.
- **Training Efficiency** The need for extensive training is reduced, leading to cost savings. AR enables faster knowledge transfer, allowing workers to become proficient more efficiently.

V ADVANTAGE AR IN INDUSTRIAL PROCESSES:

Improving Human Performance The main advantages of AR, particularly in maintenance operations, are highlighted. AR has the potential to significantly enhance human performance by providing real-time information, reducing errors, and offering guided assistance. **Economic and Operational Benefits** The article emphasizes the benefits of AR.

- **Economic considerations.** While it can lead to cost savings, it also contributes to higher reliability in maintenance operations.
- **Challenges in AR Implementation: Ongoing Issues:** The conclusion hints at the persistence of serious problems affecting AR implementation in industrial settings.
- **Implementation Challenges** Despite its promise, the article

acknowledges that AR faces serious

Challenges that currently hinder its widespread adoption in industrial environments.

In aeronautical maintenance the paper authored by Mauricio Chincapin, Andrea Caponio, Horacio Rios, and Eduardo González

- Mendivil provides a comprehensive overview of Augmented Reality (AR) and its applications in aeronautical maintenance
- Significantly easing the execution of complex operations. By merging virtual and actual reality, AR
- Introduces new tools that enhance efficiency in knowledge transfer across various processes and environments.
- The authors define AR as a variation of Virtual Reality Technology (VR), highlighting the fundamentals.
- Difference between the two. While VR immerses users in a computer-generated virtual world, replacing.
- Reality entirely, AR supplements the real world by overlaying virtual objects onto it.
- Coexistence of real and virtual objects in AR creates a unique environment where users can interact with both seamlessly.
- **Human-Machine Interaction:** perceptual-motor skills in the real world are introduced to explain the special type.
- Human-machine interaction facilitated by AR. This interaction is exemplified.
- Depicting an AR application for the maintenance of an airplane fuel filter. Notably, the virtual filter.
- Registered to the real scene through visible markers, ensuring accurate alignment.
- The paper illustrates how the AR application not only registers virtual objects but also guides users.
- Through maintenance tasks by superimposing instructions on which tools to use and how to operate the AR. The AR system enhances user proficiency. Furthermore, the inclusion of warnings for potential
- Risks ensure user safety during maintenance operations.
- The authors present examples of AR applications and emphasize the feasibility of AR.
- Maintenance tasks, underscoring the advantages they could introduce.
- Acknowledges that AR is an extremely demanding technology, still affected by serious flaws that hinder.
- Its implementations in the industrial context.

- The authors contribute to the understanding of AR applications and challenges.
- Advancements in this transformative technology. An excellent style manual for science writers.

VI. RESULT AND ANALYSIS

Certainly! Augmented Reality (AR) APIs play a crucial role in enabling developers to integrate AR functionality into their applications. AR APIs provide a set of tools, libraries, and frameworks that make it easier for developers to create immersive AR experiences. Here are some key points about Augmented Reality.

1. Platform-specific APIs: Different AR platforms (iOS, Android, Windows, etc.) have their own AR APIs. For example, ARKit for iOS and ARCore for Android are popular platform-specific APIs. These APIs leverage the specific hardware and capabilities of each platform to deliver optimal AR experiences. Cross-platform APIs:

2. Some AR APIs: aim to provide a cross-platform solution, allowing developers to build AR applications that can run on multiple devices and operating systems. Unity's AR

Improved Ant Colony Optimization Algorithm and its Applications

Bora Uma Reddy
 23DSC21,
 M.Sc. (Computational Data
 Science)
 Dept. of Computer Science P.B.
 Siddhartha College of Arts &
 Science Vijayawada, A.P, India
 umakrishna7620@gmail.com

Sandhya Naidu
 23DSC29,
 M.Sc. (Computational Data
 Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts
 & Science Vijayawada, A.P,
 Sandhyanaidu879@gmail.com

Rasani Sunandini
 23DSC13,
 M.Sc. (Computational Data
 Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P.
 rasanisunandini@gmail.com

Abstract—We made a better version of the Ant Colony Optimization (ACO) algorithm called Intelligent and Efficient Ant Colony Optimization (IMVPACO). It's designed to solve problems faster and find the best solutions. We tweaked how ants communicate (pheromones) to make sure they follow better paths. We also added a smart way to balance speed and accuracy in finding solutions, avoiding getting stuck in bad solutions.

We improved how the ants move to solve big problems more effectively. We also introduced a special way to make changes in our process that helps us find better solutions. We tested our new method on a common problem called the Traveling Salesman Problem, and it performed really well, finding the best solutions faster than other methods.

I INTRODUCTION

Since a review in the international academic journal Nature in 2000, the Ant Colony Optimization (ACO) algorithm, inspired by how real ants behave, has become a popular method for solving complex problems. It's widely used in different areas like the Traveling Salesman Problem (TSP), job scheduling, and more, thanks to its ability to find good solutions.

Despite its strengths, the ACO algorithm has some drawbacks like taking too long to search, slow convergence, and getting stuck easily. To overcome these issues, many researchers have come up with improved versions. For instance, some incorporate dynamic updates, while others use hybrid approaches like combining ACO with other algorithms.

In this paper, a new and smarter ACO algorithm, called Intelligent and Efficient Ant Colony Optimization (IMVPACO), is introduced to make global problem-solving faster and prevent getting stuck in bad solutions. They tested it on the Traveling Salesman Problem, and it showed promising results.

II BASIC ANT COLONY OPTIMIZATION ALGORITHM

Ant Colony Optimization (ACO) is a metaheuristic algorithm inspired by the foraging behavior of real ants. It was first introduced by Marco Dorigo in the early 1990s and has since gained popularity for solving combinatorial optimization problems. The basic concept behind ACO lies in simulating the way ants find the shortest path between their nest and a food source.

Overview of the Basic Ant Colony Optimization Algorithm:

1. Problem Formulation:

ACO is particularly effective in solving problems where the goal is to find the best solution from a finite set of possibilities. This often involves searching through various combinations to optimize a specific objective function. Examples include the Traveling Salesman Problem (TSP), Job Scheduling, and Routing Problems.

2. Ant Agents:

The algorithm involves a population of artificial ants that collectively search for good solutions. Each ant represents a potential solution to the optimization problem. These ants move through the solution space and deposit a substance called pheromone on the paths they traverse.

3. Pheromone:

Pheromone serves as a form of communication among the ants. It is a chemical substance that ants deposit on the ground as they move. The intensity of the pheromone trail influences the probability of other ants choosing the same path. Over time, paths with higher pheromone concentrations become more attractive to ants.

4. Solution Representation:

The problem's solutions are represented as paths in a graph, where nodes represent decision variables or problem components, and edges represent possible connections between these components. The ants construct solutions by moving from one node to another.

5.Solution Construction:

An ant begins its journey by selecting an initial solution component. Then, it probabilistically chooses the next component based on both the amount of pheromone on the connecting edge and a heuristic value. The heuristic guides the ants toward potentially better solutions. This process is repeated until a complete solution is constructed.

6.Objective Function Evaluation:

Once a solution is constructed, its quality is evaluated using the objective function associated with the optimization problem. The objective function measures how well a particular solution satisfies the problem requirements.

7.Pheromone Update:

The pheromone levels on the edges of the paths are updated based on the quality of the solutions. Better solutions contribute more to the pheromone levels, reinforcing the attractiveness of the corresponding paths. Over time, less attractive paths are gradually explored less frequently.

8.Evaporation:

To mimic the natural decay of pheromones, an evaporation process is introduced. Pheromone levels on all paths decrease over time, preventing the algorithm from converging too quickly to suboptimal solutions. This balance between exploration and exploitation is crucial for the algorithm's success.

9.Iteration:

The process of solution construction, evaluation, and pheromone update is repeated over multiple iterations. Ants collectively explore the solution space, and the pheromone trail evolves accordingly. The algorithm continues until a stopping criterion is met, such as a predetermined number of iterations or a satisfactory solution quality.

10.Solution Output:

The final solution is typically the path with the highest concentration of pheromone. This path represents the optimized solution to the given problem.

IMPROVED ANT COLONY OPTIMIZATION (IMVPACO)

A.ALGORITHM

The Improved Ant Colony Optimization (IMVPACO) algorithm is an enhanced version of the traditional Ant Colony Optimization (ACO) algorithm designed to overcome its limitations and improve overall performance. Some key aspects of the IMVPACO algorithm includes.

B.Dynamic Movement Rules of Ants

In dynamic ant movement rules, ants adapt their paths based on evolving information during optimization. The probability $P_{ij}(t)$ of an ant moving from component i to j involves pheromones τ_{ij} and heuristics η_{ij} . Dynamic

adjustments enhance the ants' adaptability and solution exploration.

C.Improved Updating Rules of Pheromones

In the Improved Ant Colony Optimization (IMVPACO) algorithm, pheromone updating rules are refined for better adaptability. The dynamic update equation, $\tau_{ij}(t+1) = (1-\rho) \cdot \tau_{ij}(t) + \Delta\tau_{ij}(t)$, enhances responsiveness to changing conditions, promoting improved convergence and solution quality

D.Adaptive Adjustment Strategy of Pheromone

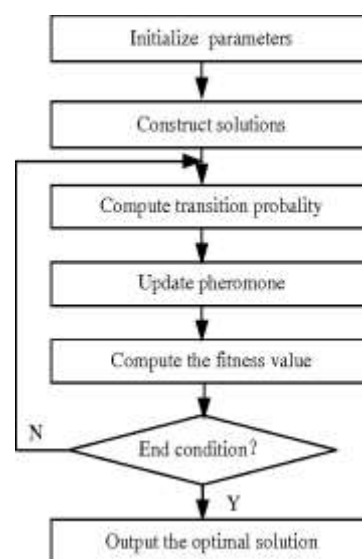
The adaptive adjustment strategy of pheromone in optimization algorithms involves dynamically fine-tuning the pheromone levels based on solution quality. This responsive approach, as seen in Improved Ant Colony Optimization (IMVPACO), ensures efficient adaptation to varying problem conditions, enhancing overall algorithm performance.

E.Dynamic Evaporation Factor Strategy

The dynamic evaporation factor strategy in optimization algorithms, like IMVPACO, dynamically adjusts the rate at which pheromones evaporate. This strategy aims to strike a balance between solution exploration and exploitation, ensuring efficient adaptation to problem dynamics and preventing premature convergence.

F.Boundary Symmetric Mutation Strategy

The Boundary Symmetric Mutation Strategy in optimization, as applied in IMVPACO, involves symmetrically mutating solutions near boundaries. This targeted mutation approach enhances the exploration of solution space, reinforcing mutation efficiency and quality, contributing to Improved Ant Colony Optimization's effectiveness in solving complex problems.



B.Flow chart of ACO Algorithm

THE STEPS OF THE PROPOSED ALGORITHM

1. Initialization:
 - Set initial parameters and create ant population.
2. Pheromone Initialization:
 - Initialize pheromone levels on edges.
3. Ant Movement:
 - Ants move between components based on pheromone and heuristic information.
4. Solution Construction:
 - Ants construct solutions based on their movements.
5. Local Search (Optional):
 - Optionally apply local search strategies to refine solutions.
6. Pheromone Update:
 - Update pheromone levels based on solution quality.
7. Dynamic Evaporation:
 - Dynamically adjust pheromone evaporation.
8. Symmetric Mutation (Optional):
 - Optionally apply boundary symmetric mutation
9. Convergence Check:
 - Check convergence criteria, if not met, repeat from step 3.
10. Output:

- Return the best-found solution

III EXPERIMENTAL ANALYSIS

A. Experiment

Introduction and Parameter

Set

In an experiment applying Ant Colony Optimization (ACO) to the Traveling Salesman Problem (TSP), the algorithm is tested on benchmark instances. Performance metrics include total tour length and convergence time. ACO's parameters are fine-tuned for optimal results, and its efficiency is compared with other optimization methods. Sensitivity analysis explores how ACO responds to changes in parameters. Statistical tests ensure result reliability. Visualizations showcase convergence curves and solution paths. This experiment aims to assess ACO's effectiveness, scalability, and adaptability, providing insights into its performance for solving real-world optimization problems

Table 1. The Initial Values of Parameters for the ACO, PACO and IMVPACO Algorithms

Ants (m)	Pheromone factor (α)	Evaporation factor (ρ)	Initial concentration ($\tau_{ij}(0)$)
100	1.0	0.05	1.5
Iteration (T_{max})	Heuristic factor (β)	Pheromone amount (Q)	initial uniform probability (q_0)
500	2.0	100	0.5

B. Experimental Results and Analysis

In the experiment, the standard ACO algorithm, IACO algorithm and IMVPACO algorithm are run on MATLAB platform. Ten TSP benchmark instances with cities scale from 51 to 14051 are performed. In here, the optimal value (Best) and the number of iterations is used to illustrate the solving ability of the proposed IMVPACO algorithm.

The experimental results are shown in Table 2.

Table 2. The Experimental Results in Solving TSP

Index	Instance	Optimum	ACO		IACO		IMVPACO	
			Best	Iterations	Best	Iterations	Best	Iterations
1	eil51	426	447	235	434	198	427	104
2	pc76	108159	109986	241	109853	205	109003	186
3	rad100	7910	8056	296	7968	276	7938	267
4	pr124	59030	59794	289	59481	253	59084	194
5	kroA150	26524	26965	368	26794	350	26683	301
6	rat195	2323	2408	358	2396	330	2353	276
7	pr299	48191	48946	423	48802	389	48662	305
8	pcb442	50778	51905	418	51843	381	51457	336
9	u724	41910	43046	445	42994	405	42859	376
10	bd14051	469385	499052	479	496034	452	490432	401

As can be seen from the Table 2, for the ten TSP instances, the optimal value (Best) and the number of iterations of the proposed IMVPACO algorithm are best in this experiment. Because the four experiment values are close to the optimal solutions. For TSP instances with eil51, pr76, rad100 and rat195, the obtained best solutions 427, 108159, 7938 and 2353 are close to the best-known solutions 426, 109003, 7910 and 2323. In the number of iterations, the proposed IMVPACO algorithm is best than the standard ACO algorithm, IACO algorithm in solving TSP with same scale. At the same time, for larger scale instances, the experiment results show that the proposed IMVPACO algorithm in the optimal value (Best) and the number of iterations is better than the ACO algorithm and IACO algorithm.

In order to further illustrate the optimization performance of the proposed IMVPACO algorithm, two best routes found by the IMVPACO algorithm are shown in Fig.2. and Fig.3

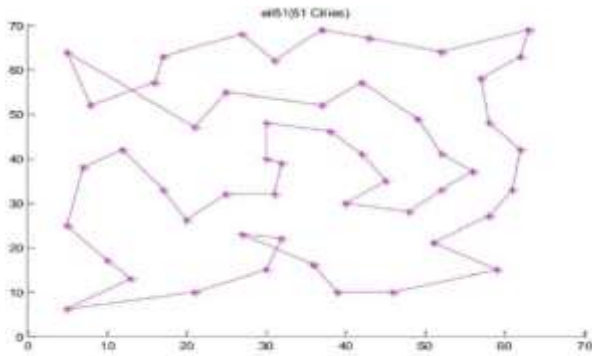


Figure 2. The Best Routes Found for eil51

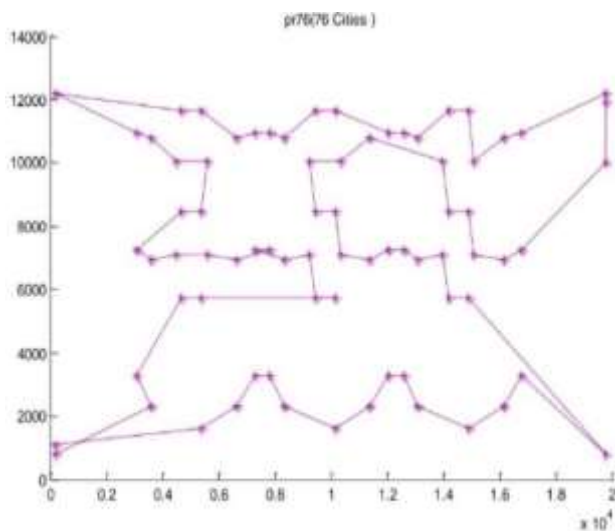


Figure 3. The Best Routes Found for pr76

IV CONCLUSION

The ACO algorithm is a stochastic, population-based, evolutionary search algorithm. It is an efficient and powerful optimization algorithm, which widely applied in scientific research and engineering field. This paper proposes an efficient and intelligent ant colony optimization (IMVPACO) algorithm. In the IMVPACO algorithm, the updating rules and adaptive adjustment strategy of pheromones are modified in order to better reflect the quality of the solution based on the increment of pheromone. The dynamic evaporation factor strategy is used to achieve the better balance between the solving efficiency and solving quality, and effectively avoid falling into local optimum for quickening the convergence speed. The movement rules of the ants are modified to make it adaptable for large-scale problem solving, optimize the path

and improve search efficiency. A boundary symmetric mutation strategy is used to obtain the symmetric mutation for iteration results, which not only strengthens the mutation efficiency, but also improves the mutation quality. The traveling salesman problem is used to test the effectiveness of the proposed IMVPACO algorithm. The simulation experiments show that the proposed IMVPACO algorithm can obtain very good results in finding optimal solution.

V FUTURE SCOPE

Improving Ant Colony Optimization (ACO) algorithms could involve enhancing their efficiency in solving complex optimization problems, expanding their applicability to diverse domains, and refining their adaptability to dynamic environments. Research might focus on refining pheromone updating mechanisms, exploring hybrid approaches with other metaheuristic algorithms, or devising strategies for handling larger-scale problems efficiently.

Additionally, integrating machine learning techniques or leveraging parallel computing could further enhance the performance and scalability of ACO algorithms.

VI REFERENCES

- [1] M. Dorigo and G. D. Caro, "ant algorithms for discrete optimization", *Artificial Life*, vol. 5, no. 3, (1999), pp. 137-172.
- [2] T. Stutzle and H. H. Hoos, "MAX-MIN ant system", *Future Generation Computer Systems*, vol. 16, no. 8, (2000), pp. 889-914.
- [3] J. Dero and P. Siarry, "Continuous interacting ant colony algorithm based on dense hierarchy", *Future Generation Computer Systems*, vol. 20, no. 5, (2004), pp. 841-856.
- [4] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem", *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, (1997), pp. 53-66.
- [5] L. M. Gambardella, E. Taillard and M. Dorigo, "Ant colonies for the quadratic assignment problem", *Journal of the Operational Research Society*, vol. 50, no. 1, (1999), pp. 167-176.
- [6] A. Colomi, M. Dorigo and V. Maniezzo, "Ant system for job-shop scheduling", *Belgian Journal of Operations Research, Statistics and Computer-Science*, vol. 34, no. 1, (1994), pp. 39-54.
- [7] S. Leng, X. B. Wei and W. Y. Zhang, "Improved ACO scheduling algorithm based on flexible process", *Transactions of Nanjing University of Aeronautics and Astronautics*, vol. 23, no. 2, (2006), pp. 154-160.
- [8] Y. Yi, Yang and J. L. Lai, "Computation model and improved ACO algorithm for p/T", *Journal of Systems Engineering and Electronics*, vol. 20, no. 6, (2009), pp. 1336-1343.

Deep SMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data

Buraga. Bhavishya Sagarika
23DSC22, M.Sc. (Computational Data
Science)
Dept. of Computer Science
P.B. Siddhartha College of Arts &
Science
Vijayawada, A.P, India
bhavishyasagarika@gmail.com

Gajjalakonda.Keerthi
23DSC25, M.Sc. (Computational Data
Science)
Dept. of Computer Science
P.B. Siddhartha College of Arts &
Science
Vijayawada, A.P, India
gajjalakondakeerthi@gmail.com

Saggurthi.Pavitra
23DSC30, M.Sc. (Computational Data
Science)
Dept. of Computer Science
P.B. Siddhartha College of Arts &
Science
Vijayawada, A.P, India
saggurthipavitra@gmail.com

Abstract - Despite progress in machine learning, dealing with imbalanced data remains a big challenge, especially in deep learning for images. To address this, we introduce Deep Synthetic Minority Oversampling Technique (Deep SMOTE), a new method for deep learning models. It creates artificial images to balance training sets, focusing on minority classes. Unlike other methods, Deep SMOTE is simple but effective, using three main parts: 1) an encoder/decoder framework 2) SMOTE-based oversampling: 3) a modified loss function. What makes Deep SMOTE unique is that it doesn't need a discriminator, producing high-quality artificial images suitable for inspection. Thus, there arises a need for an inspection. A notable advantage of Deep SMOTE over generative adversarial network (GAN)-based oversampling methods is its independence from a discriminator. Moreover, Deep SMOTE proposed approach demonstrates simplicity, effectiveness, and a valuable synergy between deep learning and SMOTE in mitigating imbalanced data challenges.

Keywords- Deep learning, SMOTE algorithm, Deep Synthetic Minority Oversampling Technique, Machine Learning, Training sets, Minority classes, Loss function, Raw images, Encoder/decoder framework

I INTRODUCTION

The fusion of deep learning techniques with various methodologies has become a area of research, driven by the desire to enhance the capabilities of machine learning models. One particularly challenging aspect is the handling of imbalanced datasets, where certain classes are underrepresented, leading to biased models.[1] In response to this, the integration of deep learning approaches with oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), has gained attention. By leveraging the strengths of deep learning models, which excel at extracting complex features, and incorporating oversampling techniques like SMOTE, we aim to enhance model generalization and performance, especially in tasks involving image data. The ensuing sections will delve into the proposed fusion methodology, discussing the design considerations, components, and potential advantages over traditional approaches. We present a concrete example, Deep Synthetic Minority Oversampling Technique (Deep

SMOTE), which integrates deep learning and SMOTE to generate artificial images that balance training sets effectively. Through this fusion, we strive to contribute to the ongoing discourse on improving the robustness and fairness of machine learning models in the face of imbalanced data challenges.[2] We introduce a pioneering oversampling technique designed explicitly for imbalanced data in deep learning models.[3] This method harnesses the benefits of the Synthetic Minority Oversampling Technique (SMOTE) seamlessly integrating it into a deep architecture capable of efficiently handling intricate data representations, particularly in the context of images. The imbalanced data problem is a well-known challenge in the field of deep learning. In the past, two main directions have been pursued to overcome this challenge: loss function modifications and resampling approaches. However, the deep learning resampling solutions are either pixel-based or use generative adversarial networks (GANs) for artificial instance generation. Both these approaches suffer from strong limitations. Pixel-based solutions often cannot capture complex data properties of images and are not capable of generating meaningful artificial images. GAN-based solutions require significant amounts of data,[5] are difficult to tune, and may suffer from mode collapse. Therefore, there is a need for a novel oversampling method that is specifically tailored to the nature of deep learning models, can work on raw images while preserving their properties, and is capable of generating artificial images that are of both of high visual quality and enrich the discriminative capabilities of deep models. Different situations can occur in confronting the imbalanced datasets, and four common cases are depicted in Figure 1, where the blue-filled circles represent the samples of the majority class;[6] in contrast, the red circles denote the minority class. It has been shown that the type of data complexity is the principal determining factor of classification performance reduction.

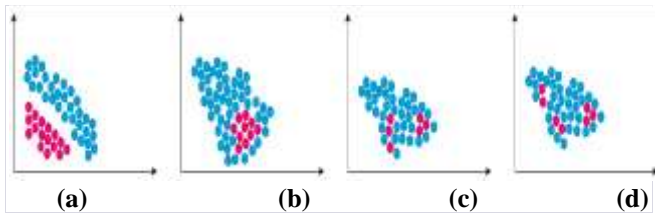


Fig.1

II RELATED WORK

The first works on imbalanced data came from binary classification problems. Here, the presence of majority and minority classes is assumed, with a specific imbalance ratio. Such skewed class distributions pose a challenge for machine learning models, as standard classifiers are driven by a 0–1 loss function that assumes a uniform penalty over both classes. Therefore, any learning procedure driven by such a function will lead to a bias toward the majority class. At the same time, the minority class is usually more important and thus cannot be poorly recognized. And low dimensional representation learning from imbalanced data streams [4]. Therefore, methods dedicated to overcoming the imbalance problem aim at either alleviating the class skew or alternating the learning procedure. The three main approaches are as follows.

A. Data-Level Approaches

Consider this solution as a preprocessing phase independent of any classifier. In this context, our focus is on achieving dataset balance [9] before initiating classifier training. Typically, this involves one of three methods: 1) decreasing the size of the majority class (under sampling); 2) increasing the size of the minority class (oversampling); or 3) a hybrid approach combining both strategies. Both random under sampling [10] and oversampling can be applied, which, although simple, may result in potential instability, such as removing crucial instances or amplifying noise. To address this, guided solutions have been proposed to intelligently select instances for preprocessing. While guided under sampling solutions are relatively limited, oversampling has garnered more attention, especially with the success of SMOTE, leading to the introduction of numerous variants. However, recent research indicates that SMOTE-based methods face challenges in handling multimodal data, instances with high intraclass overlap, or noise. Consequently, innovative approaches that do not rely on k-nearest neighbors have been successfully developed

B. Algorithm-Level Approaches

Contrary to the previously discussed approaches, algorithm level solutions work directly within the training procedure of the considered classifier. Therefore, they lack the flexibility offered by data-level approaches, but compensate with a more direct and powerful way of reducing the bias of the learning algorithm. They also require an in-depth

understanding of how a given training procedure is conducted and what specific part of it may lead to bias toward the majority class. The most commonly addressed issues with the algorithmic approach are developing novel skew-insensitive split criteria for decision trees, using instance weighting for support vector machines, or modifying the way different layers are trained in deep learning. Furthermore, cost-sensitive solutions and one-class classification can also be considered as a form of algorithm-level.

C. Loss Function Adaptation

One of the most popular approaches for making neural networks skew-insensitive is to modify their loss function. This approach successfully carried over to deep architectures and can be seen as an algorithm-level modification. The idea behind modifying the loss function is based on the assumption that instances should not be treated uniformly during training and that errors on minority classes should be penalized more strongly, making it parallel to cost-sensitive learning. Mean False Error and Focal Loss are two of the most popular approaches based on this principle. The former simply balances the impact of instances from minority and majority classes, while the latter reduces the impact of easy instances on the loss function. More recently, multiple other loss functions were proposed, such as Log Bilinear Loss, Cross Entropy Loss, and Class-Balanced Loss.

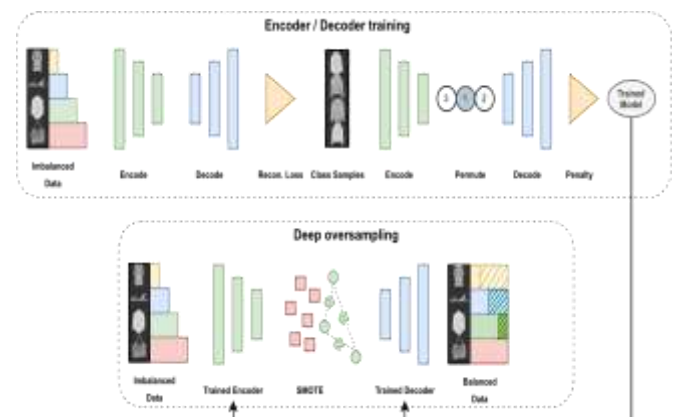


Fig.2

Deep SMOTE consists of an encoder/decoder framework, a SMOTE-based oversampling method, and a loss function with a reconstruction loss and a penalty term. Each of these features is discussed below, with Fig. 1 depicting the flow of the Deep SMOTE approach, while the pseudo-code overview of Deep SMOTE is presented in Algorithm 1.

As discussed above, oversampling is a proven technique for combating class imbalance; however, it has traditionally been used with classical machine learning models. Several attempts have been made to extend oversampling methods, such as SMOTE, to deep learning models, although the

results have been mixed. In order for an oversampling method to be successfully applied to deep learning models, we believe that it should meet three essential criteria.

- 1) It should operate in an end-to-end manner by accepting raw input, such as images (i.e., similar to VAEs, WAEs, and GANs).[8]
- 2) It should learn a representation of the raw data and embed the data into a lower dimensional feature space, which can be used for oversampling.
- 3) It should readily generate output (e.g., images) that can be visually inspected, without extensive manipulation.

Algorithm 1: DEEPSMOTE

Algorithm 1 DEEPSMOTE

Data: B: batches of imbalanced training data
 (D) $B = \{b_1, b_2, \dots, b_n\}$
Input: Model parameters: $\Theta = \{\Theta_0, \Theta_1, \dots, \Theta_j\}$; Learning Rate: α
Output: Balanced training set.
Symbols: R_L - Reconstruction loss; P_L - Penalty loss;
 T_L - Total loss;
 C - Set of classes in D;
 C_M - Set of minority classes in D;
 G - Set of generated and encoded examples;
 S - Set of generated and decoded data (balanced).
Train the Encoder / Decoder:
 for $e \leftarrow epochs$ do
 for $b \leftarrow B$ do
 $E_b \leftarrow encode(b)$
 $D_b \leftarrow decode(E_b)$
 $R_L = \frac{1}{n} \sum_{i=1}^n (D_{bi} - b_i)^2$
 $C_D \leftarrow randomly\ sample\ a\ class\ from\ C$
 $C_b \leftarrow randomly\ sample\ |b|\ instances\ from\ C_D$
 $E_S \leftarrow encode(C_b)$
 $P_E \leftarrow permute\ order(E_S)$
 $D_P \leftarrow decode(P_E)$
 $P_L = \frac{1}{n} \sum_{i=1}^n (D_{Pi} - C_{Di})^2$ $T_L = R_L + P_L$
 $\Theta := \Theta - \alpha \frac{\partial T_L}{\partial \Theta}$
Generate Samples:
 foreach $m \leftarrow minority\ class\ (C_M)$ do
 $C_{md} \leftarrow select\ (C_m\ imbalanced\ data)$
 $E_m \leftarrow encode(C_{md})$
 $G_m \leftarrow SMOTE(E_m)$
 $S_m \leftarrow decode(G_m)$

Encoder/decoder framework. It is based on the DC-GAN architecture established by Radford et al (Radford et al., 2015). Radford et al. employ a discriminator / generator in a GAN,[7] which is fundamentally similar to an encoder / decoder because the discriminator effectively encodes input (absent the final, fully connected layer) and the generator (decoder) produces output. The encoder and decoder are trained in an end-to-end fashion. During Deep SMOTE training, an imbalanced dataset is fed to the encoder / decoder in batches. A reconstruction loss is computed on the batched data. All classes are used during training so that the encoder decoder can learn to reconstruct both majority and minority class images from the imbalanced data.[11] Because there are few minority class examples, majority class examples are used to train the model to learn the basic

reconstruction patterns inherent in the data. This approach is based on the assumption that classes share some similar characteristics (e.g., all classes represent digits or faces).

III. EXPERIMENTAL STUDY

To have a more in-depth understanding of the quality of SMOTE sampling and its influencing factors, we have performed a simulation study. In the first group of experiments, we generate artificial datasets from multivariate Gaussian distributions, apply SMOTE oversampling, then estimate the distribution of the examples generated by SMOTE and compare it to the original distribution.

The advantage of using artificial data is that in this case we know the ground truth distribution and the accuracy of SMOTE. We observe that the experimental results agree to a reasonable extent with the approximate theoretical results, especially in the direction of changes in relation to the variables of influence. Moreover, the results of the experiments with real data also agree to a good extent with those of the artificial datasets. We can summarize the findings as follows: We find that the TVD is always negative, indicating the contractive nature of SMOTE. This means that SMOTE samples are more shrunk inwards. It appears that you've provided information on a simulation study investigating the quality of Synthetic Minority Oversampling Technique (SMOTE) sampling, particularly in the context of generating artificial datasets from multivariate Gaussian distributions.[14] The study aims to compare the distribution of examples generated by SMOTE with the original distribution and evaluate the accuracy of SMOTE in various scenarios.

Artificial Datasets from Gaussian Distribution

Artificial datasets were generated from multivariate Gaussian distributions. SMOTE oversampling was applied to these datasets. The distribution of examples generated by SMOTE was estimated and compared to the original distribution.

Advantages of Artificial Data:

The use of artificial data provides a known ground truth distribution for comparison.

Accuracy of SMOTE:

Experimental results reasonably align with approximate theoretical results. Results show consistency with the direction of changes in relation to the variables of influence.

Comparison with Real Data:

Results from experiments with real data align well with those from artificial datasets.

Summary of Findings:

Total Variation Distance (TVD) is consistently negative, indicating the contractive nature of SMOTE. The negative TVD suggests that SMOTE samples are more shrunk inwards. It seems like the study has provided insights into the behavior of SMOTE in terms of its impact on the distribution of generated samples. The agreement between theoretical and experimental results, as well as the

consistency across artificial and real datasets, adds credibility to the findings. The observed contractive nature of SMOTE, as indicated by the negative TVD, suggests that the generated samples tend to be more concentrated inward.

A. Setup

1) Overview of the Datasets: Five popular datasets were selected as benchmarks for evaluating imbalanced data oversampling: [13] Modified National Institute of Standards and

TABLE I

CLASS DISTRIBUTIONS OF FIVE BENCHMARK DATASETS USED IN EXPERIMENTAL EVALUATION

Class	MNIST/FMNIST			CIFAR/SVHN			CELEBA		
	Train	Bal. Test	Imbal. Test	Train	Bal. Test	Imbal. Test	Train	Bal. Test	Imbal. Test
0	4000	1200	1000	4500	1000	1000	9000	900	1000
1	2000	1200	500	2000	1000	500	4500	900	500
2	1000	1200	250	1000	1000	250	1000	900	111
3	750	1200	187	800	1000	187	500	900	55
4	500	1200	125	600	1000	125	160	900	17
5	350	1200	87	500	1000	87			
6	200	1200	50	400	1000	50			
7	100	1200	25	250	1000	25			
8	60	1200	15	150	1000	15			
9	40	1200	10	80	1000	10			

Technology dataset (MNIST), Fashion-MNIST dataset (FMNIST), CIFAR-10, the street view house numbers (SVHNs), and Large-scale CelebFaces Attributes (CelebA). Below we discuss their details, while their class distributions are given in Table I.

1. MNIST/FMNIST: The MNIST dataset consists of handwritten digits and the FMNIST dataset contains Zalando clothing article images. Both training sets have 60000 images. Both datasets contain gray-scale images ($1 \times 28 \times 28$), with ten classes each.

2. CIFAR-10/SVHN: The CIFAR-10 dataset consists of images, such as automobiles, cats, dogs, frogs, and birds, whereas the SVHN dataset consists of small, cropped digits from house numbers in Google Street View images. CIFAR-10 has 50000 training images. SVHN has 73257 digits for training. Both datasets consist of color images ($3 \times 32 \times 32$), with ten classes each.

3. CelebA: The CelebA dataset contains 200000 celebrity images, each with 40 attribute annotations (i.e., classes). The color images ($3 \times 178 \times 218$) in this dataset cover large pose variations and background clutter. For purposes of this study, the images were resized to $3 \times 32 \times 32$ and five classes were selected: black hair, brown hair, blond, gray, and bald.

Robustness and Stability Under Varied Imbalance Ratios

1. Robustness to Varying Imbalance Ratios: One of the most challenging aspects of learning from imbalanced data [12] lies in creating robust algorithms that can manage various data level difficulties. Many existing resampling methods return very good results only under specific conditions or under a narrow range of imbalance ratios. Therefore, in order to obtain a complete picture of the performance of Deep SMOTE, we analyze its robustness to

varying imbalance ratios in the range of [20, 400]. Fig. 9 depicts the relationship between the three-performance metrics and increasing imbalance ratio on five used benchmarks. This experiment allows us not only to evaluate Deep SMOTE and the reference methods under various skewed scenarios, but also offers a bird-eye view on the characteristics of the performance curves displayed by each examined resampling method. An ideal resampling algorithm should be characterized by a high robustness to increasing imbalance ratios, and skew-insensitive [15] display stable, or small, performance degradation with increased class disproportions. Sharp and significant performance declines indicate breaking points for resampling methods and show when a given algorithm stops being capable of generating useful instances and countering class imbalance.

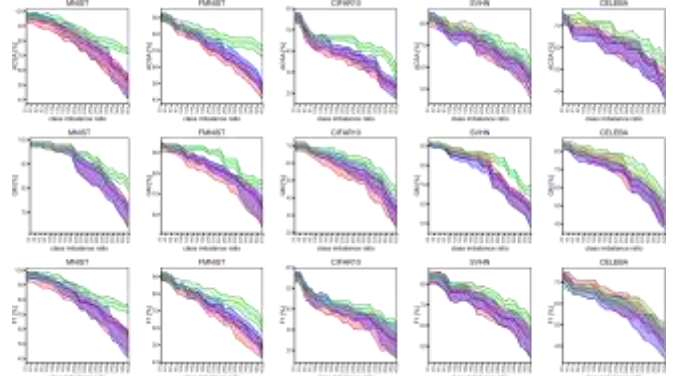


Fig.3 shows the relationship between imbalance ratio and model stability.

Under higher imbalance ratios, showing that those approaches cannot be considered as stable models for challenging imbalanced data problems. Deep SMOTE returned the lowest variance within those metrics, showcasing the high stability of our resampling algorithm. This information enriches our previous observation regarding the robustness of Deep SMOTE. Joint analysis of Figs. 9 and 10 allows us to conclude that Deep SMOTE can handle extreme imbalance among classes, while generating stable models under challenging condition

IV. CONCLUSION & FUTURE WORK

Summary: We proposed Deep SMOTE, a novel and transformative model for imbalanced data, that fuses the highly popular SMOTE algorithm with deep learning methods. Deep SMOTE is an efficient oversampling solution for training deep architectures on imbalanced data distributions. It can be seen as a data-level solution to class imbalance, as it creates artificial instances that balance the training set, which can then be used to train any deep classifier without suffering from bias. DeepSMOTE uniquely satisfies three crucial characteristics of a successful resampling algorithm in the domain of learning from images: ability to operate on raw images, creation of efficient low-dimensional embeddings, and generation of high-quality artificial images. This was made possible by a

novel architecture that combined an encoder/decoder framework with SMOTE-based oversampling and an enhanced loss function. Extensive experimental studies show that Deep SMOTE not only outperforms state-of-the-art pixel-based and GAN-based oversampling algorithms, but also offers unparalleled robustness to varying imbalance ratios with high model stability, while generating artificial images of excellent quality.

V FUTURE WORK

Our next efforts will focus on enhancing Deep SMOTE with information regarding class-level and instance-level difficulties, which will allow it to better tackle challenging regions of the feature space. We plan to enhance our dedicated loss function with instance-level penalties for focusing the encoder/decoder training on instances that display borderline/overlapping characteristics, while discarding outliers and noisy instances. Such a compound skew-insensitive loss function will bridge the worlds between data-level and algorithm-level approaches to learning from imbalanced data. Furthermore, we want to make DeepSMOTE suitable for continual and lifelong learning scenarios, where there is a need for handling dynamic class ratios and generating new artificial instances. We envision that Deep SMOTE may not only help to counter online class imbalance, but also help increase the robustness of lifelong learning models to catastrophic forgetting. Finally, we plan to extend Deep SMOTE to incorporate other data modalities, such as graphs and text data.

In the future, planning to extend our technique to deal with multi-class problems. Moreover, we shall try to exploit different versions of SMOTE (e.g., SMOTE-Cov) with generative DL models, including Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN), to explore their effectiveness in comparison to our model on other real datasets. Our model's major trade-off/limitation is that it has a longer training time and higher computational cost than traditional ML techniques.

VI REFERENCES

[1]. B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[2]. A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Switzerland: Springer, 2018, Doi: 10.1007/978-3-319-98074-4.

[3]. L. Korycki and B. Krawczyk, "Concept drift detection from multi-class imbalanced data streams," in *Proc. IEEE 37th Int. Conf. Data. Eng. (ICDE)*, Chania, Greece, Apr. 2021, pp. 1068–1079.

[4]. L. Korycki and B. Krawczyk, "Low-dimensional representation learning from imbalanced data streams," in *Proc. Adv. Knowl. Discovery Data Mining, 25th Pacific-*

Asia Conf. (PAKDD), in *Lecture Notes in Computer Science*, vol. 12712. Researchgate.net, 2021, pp. 629–641.

[5]. C. Wu and H. Li, "Conditional transferring features: Scaling GANs to thousands of classes with 30% less high-quality data for training", *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1-8, Jul. 2020.

[6]. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321-357, Jun. 2006.

[7]. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved techniques for training GANs" in *arXiv:1606.03498*, 2016.

[8]. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville, "Improved training of wasserstein GANs" in *arXiv:1704.00028*, 2017.

[9]. C. Huang, Y. Li, C. C. Loy and X. Tang, "Deep imbalanced learning recognition and attribute ", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2781-2794, Nov. 2020.

[10]. M. Koziarski, " under sampling for imbalanced data classification", *Pattern Recognit.*, vol. 102, Jun. 2020.

[11]. W.-C. Lin, C.-F. Tsai, Y.-H. Hu and J.-S. Jhang, "Clustering-based under sampling in class-imbalanced data", *Inf. Sci.*, vol. 409, pp. 17-26, Oct. 2017.

Safeguarding Digital Landscapes: Cyber Threat Management

Dirisala Pooja Sri
 23DSC23, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 poojasridirisala@gmail.com

Dr.T.Srinivasa Ravi Kiran
 HoD & Associate Professor
 Dept. of Computer Science
 P.B.Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 tsravikiran@pbsiddhartha.ac.in

Dr.N.Lakshmi Prasanna
 Professor
 Dept of CSE,
 Vasireddy Venkatadri Institute of Technology,Nambur
 nlakshmi@vvit.net

Abstract-Cyberspace is the interconnected virtual environment formed by computer systems and networks, encompassing the entirety of the internet. In this digital realm, information is exchanged, and various online activities occur, including communication, commerce, and entertainment. Cyberspace is characterized by its intangible nature, transcending physical boundaries, and has become an integral aspect of modern life, influencing societal, economic, and communication dynamics on a global scale. This article discusses different types of attacks that intruders or hackers can carry out to gain unauthorized access to Cyber Space. It also presents measures to reduce these attacks on Cyber Space. The article conducts a thorough examination of the likelihood of security threats and explores various ways to minimize the risks of hacking, providing recommendations to enhance security in Cyber Space.

Keywords-Cyber Space, Algorithm, Protocol, Attacks, Network, Malware

I INTRODUCTION

Cyberspace, an interconnected virtual environment spanning the internet, facilitates diverse online activities from communication to commerce. Its intangible nature transcends physical boundaries, profoundly shaping modern life globally. This discussion delves into various Cyber Attacks by intruders seeking unauthorized access, offering insights into their likelihood and impact. Furthermore, it presents measures to curtail such threats, emphasizing the importance of bolstering cyber security in Cyberspace. By examining security vulnerabilities, this article advocates for robust strategies to safeguard against hacking incidents. Its comprehensive approach aims to fortify Cyberspace, ensuring its resilience amidst evolving digital landscapes [1].

II RELATED WORK

In this section, we exemplify some important Cyber Space from four aspects:

Malware Attack: It is an attack where a computer system or network is infected with a computer virus or other type of malware. It is an all-encompassing term for a variety of cyber-attacks including trojan viruses. It is defined as code with malicious intent that typically steals data or destroys something on the computer.

Social Engineering Attacks: A social engineering attacker is a person who wants access to sensitive information or money. The attacker will cause discomfort to bypass, notifying the victim's vengeful objective when was manipulating the victim [2]. Based on The National Institute of Standards and Technology (NIST), social engineering is an attempt to trick someone into revealing information (e.g., a password)



Fig. 1. Dynamics of Cyber Space

to attack systems or networks [3]. Successful social engineering attacks depend on a target being manipulated or tricked into disclosing personal information [4].

Advanced Persistent Threats (APT): Advanced Persistent Threat, as the name itself implies, is not like a regular attack or attack done by a regular hacker. APTs are achieved often by a group of advanced attackers that are well-funded by an organization or government to gain crucial information about their target organization or government. APT is a military term adapted into the information security context that refers to attacks carried out by nation-states. APT is defined by the combination of three words, [6], which are:

Advanced: APT attackers are usually well-funded with access to advanced tools and methods required to perform an APT attack. These advanced methods include the use of multiple attack vectors to launch as well as to keep the attack going.

Persistent: APT attackers are highly determined and persistent and they do not give up. Once they get into the system, they try to stay in the system for as long as they can.

They plan for the use of several evasive techniques to elude detection by their target's intrusion detection systems. They follow "low and slow" approach to increase the rate of their success.

Threat: The threat in APT attacks is usually sensitive data loss or impediment of critical components or mission. These are rising threats to many nation entities and organizations that have advanced protection systems guarding their missions and/or data.

DDoS Attack: The nature of DDoS attacks is such that Information Technology (IT) corporation, Cisco, admits that DDoS attacks commonly target "corporate assets" [7]. By inundating network resources with fake requests, DDoS attacks manage to divert the target's IT facilities from their prescribed functions, which results in unprecedented downtimes and service outage.

2.1. Factors that motivate malicious attacks through DDoS

- **Malice:** DDoS attacks are an effective means of denying their victims of their computing resources. Hence, they are an apt tool for individuals planning to inflict damage based on their malicious intent.
- **Financial gain:** companies that experience the effects of a large-scale DDoS attacks are susceptible to pay the attackers monetarily in order to regain their operational capacities.
- **Activism:** individuals who wish to make a political statement can exploit the power of DDoS to coerce their opponents or authorities towards a particular objective.
- **To gain 'hacking' credibility:** 'power' computer users gain popularity in their circles when they can prove that they can initiate and sustain a DDoS attack on a prescribed target.

2.2. Factors that facilitate the perpetration of DDoS attacks

- **Ability to escape identification:** the ability to spoof IP addresses affords attackers the capability of evading the unmasking of their identities [8].
- The factor is both an aid to the actual attack and a protective utility for the attacker on conclusion of the DDoS attack.
- **Lack of a unified Internet security policy:** the proliferation of diverse security approaches impacts the Internet with varying degrees of immunity from DDoS attacks [9].
- The lack of a singular body to enforce best practices across the interconnected networks provides a loop hole for DDoS attackers to exploit the weaker defense systems.
- **Skewed allocation of network resources:** the infrastructure that connects small networks to larger ones is usually of a higher bandwidth. The feature provides attackers with the capability to 'flood' the less endowed targets through the high-capacity infrastructure [9].

- The limited nature of network resources: since target networks have a certain limit, which serves its requirements, DDoS attacks can force the network to reach that limit and deny the users of their deserved access to services [8].

Software supply Chain Attacks: The overarching goal of delivering a software product or service to end users (i.e., on a Platform-as-a-Service or Software-as-a-Service basis).

- A complex set of relationships between different organizations, such as developers, logistic centers, and distribution and assembly centers, in which each actor in the chain can operate as a supplier and/or a customer.
- The existence of two material/service streams. The upstream connects with the product creation using third-party components. The downstream is associated with the product delivery to the end user through a distribution network. These intricate systems of interaction increase the risk of impaired transparency as the end user has little insights into the quality of the delivered products and services across the entire supply chain [5].

MITM: Man-in-the-middle attacks are one of the most commonly used network attacks. This attack happens when the attacker manages to get in the middle between two parts of communication: the sender and the receiver. MITM attack using ARP Poisoning is the most commonly used technique to perform MITM attacks and this is because of the poor security of ARP protocol and also because it is the simplest way to perform the attack. Address Resolution Protocol (ARP) is a protocol that creates a mapping between MAC address and the IP address. These protocols work by using two types of messages: request and reply. The communication contains two parts: source host and destination host. ARP Request is broadcasted and is used to find which MAC address maps a certain IP. All the hosts get this request but only the host whose IP address matches the IP address in the header of the ARP Request responds to the request. To lower network traffic flow, every host has an ARP cache, which is a table that maps IP addresses with MAC addresses of every host connected to the network. ARP Poisoning means the 'the poison' of ARP cache using the main vulnerability of ARP protocol. The vulnerability of ARP protocol is that is a non-state protocol and the hosts will accept ARP reply even if they haven't sent any ARP request. This means that they will update their ARP caches every time there is an ARP reply. Because the ARP requests are broadcasts, every host connected to the network can get the requests. The attacker sends a response using a copied MAC address, and he attacks the two parts of the communication. He attacks the source host and sends him an ARP Reply where he tricks the source to believe that the IP address of the destination host maps the MAC address of the attacker, and he sends an ARP Reply to the destination host where he tricks the destination to believe that IP address of source host maps the MAC address of the attacker. After this, the source thinks that the attacker is the destination and the destination thinks

that the attacker is the source. So, every information that source and destination hosts send to each other firstly passes to the attacker, and then he forwards the packets to them. This type of attack is performed on switches and access points but not on routers because the router will not pass ARP packets to other routers [10].

Password Attacks: It is an attempt to obtain or decrypt a user's password for illegal use. Hackers can use cracking programs, dictionary attacks, and password sniffers in password attacks. Defense against password attacks is rather limited but usually consists of a password policy including a minimum length, unrecognizable words, and frequent changes. This attack can be done for several reasons but the most malicious reason is to gain unauthorized access to a computer without the computer's owner's awareness not being in place; so, this results in cybercrime such as stealing passwords to access bank information. There are three common methods used to break into a password protected system.

- **Brute-force attack:** In this, a hacker uses a computer program or script to try to log in with possible password combinations usually starting with the easiest to guess password.

Dictionary attacks: In this, a hacker uses a program or script and tries to log in by cycling through the combinations of common words. This attack tries only those possibilities which are most likely to succeed;

- typically derived from a list of words; for example, dictionary.
- These attacks are more successful because people tend to choose easy passwords like their names, birthdates, etc.
- **Keylogger attacks:** - In this, the hacker uses a program to track all of the user's keystrokes; so, at the end of the day, everything the user has typed including the login IDs and passwords has been recorded [1].

III PROPOSED WORK

We propose the following security methods to safeguard the Cyber Space from various security attacks.

1. Use Strong and Complex Passwords: Create strong passwords by using a combination of letters, numbers, and special characters (where allowed). Avoid passwords that are based on personal information that can be easily accessed or guessed. Use numbers and symbols to create words that can't be found in any dictionary of any language.

2. Keep Your Software Up to Date: To keep your software up to date because updates enhance existing features, patch security flaws, add new security features, fix bug issues and improve performance for devices. Continue reading to learn more about software updates and how you It's important can check if your software is up to date.

3. Use Anti-Virus Protection: Antivirus software protects your device from viruses that can destroy your data, slow down or crash your device, or allow spammers to send email

through your account. Antivirus protection scans your files and your incoming email for viruses, and then deletes anything malicious.

4. Authentication: Authentication is a process that verifies that someone or something is who they say they are. Technology systems typically use some form of authentication to secure access to an application or its data.

5. Use Two-Factor or Multi-Factor: Using two-factor or Multi-factor in Cyber Space enhances security by requiring users to provide multiple forms of verification before accessing an account or system. This typically involves a combination of something you know, something you have, something you are. Implementing these measures adds an extra layer of protection against unauthorized access and strengthens overall Cyber Space.

6. Access Control to Data and System: Access control is a data security process that enables organizations to manage who is authorized to access corporate data and resources. Secure access control uses policies that verify users are who they claim to be and ensures appropriate control access levels are granted to users.

7. Put up a Firewall: A firewall is a network security device that monitors incoming and outgoing network traffic and decides whether to allow or block specific traffic based on a defined set of security rules.

8. Use Security Software: Security Software aims to prevent unauthorized access to data that is stored electronically. This type of software protects businesses from data theft, malicious data, and system usage by third parties.

9. Programs and Systems Regularly: Regularly updating programs and systems in Cyber Space is crucial for maintaining a secure digital environment. Software updates often include patches for vulnerabilities, security enhancements, and bug fixes. Failing to update can leave systems exposed to exploitation by Cyber threats. Establishing a routine for applying patches and staying informed about the latest security updates helps safeguard against potential risks and ensures that your digital infrastructure remains resilient to evolving Cyber threats.



Fig. 2. Various ways to protect from

Algorithm:

1. Begin
2. Identify Cyber Security Threats.
3. Focus on the Most Probable Threats That Could Harm Cyber Space.
4. Determine Security Measures to Protect Cyber Space.
5. Put in Place Measures to Effectively Protect Cyber Space.
6. Assess the Level of Security to Prevent Unauthorized Access.
7. End

IV RESULT & ANALYSIS

First, we focus on types of threats that are possible in Cyber Space and Percentage of Vulnerability because of each threat.

S.No.	Types of Threats possible on Cyber Space	Percentage of Vulnerability
1	Malware Attack	12
2	Social Engineering Attacks	17
3	Software Supply Chain Attacks	17
4	Advanced Persistent Threats (APT)	15
5	Distributed Denial of Service (DDoS)	9
6	Man-in-middle Attack (MitM)	12
7	Password Attacks	18
Vulnerability before the implementation of Proposed Security Measures		100

Table 1. Vulnerability in Cyber Space before implementing Security Measures.

The above table shows the statistics of cyber space threats and vulnerability percentage because of each threat. To clearly understand these threats effect on cyber space, the pie chart is given below.

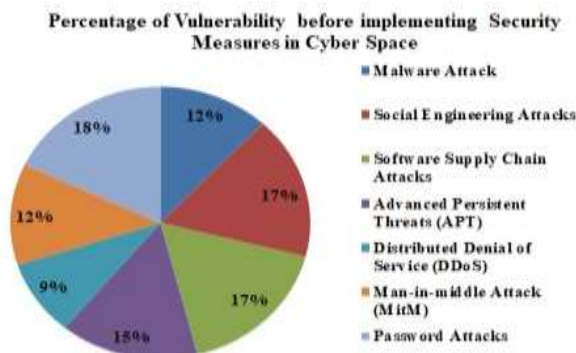


Fig. 3. Percentage of vulnerability before applying the security features in cyber space.

After implementing the proposed methods, the percentage of vulnerability is as below in Table 2.

To

S.No.	Types of Threats possible on Cyber Space	Percentage of Vulnerability
1	Malware Attack	4
2	Social Engineering Attacks	7
3	Software Supply Chain Attacks	6
4	Advanced Persistent Threats (APT)	2
5	Distributed Denial of Service (DDoS)	5
6	Man-in-middle Attack (MitM)	3
7	Password Attacks	1
Vulnerability after implementation of Proposed Security Measures		28

Table 2. Vulnerability in Cyber Space after implementing Security Measures.

clearly understand the difference before and after, pie chart is given below:

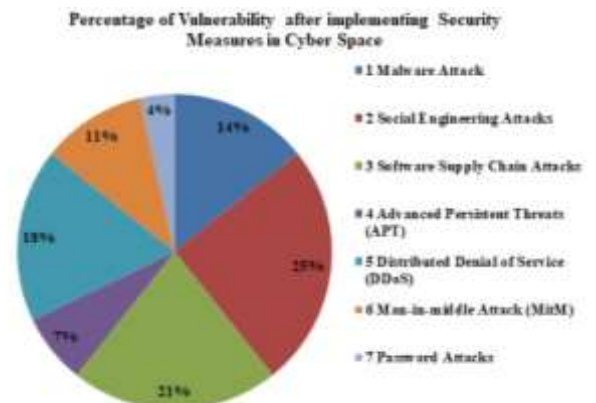


Fig. 4. Percentage of vulnerability after applying the security features in cyber space.

We observe that the percentage of vulnerability has decreased to 28.

V CONCLUSION & FUTURE WORK

Even though several security measures are implemented using security protocols / firewalls, these are unable to protect the vulnerabilities of Cyber Space. Hackers / intruders are continuously making attempt to gain the unauthorized access in Cyber Space using various attacks. As Cyber Space usage has increased privacy and security challenges, this will have an effect on their usage. In order to protect the security and integrity of Cyber Space, several new security measures, protocols and firewalls are needed to be developed and deployed effectively to challenge unauthorized access.

VI REFERENCES

- [1] Aishwarya Pradeep Zope & Rashmi Ravindra Chaudhari, "A Review Paper on Cyber-Security", International Research Journal of Engineering and Technology (IRJET), August 2022, Volume: 09 Issue: 08, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Page 1561-1566, <https://www.irjet.net>.
- [2] Wenni Syafitri et al., "Social Engineering Attacks Prevention: A Systematic Literature Review", April 2022, DOI:10.1109/ACCESS.2022.3162594, IEEE.
- [3] K. Scarfone, M. Souppaya, A. Cody, and A. Orebaugh, "Technical guide to information security testing and assessment recommendations of the National Institute of Standards and Technology (SP 800-115)", NIST Special Publication, 2008, pp.1-80, vol. 800.[Online]. Available: <https://csrc.nist.gov/publications/detail/sp/800-115/final>
- [4] M. Junger, L. Montoya, and F.-J.Overink, "Priming and warnings are not effective to prevent social engineering attacks", Comput. Hum. Behav., vol. 66, pp. 75-87, Jan. 2017.
- [5] Sean Cordey, "Software Supply Chain Attacks An Illustrated Typological Review" Center for Security Studies (CSS), ETH Zürich, January 2023, DOI: 10.3929/ethz-b-000584947, <https://css.ethz.ch/en/publications/risk-and-resilience-reports.html>.
- Engineering and Technology (IRJET), August 2022, Volume: 09 Issue: 08, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Page 1561-1566, <https://www.irjet.net>.
- [6] R. S. Ross, "Managing information security risk: Organization, mission, and information system view," Special Publication (NIST SP)- 800-39, 2011.
- [7] Cisco, 'DDoS attack prevention', Cisco, 2015. [Online]. Available: <http://www.cisco.com/c/en/us/solutions/enterprise-networks/ddos-attack-prevention/index.html>. [Accessed: 14- Sep - 2015].
- [8] J. Mirković, G. Prier and P. Reiher, 'Attacking DDoS at the source', in Network Protocols, 2002. Proceedings. 10th IEEE International Conference, 2002, pp. 312-321.
- [9] F. Freiling, T. Holz and G. Wicherski, 'Botnet tracking: Exploring a root-cause methodology to prevent distributed denial-of-service attacks', in Proceedings of the 10th European Symposium on Research in Computer Security, Milan, Italy, 2005, pp. 319-335.
- [10] Enkli Yllia , Dr. Julian Fejzajb, Man in the Middle: Attack and Protection, Proceedings of RTA-CSIT 2021, May 2021.

The Relation of Big Data Analytics with CRM

Sri Lekha Ejnavarjala
 23DSC24, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts and Science
 Vijayawada, A.P.,India
 srilekha.ejnavarjala03@gmail.com

Gayathri Marrivada
 23DSC10, M.Sc.(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts and Science
 Vijayawada, A.P., India
 marrivadagayathri@gmail.com

Patnala Meghana Durga
 23DSC04, M.Sc(Computational Data Science)
 Dept. of Computer Science
 P.B.Siddhartha College of Arts and Science
 Vijayawada, A.P., India
 meghanapatnala786@gmail.com

Abstract— Big Data Analytics is a pivotal in modern business, particularly in enhancing Customer Relationship Management (CRM). Managing vast data, categorized by Velocity, Volume and Variety, poses a significant challenge. Leveraging techniques like Data mining Frameworks is essential to integrate big data into CRM, enabling better decision-making through extensive database analysis. In the globalized economy and rapid e-commerce growth, CRM has emerged as a cornerstone for company expansion. Big Data fuels personalized customer experiences and tailored sales/ services, demanding new tools for its capture, storage, and analysis. E-Commerce sets a precedent for personalized consumer interactions, raising expectations across industries like banking, retail, and media. Despite recognizing Big Data's transformative potential, businesses often grapple with identifying suitable cases. This research scrutinizes Big Data Analytics, Data Mining techniques, and analytical frameworks pertinent to CRM. Additionally, it delves into applications of key data mining methods- clustering, classification and association rules- in CRM.

Keywords—Customer Relationship Management, Volume, Velocity, Variety, Data Mining, Customer Relationship Management Performance, Customer Orientation.

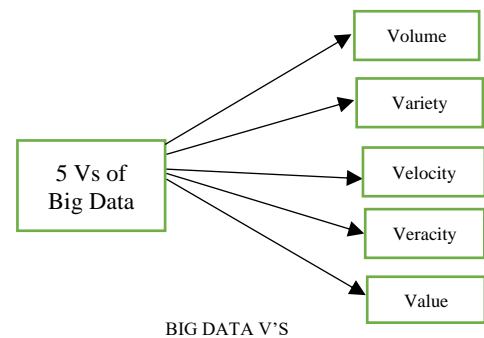
I INTRODUCTION

1.1 Big Data

Big Data is a word that is used to denote huge amounts of data that cannot be handled in normal methods. Big data analytics refers to examining and analyzing large and varied data sets with the purpose of discovering hidden patterns, trends and different customer preferences etc. This information will be a competitive advantage to face the business world. With the development of Networking, Internet and computers Big Data Analytics have become a major trend in the business world [1]. The patterns identified using Big Data Analytics can be used for purposes such as advertising and even for new product innovations. But different techniques and methods should be adapted to process Big Data.

Big data creates opportunities for business that can use it for generating business value. The purpose is intended to gain value from volumes and a variety of data by allowing velocity of analysis. It is known as 5 Vs model; volume, velocity, and variety, value, and veracity. Volume means processing massive data scale from any data type gathered. The

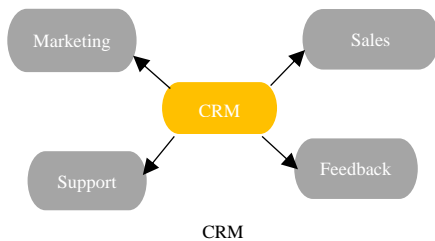
explosive of data volumes improve a knowledge sharing and people awareness. Big data is a particularly massive volume with a large data sets, and those data cannot be analyzed its content using traditional database tools, management, and processing. Velocity means real time data processing, specifically data collection and analysis. Velocity processes very large data in real-time processing. In addition, big data escalates its speed velocity surpassing that of old methods of computing. Variety is any types of data from various channels including structured and unstructured data like audio, video, image, location data for example Google Map, webpage, and text, as well as traditional structured data. Veracity refers to data authenticity with the interest in the data source of Web log files, social media, enterprise content, transaction, data application. Data needs a valid power of information to ensure its authenticity and safety.



1.2 CRM

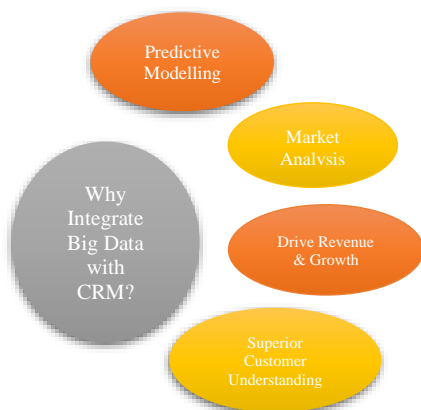
Customer Relationship Management refers to managing of customer interactions and relationships with a business. It handles attracting new customers, retaining existing customers and making a bond between customers stronger [3]. In Customer Relationship Management different techniques such as Data Mining are used to extract information about customers and understand their preferences. Thereby provide a better service to the customer. At present, businesses use different software systems to manage customer relationships. Because of the development of internet and wireless technologies, there is a large flow of information about customers. And many successful

companies are interested in using those data to manage their customer relationships effectively. the economic globalization and rapid development of electronic commerce is changing the original competition rules between the enterprises with an unprecedented scope and depth. More and more enterprises start to focus on a new marketing management theory—customer relationship management (CRM), which places the retaining and upgrading of customer value as the core of enterprise organic development [1]. With a mass of business data and customer information, how to pick up useful information and maximize customer value becomes the key customer relationship management problem needs to be solved urgently for modern enterprises. Data Mining is just the very tool suitable to solve this problem.



1.3 Big Data in CRM

Big Data Analytics for Management of Customer Relationships is a major trend in the business world because of the large amounts of information that flows over the internet and networks. Big Data is a useful tool to identify what customers actually expect from companies and predict their future demands. Thereby analyzing big data can be used to provide a better service to the customers and manage their relationships effectively. This research focuses on how to utilize Big Data Analytics to maintain customer relationships effectively and successfully.



INTEGRATION OF BDA WITH CRM

Objective of the Study:

1. To study the relation of how Big Data Analytics is interlinked with CRM.
2. To know about the current trends of Big data analytics applied to CRM.

II RELATED WORK

Usage of Big Data Analytics in CRM

Big Data can be identified as large data sets that cannot be handled by normal computing techniques. It consists of, Structured data (easy to use, handle, model and extract information. Have a specific type and size. Can be presented

in relational database or spreadsheet.), Semi-structured data (Has meta model (tags and markers),for an example XML), Unstructured data (No pre-defined format or structure). To use we need methods such as NOSQL.)

Big data enable large opportunities in business growth and development, government planning, innovations, new market identification and healthcare services etc.

Big Data cannot be handled by normal methods. Complex methodologies are needed for processing and managing Big data. These tools are called Big Data analytics.

The first method is the verification of assumption. In this method data analysts make an initial assumption and verify the assumption through data analysis. The next method is identification. In this mode data analysts collect data through different sources and make an effort to find hidden patterns and information in them.

Big data also reduces the maintenance costs for instance, organizations deploy cloud computing approach where data are stored in the cloud [3]. The emergence of cloud computing has enabled big data analytics to be cost efficient, easily accessed, and reliable. Cloud computing is robust, reliable and responsive when there are issues because it is responsible of cloud service provider. Since, service outages are unacceptable at the business. Whenever data analytic goes down impacting marketing activities are disrupted and customers have to question whether to trust such a system. Therefore, reliability is competitive advantage of cloud computing in big data application [3].

C. Descriptive analytics

Descriptive analytics is the set of techniques which are used to describe and report on the past. Retailers can use descriptive analytics to describe and summarize sales by region and inventory levels. Examples of techniques include data visualization, descriptive statistics and some data mining techniques [5].

D. Predictive analytics

Predictive analytics consists of a set of techniques which use statistical models and empirical methods on past data in order to create empirical predictions about the future or determine

the impact of one variable on another. In their retail industry, predictive analytics can extract patterns from data to make predictions about future sales, repeat visits by customers and likelihood of making an online purchase. Examples of predictive analytic techniques which can be applied to big data include data mining techniques and linear regression.

E. Prescriptive analytics

Prescriptive analytics uses data and mathematical algorithms in order to determine the best course of actions to take based on a set of requirements and with the objective of improving business performance. Retailers can use prescriptive analytics to determine price mark down models to aid in setting discount levels to maximize revenue. Examples of prescriptive analytic techniques which can be applied to big data include optimization methods.

F. Scalability of analytic algorithms

As big data is concerned with high volumes, one of the challenges to data analysis is the scalability of the algorithms which are used for analyzing big data. The scalability of the algorithm is the ability of the algorithm to scale rapidly with increasing dataset volumes. As the size of data grows, the time taken to access that data becomes less and less efficient. Analytical algorithms should therefore be developed to ensure scalability as datasets grow in volume. The challenge for retailers was in selecting analytical algorithms which would allow them to cope with large volumes of data.

Customer Relationship Management is a concept in which the three aspects People, Process and Technology are combined together and is used to understand and maintain customer relationships with the company. Techniques such as Text Analytics and other Big Data Analytic techniques play an important role in modern CRM systems.

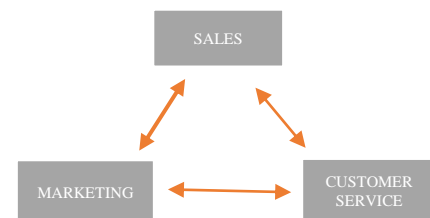
Big data analytics helps the CRM by following key points:

- Enhanced Customer Insights- This includes demographics, behaviors, preferences, social media interactions, and more. CRM integrated with Big Data enables businesses to gain deep insights into customer behavior, needs, and preferences.
- Personalized Customer Experiences- This involves targeted marketing campaigns, tailored product recommendations, and personalized communication, leading to better engagement and customer satisfaction.
- Predictive Analytics- It allows businesses to anticipate customer behavior, identify trends, forecast sales and make data driven decisions about future strategies.
- Improved Customer Service- Analyzing customer data in real time allows for quicker issue resolutions, personalized support and the ability to address concerns before they escalate.

III PROPOSED WORK

Business realizes that their most valuable assets are relationships with customers and all stakeholders. In fact, building personal and social relationships become important area in marketing. The importance of relationships as market-based assets that contribute to customers' value. With the amount of data increase, some business organizations use advanced powerful computers with a huge storage to process big data analytics and to increase their performance resulting in tremendous cost saving [2]. Businesses manage structured and unstructured data sources such as social marketing, retail databases, recorded customer activity, logistics, and enterprise data to establish a quality level of CRM strategies by having the abilities or knowledge on how to recognize big data and its advantage. While, big data analytics is a process to reveal the variety of data types in big data itself. There are some CRM strategies that can happen through big data and big data analytics.

CRM with big data brings a promise of big transformation that can affect organization in delivering CRM strategies. There were many benefits for using big data in CRM and the following were just some of the benefits such as accurate and update in profiling of target costumers, predicting trend on customer reaction toward marketing messages and product offerings, create personalize message that create emotional attachment and product offering, maximizing value chain strategies, producing accurate assessment measures, effective digital marketing and campaign-based strategies, customers retention which was a cheaper option, and create tactics and getting product insights.



CRM SCOPE

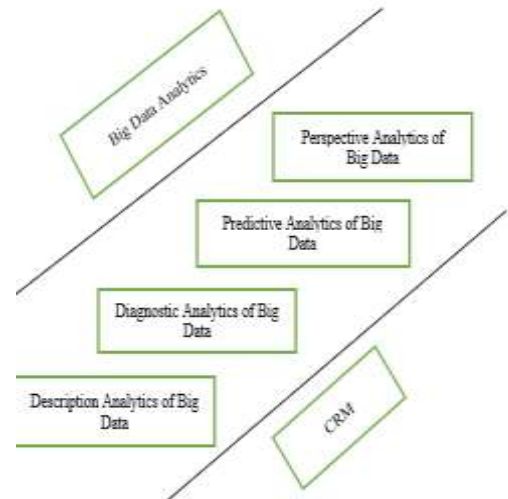
CRM's performance in Big Data Analytics

Customer Relationship Management (CRM) systems play a vital role in utilizing big data analytics to enhance performance and drive business growth. Here's a detailed breakdown:

1. Data Accumulation- CRM systems serve as repositories for diverse customer-related data, demographics, purchase history, interactions, preferences, social media engagements, etc. Big data analytics within CRM leverages this wealth of information, integrating it from various sources to create comprehensive customer profiles.
2. 360 Degree Customer View- Big data analytics in CRM consolidates and analyzes data from multiple touch points, offering a holistic view of each

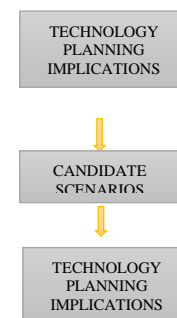
customer. This enables businesses to understand behavior patterns, preferences, and predict future actions, fostering more personalized interactions.

3. **Predictive Analytics-** By employing machine learning algorithms and predictive modeling, CRM systems can anticipate customer behavior and trends. This enables proactive decision-making, personalized marketing strategies, and tailored offerings based on anticipated needs.
4. **Enhanced Customer Segmentation-** BDA helps segment customers more precisely based on behavior, preferences and buying patterns. This segmentation allows for targeted marketing campaigns, personalized recommendations, and tailored services, thereby improving customer satisfaction and retention.
5. **Improved Customer Services-** Analytics tools within CRM systems enable sentiment analysis, allowing companies to gauge customer feedback and sentiment from various channels. This real time analysis helps in identifying potential issues, improving service delivery, and addressing customer concerns promptly.
6. **Optimized Sales and Marketing Efforts-** Big data analytics in CRM assists in optimizing sales and marketing strategies. It helps identify the most profitable customer segments, the most effective communication channels, and the best times to engage, leading to more efficient lead generation and conversion.
7. **Data- Driven Decision Making-** With the insights derived from big data analytics, businesses can make data-driven decisions. It enables them to allocate resources more effectively, prioritize opportunities, and fine-tune strategies based on empirical evidence rather than assumptions.
8. **Business Process Optimization-** CRM analytics provides insights into operational inefficiencies and bottlenecks, allowing organizations to streamline processes, enhance workflows and improve overall operational performance.
9. **Customer Retention and Loyalty-** By understanding customer behavior and preferences, CRM systems help in designing loyalty programs and retention strategies tailored to individual customers, thus fostering long-term relationships.
10. **Continuous Improvement-** Analytics in CRM systems facilitate continuous improvement by providing feedback loops. Businesses can track the effectiveness of their strategies, campaigns and customer interactions, allowing for adjustments and refinement to enhance performance continually.



PERFORMANCE OF BDA WITH CRM

Since big data can provide a pattern of customers' information, businesses can predict and assume what are the needs of their customers now-a-days. Big data had helped shaped many industries and changed the way businesses operated now-a-days. Big companies definitely benefited from this shift especially companies such as technology giants such as Amazon and google and would continue to serve these giants from the sheer volume of data they generated. Data Velocity showed how marketers could have access to real-time data, for example real time analytics of interactions on internet sites and also social media interactions



FRAME WORK FOR STRATEGIC PLANNING

Data Mining

Data mining, which is also called KDD (Knowledge Discovery in Database), is the process of abstracting unaware, potential and useful information and knowledge from plentiful, incomplete, noisy, fuzzy and stochastic actual data, it is a process to pick up the information and knowledge which

cannot be discovered directly but with potential value from a mass of data. Data Mining is an interdisciplinary with extensive knowledge scope, which mainly involved into three areas: database, artificial intelligence and probability and mathematical statistics. For the detailed application, the data mining has powerful data analysis and process capability, which can discover knowledge as useful rules, regulations, relationship and modes from a mass of data, by means of concept description, clustering analysis, classification and forecast, similar pattern analysis, abnormal pattern analysis, association analysis, time sequence analysis and regression analysis, etc.

Techniques and technologies for big data analytics

Techniques and technologies have been developed which can help capture, store and analyze big data to obtain information which is of value for decision making.

Neural networks-Neural networks are computational models which are based on biological neural networks and are used for detecting patterns in data Retailers apply neural networks to: identify high-value customers which are at risk of leaving a particular company and detect fraudulent insurance claims.

Machine Learning- Machine learning is a part of the field of artificial intelligence and involves the design of algorithms which allow computers to adapt behavior based on empirical data

Cluster analysis- Cluster analysis uses techniques to break down a diverse group into smaller groups of objects with similar characteristics.

IV EXSISTING SYSTEMS

Analyzing Big Data is a major trend in the present society. At present Big Data Analytics are used for many applications [1]. Applications of Big Data Analytics related to customer relationship management are as follows,

A. Intelligent Systems

The Intelligent system is a software mechanism used for interacting and cooperating with intelligent agents in a framework. It has capability to handle huge volumes of data and can give valuable insights to the users.

B. Data Mining and Big Data with CRM

Data mining has the ability to identify the patterns in customer behavior, which helps the Business to understand customer behavior and their preferences. Large CRM software analyzes big data to find patterns which are available over the internet and other networks.

C. Online Government Data Analysis

Processing open government data to analyze customer demands and patterns is a profitable area in Big Data Analytics which prevail in many countries. With the

development of technology society is more dependent on internet, social networks and mobile technology and then OGD will accelerate economic growth and new business initiatives. It is generally used in banking sector and other business organizations.

D. Micro-Segmenting Customers based on Big data

Big Data can be used to segment and micro-segment existing and potential customers and then optimize diversity in retail sector. This is done through correlations between items bought as well as location and time dependent purchase patterns.

E. Big Data based Promotions

Use of Big Data Analytics in promotion decisions is another application. By identifying proper customer demands and behavior patterns these promotion decisions will become efficient and effective. Amazon uses big data analytics to predict demand in different geographical locations.

F. Social CRM

It enables a better social CRM by Centralization of related knowledge, Quick compilation, recording history etc. It uses Big Data. Social Network Analysis is a Big Data Analytic technique used in Social CRM which is explained under major researches in this research paper.

G. Customization and Personalization

These data are used by company giants such as Amazon and Netflix to build personal and social relationships with customers.

V CHALLENGES

Big Data can be defined as huge volumes of facts and data that cannot be handled in normal methods [1]. There are many issues that will arise when analyzing big data [4].

H. Limited Storage

First one is Storage. Data storage is the basis of Big Data Networking aspect [1]. And to store such large volumes of data high-capacity media are needed which are not found in normal context.

I. Lack of Quality of Data Collected

Second one is Quality of Data. Big Data may include a vast amount of data that flow over the internet. Most of these data might not be accurate and a proper method is needed to extract correct data.

J. Abundance of many Irrelevant Data

Next point is Relevance of the data. Most of the data we extract in Big Data Analytics would not be relevant to our subject. Most of the researches that were mentioned in this paper have tried to address this problem and find the most relevant data.

K. High Cost

Last point is the Cost. Because of the above reasons the cost of handling big data analytics would be very high and proper methods are needed to handle maintain an efficient procedure when analyzing big data

VI CONCLUSION

The intended outcome of this study is to analyze Big Data Analytics techniques that can be used in Customer Relationship Management. In this research we have identified major issues in Big Data Analytics in Customer Relationship Management such as storage issues, quality issues, process issues and cost issues. And discussed how we can use Big Data Analytics in Customer Relationship Management successfully overcoming those problems. This study discusses about frameworks adapted to use big data to be used in customer relationship management and applications of Big Data Analytics in Customer Relationship Management. Finally, we can identify that Big Data Analytics is a difficult aspect to be used in Customer Relationship Management, but by using a proper framework identified it is possible to use Big Data Analytics effectively in Customer Relationship Management.

VII FUTURE SCOPE

Many researchers are developing frameworks and techniques to use Big Data Analytics properly for Customer Relationship Management.

- There will be researches that focus on an evaluation with comparisons for the suitable algorithms that analyze online customers Future will direct attention to use big data analytics to segment customers and build up better relationships.
- There will be researches in the fields of pricing of products, availability of resources and other matters, diversity of customers and planning the layout of products especially in retail sector.
- Predictions to identify customer demands and improve customer satisfaction.
- These data can be used to address Customer Relationship Management issues and world will point to usage of Enhanced Knowledge Management Techniques for Customer Relationship Management.

VIII REFERENCES

- [1] W.K.R.Perera Faculty of Information Technology, University of Moratuwa, Moratuwa, Sri Lanka, K.A.Dilini, Department of Computational Mathematics, University of Moratuwa, Sri Lanka
- [2] Kun Wu School of management Tianjin Polytechnic University Tianjin, Feng-ying Liu School of management Tianjin Polytechnic University Tianjin, China
- [3] Matthew Ridge, Kevin Allan Johnston and Brian O'Donovan, Vol. 9(19), pp. 688-703, 14 October, 2015 DOI: 10.5897/AJBM2015.7827 Article Number: 678BD4555626 ISSN 1993-8233

- [4] P. S. Hema Latha P. S. Lalitha Assistant Professor, Department of MCA Assistant Professor, Department of MBA KMM IPS, Tirupati KMM ITS, Tirupati.
- [5] Muhmmad Anshari, Mohammad Nabil Almunawara, Syamimi Ariff Lima, Abdullah Al-Mudimigh b aUniversiti Brunei Darussalam, Brunei Darussalam bDar Al Uloom University, Saudi Arabia.
- [6] Romika Yadav, Monika, Tarun kumar, Garirma, "Usage of BDA with CRM " Management, "International Journal of Advanced Research in Computer Science"
- [7] Injazz J.Chen and karen popovich, "Understanding customer relationship management people, process and technology," Business process management journal
- [8] C, -h.Liu, "A conceptual framework of analytical CRM in big data age", (IJACSA) International Journal of Advanced Computer Science and Applications, vol 6, no.6,p.4,2015.

An Optimal Algorithm for Finding Champions in Tournament Graph

Gajjalakonda Keerthi,
 23DSC25, M.Sc.
 (Computational Data Science)
 Dept. of Computer Science, P.B.
 Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 gajjalakondakeerthi@gmail.com

Penumudi Bhargavi,
 23DSC05, M.Sc. (Computational
 Data Science)
 Dept. of Computer Science, P.B.
 Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 bhargavibharu741@gmail.com

Singavarapu Bhanu Sri
 23DSC32, M.Sc. (Computational Data
 Science)
 Dept. of Computer Science, P.B.
 Siddhartha College of Arts & Science
 Vijayawada, A.P, India
 bhanusrisingavarap@gmail.com

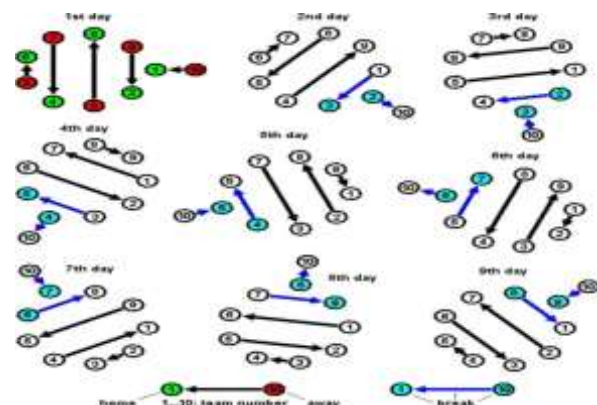
Abstract: This paper proposes an algorithm to find the Copeland winner in a tournament graph with n players. The Copeland winner is the player who wins the highest number of matches in the tournament. The algorithm can speed up several Information Retrieval and Recommender System applications, including question answering, conversational search, etc. The authors prove that any deterministic or randomized algorithm finding a champion with constant success probability requires $\Omega(\ell n)$ comparisons, where ℓ is the number of matches lost by the champion. They then present an asymptotically optimal deterministic algorithm matching this lower bound without knowing ℓ , and they extend their analysis to three variants of the problem. Lastly, the authors conduct a comprehensive experimental assessment of the proposed algorithms on a question answering task on public data. Results show that their proposed algorithms speed up the retrieval of the champion up to 13 \times with respect to the state-of-the-art algorithm that performs the full tournament.

Keywords: *Tournament Graph, Round-Robin Tournament, Copeland Winner, Minimum Selection, Pairwise Ranking*

I INTRODUCTION

A tournament graph is a complete directed graph $T = (V, E)$, where V and E are the sets of nodes and arcs, respectively. The orientation of an arc tells the winner of the match, i.e., we have the arc $(u, v) \in E$ iff u beats v in their match. In the following, we call arc lookup or arc unfold the operation of looking at the direction of an arc between two nodes. The problem of finding a champion of the tournament, also known as Copeland winner, can be addressed by finding a vertex in V with the maximum out-degree, i.e., a player that wins the highest number of matches. Our goal is to find a champion by minimizing the number of arc lookups, i.e., the number of matches played. Note that a tournament graph may have more than one champion. In this case, we aim at finding any of them, even if all the proposed algorithms can find all of them without

increasing the complexity. If the tournament is transitive—whenever u wins against v and v wins against w , then u wins against w —we can trivially identify the unique tournament champion with $\Theta(n)$ arc lookups. Indeed, the champion is the only vertex that wins all its matches and, thus, we can perform a knockout tournament where the loser of any match is immediately eliminated. However, finding the champion of general tournament graphs requires $\Omega(n^2)$ arc lookups, and thus, there is nothing better to do than to play all the matches. This means that the structure of the underlying tournament graph heavily impacts the complexity of the problem. The authors present an optimal algorithm for finding champions in tournament graphs. They prove that any deterministic or randomized algorithm finding a champion with constant success probability requires (n) comparisons, where n is the number of players. They then present an asymptotically optimal deterministic algorithm matching this lower bound without knowing the number of matches lost by the champion, and they extend their analysis to three variants of the problem. Lastly, they conduct a comprehensive experimental assessment of the proposed algorithms on a question answering task on public data. Results show that their proposed algorithms speed up the retrieval of the champion up to 13 with respect to the state-of-the-art algorithm that performs the full tournament. This parametrization of the problem with the number ℓ of matches lost by the champion is motivated by many applications in information retrieval and recommender systems that exploit pairwise ranking.



Example of a round-robin tournament with 10 participants

II RELATED WORK:

Tournament graphs, a widely utilized model across sociology, psychology, statistics, and computer science, have diverse applications, including round-robin tournaments, paired-comparison experiments, majority voting, and communication networks. Two primary research lines emerge: one focuses on determining the tournament winner, while the other centers on ranking candidates using pairwise methods. The identification of a tournament winner lacks a universal definition, but commonly involves recognizing a Condorcet winner, a candidate defeating all others. The complexity of finding a winner varies based on the tournament's characteristics. For transitive tournament graphs, the task is straightforward, achieved in linear time through a knock-out tournament. However, for general tournaments, the complexity depends on the specific definition of a winner. Banks proposes a solution based on maximal transitive sub-tournaments, with the computation of just one winner in $\Theta(n^2)$ arc lookups, while finding all winners becomes an NP-hard problem. Slater defines a winner through a total order on vertices, and despite being NP-hard, Ailon et al. introduce a Quicksort-based approximation method. Shen et al. and Ajtai et al. introduce rankings based on the notion of a king, defined as a vertex connected to every other vertex by a directed path of length at most 2. Jian et al. provide an algorithm for a sorted sequence of vertices with a complexity of $O(n^{3/2})$. Despite progress, the deterministic complexity of finding a king remains unknown. However, the definition of king is considered weaker than that of a Copeland winner. The ranking problem under persistent comparison errors involves scenarios where random noise affects queries, resulting in a transitive order on vertices. Geissmann et al. establish lower bounds on dislocation and settle the problem in $O(n \log n)$ time. While this model does not guarantee a strong champion quality, it places the champion within the top $O(\log n)$ candidates. Non-persistent comparison errors, where comparisons can be queried more than once, have been studied. Recent advancements in approximate and minimum-selection problems, such as settling the complexity of minimum-selection by Leuci and Liu, are relevant to this model. Numerous winner definitions exist, with Copeland's solution ranking vertices based on match wins. Geisman et al. consider a model with bounded errors and propose an algorithm for finding the Copeland winner in $O(n\sqrt{e})$ comparisons and time.

LOWER BOUNDS:

Within this section, we present a proof for the lower bound of the Copeland winner problem. Gutin et al. [16] utilized an adversarial approach to demonstrate that the task of identifying a champion mandates a minimum of $\Omega(n^2)$ arc lookups under the most challenging conditions. The Copeland winner problem, defined by Copeland [12], involves ranking vertices by their match victories, with the champion being the candidate securing the highest number

of wins. The lower bound proof offered here builds upon Gutin et al.'s adversarial argument. This argument seeks to establish that, in scenarios where the challenges are maximized, the process of pinpointing a Copeland winner requires, at minimum, $\Omega(n^2)$ arc lookups. This implies that any algorithm devised for this purpose must, in the worst-case scenario, examine a quadratic number of arcs within the tournament graph.

Through this adversarial reasoning, this section contributes to a deeper understanding of the inherent complexity involved in the Copeland winner problem, providing insights into the computational intricacies associated with ranking candidates based on their performance in a tournament

III OPTIMAL DETERMINISTIC ALGORITHM:

Within this section, we unveil an uncomplicated, deterministic, and asymptotically optimal algorithm crafted to discover all champions using $\Theta(n)$ arc lookups and time. The algorithm is introduced, its correctness is proven, and a constraint on the number of arc lookups is established. Furthermore, we explore implementation specifics, showcasing that the algorithm's operational complexity is $\Theta(\ell n)$, with linear space requirements.

Algorithm 1

```

1: procedure FINDCHAMPION( $T = (V, E)$ )
2:   for ( $\alpha = 1$ ; true;  $\alpha = 2\alpha$ ) do
3:      $A = V$ 
4:      $S = \{(u, u) \mid u \in V\}$ 
5:      $\forall u \in V \text{ lost}[u] = 0$ 
6:     while  $|A| > 2\alpha$  do
7:       choose a pair of vertices  $u, v$  in  $A^2 \setminus S$ 
8:        $S = S \cup \{(u, v), (v, u)\}$ 
9:       loser = if  $(u, v) \in E$  then  $v$  else  $u$ 
10:      ++lost[loser]
11:      if lost[loser]  $\geq \alpha$  then
12:         $A = A \setminus \{\text{loser}\}$ 
13:      end if
14:    end while
15:     $c, \text{lost}_c = \text{FINDCHAMPIONBRUTEFORCE}(A, E)$ 
16:    if  $\text{lost}_c < \alpha$  then return  $c$ 
17:    end if
18:  end for
19: end procedure

```

ALGORITHM DESCRIPTION:

Algorithm 1 is elaborated as follows. The algorithm faces the challenge of an unknown number ℓ of matches lost by the champion. It addresses this uncertainty by conducting an exponential search (line 2) to identify the suitable value of α , ensuring that $\alpha/2 \leq \ell < \alpha$. The problem

is then approached under the assumption that the champion loses fewer than α matches. In each iteration, the algorithm maintains a set A of "alive" vertices initially equivalent to V . It conducts an elimination tournament among the vertices in A , eliminating a player each time it loses α matches (line 12) until only 2α vertices remain viable (line 6). This stopping condition ensures the algorithm's convergence. Matches are chosen arbitrarily to prevent repeated play of the same match (line 7). Upon completion of the elimination tournament, the E procedure is employed to exhaustively identify a candidate champion. This candidate, denoted as c , is the vertex in A with the maximum out-degree in T . If c loses at least α matches indicating a potential error in the elimination process, the algorithm proceeds to the next α value.

PARALLEL (BATCHED) VERSION:

In contemporary architectures, such as GPUs, the simultaneous execution of multiple arc lookup operations in parallel is achievable. An important question arises: can we fully leverage this parallelism to reduce the complexity of Algorithm 1? In this subsection, we introduce Algorithm 2, operating under the assumption that we can concurrently process a batch of B arcs. Specifically, Algorithm 2 performs $O(1)$ Algorithm 2 represents a minor modification of Algorithm 1. Similar to the previous algorithm, it conducts an exponential search for ℓ by doubling the parameter α . For each α , it assumes the champion resides within the set of alive vertices, A , and conducts an elimination tournament among the vertices of A , eliminating players who lose α matches. Notably, the elimination step is now performed in batches (line 12), terminating when the number of alive players becomes sufficiently small (line 7). The FIND CHAMPION BRUTEFORCE PAR method (line 18) can be parallelized effortlessly by unfolding all $O(6\alpha n)$ arcs in batches of B arcs at a time. However, our focus is on the elimination step.

The key distinction from Algorithm 1 lies in the BUILD BATCH procedure, which determines the B arcs to be looked up in parallel. It creates local copies, A_{loc} and $lost_{loc}$, of the set A and the vector $lost$. The procedure selects matches in A_{loc}

A_{loc} that hasn't been played yet and assigns a loss to both opponents. If the batched games were played sequentially, $lost_{loc}$ would provide an upper estimate of $lost$, and A_{loc} would be a subset of A . Consequently, every insertion in a batch is guaranteed to produce a match loss for a player that would still be alive if the batch were unfolded sequentially. This guarantee ensures that $lost[u] \leq \alpha$ for each $u \in V^2$. While BUILD BATCH might not always produce a batch of size B , enforcing a condition where A has at least $2B + 2\alpha$ elements ensures this (line 8). Importantly, halving the batch size when this condition is not met does not compromise the complexity of Algorithm 2. The elimination step intuitively involves two epochs: the first

unfolds arcs in B -sized batches until $|A| \geq 2B + 2\alpha$, and the second processes smaller batches until $|A|$ is sufficiently small (line 7).

Algorithm 2

```

1: procedure FINDCHAMPIONPARALLEL( $T = (V, E)$ ,  $B$ )
2:   for ( $\alpha = 1$ ; true;  $\alpha = 2\alpha$ ) do
3:      $A = V$ 
4:      $S = \{(u, u) \mid u \in V\}$ 
5:      $\forall u \in V \text{ } lost[u] = 0$ 
6:      $B' = B$ 
7:     while  $|A| > 6\alpha$  do
8:       while  $|A| < 2B' + 2\alpha$  do
9:          $B' = B'/2$ 
10:      end while
11:       $batch = \text{BUILD BATCH}(A, S, B', lost, \alpha)$ 
12:      UNFOLDINPARALLEL( $batch$ )
13:      for ( $u, v$ ) in  $batch$  do
14:         $loser = \text{if } (u, v) \in E \text{ then } v \text{ else } u$ 
15:        INCREASELOSS( $A, lost, \alpha, loser$ )
16:      end for
17:    end while
18:     $c, lost_c = \text{FINDCHAMPION BRUTEFORCE PAR}(A, E, B)$ 
19:    if  $lost_c < \alpha$  then return } c
20:  end if
21: end for
22: end procedure
23:
24: procedure BUILD BATCH( $A, S, B', lost, \alpha$ )
25:    $batch = \emptyset$ 
26:    $A_{loc} = A$ 
27:    $lost_{loc} = lost$ 
28:   while  $|batch| < B'$  do
29:     Choose  $(u, v) \in A_{loc}^2 \setminus S$ 
30:      $S = S \cup \{(u, v), (v, u)\}$ 
31:      $batch = batch \cup \{(u, v)\}$ 
32:     INCREASELOSS( $A_{loc}, lost_{loc}, \alpha, u$ )
33:     INCREASELOSS( $A_{loc}, lost_{loc}, \alpha, v$ )
34:   end while
35:   return } batch
36: end procedure
37:
38: procedure INCREASELOSS( $A, lost, \alpha, v$ )
39:    $++ lost[v]$ 
40:   if  $lost[v] \geq \alpha$  then
41:      $A = A \setminus \{v\}$ 
42:   end if
43: end procedure

```

In this section, we conduct a thorough experimental evaluation of the proposed algorithms within the context of a Question Answering task. Specifically, our focus is on passage ranking, where the objective is to choose the most relevant textual passage from a given set in response to a question. For this purpose, we utilize an existing state-of-the-art pairwise model, which functions by comparing results in pairs and determining winners in a round-robin tournament fashion. The proposed algorithms are designed to identify the tournament champions, thereby reducing the number of pairwise comparisons (arc

lookups) executed using the machine learning (ML) model.

In the subsequent sections, we outline the experimental setup, followed by an evaluation of the proposed algorithms, measured in terms of the number of pairwise comparisons required by the ML model.

TABLE 1
 Average number of inferences of different implementations of Algorithm 1 when applied to duoBERT to retrieve the top-1 result on the MS MARCO dataset. Columns identify whether the implementation exploits the input order, while rows identify whether it exploits the past lookups to avoid multiple unfolds of a same arc.

	Ignore input order	Exploit input order
Ignore past lookups	126.09	125.81
Exploit past lookups	76.58	64.62

TABLE 2
 Efficiency-Effectiveness performance achieved by monoBERT, duoBERT, and duoBERT & Alg. 1 when retrieving the top-1 result on the MS MARCO dataset.

Method	Recall@1	Inferences	Time (s)
BM25 + monoBERT	0.251	1000	65.91
+ duoBERT _{BINARY}	0.269	870	57.34
+ duoBERT _{BINARY} & Alg. 1	0.269	65	4.26

In the given section, two crucial implementation aspects of the algorithm are discussed. The first aspect involves the consideration of the input order, suggesting that it might be advantageous to commence comparisons among more relevant passages from the second stage. The second aspect addresses the prevention of multiple unfolding of the same arc by storing the arc lookups performed during the tournament, thereby saving time at the expense of a small additional space requirement.

An assessment of the impact of these implementation aspects is provided, resulting in four distinct implementations. Table 1 displays the average number of inferences for different implementations of Algorithm 1 when applied to duoBERT for retrieving the top-1 result on the MS MARCO dataset. As anticipated, both aspects contribute to a reduction in the average number of inferences. Notably, the implementation leveraging input order is more efficient when combined with the hash table, and their combination nearly halves the number of inferences compared to the implementation ignoring both aspects.

Table 2 further presents the performance of the best implementation, combining input order and past lookups, within the ranking pipeline proposed by Nogueira et al. Metrics such as Recall@1, the number of inferences, and inference time for all ranking stages are reported. The results indicate a significant improvement in the ranking

process, with a considerable reduction in the time cost of the third stage compared to previous configurations.

The average number of inferences required by this approach is approximately 65, closely approaching the minimum necessary when the champion wins all comparisons ($29 \times 2 = 58$ inferences). A noteworthy observation is that 95% of queries are resolved with only 50 comparisons or fewer, equivalent to less than 100 model inferences. Additionally, it is emphasized that if the algorithm were applied to a symmetric model, where two inferences per comparison are not required, the algorithm would perform just a few inferences per item.

Efficiency of parallel (batched) implementations of monoBERT, duoBERT, and duoBERT & Alg. 2 when retrieving the top-1 result on the MS MARCO dataset.

Method	Metric	Batch Size							
		2	4	8	16	32	64	128	256
BM25 + monoBERT	Inferences	300	290	123	63	32	16	8	4
	Time (s)	52.95	16.48	8.24	4.15	2.11	1.05	0.53	0.26
+ duoBERT _{BINARY}	Inferences	435	218	109	55	28	14	7	4
	Time (s)	28.67	14.37	7.18	3.62	1.83	0.92	0.46	0.26
+ duoBERT _{BINARY} & Alg. 2	Inferences	33	33	14	8	5	4	4	4
	Time (s)	2.14	1.54	0.93	0.55	0.31	0.28	0.26	0.25
	Speedup	(15.4x)	(9.3x)	(7.7x)	(6.6x)	(5.8x)	(5.3x)	(4.7x)	(4.8x)

PARALLEL (BATCHED) VERSION:

Table 3 reports the performance of Algorithm 2 in the parallel setting where the algorithm can unfold a batch of multiple arcs in parallel. The table reports the number of inferences and the inference time of all ranking stages, for values of batch size between 2 and 256 when retrieving the top-1 result on the MS MARCO dataset. The Recall@1 metric is not reported as the correctness of the algorithm guarantees that the effectiveness does not change with the batch size. Indeed, Recall@1 is always close to 27% as in the non-parallel setting. The first row shows the performance of the first two stages of the ranking pipeline, i.e., BM25 + monoBERT, while the second row shows the performance of the third stage, i.e., duoBERT_{BINARY}. The number of batch inferences linearly decreases when increasing the batch size for both configurations, as we can unfold more arcs in parallel per batch. For instance, with a batch size of 64, we can perform 64 inferences at a time and the full round-robin tournament requires only $\lceil 870/64 \rceil = 14$ rounds to perform all inferences. The third row shows the performance of duoBERT_{BINARY} used as third stage when employing Algorithm 2 to perform the (batched) tournament among the top-30 results of each query. Our algorithm speeds up the ranking from $13 \times$ to $3 \times$ for batch size ranging from 2 to 64. As expected, the speedup decreases when increasing the batch size as the number of results involved in the tournament is very limited. Indeed, the algorithm can accurately unfold only one arc for each alive vertex (Algorithm 2, set A); it then fills the batch with a simple heuristic that explores all arcs of just a few promising vertices (as described in the "Implementation Details" subsection of Section 5.3). Therefore, as the batch size becomes bigger than the number of results, i.e., 30 in our setting, the choices of the algorithm become less oriented. Nevertheless, Algorithm 2 speeds up the ranking of duoBERT_{BINARY} for all the values of batch size tested.

inference time for all ranking stages are reported. The results indicate a significant improvement in the ranking process, with a considerable reduction in the time cost of the third stage compared to previous configurations.

The average number of inferences required by this approach is approximately 65, closely approaching the minimum necessary when the champion wins all comparisons ($29 \times 2 = 58$ inferences). A noteworthy observation is that 95% of queries are resolved with only 50 comparisons or fewer, equivalent to less than 100 model inferences. Additionally, it is emphasized that if the algorithm were applied to a symmetric model, where two inferences per comparison are not required, the algorithm would perform just a few inferences per item.

Efficiency of parallel (batched) implementations of monoBERT, duoBERT, and duoBERT + Alg 2 when retrieving the top-1 result on the MS MARCO dataset.

Method	Metric	Batch Size							
		2	4	8	16	32	64	128	256
BM25 + monoBERT	Inferences	300	290	125	65	32	16	8	4
	Time (s)	52.95	16.48	8.24	4.15	2.11	1.05	0.53	0.26
+ duoBERT _{naive}	Inferences	435	218	109	55	28	14	7	4
	Time (s)	28.67	14.37	7.18	3.62	1.83	0.92	0.46	0.26
+ duoBERT _{naive} & Alg 2	Inferences	33	33	14	8	5	4	4	4
	Time (s)	2.14	1.54	0.93	0.55	0.31	0.28	0.26	0.25
	Speedup	(13.4x)	(18.7x)	(27.6x)	(36.6x)	(59.1x)	(13.7x)	(17.7x)	(10.6x)

PARALLEL (BATCHED) VERSION:

Table3 reports the performance of Algorithm 2 in the parallel setting where the algorithm can unfold a batch of multiple arcs in parallel. The table reports the number of inferences and the inference time of all ranking stages, for values of batch size between 2 and 256 when retrieving the top-1 result on the MS MARCO dataset. The Recall@1 metric is not reported as the correctness of the algorithm guarantees that the effectiveness does not change with the batch size. Indeed, Recall@1 is always close to 27% as in the non-parallel setting. The first row shows the performance of the first two stages of the ranking pipeline, i.e., BM25 + monoBERT, while the second row shows the performance of the third stage, i.e., duoBERTBINARY. The number of batch inferences linearly decreases when increasing the batch size for both configurations, as we can unfold more arcs in parallel per batch. For instance, with a batch size of 64, we can perform 64 inferences at a time and the full round-robin tournament requires only $\lceil 870/64 \rceil = 14$ rounds to perform all inferences. The third row shows the performance of duoBERTBINARY used as third stage when employing Algorithm 2 to perform the (batched) tournament among the top-30 results of each query. Our algorithm speeds up the ranking from 13x to 3x for batch size ranging from 2 to 64. As expected, the speedup decreases when increasing the batch size as the number of results involved in the tournament is very limited. Indeed, the algorithm can accurately unfold only one arc for each alive vertex (Algorithm 2, set A); it then fills the batch with a simple heuristic that explores all arcs of just a few promising vertices (as described in the “Implementation Details” subsection of Section 5.3). Therefore, as the batch size becomes bigger than the number of results, i.e., 30 in our setting, the choices of the algorithm become less oriented. Nevertheless, Algorithm 2 speeds up the ranking of duoBERTBINARY for all the values of batch size tested.

USING THE TEMPLATE

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

IV CONCLUSION & FUTURE WORK :

In this study, we tackled the challenge of efficiently retrieving the top-1 result when employing pairwise machine learning classifiers. This problem was mapped to finding champions in tournament graphs with the goal of minimizing the number of arc lookups, representing the comparisons done through the classifier. Key contributions and findings include:

A lower bound of $\Omega(\ell n)$ arc lookups for finding a champion when the number ℓ of matches lost by the champion is known, with a generalization to randomized algorithms.

Presentation of an asymptotically optimal deterministic algorithm that matches the established lower bound without requiring knowledge of ℓ .

Addressing three natural variants of the original problem:

Solving the problem of finding all top-k players simultaneously.

Extending the approach to a probabilistic tournament with probabilities in the adjacency matrix.

Achieving linear speedup by probing B adjacency matrix cells in parallel.

Experimental evaluation of the proposed algorithms in the context of ranking, demonstrating up to a 13x speedup for the retrieval of the top-1 result in the classic binary setting. Variants of the original problem were also evaluated, showcasing speed-ups ranging from 13x to 2x for different values of k in the binary setting, and from 6x to 2x in the probabilistic setting. In the parallel setting, consistent speed-ups were observed for various batch sizes.

As future research directions, three main areas are identified:

Theoretical exploration to characterize the leading constant in the complexity of finding the Copeland winner, enhancing the comparison between lower bounds and proposed algorithms.

Applied research on heuristics to further improve the speed-up of the algorithms while maintaining theoretical performance.

Investigation of the dependency between the number of arc lookups and the probability distribution of graph arcs, aiming to establish a link between complexity and the data characteristics.

V REFERENCES:

- [1] Qingyao Ai, Xuanhui Wang, Nadav Golbandi, Mike Bendersky, and Marc Najork. Learning groupwise scoring functions using deep neural networks. 2019
- [2] Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. In 21st Annual Conference on Learning Theory (COLT 2008), pages 87–98. Omnipress, 2008.
- [3] Nir Ailon and Mehryar Mohri. Preference-based learning to rank. *Machine Learning*, 80(2-3):189–211, 2010.
- [4] Miklos Ajtai, Vitaly Feldman, Avinatan Hassidim, and Jelani Nelson. Sorting and selection with imprecise comparisons. *ACM Trans. Algorithms*, 12(2):19:1–19:19, 2016.
- [5] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- [6] Jeffrey S Banks. Sophisticated voting outcomes and agenda control. *Social Choice and Welfare*, 1(4):295–306, 1985. The template will number citations consecutively within brackets
- [7] Lorenzo Beretta, Franco Maria Nardini, Roberto Trani, and Rossano Venturini. An optimal algorithm to find champions of tournament graphs. In *String Processing and Information Retrieval - 26th International Symposium, (SPIRE 2019)*, volume 11811 of *Lecture Notes in Computer Science*, pages 267–273. Springer, 2019.
- [8] Arindam Biswas, Varunkumar Jayapaul, Venkatesh Raman, and Srinivasa Rao Satti. Finding kings in tournaments. *Discrete Applied Mathematics*, 322:240–252, 2022.
- [9] Felix Brandt, Markus Brill, and Paul Harrenstein. Tournament solutions. In *the Handbook of Computational Social Choice*, pages 57–84. Cambridge University Press, 2016.
- [10] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In Shang-Hua Teng, editor, *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, California, USA, January 20-22, 2008, pages 268–276. SIAM, 2008

Survey on Data Pre Processing Methods for Improved ML Model Performance

Neelima Kolli
23DSC26, M.Sc. (Computational Data Science), Dept. of Computer Science
P.B. Siddhartha College of Arts & Science
Vijayawada, Andhra Pradesh, India
iamkollineelima@gmail.com

Sri Latha Kolli
Masters in Artificial Intelligence
Dept. of Electrical Engineering & Computer Science
Florida Atlantic University
Boca Raton, Florida, United States
iamsrikolli@gmail.com

Mr.P.R. Krishna Prasad
Associate Professor
Dept of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur
prkrishnaprasad@vvit.net

Abstract — As machine learning (ML) continues to permeate various domains, the importance of data quality and preprocessing techniques in enhancing model performance has become increasingly evident. This paper presents a comprehensive survey of state-of-the-art data pre-processing methods aimed at optimizing ML model outcomes. The study explores various techniques such as dimensionality reduction, normalization, discretization and J48 classifier among others. Through an in-depth analysis of each method's strengths, limitations, and applicability, this survey provides a nuanced understanding of their impact on model performance. The insights gathered from this survey aim to serve as a valuable resource for researchers, practitioners, and enthusiasts seeking to navigate the intricate landscape of data pre-processing for the purpose of achieving superior ML model performance.

Keywords— *Dimensionality Reduction, Normalization, Discretization, J48 Classifier.*

I INTRODUCTION

In the rapidly evolving field of machine learning (ML), the pivotal role of data pre-processing in enhancing model performance cannot be overstated. As ML applications span diverse domains, from healthcare to finance and beyond, the quality of input data significantly influences the efficacy of predictive models. This paper addresses the critical need for a comprehensive understanding of data pre-processing methods and their impact on ML model performance.

The objective of this survey is to delve into the myriad techniques employed for data pre-processing, with a particular focus on optimizing ML outcomes. The survey covers a spectrum of methods, including but not limited to dimensionality reduction, normalization, discretization and J48 classifier among others by meticulously examining the strengths, limitations, and applicability of each technique, this study aims to provide a nuanced perspective on their role in refining model performance.

By synthesizing this wealth of information, our intention is to equip researchers, practitioners, and enthusiasts with a valuable resource to navigate the intricate terrain of data pre-processing, fostering improved ML model performance

DIMENSIONALITY REDUCTION

The primary goal of dimensionality reduction is to express a high-dimensional (HD) dataset in a low-dimensional (LD) s

space while retaining the inherent structures, such as outliers and clusters, *present in the HD* data [1]. Dimension reduction holds a distinct advantage in visualizing high-dimensional datasets, offering scalability; however, this method is not without its drawbacks. One notable limitation lies in the inevitable loss of information during the transformation of data into the LD projection. The subsequent section explores two widely used dimension reduction algorithms, namely PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis).

L. Principal Component Analysis

Principal Components Analysis (PCA) is a feature extraction method that transforms the original features of a dataset into new linear combinations [2]. In this process, each example in a given dataset, initially existing in a d -dimensional space, is mapped to a k -dimensional subspace, where k is less than d . The newly generated set of k dimensions is referred to as the Principal Components (PC). These Principal Components capture the maximum variance in the data, with each PC being directed towards the highest remaining variance, excluding the variance already accounted for by its preceding components.

This method effectively reduces the dimensionality of the data while retaining the most significant information, facilitating improved computational efficiency and interpretability in subsequent analyses.

M. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) serves as a valuable dimensionality reduction method, particularly when the objective is to mitigate computation time complexity and enhance classification performance [3]. Primarily employed in image processing, signal processing, as well as in addressing challenges such as bankruptcy prediction and market analysis, LDA offers distinct advantages in discrimination tasks.

While Principal Component Analysis (PCA) excels in extracting efficient features from datasets, it may not be optimally suited for discrimination-focused objectives. In

contrast, LDA excels in transforming a high-dimensional feature space into a lower-dimensional counterpart by selecting an optimal projection matrix. This transformation preserves essential information crucial for accurate data classification.

The LDA process involves defining two essential scatter matrices. The first matrix, denoted as SB, represents the inter-class scatter matrix, capturing variations between different classes. The second matrix, SW, represents the intra-class scatter matrix, encapsulating variations within the same class.

Through these defined matrices, LDA orchestrates an effective transformation, striking a balance between dimensionality reduction and the preservation of critical information essential for robust data classification.

NORMALIZATION

Attribute normalization is a crucial process involving the scaling of values to fit within a predefined range. This procedure holds particular significance in classification frameworks, especially those utilizing distance measurements or neural networks [4]. In the context of classification employing the neural network backpropagation algorithm, normalizing input values for each measured attribute in the training set accelerates the learning phase. This acceleration contributes to an enhanced efficiency of the classification process.

In the realm of distance-based techniques, normalization plays a pivotal role in ensuring that attributes with initially limited ranges do not get overshadowed by those with larger ranges. Several contemporary data normalization techniques are employed for this purpose, including but not limited to decimal scaling, Z-score normalization, and min-max normalization. These techniques collectively contribute to refining the input data, thereby optimizing the performance and reliability of classification model.

N. Min-Max

Min-max normalization aims to linearly transform original data [5]. Preserving the relationship between original values, min-max normalization may lead to "out-of-bounds" errors if future normalization input cases fall outside the original data range for A.

O. Z-Score

Z-score normalization normalizes attribute values based on the mean (μ_A) and standard deviation (σ_A) of the attribute [5]. Z-score normalization is particularly useful when outliers dominate the min-max normalization process or when the actual minimum and maximum values of attribute A are unknown.

P. Decimal Scaling

Decimal scaling involves moving the decimal point of attribute values based on the absolute maximum value of the attribute [5].

DISCRETIZATION

Discretization involves the transformation of continuous variables into discrete counterparts by dividing the range of values into finite intervals, commonly referred to as intervals, buckets, or bins [6]. The discretization algorithm partitions continuous variables into distinct intervals, treating them as categorical entities. To minimize biases, a logarithmic transformation to the base 2 is applied to continuous-valued features before converting the result into an integer [3].

J48 CLASSIFIER

The J48 classifier, a simplified version of the C4.5 decision tree, is primarily employed for classification tasks [7]. While generating a binary tree, the decision tree method is widely utilized in addressing classification problems. This classifier constructs a tree to represent the classification process, applying the resulting tree to each tuple in the database for classification. Notably, the J48 classifier does not account for missing values during tree construction, allowing for the prediction of the missing item's value based on information gained from other attributes. The primary objective is to partition the data into ranges based on attribute values from items in the training set. The J48 classifier offers the flexibility of classifying attributes through either decision trees or the rules derived from them.

II HANDLING CATEGORICAL DATA

Dealing with categorical data is a crucial aspect of data preprocessing, particularly when working with diverse datasets that include non-numeric features. The representation of categorical variables plays a pivotal role in the performance of machine learning models. This section explores various techniques for handling categorical data, including one-hot encoding, label encoding, and the use of embeddings.

Q. One-Hot Encoding

One-hot encoding is a widely employed technique for transforming categorical variables into a binary matrix. Each category is represented by a binary value (0 or 1) in a separate column, effectively creating a binary vector for each category. While one-hot encoding ensures that categorical variables do not impose an ordinal relationship, it introduces sparsity in the dataset, especially when dealing with a large number of unique categories. The resulting binary matrix serves as input for machine learning models, enabling them to interpret and utilize categorical information effectively.

R. Label Encoding

Label encoding is an alternative technique for encoding categorical variables, assigning a unique numerical label to each category. Unlike one-hot encoding, label encoding introduces ordinal relationships between the encoded values. While this may be suitable for certain algorithms, such as decision trees, it might mislead models that interpret numerical values as having inherent order. Label encoding is particularly beneficial when dealing with ordinal categorical data, where the order of categories holds significance.

S. *Embeddings for Categorical Variables*

In recent years, the use of embeddings has gained prominence, especially in the context of categorical variables with high cardinality. Embeddings involve representing categorical variables as dense vectors in a lower-dimensional space, capturing meaningful relationships between categories. This technique is particularly advantageous when dealing with large and complex categorical features, such as user IDs or product IDs in recommendation systems. Embeddings leverage neural network architectures to learn and represent intricate patterns within categorical variables, providing a more nuanced and efficient representation for machine learning models.

The choice between these techniques depends on the nature of the data, the machine learning algorithm's requirements, and the specific characteristics of the categorical variables under consideration. Each method has its advantages and considerations, and understanding their implications is crucial for making informed decisions during the preprocessing phase.

HANDLING IMBALANCED DATASETS

In real-world scenarios, imbalanced datasets, where certain classes have significantly fewer instances than others, present a common challenge. Addressing this issue is crucial to prevent models from being biased toward the majority class and to improve the overall accuracy and fairness of the predictions. Several techniques are employed for handling imbalanced datasets, including oversampling, under sampling, and synthetic data generation methods.

T. *Oversampling*

Oversampling involves increasing the number of instances in the minority class to balance the class distribution. This technique ensures that the model is exposed to a more equitable representation of both classes during training. Methods such as random oversampling, where instances of the minority class are duplicated, or SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic instances along the line segments connecting existing minority class instances, are widely utilized. While oversampling can be effective in mitigating class imbalance, it is essential to be cautious about potential overfitting and increased computational demands.

U. *Undersampling*

Under sampling aims to reduce the number of instances in the majority class to match that of the minority class, achieving a more balanced distribution. This technique helps prevent the model from being biased toward the majority class, enhancing its sensitivity to the minority class. Common under sampling methods include random under sampling, where instances from the majority class are randomly removed, and cluster-based under sampling, which involves clustering the majority class instances and keeping only representative samples from each cluster. Under sampling can be computationally efficient but may risk losing valuable information present in the majority class.

V. *Synthetic Data Generation (Smote)*

Synthetic data generation methods, such as SMOTE (Synthetic Minority Over-sampling Technique), address class imbalance by creating synthetic instances of the minority class. SMOTE works by interpolating between existing minority class instances, effectively generating new instances in regions of the feature space where minority class examples are sparse. This method helps the model better generalize to the minority class and reduces the risk of overfitting associated with simple oversampling. SMOTE has proven effective in various machine learning applications and is widely adopted for enhancing the predictive performance of models trained on imbalanced datasets.

Incorporating these techniques for handling imbalanced datasets is crucial for ensuring that machine learning models provide fair and accurate predictions across all classes, contributing to the robustness and reliability of the overall model.

III TEXT DATA PREPROCESSING

For datasets containing textual information, a specialized set of preprocessing techniques is essential to transform unstructured text into a format suitable for machine learning models. The following methods play a crucial role in extracting meaningful features from text data:

W. *Tokenization*

Tokenization involves breaking down the text into smaller units, typically words or phrases referred to as tokens. This process facilitates the analysis of the frequency and distribution of words in a document. Advanced tokenization techniques, such as n-gram tokenization, capture sequences of adjacent words, providing additional context for the model.

X. *Stemming*

Stemming is a text normalization technique that aims to reduce words to their root or base form. This process involves removing suffixes to achieve a common linguistic root. While stemming may result in the generation of non-real words, it helps reduce dimensionality and captures the essence of related words.

Y. *Lemmatization*

Lemmatization is another form of text normalization that goes beyond stemming. It involves reducing words to their base or dictionary form (lemma) by considering the context and meaning of the word. Lemmatization results in linguistically valid words, contributing to a more refined representation of the text.

Z. *Removing Stop Words*

Stop words are common words (e.g., "the," "and," "is") that carry little semantic meaning and may introduce noise into the analysis. Removing stop words during preprocessing helps focus on content-carrying words, enhancing the relevance of features and improving the efficiency of subsequent analyses.

AA. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a numerical statistic that quantifies the importance of a term in a document relative to a collection of documents. It combines term frequency (TF), representing the frequency of a term in a document, and inverse document frequency (IDF), which measures how rare or common a term is across documents. TF-IDF assigns higher weights to terms that are frequent in a document but rare in the entire collection, emphasizing their significance.

BB. Word Embeddings

Word embeddings represent words as vectors in a continuous vector space, capturing semantic relationships between words. Techniques such as Word2Vec and GloVe generate embeddings by considering the context in which words appear. These embeddings encode semantic similarities and relationships, enabling machine learning models to grasp the contextual nuances of the text.

Incorporating these text preprocessing techniques is crucial for extracting meaningful information from textual data, enabling machine learning models to derive valuable insights from unstructured information. The careful application of these methods contributes to the success of natural language processing tasks and ensures that textual data is effectively leveraged in the machine learning pipeline.

HANDLING MISSING DATA

Addressing missing data is a critical preprocessing step to ensure the robustness and accuracy of machine learning models. Missing data can introduce biases and hinder the performance of algorithms. Various techniques are employed to handle missing values effectively.

CC. Mean Or Median Imputation

Mean or median imputation involves replacing missing values with the mean or median of the observed values for that particular feature. This method is simple, computationally efficient, and suitable for datasets with small to moderate amounts of missing data. However, it may lead to biased results if the missing data is not missing completely at random, as it does not consider the underlying relationships between variables.

DD. Regression Imputation

Regression imputation utilizes the relationships between variables to predict missing values based on the observed data. A regression model is trained using the observed values, and the model is then used to predict the missing values. This method is particularly useful when there is a pattern or correlation between the missing values and other variables. While more sophisticated than mean or median imputation, regression imputation assumes a linear relationship between variables, and its accuracy relies on the validity of this assumption.

EE. Multiple Imputation

Multiple Imputation is an advanced technique that involves generating multiple datasets with imputed values, each reflecting the uncertainty associated with missing data. The imputed datasets are then analyzed separately, and the results are combined to provide more accurate estimates and standard errors. Multiple Imputation accounts for the variability introduced by imputing missing values and yields more robust and reliable results compared to single imputation methods.

This technique is particularly beneficial when the missing data mechanism is non-random or when imputed values have uncertainty. Multiple Imputation can be implemented using various statistical methods, such as the Markov Chain Monte Carlo (MCMC) method.

The choice of imputation method depends on the nature of the data and the assumptions that can be reasonably made about the missing data mechanism. Careful consideration and validation are necessary to ensure that the chosen method aligns with the characteristics of the dataset, promoting accurate and unbiased results in subsequent machine learning analyses.

TIME SERIES DATA PREPROCESSING

Time series data, characterized by a sequential order of observations indexed by time, often presents unique challenges and opportunities in the realm of data preprocessing. Addressing these challenges is crucial for extracting meaningful patterns and building accurate predictive models. This section delves into specific techniques tailored for time series data preprocessing.

FF.. Handling Irregular Time Intervals

Time series datasets may exhibit irregular time intervals, requiring careful handling to ensure accurate analysis and modeling. Resampling techniques, such as interpolation or aggregation, can be employed to regularize time intervals. This involves aligning observations to a consistent time grid, providing a uniform structure for analysis while preserving the temporal order of the data.

GG. Smoothing Techniques

Smoothing techniques are vital for reducing noise and highlighting underlying patterns in time series data. Methods like moving averages or exponential smoothing can help in achieving a clearer representation of trends by attenuating short-term fluctuations. These techniques contribute to a more stable foundation for subsequent analysis and modeling.

HH. Feature Engineering Based On Temporal Patterns

Feature engineering plays a pivotal role in extracting relevant information from time series data. Creating new features based on temporal patterns can enhance the model's ability to capture underlying structures. For instance, introducing features such as day of the week, month, or quarter, can help the model discern periodic trends and seasonality within the data.

II. Lag Features

Lag features involve introducing time-shifted versions of a variable as new features. By incorporating past observations, the model can capture temporal dependencies and memory effects within the data. Lag features are particularly useful for modeling phenomena where the current state is influenced by previous states, such as stock prices or weather conditions.

JJ. Rolling Statistics To Capture Time-Dependent Relationships

Applying rolling statistics involves calculating metrics, such as moving averages or standard deviations, over a window of consecutive observations. This technique aids in capturing time-dependent relationships and identifying trends or patterns that may not be apparent in individual data points. Rolling statistics contribute to a more comprehensive understanding of the temporal dynamics inherent in the dataset.

Incorporating these time series data preprocessing techniques enhances the robustness and predictive power of machine learning models when applied to sequential data. The careful consideration of irregular time intervals, smoothing, feature engineering, lag features, and rolling statistics collectively contributes to a more refined analysis and interpretation of time series datasets.

BATCH NORMALIZATION

Batch normalization is a technique commonly applied in neural networks to improve training stability and convergence. It involves normalizing the input of each layer by subtracting the batch mean and dividing by the batch standard deviation. This can accelerate the training process and contribute to better model generalization.

IV CONCLUSION

In conclusion, this survey comprehensively explores the landscape of data preprocessing methods for improving machine learning model performance. The varied techniques discussed, ranging from dimensionality reduction to normalization, feature engineering, and beyond, underscore the multifaceted nature of data preprocessing in the realm of machine learning.

The nuanced analysis of each method's strengths, limitations, and applicability serves as a valuable resource for researchers, practitioners, and enthusiasts navigating the intricate terrain of data preprocessing. By synthesizing insights from diverse techniques, this survey aims to empower the community with knowledge that fosters not only a deeper understanding but also the practical implementation of these techniques for achieving superior machine learning outcomes.

As the field continues to evolve, it is our hope that this survey contributes to ongoing discussions, sparks new avenues of research, and inspires further innovations in the dynamic intersection of machine learning and data preprocessing.

VI REFERENCES

- [1] I. C. N. R. John Wenskovitch, "Towards a Systematic Combination of Dimension Reduction," *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, vol. 24, 2018.
- [2] B. S. K. Keerthi Vasan, "Dimensionality reduction using Principal Component Analysis for network intrusion detection," *Elsevier*, vol. 8, p. 510-512, 2016.
- [3] G. D. S. H. Hamed Haddad Pajouh, "Two-tier network anomaly detection model: a machine learning approach," *Journal of Intelligent Information Systems*, Springer, vol. 48, p. 61-74, 2015. 1989.
- [4] Z. S. a. B. K. Luai Al Shalabi, "Data Mining: A Preprocessing Engine," *Journal of Computer Science*, pp. 735-739, 2006.
- [5] M. K. a. J. P. Jiawei Han, *Data Mining Concepts and Techniques*, Elsevier, 2012.
- [6] F. Y. Zeynel Cebeci, "Comparison of Chi-square based algorithms for discretization of continuous chicken egg quality traits," *Journal of Agricultural Informatics*, vol. 8, 2017.
- [7] M. S. S. S. Tina R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *International Journal of Computer Science and Applications*, vol. 6, 2013.

YOLOv8-Based Helmet Detection with Integrated E-mail Notifications

Muddamsetty Sriya
23DSC28, M.Sc.(Computational Data
Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts &
Science
Vijayawada, A.P, India
muddamsetty.sriya@gmail.com

Shaik Obaid
23DSC17, M.Sc.(Computational Data
Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts &
Science
Vijayawada, A.P, India
Obaidsk7865@gmail.com

Mr.Ch.Hari Prasad
Associate Professor
Dept of CSE,
Vasireddy Venkatadri Institute of
Technology,Nambur
hari.chandika@vvit.net

Abstract- Two-wheeler riders are at a higher risk of head injuries in the event of accidents, and helmets play a critical role in mitigating the severity of such injuries. The project focuses on improving road safety by developing a deep learning-based system for detecting individuals riding two-wheelers without helmets and identifying their number plates. The system utilizes the YOLOv8 algorithm for processing video feeds and detecting the presence of helmets and number plates in the frames. The system can accurately identify individuals riding without helmets and send them email notifications, encouraging them to wear helmets while riding. The project has significant implications for improving road safety and reducing the number of head injuries caused by two-wheeler accidents. The use of YOLOv8 algorithm helped in achieving high accuracy and fast processing speed, making the system efficient and effective for detecting helmets and number plates in the video frames.

Keywords-Deep Learning, YOLOv8, Object Detection, Image Processing, Computer Vision, Flask

I INTRODUCTION

The use of helmets while riding a two-wheeler is crucial for ensuring road safety. However, many riders fail to wear helmets, leading to a high number of accidents and fatalities. Road safety is a critical issue worldwide, with millions of accidents and fatalities occurring every year. One of the primary causes of these accidents is the failure of individuals to wear helmets while riding two-wheelers. The use of helmets while riding a two-wheeler is crucial for ensuring road safety. Helmets protect the head and reduce the risk of fatal injuries in case of accidents [6]. Studies show that wearing helmets can reduce the risk of head injury by up to 70% and the risk of death by up to 40%. Despite these statistics, many riders fail to wear helmets, leading to a high number of accidents and fatalities. People often neglect helmets due to various reasons such as discomfort, inconvenience, and the perception that helmets are unnecessary for short distances.

In addition, riders often forget or simply choose not to wear helmets due to lack of awareness of the importance of helmets in preventing accidents and fatalities. *NEGLECTING HELMETS CAN*

LEAD TO SERIOUS CONSEQUENCES, NOT ONLY FOR THE RIDERS BUT ALSO FOR THEIR FAMILIES AND SOCIETY.

In this context, developing an automated system for detecting individuals riding two-wheelers without helmets can contribute significantly to improving road safety. This project aims to develop such a system using deep learning techniques [1]. The system uses a video feed taken from traffic security cameras as input and processes the frames using the YOLOv8 algorithm to detect the presence of helmets and number plates. The system can accurately classify frames based on whether the rider had a helmet on or not and identify the number plate of the vehicle [2]. The use of deep learning techniques can make the system efficient and effective in detecting helmets and number plates in video frames. The project has the potential to contribute towards improving road safety by encouraging helmet usage among two-wheeler riders. The neglecting of helmets is a serious issue that needs to be addressed, and the project can raise awareness among riders about the importance of wearing helmets while riding two-wheelers. By promoting safe riding practices, the project can contribute towards reducing accidents and fatalities on roads. The system is integrated with an email notification feature that automatically sends emails to individuals detected without helmets, informing them about the importance of wearing a helmet while riding. This can help in improving awareness and encouraging riders to wear helmets while riding, thus contributing towards reducing accidents and fatalities.

II RELATED WORK

Currently, there are several existing systems for detecting helmets and number plates in two-wheeler riders. Some of these systems use computer vision techniques such as object detection and image segmentation to analyse images and videos and detect helmets and number plates. However, most of these existing systems are manual and require human intervention to analyze images and videos, which can be time-consuming and inefficient. Moreover, these systems may not be accurate enough to detect helmets and number plates in real-time or under different lighting and environmental conditions. There are also some existing

systems that use machine learning algorithms such as support vector machines and decision trees to classify helmets and number plates in images and videos [3]. However, these systems may require significant manual intervention in the training and testing of the machine learning model, and may not be flexible enough to adapt to different settings and environments. Therefore, there is a need for an automated system that can accurately detect helmets and number plates in real-time using deep learning techniques and can integrate with other features such as email notifications to promote safe riding practices among two-wheeler riders.

III PROPOSED WORK

Our proposed system is made by using deep learning based object detection algorithms for helmet detection and number plate recognition to detect the people riding two wheelers without a helmet. The system consists of two main components: a helmet detection module and a number plate detection module. The helmet detection module uses a deep learning-based object detection algorithm to detect helmets in real-time. The algorithm is trained on a large dataset of images containing helmets and non-helmets and can accurately identify helmets even in challenging situations. The helmet detection module runs on a video stream captured from a camera installed on a road or a traffic signal. The number plate recognition module uses another deep learning-based algorithm to recognize the number plate of the vehicle. The algorithm is trained on a large dataset of images containing different types of number plates and can accurately recognize number plates even in challenging situations, such as varying lighting conditions, different fonts, and sizes. The number plate recognition module runs on the same video stream captured from the camera installed on the road or the traffic signal. Once the helmet detection and number plate recognition modules have processed the video stream, the system checks if the two-wheeler riders are wearing helmets and if their vehicles have a valid number plate. If a rider is not wearing a helmet or if the vehicle does not have a valid number plate, the system sends an email alert to the concerned authorities with the location and time of the violation.

This project is developed by using the following soft wares, libraries, IDEs, languages etc. which are explained clearly below:

Python is a popular programming language that is easy to learn and widely used in the field of data science and machine learning. It has a simple syntax, which makes it a great language for beginners and experts alike. One of the most popular applications of Python is deep learning, which is a type of machine learning that involves training neural networks to perform complex tasks such as image recognition, natural language processing, and speech recognition. Deep learning is becoming increasingly important in a wide range of fields, including healthcare, finance, and manufacturing. Object detection is a specific

application of deep learning that involves identifying and locating objects within an image or video. It is used in many real-world applications, such as self-driving cars, surveillance systems, and medical imaging. Python provides a wide range of libraries and tools for deep learning and object detection, such as TensorFlow, Keras, PyTorch, and OpenCV. These libraries make it easy to build and train neural networks, perform image processing, and analyze data.

- **Computer Vision:** Computer vision is a field of study in artificial intelligence and computer science that focuses on enabling machines to interpret and understand visual data from the world around them. The goal of computer vision is to teach machines to recognize and process images and video in a way similar to how humans do. This includes tasks such as object recognition, face detection, image segmentation, and tracking [4]. Computer vision relies on various techniques such as image processing, machine learning, and deep learning to analyze visual data. These techniques involve the use of algorithms and mathematical models to extract features and patterns from visual data [5].
- **YOLOv8:** YOLOv8, or You Only Look Once version 8, is a state-of-the-art object detection algorithm that uses a deep neural network to detect objects within an image or video. It is considered to be one of the most accurate and efficient object detection algorithms available, making it an important tool for a wide range of applications, such as self-driving cars, surveillance systems, and robotics. One of the main advantages of YOLOv8 is its speed and efficiency. Unlike other object detection algorithms that require multiple passes through an image or video, YOLOv8 uses a single neural network to perform object detection in real-time. This makes it well-suited for applications where speed is critical, such as self-driving cars or drones. YOLOv8 is also highly accurate, with state-of-the-art performance on benchmark datasets such as COCO and PASCAL VOC. It achieves this high level of accuracy by using a combination of advanced techniques, such as anchor-based object detection, multi-scale feature extraction, and spatial pyramid pooling. Another important feature of YOLOv8 is its ability to detect multiple objects within the same image or video. It can detect objects of different classes, such as people, cars, and animals, and can even detect small or partially obscured objects. Overall, YOLOv8 is an important tool for object detection, thanks to its speed, accuracy, and versatility. Its ability to detect multiple objects of different classes within a single pass makes it well-suited for real-time applications, while its high level of accuracy makes it useful for tasks where precision is critical.
- **OpenCV:** OpenCV, or Open-Source Computer Vision Library, is an open-source computer vision and machine learning software library that provides a wide range of tools and functions for image and video processing. It is

written in C++, but also provides bindings for Python, Java, and other languages [4].

- **LabelImg:** LabelImg is a graphical image annotation tool used to label object bounding boxes in images for object detection tasks. It is a free and open-source software tool that can be used to create annotations for training machine learning models.
- **Pytesseract:** Pytesseract is a Python library used for Optical Character Recognition (OCR). OCR is a technology that enables computers to recognize printed or handwritten text characters within images or scanned documents.

SMTPLIB: The smtplib is a Python library used for sending email messages using the Simple Mail Transfer.

- **Protocol (SMTP).** It provides a simple and easy-to-use interface for sending emails from a Python application [8].

Implementation of the project takes place as explained in the flow chart below:

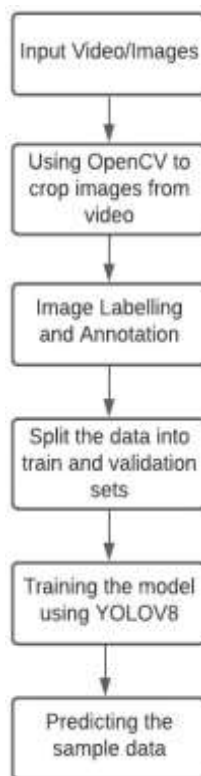


Fig. 1. Implementation of the System

- **Data Collection and Preprocessing:**

Data preprocessing is ensuring the quality of data which is crucial in obtaining an accurate model. In this project, the images were collected from various sources, including online repositories and publicly available datasets. Before feeding into algorithm, several preprocessing steps were

included to obtain the images with uniform size, orientation and quality. All the images were resized to a uniform size of 640 x 640 px. This was done to ensure that all the images are of same size, which is a requirement for YOLO algorithm. The images were cropped to focus on the regions of interest, i.e., the helmet and number plate regions. This step is essential to remove any unnecessary background noise and improve accuracy of algorithm. Data augmentation techniques like blur, noise and brightness are used to increase the size of dataset and also to prevent over fitting. By doing this, the dataset size has increased by 2 times. Finally, the processed data is splitted into training and validation sets 70:30 respectively. The training set is used to train the algorithm and validation set was used to tune the hyper parameters and prevent over fitting. Further sample data is used to do the predictions. This ensured that the algorithm was trained on a diverse set of data and was able to generalize well to new, unseen data.

- **YOLOv8 Algorithm implementation:**

Before training the model, various hyper parameters were tuned to achieve the best results. These hyper parameters include the learning rate, batch size, and number of epochs. The learning rate determines how much the weights of the algorithm are adjusted during training. The batch size determines the number of images processed in each training iteration, and the number of epochs determines the number of times the entire dataset is passed through the algorithm during training. To train the model, YOLOv8 algorithm is used which is imported from ultralytics. The model is trained using 150 epochs and with default optimizer (SGD) and loss function. To train the model using this algorithm, GPU is required. So, NVIDIA GeForce GTX 1050 Ti with 4 GB VRAM is used.

- **Helmet and Number Plate Detection:**

The YOLOv8 algorithm is capable of detecting and classifying objects in real-time. For helmet detection, the predicted bounding boxes are filtered based on the class probability, and only those with a high probability of containing a helmet are retained. This step is important in ensuring that only relevant information is passed on to the next stage of processing. The retained boxes are then passed through a non-maximum suppression algorithm to eliminate duplicate boxes. This algorithm is designed to select only the most relevant bounding boxes and discard the rest. For number plate recognition, the predicted bounding boxes are first cropped from the input image and resized to a fixed size. This step is important in ensuring that the input images have a consistent size, making it easier for the OCR model to recognize the characters on the license plate. The cropped images are then passed through a pre-trained OCR model, which is specifically designed for recognizing characters on license plates. The output of the OCR model is the recognized characters on the license plate. Finally, if a helmet is not detected in the input image, an email notification is sent to the specified email address using the Python's built-in SMTP

library. The email notification includes the timestamp of the image and a message indicating that a helmet was not detected in the input image.

- Email Sending Implementation:** The email sending functionality of the system was implemented using the Simple Mail Transfer Protocol (SMTP), which is a standard protocol for sending emails over the internet. When the system detects an event where a helmet is not detected or the number plate is not recognized, it checks the predefined conditions and sends an email alert to a specified email address. The email alert includes a snapshot of the image where the violation occurred, along with the date and time of the event. The system also sends a brief message to inform the recipient about the violation and the location where it occurred.

IV RESULT & ANALYSIS

The software used in this project included the following:

- Operating System:** Windows 10, 64-bit
- Programming Language:** Python 3.10.7
- Libraries:** OpenCV 4.5.2, NumPy 1.20.3, Pillow 8.3.1, PyTesseract 5.0.0- alpha.20201127, smtplib, imghdr
- Integrated Development Environment:** Visual Studio Code
- The YOLOv8 algorithm was trained using following Parameters:
- Batch size:** 16
- Learning rate:** 0.01
- Weight Decay:** 0.0005
- Number of epochs:** 100
- Input size:** 3401
- Number of classes:** 4

The system works as shown in the below flow chart:

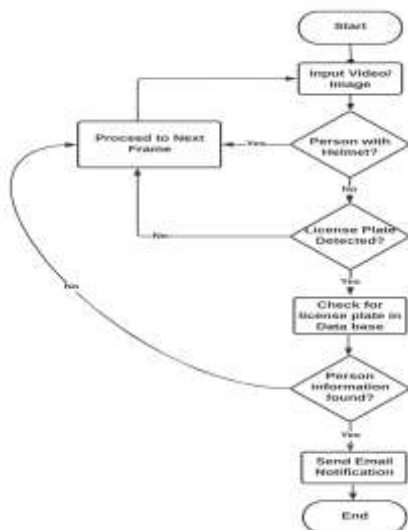


Fig. 2. System Workflow

Precision, Recall and Loss:

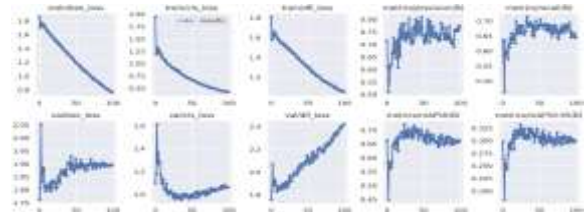


Fig. 3. Precision, Recall and Loss Comparative

Precision-Confidence Curve:

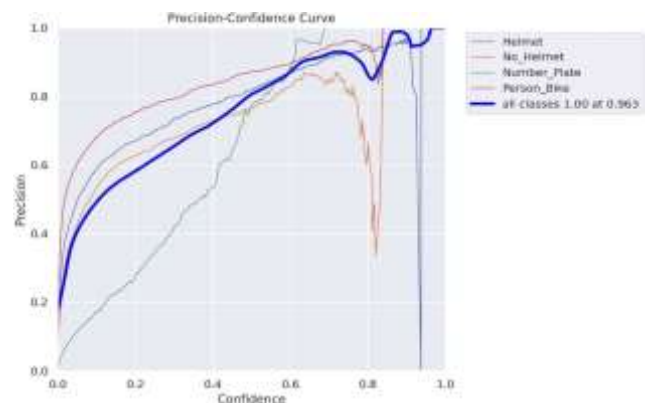


Fig. 4. Precision-Confidence Curve

A precision-recall curve (also called a precision-confidence curve) is a graph that represents the relationship between precision and recall at different probability thresholds used by a machine learning model. In the context of binary classification, a machine learning model assigns a probability score to each prediction that represents the likelihood of that prediction being positive or negative. The precision-recall curve plots the precision and recall of the model at different probability thresholds. The precision-recall curve is a useful tool for evaluating the performance of a model, especially in imbalanced datasets where there are many more negative instances than positive instances. The precision-recall curve can help us understand the trade-off between precision and recall, and choose the optimal probability threshold that maximizes both metrics. Interpreting a precision-recall curve involves looking at the shape and slope of the curve. A good model will have a precision-recall curve that closely hugs the upper right corner of the plot, indicating high precision and high recall at all probability thresholds. A poor model will

have a precision-recall curve that hugs the lower left corner, indicating low precision and low recall.

Detection:



Fig. 5. Classes Detection by System

Flask Application:



Fig. 6. Application Home Page



Fig. 7. Results Page

Email-notification:



Fig. 8. Email Notification

V FUTURE SCOPE

The current system can analyze videos in near real-time, but there is scope for further optimizing the model to achieve even faster analysis times. This can be done by implementing a distributed computing model that can distribute the workload across multiple GPUs, or by using a more efficient deep learning architecture that can perform faster object detection. The current system operates independently, but future versions of the system can be integrated with traffic signals to detect and alert riders violating traffic signals. This can help improve road safety and reduce the number of accidents caused by reckless driving. While the current system focuses on detecting individuals without helmets, future versions of the system can incorporate additional safety features such as detecting individuals with improperly fastened helmets or detecting individuals without protective gear such as knee and elbow pads. The current system is designed to send email notifications to riders without helmets, but future versions of the system can be integrated with law enforcement agencies to automatically send fines to riders violating traffic rules. This can help improve compliance with traffic rules and reduce accidents caused by reckless driving. In conclusion, the proposed system has a lot of potential for future development and improvement, with several possible features that can be added to enhance its capabilities and effectiveness. The system can be improved to achieve faster analysis times, higher detection accuracy, and more advanced safety features, and can be integrated with other systems and agencies to further improve road safety.

VI CONCLUSION

In conclusion, the deep learning project on helmet detection has successfully developed an automated system for detecting individuals riding two-wheelers without helmets using the YOLOv8 algorithm. The system can accurately detect the presence of helmets and number plates in video feeds and send email notifications to individuals detected without helmets, promoting safe riding practices among two-wheeler riders. The project's objectives were achieved by developing a deep learning model for object detection and integrating it with an email notification feature. The system's performance was optimized by fine-tuning the model and testing it on different video feeds. The results show that the system can accurately detect helmets and number plates in real-time with high accuracy and efficiency. The project's scope included developing a user-friendly interface for the system,

documenting its architecture and implementation details, and raising awareness among two-wheeler riders about the importance of wearing helmets. The project's objectives were achieved by delivering a well- documented and optimized system that can contribute towards improving road safety and promoting safe riding practices. The project's contribution to the field of computer vision and road safety is significant, as it provides an automated and accurate system for detecting individuals riding two-wheelers without helmets, which can help reduce accidents and fatalities caused by neglecting helmets. The project also highlights the importance of using deep learning techniques and advanced algorithms for object detection in real-time applications. Overall, the deep learning project on helmet detection is a successful application of advanced computer vision techniques and machine learning algorithms to promote safe riding practices and improve road safety. The system's accuracy, efficiency, and automated features make it a valuable tool for promoting safe riding practices and reducing accidents caused by neglecting helmets.

[8] P. Tzerefos et al., "A comparative study of Simple Mail Transfer Protocol (SMTP), Post Office Protocol (POP) and X.400 Electronic Mail Protocols", 1997 IEEE Proceedings of 22nd Annual Conference on Local Computer Networks, Minneapolis, MN, USA, Print ISBN:0-8186-8141-1, Print ISSN: 0742-1303.

VII REFERENCES

- [1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.
- [2] Redmon, Joseph & Farhadi, Ali. (2017). YOLO9000: Better, Faster, Stronger. 6517-6525. 10.1109/CVPR.2017.690.
- [3] Redmon, Joseph & Farhadi, Ali. "YOLOv3: An Incremental Improvement" IEEE Computer Vision and Pattern Recognition, 2018. doi.org/10.48550/arXiv.1804.02767.
- [4] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.
- [5] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich feature Hierarchy for Accurate object Detection" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580-587
- [6] R. R. V. e. Silva, K. R. T. Aires, and R. d. M. S. Veras, "Helmet Detection on Motorcyclists Using Image Descriptors and Classifiers," 2014 27th SIBGRAP Conference on Graphics, Patterns and Images, Rio de Janeiro, Brazil, 2014, pp. 141- 148, doi: 10.1109/SIBGRAP.2014.28.
- [7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, "SSD: Single Shot MultiBox Detection", IEEE Computer Vision and Pattern Recognition, 2016, doi.org/10.48550/arXiv.1512.02325.

Integrating Notifications And Manual Approval Workflows In Aws CDK Pipelines For Enhanced Devops Automation

Sandhya Naidu
23DSC29, M.Sc.(Computational Data Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts & Science
Vijayawada, A.P, India
Sandhyanaidu879@gmail.com

Rasani sunandini
23DSC28, M.Sc.(Computational Data Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts & Science
Vijayawada, A.P, India
rasanisunandini @gmail.com

Bora Uma Reddy
23DSC28,M.Sc.(Computational Data Science)
Dept. of Computer Science
P.B.Siddhartha College of Arts & Science
Vijayawada, A.P, India
umakrishna7620@gmail.com

Abstract—This paper proposes enhancing DevOps automation by integrating notifications and manual approvals within AWS CDK Pipelines, ensuring a streamlined CI/CD process with transparency. It examines CDK Pipelines' capabilities, identifies relevant notification tools, and implements human approval processes, analyzing real-world cases for impact on deployment speed and consistency. Results show improved process visibility and a balanced approach between automation and human input, addressing security needs. Leveraging Amazon SNS and AWS Step Functions aids in stakeholder alerts and crucial decision-making during pipeline execution. The study offers insights, guidance, and implementation steps for DevOps and AWS CDK users, fostering advancements in CI/CD pipelines and future AWS DevOps automation.

Keywords—AWS CDK Pipelines, DevOps Automation Notifications, Manual Approval, Continuous Integration (CI) Continuous Deployment (CD)

I INTRODUCTION

It has become increasingly apparent that the use of DevOps aptitudes, and the speed of delivery, has become an increasingly critical part of a successful software development process as the pace of software development has become ever more rapid. A DevOps model is a combination of the words "Advancement" and "Operations" in the modern world. Essentially, it is a social movement and a set of skills that are designed to bridge the gap between IT operations and development, encourage cooperation, as well as automate processes, so that the pipeline of program delivery can be accelerated. Through the use of continuous integration/continuous arrangement (CI/CD) pipelines, which serve as the backbone of sophisticated computer programme development, it has

become increasingly common for groups to develop, test, and deliver computer programme updates in an efficient and consistent manner. There is a possibility that Amazon Web Services (AWS) Cloud Development Kit (CDK) can serve as a powerful and adaptable tool for promoting Foundation as Code (IAC) within the AWS environment in a powerful and adaptable way by acting as an effective and adaptable tool. With the AWS CDK, developers can characterize cloud assets and foundations using standard programming dialects, such as TypeScript, Python, and Java, thereby simplifying the process of managing and expanding AWS assets. It streamlines asset provisioning, allowing teams to focus on development rather than framework management. Having a CDK provides a foundation for creating and sending frameworks as part of CI/CD pipelines, which is important for cutting-edge DevOps training. Regardless of what the case may be, the AWS CDK Pipelines provide a simple method of automating organizations, but there are fundamental viewpoints that need to be taken into account for these pipelines to succeed. Notices and manual approval procedures are two such perspectives that are crucial in the CI/CD management

A.AWS CDK's Importance in Advanced DevOps Hones
It is important to understand the relevance of AWS CDK in the context of advanced DevOps training before we delve deeper into the role of notifications and manual approvals. A typical foundation provisioning process in a traditional IT environment was often time-consuming and labor-intensive, resulting in bottlenecks and unproductive characteristics that hindered programme delivery. In recent years, with the advent of cloud computing, there has been a shift toward Framework as Code (IAC), an approach to framework management based on code. It has been reimagined how AWS assets are provisioned with AWS CDK, which is one of the components of the IAC development. AWS CDK enables developers to use

standard programming languages to characterise cloud assets.

B.CI/CD Pipelines Require Notices and Manual Endorsement Workflows There are a number of situations where human interaction is required in order to speed up the delivery process of software changes. CI/CD pipelines for software development and testing are primarily designed to automate the process of creating, testing, and communicating changes to software.

Critical Deployments: In order to ensure compliance and security requirements are met, some organizations, especially those that deploy security patches or make significant modifications to frameworks, require scrutiny to ensure that security and compliance requirements are met.

Regulatory Compliance: Organisations operating in restricted industries, such as healthcare, are often required to adhere to stringent compliance requirements. Following recent communication of changes, manual endorsements are essential for confirming **compliance**.

Business Critical Changes: Every time a change will have the potential to impact the business in a significant way, such as a new feature release or a rework of large forms, manual endorsement can be used to ensure alignment with the goals of the organization.

Emergency Rollbacks: During an automated test failure or a problem after an implementation is triggered by manual endorsements, an immediate rollback can be initiated by using manual endorsements.

The notices, on the other hand, serve as a vital communication channel between robotized pipeline stages and their partners in order to maintain transparency and respond swiftly to any potential problems. They provide a critical link between all parties involved in the pipeline.

Even though coordination notifications and manual approval procedures play an important role in AWS CDK Pipelines, consistency between them and AWS CDK Pipelines is not always evident despite their importance.

Finding the right balance between automation and human mediation requires a deliberate strategy that utilizes AWS administrations and best practices. As part of this term paper, we will examine the integration of notifications and human approval processes into the AWS CDK Pipelines and provide a detailed analysis of how these features are implemented.

II LITERATURE SURVEY

In the realm of secure software development, Stephen R. Schach's study in 2007 sheds light on the critical task of classifying and selecting high-integrity components. Emphasizing the paramount importance of human approval procedures within security-oriented environments, Schach's work offers insights into enhancing the safety and robustness of software [1]. In 2010, Junjie Ma, Yansheng Lu, and Yinhua Yang delved into the integration of CDK (Chemoinformatics Development Kit) with MapX, focusing on notification methods and processes for

chemical informatics applications. Their research not only contributes to the field of chemoinformatics but also underscores the significance of effective notification mechanisms in software integration [2]. A notable contribution in 2011 comes from Dorina C. Petriu and Ewa G. Boryczko, who proposed an agile approach to developing safety-critical software. Highlighting manual approval procedures as a crucial element in ensuring safety compliance, their work addresses the evolving landscape of software development methodologies [3]. In 2013, Bo Zhou, Zongxing Xie, Zonghua Gu, and Hong Mei presented a knowledge-based approach to automatic code review, discussing the roles of alerts and manual approval in code review procedures. Their work not only contributes to improving code quality but also underscores the importance of human intervention in the software development lifecycle [4]. Moving forward to 2016, Dan Sommerfield, Roger Longbotham, and Ron Kohavi offer a practical guide to controlled experiments in software engineering. Their document not only provides insights into experimental methodologies but also recognizes the roles of alerts and manual approval protocols in ensuring the validity of research outcomes [5].

In 2018, Hien Nguyen, Tung Nguyen, and Tung Do propose a framework for evaluating continuous delivery processes. Taking into account manual approval routines as part of the assessment process, their work contributes to the ongoing discourse on optimizing software delivery practices [6]

Elif Kocaoglu, Eric Bodden, and Mira Mezini, in 2019, addressed secure cloud-edge integration. Their article not only explores safe integration practices but also underscores the pivotal role of alerts and approval protocols in guaranteeing security in the dynamic landscape of cloud-edge systems [7].

Christian Jansen, Pascal Kaufmann, and Kurt Schneider's 2019 survey investigates the needs for DevOps tooling, emphasizing effective notification methods and approval procedures. The survey provides valuable insights into the evolving requirements of developers in the DevOps ecosystem [8].

In 2020, Lucas Freire, Daniel Oliveira, and Uirá Kulesza conducted a survey on DevOps practices for mobile 2022 applications. Emphasizing the significance of alerts and approval procedures in mobile app development, their work provides a comprehensive overview of the DevOps landscape specific to the mobile application domain [9].

Finally, in the same year, Claudio Menghi, Elisabetta Di Nitto, and Carlo Ghezzi explored DevOps practices for high-integrity systems. Their article delves into the unique challenges of maintaining high integrity in systems, emphasizing the role of manual approval protocols in ensuring the robustness of such systems [10]

III PROBLEM STATEMENT

Addressing the challenges in AWS CDK Pipelines and achieving the outlined objectives requires a comprehensive

approach to enhance the overall efficiency and reliability of DevOps automation. One crucial aspect is the implementation of real-time notifications, which involves incorporating instant alert mechanisms and updates throughout the pipeline. Real-time notifications ensure that the development and operations teams are promptly informed about any issues, progress, or successful deployments. This proactive communication fosters agility and responsiveness, allowing teams to quickly identify and address potential bottlenecks, resulting in faster issue resolution and an overall streamlined development lifecycle. The integration of flexible manual approval workflows is another pivotal objective. While automation is fundamental, there are scenarios where human intervention is necessary, especially for critical decision-making or compliance checks. By incorporating manual approval steps into the AWS CDK Pipelines, stakeholders can review and approve changes before they are deployed to production. This not only enhances control over deployments but also ensures that regulatory and compliance requirements are met. The inclusion of manual approval workflows adds a layer of governance to the pipeline, contributing to a more secure and compliant DevOps process. Enabling customizable notifications further enhances the adaptability and user-friendliness of the AWS CDK Pipelines. Recognizing that different teams and stakeholders have unique communication preferences, providing the ability to tailor notifications becomes essential. Teams can choose specific channels, formats, or frequencies based on their workflow and preferences. This customization fosters better collaboration among team members and stakeholders, contributing to a more adaptable and user-centric DevOps environment. Additionally, customizable notifications play a crucial role in keeping everyone informed in a way that aligns with their specific needs.

In summary, achieving these objectives not only addresses the current challenges in AWS CDK Pipelines but also brings about a transformative impact on DevOps automation.

FLOW DIAGRAM

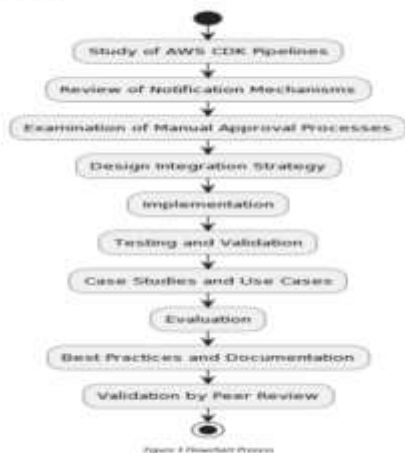


Figure 2 Research Process

IV PROPOSED SOLUTION

The proposed solution for integrating notifications and manual approval workflows into AWS CDK pipelines is based on the following approach:

1. Create a custom AWS CDK construct that implements the desired notification and manual approval workflow.
2. Use the AWS CDK Pipeline to deploy the custom construct to AWS.
3. Configure the AWS CDK Pipeline to trigger the notification and manual approval workflow when necessary.

It is possible to implement a custom AWS CDK construct in any programming language supported by the AWS CDK. The construct should expose methods for sending notifications and requesting manual approvals.

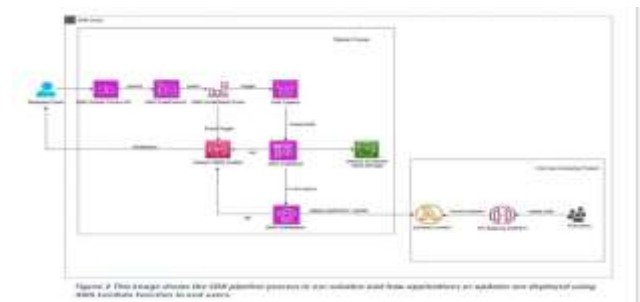
Any point in the AWS CDK Pipeline can be configured to trigger a notification and manual approval workflow. In addition to sending a notification to a group of approvers before deploying a new change to production, the pipeline could also be configured to

wait for manual approval before deploying a new change.

The team has implemented a CDK Pipeline for their web project, which consists of a static web page served by an AWS Lambda function through Amazon API Gateway. The CI/CD process initiates when a developer pushes changes to the repository, triggered by a CloudWatch event. The pipeline has two key stages: Pre-production for testing and Production for the end product. Upon reaching the pre-production stage, the CI/CD process pauses due to a manual approval gate. A CloudWatch event notifies stakeholders for their review, and the pipeline includes an SNS notification to alert them when manual approval is required. Once changes are approved, the CI/CD process advances to the production stage, making the updated website available to users. In case of rejection, the process concludes at the pre-production stage with no impact on end users. This workflow ensures thorough testing and stakeholder approval before deploying changes to the production environment.

SOLUTION ARCHITECTURE

The following diagram illustrates the solution architecture



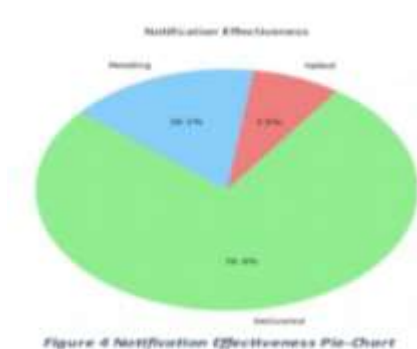
V. IMPLEMENTATION RESULTS

A. Comparison of Arrangement Times:

As shown in the bar chart above, a comparison of setup times across organizations allows us to gain some insight into the efficiency of various phases of the pipeline process. There are significant disparities between the Construct and Test phases, suggesting that there are differences in the construction and testing methodologies. A balanced setup time is observed in the Convey phase, suggesting consistency in the conveying process. In the Manual Endorsement phase, setup time has been significantly reduced in the last few years, particularly after manual coordination endorsements, which indicates that coordination protocols have been improved as well. In the generation phase, setup times are convergent, highlighting the similarity of efficiency levels. Overall, the analysis provides a comprehensive overview that enables organizations to identify areas for improvement and capitalize on positive trends to enhance overall operational efficiency

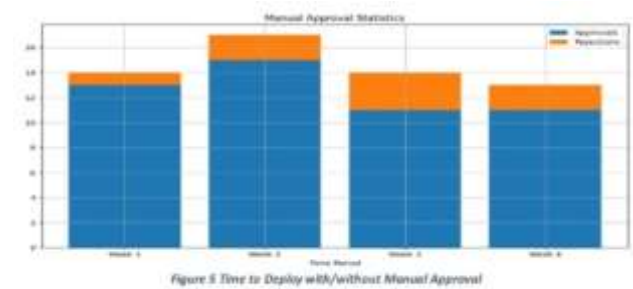
Notification Efficiency:

The Pie Chart analysis of notification conveyance statuses provides a concise evaluation of the efficiency of a notification instrument. This chart presents a visual snapshot of the conveyance status in terms of Conveyed, Fizzled, and Pending conveyances, which provides a good understanding of the performance of the instrument. As a key metric for determining the overall effectiveness, the Adequacy Index, derived from the size of the Conveyed segment, is an important one. A larger conveyed segment indicates a higher success rate in delivering notifications, reflecting the instrument's efficiency. A minimal Fizzled segment suggests reliability, while the Pending segment highlights notifications awaiting delivery. This visual representation enables quick identification of strengths and areas for improvement, empowering organizations to optimize their notification instruments for effective communication.



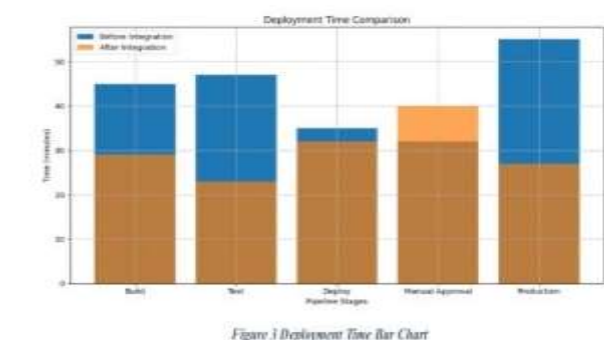
Statistics on Manual Approval:

The analysis of manual approval activities over time, visualized through a Stacked Bar Chart, provides a comprehensive view of the dynamics in decision-making processes. The chart, with time periods on the X-axis and manual approval counts on the Y-axis, distinguishes between approvals and rejections through segmented bars. Monitoring fluctuations in Manual Approvals and Rejections segments allows for insights into the overall volume and trends. Patterns, peaks, and consistency in distribution across time periods offer valuable information, aiding in the identification of potential challenges or areas for improvement in the manual approval workflow. Additionally, analyzing the ratio between approvals and rejections provides a nuanced understanding of the approval process. Overall, this visual representation is a powerful tool for stakeholders to refine processes, enhance efficiency, and maintain transparency in manual approval systems.



VI FUTURE SCOPE

In terms of potential directions, the paper suggests that the integration of notifications and approval workflows into AWS CDK Pipelines is just the beginning.



Using machine learning and artificial intelligence in conjunction with deployment automation to enhance deployment automation could be the next step in future research. The paper also suggests future research could further optimize the balance between automation and

human intervention. Further research might be conducted in order to explore the possibility of integrating other tools and technologies into AWS CDK Pipelines so as to further enhance the overall DevOps process.

VII CONCLUSION

In an era of rapid program development and arrangement, the inclusion of notifications and human approval procedures into AWS CDK Pipelines has emerged as a critical technique for enhancing DevOps computerization in a time of rapid change. This investigation has delved into the key components of this integration, its recommendations, as well as its applicability in the real world. The use of AWS CDK Pipelines as the foundation for DevOps forms has enabled organizations to simplify their program conveyance pipelines by using AWS CDK Pipelines as a base. Nevertheless, as arrangements became more complex and administrative requirements became more stringent, it became increasingly necessary for discharge forms to be regulated and auditable in order to meet compliance requirements. The purpose of this thought was to provide a solution that consistently connects notifications to manual endorsement steps within AWS CDK pipelines in order to address this need. It is important for firms to integrate notifications and manual endorsements into their DevOps development as a reference point for improved mechanization as they make progress in their DevOps journey. In addition to addressing fundamental concerns about compliance, security, and controlled discharges, it also adapts to industry best practices as well.

The conclusion of this study is that it has provided a path to a more robust and effective DevOps biological system by combining notifications with manual endorsement procedures within AWS CDK Pipelines, in addition to paving the way for better DevOps performance. Organizations are able to take advantage of this technique in order to overcome the challenges associated with modern computer program delivery with confidence, knowing they have the tools to ensure speed and control of the delivery process. DevOps' long-term sustainability lies within the cohabitation of mechanization and supervision that is acceptable, and it is this integration that sets the course for a future that is harmonious between the two.

VIII REFERENCES

- [1] "What is the AWS CDK? - AWS Cloud Development Kit (AWS CDK) v2,"
- [2] G. Kim, P. Debois, J. Willis, J. Humble, and J. Allspaw, *The Devops Handbook*

How to Create World-class Agility, Reliability, and Security in Technology Organizations. *It Revolution Pr*, 2015.

[3] J. Humble and D. Farley, *Continuous Delivery*. Upper Saddle River, Nj: AddisonWesley, 2011.

[4] "What is the AWS CDK? - AWS Cloud Development Kit (AWS CDK) v2,"

[5] J. Davis and R. Daniels, *Effective DevOps*. "O'Reilly Media, Inc.," 2016.

[6] G. Kim, K. Behr, and G. Spafford, *The Phoenix Project: a novel about IT, DevOps, and helping your business win*. Portland, Oregon: It Revolution Press, 2018.

[7] C. Steiner and X. Li, *算法帝国 = Automate this: how algorithms came to rule our world /Suan fa di guo = Automate this: how algorithms came to rule our world*. 人民邮电出版社, Beijing: Ren Min You Dian Chu Ban She, 2014.

[8] "What is AWS Step Functions? - AWS Step Functions," docs.aws.amazon.com.

[9] Veselin Kantsev, *Implementing DevOps on AWS*. Veselin Kantsev, 2016.

[10] "AWS Well-Architected Framework - AWS Well-Architected Framework,"

Swarm Intelligence Unleashed: Unraveling Nature's Algorithms for Computational Efficiency

Sagurthi Pavitra

23DSC30, M.Sc. (Computational Data Science)

Dept. of Computer Science P.B.
Siddhartha College of Arts & Science

Vijayawada, A.P, India

sagurthipavitra@gmail.com

Bhavisya Sagarika

23DSC22, M.Sc. (Computational Data Science)

Dept. of Computer Science P.B.
Siddhartha College of Arts & Science

Vijayawada, A.P, India

bhavisyasagarika@gmail.com

Muddamsetty Sriya

23DSC28, M.Sc. (Computational Data Science)

Dept. of Computer Science P.B.
Siddhartha College of Arts & Science, Vijayawada, A.P, India

muddamsetty.sriya@gmail.com

Abstract- Swarm intelligence, inspired by the collective behavior observed in natural systems, has emerged as a powerful paradigm for solving complex computational problems. This paper delves into the exploration of various nature-inspired algorithms that mimic the collaborative and decentralized strategies observed in social insects, birds, and other natural entities. Our study reviews key swarm intelligence algorithms, such as Ant Colony Optimization, Particle Swarm Optimization, Bee Algorithm, Firefly Algorithm, Bat Algorithm, Cuckoo Search, and Artificial Bee Colony. We examine the foundational works that introduced these algorithms and delve into their underlying principles. Through this exploration, we unveil the mechanisms that govern the dynamics of swarm intelligence and highlight their relevance in achieving computational efficiency. The review encompasses not only the original formulations of these algorithms but also their subsequent extensions, variations, and hybridizations, showcasing the continuous evolution of swarm intelligence research. Furthermore, we discuss real-world applications across diverse domains where swarm intelligence algorithms have demonstrated their efficacy. The adaptability, scalability, and self-organizing capabilities of these algorithms make them particularly well-suited for addressing optimization challenges in fields ranging from engineering and logistics to finance and telecommunications. The paper concludes by emphasizing the ongoing research trends, challenges, and potential future directions in the realm of swarm intelligence. By unraveling nature's algorithms, this study contributes to a deeper understanding of the computational principles that govern swarm intelligence and provides insights into harnessing these algorithms for enhanced computational efficiency in solving intricate problems across various domains.

I INTRODUCTION

The fusion of scalable computing and artificial intelligence has led to remarkable strides in the development of swarm

intelligence approaches. At its core, swarm intelligence algorithms emulate the cooperative and group behaviors observed in social organisms within the natural world. This introductory segment sets the stage for a comprehensive exploration of swarm intelligence, providing insight into how advancements in scalable computing have paved the way for innovative solutions inspired by the collaborative nature of social organisms. Unveiling Swarm Intelligence Delving deeper, this section unravels the essence of swarm intelligence, shedding light on how algorithms mimic the cooperative and group behaviours found in nature. By dissecting the principles behind swarm intelligence, readers gain a foundational understanding of how these algorithms tap into the collective wisdom of decentralized systems. The Symbiosis of Technology and Nature Exploring the intricate relationship between technology and nature, this part of the tutorial examines how swarm intelligence leverages scalable computing and artificial intelligence. By elucidating this symbiotic connection, we underscore the transformative impact of swarm intelligence on problem-solving paradigms. This sets the stage for an exploration of real-world applications, showcasing the versatility of swarm intelligence in various domains. Real-World Applications Transitioning from theory to practicality, we delve into diverse real-world examples and applications of swarm intelligence algorithms. From optimization challenges to dynamic problem-solving scenarios, readers gain insights into how these algorithms are reshaping industries and addressing complex problems [1]. **The Power of Swarm:** Demonstrating Importance Through Optimization The tutorial culminates by spotlighting one of the most popular swarm intelligence-based optimization algorithms. Through a detailed example, we illustrate the practical significance of the swarm intelligence approach, showcasing its prowess in optimizing complex systems and addressing real-world challenges.

Swarm intelligence, an influential technique within both artificial and natural intelligence, finds its roots in the study of collective behavior observed in decentralized and self-organized systems. The term was officially introduced by Gerardo Benny and Joon Wang in 1989, with a specific focus on cellular robotics systems. This section unravels the core

principles of swarm intelligence, shedding light on its dual identity as an artificial and natural intelligence phenomenon. By exploring the intricacies of collective behavior, the tutorial delves into the foundational aspects that define swarm intelligence, emphasizing its reliance on decentralized and self-organized systems. As we journey through this exploration, the tutorial aims to provide a comprehensive understanding of swarm intelligence, showcasing its significance in the realms of artificial and natural intelligence. The focus remains on elucidating the transformative potential of studying collective behaviors as demonstrated by Benny and Wang's pioneering work in the context of cellular robotics systems [18].

Principles of Swarm Intelligence In dissecting the foundations of swarm intelligence, several core principles govern the behavior and functionality of the collective. Understanding these principles is crucial for unveiling the intricate dynamics that drive swarm intelligence systems.

Awareness: Each member within the swarm must possess a heightened awareness of both their immediate surroundings and individual capabilities. This heightened awareness enables adaptive responses to changes in the environment.

Autonomy: A fundamental tenet of swarm intelligence, autonomy ensures that each member operates as an autonomous entity, not subordinate to a centralized authority. This autonomy fosters self-coordination and empowers individual members to act as masters, contributing to the collective without being directed as slaves.

Solidarity: Upon completing a task, swarm intelligence emphasizes the importance of solidarity among members. Individual entities should autonomously seek out new tasks, fostering a seamless transition between different objectives and maintaining the collective momentum [2].

Expandability: The system's architecture must facilitate dynamic expansion, allowing for the seamless integration of new members into the swarm. This adaptability is crucial for scalability and responsiveness to varying task complexities.

Resilience In the face of member removal or system perturbations, swarm intelligence systems must exhibit resiliency. The system should be inherently self-healing, adapting to changes in the swarm's composition without compromising overall functionality [3].

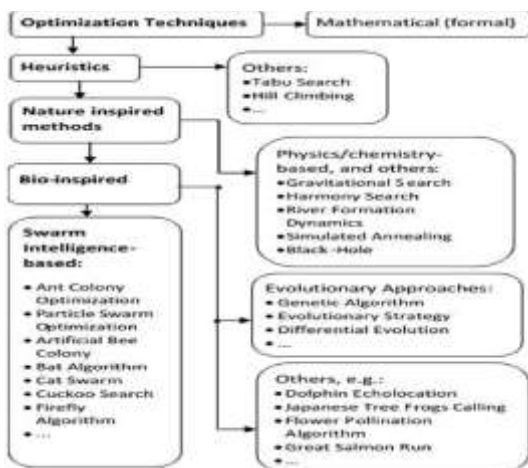
This section delves into the fundamental principles that govern swarm intelligence, highlighting the interplay of awareness, autonomy, solidarity, expandability, and resiliency. Mastery of these principles is essential for comprehending the robust and adaptive nature of swarm intelligence systems.

Real-World Examples: Certainly! Swarm intelligence is a fascinating field with numerous real-world examples and applications. Here are additional examples and areas of development in swarm intelligence:

- Robotics and Drone Swarms:** Swarm robotics involves the coordination of multiple robots to perform tasks collectively. Drone swarms are a prominent example where multiple drones collaborate to achieve complex tasks such as search and rescue operations, environmental monitoring, or surveillance. These systems mimic the collaborative behavior seen in natural swarms [4].
- Traffic Management:** Swarm intelligence principles can be applied to optimize traffic flow in urban areas. By mimicking the self-organization observed in ant colonies or bird flocks, traffic signals and flow can be dynamically adjusted to minimize congestion and improve overall efficiency.
- Supply Chain Optimization:** Swarm intelligence algorithms can be employed to optimize supply chain logistics, helping in route planning, inventory management, and distribution. The decentralized and self-organizing nature of swarm intelligence is well-suited for handling the complexity of supply chain networks.
- Medical Applications:** In the field of medicine, swarm intelligence is being explored for tasks such as medical imaging analysis, drug discovery, and treatment planning. Algorithms inspired by swarm behavior can help in identifying patterns and making predictions based on large datasets.
- Environmental Monitoring:** Swarm intelligence can be applied to monitor environmental parameters by deploying a network of sensors. The sensors can communicate and adapt collectively to changes in the environment, providing a more comprehensive and efficient monitoring system [5].
- Internet of Things (IoT):** In the context of IoT, swarm intelligence algorithms can enhance the efficiency of connected devices by enabling them to work together seamlessly. This is particularly useful in smart homes, industrial automation, and other IoT applications where devices need to coordinate their actions.
- Financial Modeling:** Swarm intelligence is explored in the financial domain for modeling and predicting market trends. The collective decision-making process can be applied to analyze market data and make predictions about stock prices or investment strategies.
- Cybersecurity:** Applying swarm intelligence principles to cybersecurity involves creating adaptive defense mechanisms [7]. Systems can learn and adapt to evolving threats collectively, enhancing overall security by responding dynamically to cyber-attacks in terms of development areas, ongoing research is focused on refining swarm intelligence algorithms, improving scalability, and adapting them to diverse applications. Additionally, efforts are being made to integrate swarm intelligence with other emerging technologies such as artificial intelligence, machine learning, and decentralized computing for more robust and intelligent systems.

Applications: Swarm intelligence transcends its roots in optimization and finds diverse

Fig. 1. Scope of Swarm Intelligence



applications across a spectrum of fields, showcasing its adaptability and efficacy in solving complex problems [6]. Some notable applications include: Optimization Issues: Swarm intelligence's traditional application lies in optimizing solutions for complex problems, ranging from mathematical optimization to logistical challenges. Library Item **Acquisition:** In library management, swarm intelligence is utilized for efficiently acquiring and organizing library items. This application enhances the retrieval and categorization of resources. Communications: Swarm intelligence aids in optimizing communication networks, ensuring efficient data transfer, and enhancing the overall performance of communication systems. Categorization of Medical Datasets: In healthcare, swarm intelligence is applied to categorize and analyze medical datasets [15]. This assists in diagnosing diseases, identifying patterns, and optimizing medical decision-making processes. Dynamic Control: Swarm intelligence plays a role in dynamic control systems, where it contributes to adaptive and responsive control mechanisms in various domains, such as robotics and manufacturing. Heating System Planning: Planning heating systems efficiently is made possible through swarm intelligence. This application optimizes energy usage and distribution for heating purposes in different environments. Tracking and Prediction of Moving Objects: Swarm intelligence is employed for tracking and predicting the movements of dynamic entities, finding applications in fields like surveillance, traffic management, and autonomous vehicle navigation.

Basic Research: In basic research, swarm intelligence serves as a valuable tool for exploring and understanding complex phenomena, contributing to advancements in fields such as biology, ecology, and physics. Engineering: Swarm intelligence is harnessed in engineering for optimizing design processes, enhancing structural efficiency, and solving complex engineering problems. Business and Social Sciences: Within business and the social sciences, swarm intelligence finds applications in areas such as market analysis, decision-making processes, and modeling social behaviors. The broad scope of swarm intelligence applications underscores its versatility and effectiveness across a multitude of disciplines. As technology advances, the influence of swarm intelligence is likely to expand, contributing to innovative solutions in both theoretical research and practical problem-solving scenarios. Certainly! Let's delve deeper into the concepts of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO).

II RELATED WORK

Swarm intelligence is a field of study that draws inspiration from the collective behavior of social insects and other natural systems to design algorithms for solving complex problems. There has been a significant amount of research in this area, and several nature-inspired algorithms have been proposed. Here are some of the key algorithms in swarm intelligence along with related work: Ant Colony Optimization (ACO): Original Work: Dorigo, M., Maniezzo, V., & Coloni, (1996). The Ant System: Optimization by a colony of cooperating agents. IEEE Transactions on Systems,

Man, and Cybernetics - Part B: Cybernetics, 26(1), 29-41. Related Work: Numerous extensions and variations of ACO have been proposed, such as Max-Min Ant System, Ant Colony System, and Rank-Based Ant System. Particle Swarm Optimization (PSO): Original Work: Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of IEEE International Conference on Neural Networks (pp. 1942-1948). Related Work: Many variants and improvements to PSO have been introduced, including Constriction Factor PSO, Adaptive PSO, and Quantum PSO. Bee Algorithm (BA): Original Work: Pham, D.T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., & Zaidi, M. (2005). The Bees Algorithm—a novel tool for complex optimization problems. In Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(11), 1273-1285. Related Work: Modified versions of the Bee Algorithm have been proposed for specific applications, such as the Enhanced Bees Algorithm and the Modified Bees Algorithm. Firefly Algorithm (FA): Original Work: Yang, X. S. (2008). Firefly algorithms for multimodal optimization. In International Symposium on Stochastic Algorithms (pp. 169-178). Related Work: Various adaptations and hybridizations of the Firefly Algorithm have been explored, including the Enhanced Firefly Algorithm and the Hybrid Firefly Algorithm. Bat Algorithm (BA): Original Work: Yang, X. S. (2010). A new metaheuristic bat-inspired algorithm. In Nature Inspired Cooperative Strategies for Optimization (pp. 65-74). Related Work: Several variants of the Bat Algorithm have been proposed, such as the Improved Bat Algorithm and the Hybrid Bat Algorithm. Cuckoo Search (CS): Original Work: Yang, X. S., & Deb, S. (2009). Cuckoo search via Lévy flights. In World Congress on Nature & Biologically Inspired Computing (pp. 210-214). Related Work: Enhanced versions like the Modified Cuckoo Search and applications in various domains have been investigated. Artificial Bee Colony (ABC): Original Work: Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm. Journal of Global Optimization, 39(3), 459-471. Related Work: Extensions and improvements, including the ABC algorithm with differential evolution, have been proposed. These are just a few examples, and there are many other swarm intelligence algorithms and hybrid approaches that researchers have developed over the years. Additionally, researchers often explore combining different swarm intelligence techniques or integrating them with other optimization methods to create more robust and efficient algorithms for solving diverse optimization problems.

PROPOSED WORK

Particle Swarm Optimization

(PSO): Overview:

- **Inspiration:** PSO draws inspiration from the social behavior of animals and insects, particularly bird flocking or fish schooling. Key Concepts: Particles: In PSO, each potential solution to an optimization problem is represented as a particle. Swarm: The

particles collectively form a swarm. Each particle adjusts its position in the solution space based on its own experience and the experience of the entire swarm. **Movement:** Particles move through the solution space, adjusting their positions based on their current best solution and the global best solution found by any particle in the swarm. Working Mechanism: **Initialization:** Randomly initialize particles in the solution space. Evaluation: Evaluate the fitness of each particle. Update Velocity and Position: Adjust the velocity and position of each particle based on its historical best position and the best position found by any particle in the swarm. Iteration: Repeat the process until a convergence criterion is met.

Algorithm 1: Pseudo code of the original version of PSO

1. initialize the swarm, it could be randomly
2. evaluate particles in the swarm – calculate their fitness
3. select the best particle G_{best} – the best fitness value in the swarm
4. while the stop criterion is not met do
5. for each particle P_i do
6. evaluate particle P_i – calculate its fitness
7. if fitness value of P_i is better than the personal best in History (P_{best}) then update P_{best} with the use of P_i
8. end if
9. if the fitness value of P_i is better than the fitness value of G_{best} then update G_{best} with the use of P_i
10. end if
11. end for
12. for each particle P_i do
13. calculate particle velocity according to equation 1
14. update particle position according to equation 2
15. end for
16. end while

III ANT COLONY OPTIMIZATION (ACO):

Indeed, Ant Colony Optimization (ACO) is a nature-inspired optimization algorithm based on the foraging behavior of real ants. It's a fundamental component of swarm intelligence and has been successfully applied to various combinatorial optimization problems. Here's a more detailed explanation of how Ant Colony Optimization works.

OVERVIEW:

Inspiration from Ant Behavior: ACO draws inspiration from the foraging behavior of real ant colonies. Ants are known for their ability to find the shortest path between their nest and a food source, even when the path involves complex terrain.

KEY CONCEPTS:

- **Ants:** In the algorithm, artificial ants represent solutions to the optimization problem. Pheromones: Ants deposit a chemical substance called pheromone on the paths they travel. The intensity of pheromones influences the probability of other ants choosing the same path.
- **Exploration:** ACO strikes a balance between exploitation (choosing paths with higher pheromone levels) and exploration (choosing new paths) to find optimal solutions. Working Mechanism: Initialization: Place a population of artificial ants at the starting point. Solution
- **Construction:** Ants construct solutions by moving through the solution space. The probability of choosing a particular path is influenced by the amount of pheromone on that path and a heuristic value based on domain-specific information. Evaluation: Evaluate the quality of each solution constructed by ants based on the objective function of the optimization problem.
- **Pheromones:** Adjust pheromone levels on solution components based on the quality of solutions. Stronger solutions leave more pheromone, influencing the choices of future ants. Iteration: Repeat the process for a predefined number of iterations or until a convergence criterion is met.

Analogy to Ant Foraging: In the real world, ants leave the nest to search for food. As they find food, they leave a pheromone trail on their way back to the colony. The intensity of the pheromone trail attracts other ants, guiding them to the food source. Over time, the shortest path becomes more attractive due to the higher concentration of pheromones, leading to the establishment of an efficient route.

BASIC CONCEPTS OF ACO:

- **Pheromones:** In ACO, artificial ants construct solutions by moving through the solution space. They leave a trail of artificial pheromones on the components of the solution they traverse. Pheromones are a form of communication among ants. In the context of optimization, they represent information about the quality of solutions.
- **Construction:** Ants construct solutions by making probabilistic decisions based on both the pheromone levels and a heuristic value. The probability of choosing a particular path is influenced by the amount of pheromone on that path and a heuristic value, which provides additional information about the desirability of the path based on the specific problem.
- **Evaporation:** Pheromones evaporate over time. This simulates the natural decay of pheromones in real ant behavior and prevents the algorithm from getting stuck in suboptimal solutions. Evaporation

ensures that paths that are not reinforced by ants over time lose their pheromone intensity.

Introduction to ACO: ACO was introduced in the early 1990s by Marco Dorigo. It is a search technique inspired by the swarm intelligence observed in ant colonies. **Swarm Intelligence:** ACO mimics the foraging behavior of real ants, where ants collectively find the shortest paths between their nest and a food source. Ants communicate indirectly through the use of pheromones, which act as chemical messengers on the ground.

Trail: When ants search for food, they initially explore the surrounding area randomly. Moving ants deposit a pheromone trail on the ground, representing the paths they have explored. Pheromones evaporate over time. **Path Selection:** Ants choose their paths probabilistically based on the concentration of pheromones on the explored paths. Higher concentrations of pheromones make a path more attractive to other ants—indirect Communication. The use of pheromones for path marking and following enables ants to find efficient routes between their nest and a food source.

Real-Life Analogy: In the real world, when ants search for food, they initially explore the area surrounding their nest randomly. The pheromone trail left by moving ants acts as a guide for other ants. Ants probabilistically choose paths with higher pheromone concentrations, leading to the discovery of efficient routes between the nest and the food source.

Algorithm 2: Pseudocode of the ACO

1. for the CO problem to be solved, derive the finite set $C = \{c_1, c_2, \dots, c_n\}$; c_i : cng of solution components (they are used to assembly solution to the CO)
2. define the pheromone values T (the pheromone model, it is a parametrized probabilistic model), the pheromone values τ_{ij} are linked with solution components
3. while termination conditions not met do
4. $s = []$ – an empty sequence of solution components (antbased solution construction starts)
5. while $N(s) \neq \emptyset$ do
6. c chooses from $(N(s))$ – add a feasible solution component at each construction step with respect to the pheromone model (equation 3 can be used)
7. determine $N(s)$, $N(s) \setminus C_n$, the specification of $N(s)$ depends on the solution construction mechanism
8. end while
9. pheromone update
10. end while

IV RESULT & ANALYSIS

Particle Swarm Optimization (PSO) is an optimization algorithm inspired by the social behavior of animals, such as bird flocking or fish schooling. In PSO, particles represent potential solutions to an optimization problem, and the algorithm iteratively adjusts their positions in the solution space to find the optimal solution.

Here's a simplified explanation of how particles move in PSO over different iterations: **Initialization:** Randomly initialize a swarm of particles in the solution space. Assign initial velocities to the particles.

Iteration Process: For each iteration, the particles move towards their optimal positions based on their current positions, velocities, and historical p ; information.

Update Velocity and Position: Adjust the velocity of each particle based on its historical best position (personal best) and the best position found by any particle in the swarm (global best). Update the position of each particle based on its new velocity.

Convergence: With each iteration, particles tend to adjust their positions in a way that brings them closer to the optimal solution. The algorithm continues iterating until a termination criterion is met, such as maximum number of iterations or achieving a satisfactory solution. Now, let's discuss the concept of particles getting closer to the goal at different iterations.

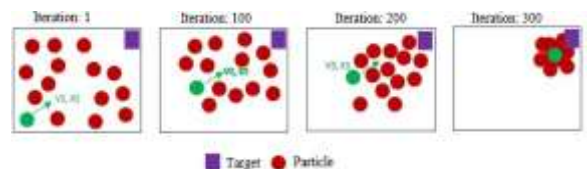


Fig. 2. PSO Iterations

Iteration 1: Particles are initialized randomly and start moving towards their optimal positions. Some particles may make significant progress towards the goal, while others might explore different regions of the solution space.

Iteration 100: After 100 iterations, particles have adjusted their positions based on their historical and global best information. Many particles have likely converged towards a common solution, and the overall swarm is closer to the optimal solution.

Iteration 200: Continued adjustment of positions and velocities. The swarm may be converging further, with particles fine-tuning their positions based on the shared knowledge within the swarm.

Iteration 300: Further refinement of particle positions. By this point, the swarm may have reached a state of convergence, with particles closely grouped around the optimal solution.

In summary, as the iterations progress in PSO, particles dynamically adjust their positions and velocities, collaboratively converging toward the optimal solution. The swarm's collective intelligence, influenced by both individual and global knowledge, helps refine the solution over time.

Here are some reasons why PSO is preferred in many engineering problems:

Ease of Implementation: PSO has a straightforward and easy-to-understand conceptual framework. Its simplicity

makes it accessible to practitioners with varying levels of expertise in optimization and engineering.

Few Parameters: PSO typically requires tuning only a small number of parameters, often just two: the inertia weight and the acceleration coefficients. This simplicity contrasts with other optimization algorithms that might have more intricate parameter configurations.

Efficient Convergence: Despite its simplicity, PSO often exhibits efficient convergence to the optimal solution, especially in cases where the solution space is not highly complex or noisy.

Global and Local Search Balance: PSO strikes a balance between global and local exploration, allowing particles to explore diverse regions of the solution space while also converging towards promising areas.

Applicability to Multidimensional Problems: PSO is effective in handling multidimensional optimization problems, where the search space involves a larger number of variables.

Versatility: PSO is versatile and has been successfully applied in various engineering domains, including control systems, signal processing, image processing, and parameter tuning for machine learning algorithms. **Parallelization Potential:** The algorithm is inherently parallelizable, making it suitable for implementation on parallel computing architectures, which can enhance its efficiency further. **No Gradients Required:** PSO is a derivative-free optimization method, meaning it doesn't require information about the derivatives of the objective function. This makes it applicable to scenarios where the analytical form of the objective function is unknown or difficult to obtain. While PSO excels in certain scenarios, it's essential to note that the algorithm might face challenges in highly complex and multimodal optimization problems. In such cases, more advanced algorithms or hybrid approaches might be considered. Nonetheless, its simplicity, ease of implementation, and ability to quickly converge make PSO a valuable tool in the engineer's optimization toolkit.

Ant Colony Optimization (ACO) is a popular algorithm in swarm intelligence inspired by the foraging behavior of ants. In ACO, artificial ants are used to find optimal solutions to combinatorial optimization problems. The algorithm is widely applied in various domains, including routing problems, scheduling, and logistics.

Results: Convergence to Optimal Solutions: ACO is known for its ability to converge to near-optimal or optimal solutions for complex optimization problems. The algorithm often outperforms traditional optimization algorithms, especially in problems with a large search space.

Robustness: ACO tends to be robust against changes in the problem instance or parameters. It can adapt well to dynamic environments, making it suitable for real-world applications where conditions may change over time.

Scalability: ACO exhibits good scalability, making it suitable for both small and large-scale optimization problems. The algorithm's performance remains competitive as the problem size increases.

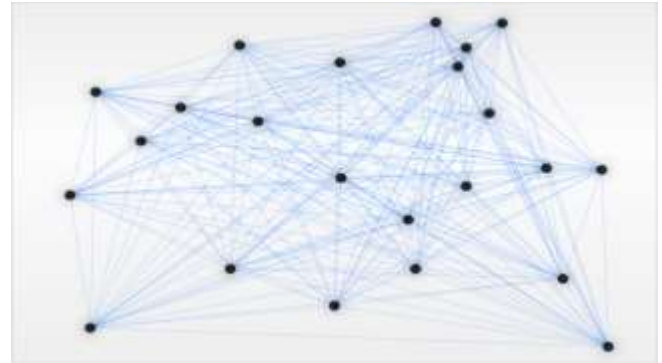


Fig. 3. Ant Colony Optimisation-1

Exploitation Balance: ACO maintains a good balance between exploration and exploitation, allowing it to discover new solutions while refining existing ones. This property is crucial for tackling complex optimization landscapes.

Parallelization: ACO can be parallelized efficiently, allowing for faster convergence and improved performance on parallel computing architectures. **Analysis: Parameter Sensitivity:** The performance of ACO is influenced by its parameters, such as the pheromone update rate, exploration probability, and the number of ants. Fine-tuning these parameters is essential to achieve optimal results, and different problem instances may require adjustments.

Local Optima Handling: ACO may struggle with getting stuck in local optima, particularly in problems with rugged landscapes. Various enhancements, such as introducing pheromone evaporation and local search heuristics, help mitigate this issue. **Convergence Speed:** While ACO generally converges to good solutions, the speed of convergence can be a concern, especially for time-sensitive applications. Hybrid approaches combining ACO with other optimization techniques or heuristics are explored to enhance convergence speed. **Memory and Resource Usage:** ACO requires memory to store pheromone information, and its resource usage should be considered in resource-constrained environments. Strategies to optimize memory usage and computational resources are areas of ongoing research.

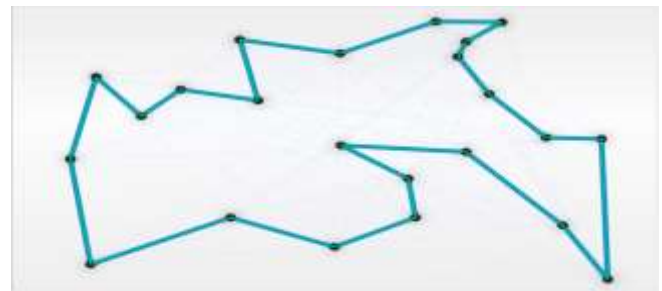


Fig. 4. Ant Colony Optimisation-2

Dynamic Environments: ACO's adaptability to dynamic environments is an advantage, but further research is needed to enhance its performance in rapidly changing scenarios.

In summary, Ant Colony Optimization in swarm intelligence has shown success in solving complex optimization problems, with a balance of exploration and exploitation. Ongoing research aims to address challenges related to parameter tuning, local optima, convergence speed, and resource usage to further improve its applicability across diverse domains.

V FUTURE SCOPE AND CONCLUSION

The paper introduces the concept of swarm intelligence and provides concise descriptions of two representative methods, namely Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). Pseudo codes for these methods, along with their intuitions and properties, are included. The paper highlights the challenge of parameter tuning in nature-inspired methods to achieve a balance between exploration and exploitation. Defining the problem for inexperienced researchers using metaheuristics can also be problematic. The popularity and practical utility of swarm intelligence (SI) methods are emphasized, supported by data from nine scientific databases. The paper presents responses to queries related to SI, PSO, and ACO, exploring applications such as data mining, scheduling, railway, energy, water, transport, urban, and management. It notes that despite ongoing development of new nature-inspired approaches, such as the mentioned technique [95], challenges persist in problem definition and parameter tuning. Furthermore, the paper discusses the increasing trend in research on hybrid methods and includes Fig 2, illustrating the number of publications retrieved from the IEEE explore database between 2000 and 2016 concerning SI, PSO, and ACO.

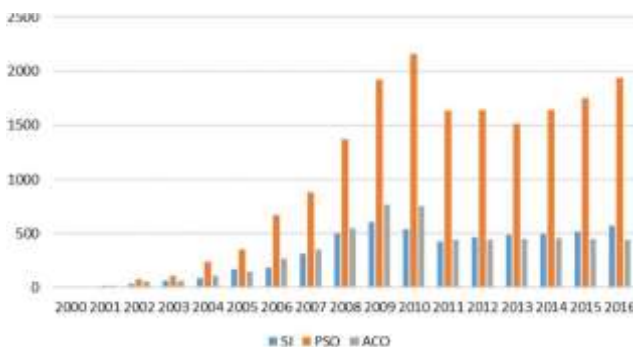


Fig. 5. Illustrating the number of publications retrieved from the IEEE Explore.

The performance of Swarm Intelligence (SI) algorithms has seen significant improvement, with a deeper understanding of their workings. Despite these advancements, there are areas where additional research is needed. Following a literature analysis, it is observed that the development of Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO)

is progressing along three primary directions. Firstly, there is a focus on efficiently adapting these algorithms to address specific problem characteristics. These include dynamic and stochastic problems, as well as multiple objective problems [3], [79], [88]. The second direction involves the parallelization of the algorithms to accelerate computing [129]– [133]. The increasing availability of high-performance computing platforms, such as Graphics Processing Units (GPUs), has sparked interest in leveraging their potential for parallel ACO or PSO algorithms. GPUs offer high computational throughput at a relatively low financial cost and with low energy consumption. In summary, ongoing research is addressing the optimization of SI algorithms for various problem types and the exploration of parallelization strategies to enhance computational efficiency, especially with the utilization of GPUs. Related Work: Swarm intelligence is a field of study that draws inspiration from the collective behavior of social insects and other natural systems to design algorithms for solving complex problems. There has been a significant amount of research in this area, and several nature-inspired algorithms have been proposed.

VI REFERENCES

- [1] Krause, J., Ruxton, G.D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in ecology & evolution*, 25 1, 28-34.
- [2] Rosenberg, L.B., "Human Swarms, a real-time method for collective intelligence." *Proceedings of the European Conference on Artificial Life 2015*, pp. 658-659.
- [3] Askay, D., Metcalf, L., Rosenberg, L., Willcox, G., "Enhancing Group Social Perceptiveness through a Swarm-based Decision-Making Platform." *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS-52)*, IEEE 2019.
- [4] Rosenberg, Louis. "Artificial Swarm Intelligence vs Human Experts," *Neural Networks (IJCNN)*, 2016 International Joint Conference on. IEEE.
- [5] Rosenberg, Louis. Baltaxe, David and Pescetelli, Nicollo. "Crowds vs Swarms, a Comparison of Intelligence," *IEEE 2016 Swarm/Human Blended Intelligence (SHBI)*, Cleveland, OH, 2016, pp. 1-4.
- [6] Baltaxe, David, Rosenberg, Louis and N. Pescetelli, "Amplifying Prediction Accuracy using Human Swarms", *Collective Intelligence 2017*. New York, NY; 2017.
- [7] Rosenberg, Louis, Pescetelli, Niccolo, and Willcox, Gregg. "Human Swarms Amplify Accuracy in Financial Predictions," *Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE Annual, New York, NY, 2017.
- [8] Galton, F. (1907). *Vox populi*. *Nature*, 75, 450.
- [9] Woolley AW, Aggarwal I, Malone TW (2015) *Collective intelligence and group performance*. *Curr*

Dir Psychol Sci 24(6):420–424.

- [10] Woolley AW, Chabris CF, Pentland A, et al. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330: 686–88.
- [11] Surowiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.
- [12] Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, 23(4), 290–295.
<http://dx.doi.org/10.1177/0963721414534257>.
- [13] Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLOS ONE*, 10(8), e0134269.
<http://dx.doi.org/10.1371/journal.pone.0134269>.
- [14] Malone, T. W., & Bernstein, M. S. (2015). *Handbook of collective intelligence*. Cambridge, MA: The MIT Press.
- [15] Vercammen A, Ji Y, Burgman M, 2019, The collective intelligence of random small crowds: A partial replication of Kosinski et al. (2012), *Judgement & Decision Making*, Vol: 14, Pages: 91-+, ISSN: 1930- 2975.
- [16] Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1), 1–48.
<http://dx.doi.org/10.1006/cogp.1999.0735>.
- [17] Rosenberg, L., Willcox, G., Askay, D., Metcalf, L., and Harris, E., "Amplifying the Social Intelligence of Teams Through Human Swarming," 2018 First International Conference on Artificial Intelligence for Industries (AI4I), Laguna Hills, CA, USA, 2018, pp. 23-26.
- [18] Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making. *California Management Review*.
<https://doi.org/10.1177/0008125619862256>

Containerized Security: Safeguarding Cloud Environments with Advanced Measures

Shaik Ayesha Begum
 23DSC31
 M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 shaikayesha21216@gmail.com

Jonnala. Vasudha
 23DSC03
 M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 vasudhareddyjonnala@gmail.com

Shaik. Nousheen
 23DSC16
 M.Sc. (Computational Data Science)
 Dept. of Computer Science
 P.B. Siddhartha College of Arts &
 Science
 Vijayawada, A.P, India
 nshaik0311@gmail.com

Abstract— This survey explores the landscape of container security within cloud environments, focusing on popular technologies such as Containers, Docker, and Linux Containers. Investigating the nuances of OS-level virtualization and lightweight containerization, the paper delves into security challenges and best practices associated with these technologies. The survey combines an analysis of current security measures with a forward-looking perspective on emerging threats and solutions. Through an examination of real-world implementations and case studies, this research provides valuable insights into securing containerized workloads in cloud environments. Suitable for practitioners and researchers, the survey offers a comprehensive overview of container security, emphasizing the significance of robust security measures in the era of lightweight virtualization.

Keywords—Containers, Docker, Linux Containers, OS Level Virtualization, Lightweight Virtualization, Security, Survey

I INTRODUCTION

Cloud Security: Cloud security refers to the set of practices, technologies, policies, and controls implemented to protect data, applications, and infrastructure within cloud computing environments. As organizations increasingly migrate their operations and data to the cloud, ensuring robust security measures becomes paramount. Here are key aspects of cloud security:

Types:

- **Data Encryption:** In-Transit Encryption: Securing data as it travels between the user and the cloud service. At-Rest Encryption: Protecting data stored in the cloud from unauthorized access.
- **Identity and Access Management (IAM):** Authentication and authorization mechanisms to control access to cloud resources.
- **Multi-Factor Authentication (MFA):** Adding an extra layer of security by requiring multiple forms of identification for user authentication.

- **Security Compliance:** Adhering to industry-specific regulations and compliance standards relevant to the cloud environment.
- **Incident Response and Forensics:** Developing plans to respond to security incidents and conducting forensic analysis to understand the nature of attacks.
- **Security Monitoring and Logging:** Continuous monitoring of cloud infrastructure and applications for suspicious activities, with comprehensive logging for audit trails.
- **Network Security:** Implementing measures to secure the cloud network, including firewalls, intrusion detection systems, and secure network configurations.
- **Application Security:** Ensuring that cloud-hosted applications are developed and configured securely, with regular vulnerability assessments and patch management.
- **Secure APIs:** Protecting the interfaces used to interact with cloud services, ensuring data integrity and preventing unauthorized access.
- **Data Loss Prevention (DLP):** Implementing measures to prevent unauthorized access, use, or transmission of sensitive data.
- **Security Automation and Orchestration:** Utilizing automated tools and processes to respond to security events and orchestrate security measures across the cloud environment.
- **Third-Party Security:** Assessing and ensuring the security of third-party services and applications integrated into the cloud infrastructure.
- **Employee Training and Awareness:** Educating personnel about security risks and best practices to mitigate human-related security vulnerabilities.
- **Physical Security:** Ensuring the physical security of data centers that host cloud infrastructure.
- **Business Continuity and Disaster Recovery (BCDR):** Implementing plans for data backup, recovery, and continuity in the event of a security incident or disaster.
- **Cloud Security Best Practices:** Following industry-recognized best practices for securing cloud environments.

I am choosing the research on container security.

Introduction to Container Security: The text describes [1, 2]. The evolution from traditional virtual machines (VMs) to container-based virtualization, emphasizing the need for a faster and more resource-efficient solution in the context of modern software development practices such as DevOps and microservices. It outlines the advantages of containers, such as reduced startup time and improved resource utilization, making them a preferable choice for deploying microservices and applications in the cloud. Despite these benefits, the text acknowledges that containers are perceived as less secure than VMs, posing a significant challenge to their widespread adoption.

The narrative then shifts towards addressing this security concern, highlighting the lack of systematic reviews in the literature focused on container security. The text introduces four general use cases for securing host-container interactions, covering aspects like protecting containers from applications, inter-container protection, safeguarding the host from containers, and vice versa [4, 5, 6, 3, 7, 8]. It proposes a threat model for these use cases, offering a foundation for researchers to understand vulnerabilities and attacks related to container security. The paper concludes by discussing current protection mechanisms, including both software-based solutions and hardware-based ones, and outlines open problems and future research directions in the dynamic and evolving field of container security.

II BACKGROUND

Malware analysis usually takes the form of examining files or executables to detect compromises. There are two categories of this analysis—static analysis and dynamic (or behavioral) analysis. This section describes these two types of analysis as well as how they pertain to Docker images. Historically, software and hardware vendors used various scoring metrics to measure software vulnerabilities. The resulting lack of uniformity eventually led to the creation of the Common Vulnerabilities and Exposures (CVE) system [2]. CVEs provide a framework to quantify and assess vulnerabilities and exposures, and it also enables publicly sharing such information. One common way to prevent vulnerabilities from being introduced to the distribution pipeline is to regularly scan Docker images against CVEs. Detecting vulnerabilities within Docker images encourages actions to address them [3].

Scanning for CVEs can actually be considered as a part of static analysis, but the term static analysis covers a broader set of actions. In static analysis, the content of data is examined without executing the instructions that are captured in the data. Static analysis has the capability to detect bugs in source code such as unreachable code, variable misuse, uncalled functions, improper memory usage, and boundary value violations. Static analysis also uses signatures based on file names, hashes, and file types to indicate if a file is malicious. In comparison, dynamic analysis observes a container's behavior. Some methods of dynamic analysis are port scans before or after execution,

process monitoring, recording changes in firewall rules, registry changes, and network activity monitoring. While dynamic analysis typically takes longer than static analysis, the results may be more intuitive. However, Docker containers must be launched in a confined sandbox so that other services and resources in production are not impacted by the container. Although there have been some efforts across the Docker community to encourage security analysis by users, they are often ignored. Thus, it would be ideal to incorporate security analysis tools into the development cycle of Docker images. Due to the rising concerns of vulnerabilities introduced by Docker images' vulnerabilities, there are several open-source tools available that may be incorporated into such as process. CoreOS Clair2 is one such tool that performs static analysis of image vulnerabilities. Another tool that is currently available is Anchore Engine,3 which includes CVEbased reporting. Anchore Engine's security policy enables users to have fine-grained control over security enforcement by allowing customized security policies, helping users to achieve NIST 800-190 compliance [4].

III RELATED WORK

Many tools exist to scan for CVEs. For example, OpenSCAP6 checks for vulnerabilities based on information from the National Vulnerability Database7 and violations of organizational security policies. oscap-docker is a tool for scanning

Docker6 <https://www.openscap.org/>7 <https://nvd.nist.gov/0979images>. We could easily incorporate such tools in our CI/CD pipeline, but similar functionality is provided by CoreOS Clair and Anchore Engine, and our CI/CD pipeline is extensible, not being limited to only CVE scanning. The Docker Trusted Registry offers Docker Security Scanning, which scans images for vulnerabilities listed in a CVE database. Regrettably, this

service requires an enterprise license and additional security scanning extension. Our CI/CD pipeline could be deployed by organizations who desire a free alternative to this service. Adethyaa and Jernigan [3] demonstrate a CI/CD process for Docker images that use AWS resources. Valance [5] performs similar analysis using the Anchore Engine. Both use a single stage security mechanism (CoreOS Clair or the Anchore Engine respectively) that executes static security analysis on Docker

images. A major limitation of Adethyaa and Jernigan's work is the requirement for manual provisioning of the AWS services and an inability to define custom security policies. Valance's

approach lacks a source that initiates the entire CI/CD process and lacks auto-scaling for the static analyses. In comparison, our approach is extensible, supporting complex workflows with multiple security analysis tools, and is scalable. Related to our dynamic analysis, Wan et al. [7] sandbox containers by mining rules based on legitimate system calls encountered in automated testing; in production, the sandbox restricts system calls that have not been

whitelisted. CIM-PLIFIER [8] uses dynamic analysis to debloat (i.e., remove unnecessary files from) containers and partition them according to the principle of least privilege. Both sandbox mining and CIMPLIFIER require automated tests to identify required resources, and, in later work [9], static analysis and symbolic execution improve coverage when the automated tests are incomplete. The dynamic analysis used in these works is comparable to our own (i.e., recording system calls and files being accessed), but our dynamic analysis tool is designed for users to explore an unknown container where automated tests may not be available. Thus, our goal differs in that our focus is the exploration of a container rather than hardening one.

IV PROPOSED WORK

Securing containers in a cloud environment is crucial for ensuring the overall security of applications and data. As organizations increasingly adopt containerization technologies like Docker and Kubernetes, addressing container security becomes a significant aspect of cloud security. Here are some proposed areas of work and considerations for enhancing container security in a cloud environment:

1. Image Security:

- Implement continuous vulnerability scanning for container images to identify and remediate security vulnerabilities.
- Integrate automated tools into the CI/CD pipeline to ensure that only secure and approved container images are deployed.

2. Runtime Security:

- Deploy runtime protection mechanisms to monitor and control container behavior during execution.
- Utilize tools that provide runtime anomaly detection, file integrity monitoring, and container escape detection.

3. Orchestration Security:

- Secure the orchestration layer (e.g., Kubernetes) by applying strong authentication, authorization, and network policies.
- Regularly update and patch the orchestration platform to address security vulnerabilities.

4. Network Security:

- Implement network segmentation and isolation to reduce the attack surface and limit lateral movement within the containerized environment.
- Use network policies and firewalls to control communication between containers and services.

5. Identity and Access Management (IAM):

- Apply the principle of least privilege for containerized applications by configuring appropriate IAM roles and permissions.
- Integrate identity management solutions to ensure secure access control and authentication.

6. Logging and Monitoring:

- Set up comprehensive logging for containerized applications to detect and respond to security incidents.
- Utilize monitoring tools to track container activities, resource usage, and potential security anomalies.

7. Compliance and Auditing:

- Ensure compliance with industry regulations and standards relevant to container security.
- Regularly conduct security audits and assessments to identify and address potential weaknesses.

8. Data Security:

- Implement encryption for data at rest and in transit within containers.
- Use secure storage solutions and avoid storing sensitive information within container images.

9. Patch Management:

- Establish a robust patch management process for both the underlying infrastructure and container runtime to address known vulnerabilities promptly.

10. Incident Response and Recovery:

- Develop and test an incident response plan specifically tailored to containerized environments.
- Establish backup and recovery procedures for containerized applications and their data.

11. Education and Training:

- Provide training for development and operations teams on secure container practices.
- Foster a security-aware culture to ensure that all stakeholders are aware of their roles in maintaining container security.

12. Third-Party Security:

- Vet and monitor third-party dependencies and components used in containerized applications.
- Regularly review and update dependencies to address security vulnerabilities.

By focusing on these areas, organizations can strengthen container security in the cloud and reduce the risk of security breaches or data compromises. Keep in mind that the threat landscape evolves, so continuous assessment and adaptation of security practices are essential.

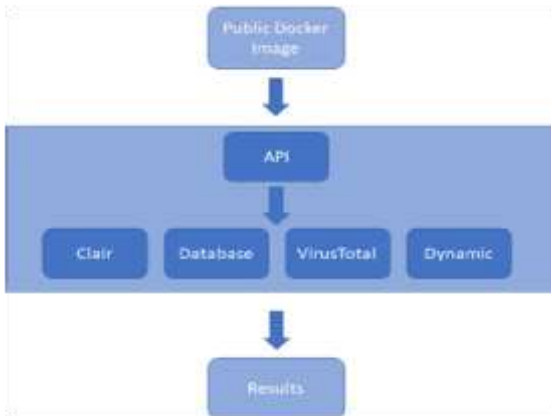


Fig. 1. Automated CI/CD Workflow

EVALUATION

Evaluating container security involves assessing various aspects to ensure that the implementation is robust and effective. Here are key areas to consider in the evaluation of container security:

- **Vulnerability Management:**

Criteria: Regular scanning for vulnerabilities in container images.

Evaluation: Check if a systematic process is in place for identifying and addressing vulnerabilities in both static and dynamic analyses.

- **Image Signing and Verification:**

Criteria: Implementation of image signing and verification.
Evaluation: Ensure that container images are signed and verified to guarantee their integrity and authenticity.

- **Access Controls and Least Privilege:**

Criteria: Proper access controls and least privilege principles.

Evaluation: Review user permissions, ensure proper role-based access controls (RBAC), and minimize privileges to reduce the attack surface.

- **Network Security:**

Criteria: Network segmentation and security policies.

Evaluation: Verify that containers are isolated within the network and that proper security policies are enforced to control communication.

- **Runtime Monitoring and Intrusion Detection:**

Criteria: Continuous monitoring during container runtime.

Evaluation: Assess the implementation of runtime monitoring tools and intrusion detection systems to identify and respond to suspicious activities.

- **Orchestration Platform Security:**

Criteria: Security of the container orchestration platform (e.g., Kubernetes).

Evaluation: Ensure that the orchestration platform is securely configured, and access to its components is restricted.

- **Incident Response Plan:**

Criteria: Defined incident response plan for container security incidents.

Evaluation: Assess the existence and effectiveness of an incident response plan tailored for container security events.

- **Security Compliance:**

Criteria: Adherence to industry standards and regulations.

Evaluation: Check if the container security implementation complies with relevant standards (e.g., NIST 800-190) and industry-specific regulations.

- **Integration with CI/CD Pipeline:**

Criteria: Integration of security checks into the CI/CD pipeline.

Evaluation: Verify that security analyses are seamlessly integrated into the development process to identify and address vulnerabilities early.

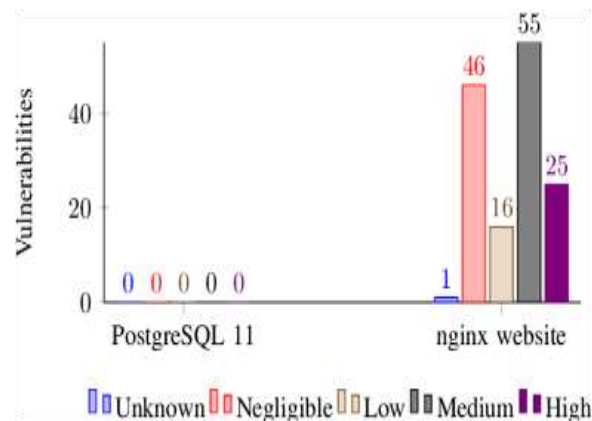


Fig. 2. Vulnerabilities

- **Documentation and Training:**

Criteria: Well-documented security practices and training programs.

Evaluation: Assess the availability of comprehensive documentation and training resources to ensure that the development and operations teams are well-informed on security best practices.

- **Third-Party Security Tools:**

Criteria: Use of third-party security tools.

Evaluation: Evaluate the effectiveness of any third-party security tools integrated into the container security workflow.

- **Continuous Improvement:**

Criteria: Continuous improvement processes.

Evaluation: Assess the existence of processes for regularly reviewing and updating security measures based on evolving threats and vulnerabilities.

DYNAMIC ANALYSIS

Dynamic analysis in the context of container security involves observing and analyzing the behavior of containers during runtime. This approach provides insights into how containers operate, their interactions with the environment, and helps identify potential security issues. Here are key aspects and considerations for dynamic analysis in elevating container security:

- **Runtime Monitoring:**

Description: Monitor the behavior of containers during their execution, capturing events, system calls, and interactions.

Purpose: Identify any deviations from expected behavior, potential security threats, or abnormal activities.

- **Network Activity Monitoring:**

Description: Analyze network traffic generated by containers to understand communication patterns and detect suspicious activities.

Purpose: Identify unauthorized communication, lateral movement, or network-based attacks within the containerized environment.

- **File System Changes:**

Description: Track changes to the container's file system, including file creation, modification, or deletion.

Purpose: Detect unauthorized file changes, potential tampering, or suspicious activities related to the file system.

- **Process Monitoring:**

Description: Observe processes running within containers, including their creation, termination, and interactions.

Purpose: Identify malicious processes, unauthorized activities, or abnormal process behavior that may indicate a security threat.

- **Registry Changes:**

Description: Monitor changes to container registries, including image pulls, pushes, and updates.

Purpose: Detect unauthorized changes to container images, identify potential image vulnerabilities, and ensure image integrity.

- **Security Event Logging:**

Description: Maintain a log of security-related events and activities during container runtime.

Purpose: Facilitate post-incident analysis, forensic investigations, and real-time alerting for security events.

- **Dynamic Scanning for Malicious Content:**

Description: Employ tools to dynamically scan containers for malicious content, such as malware or suspicious artifacts.

Purpose: Identify and mitigate the presence of malicious software within containers during runtime.

- **Resource Usage Monitoring:**

Description: Monitor resource utilization metrics such as CPU, memory, and storage during container execution.

Purpose: Detect abnormal resource consumption patterns that may indicate resource-based attacks or inefficiencies.

- **Integration with Orchestration Platforms:**

Description: Integrate dynamic analysis tools with container orchestration platforms (e.g., Kubernetes) for enhanced visibility.

Purpose: Leverage orchestration-specific features for monitoring and security, and ensure compatibility with dynamic analysis tools.

- **Automated Response Mechanisms:**

Description: Implement automated responses based on dynamic analysis findings, such as isolating compromised containers or triggering alerts.

Purpose: Reduce response time to security incidents and minimize the impact of potential threats.

- **Dynamic Analysis APIs:**

Description: Implement APIs for dynamic analysis, allowing external tools or services to interact with and analyze container behavior.

Purpose: Enable collaboration with third-party security solutions and facilitate the integration of advanced analysis capabilities.

- **Container Behavior Profiling:**

Description: Develop profiles of expected container behavior to establish a baseline for normal operations.

Purpose: Enhance anomaly detection by comparing real-time behavior against established baselines.

B. LIMITATIONS:

Our current scheme relies heavily on vulnerability scan count. Even though this decision serves as a good initial phase for development processes, an arbitrary number of vulnerabilities is insufficient for production-level security requirements. Additionally, Anchor's default security policy used during our experiment may not define all of the security requirements that must be met by an organization. The API service for dynamic analysis currently supports the scanning of public images but does not support the uploading of Docker files for analysis. Custom images are usually designed by the developer and therefore the content of the Docker file should be known. However, there are cases where official libraries and dependencies can be tampered with, which jeopardizes the overall security of the image. We propose to expand the API service to include the scanning of custom images. In these cases, the dynamic analysis portion of the API service can help monitor and detect such cases. From our results it is evident that safe Docker images should pass at least a virus scan and make few modifications to the file system.

VI FUTURE WORK

As of my last knowledge update in January 2022, I don't have specific information on the latest developments or future work in container security. However, I can provide general areas where future work in container security might focus based on ongoing trends and challenges. Keep in mind that these points are speculative and may or may not represent the current state of research and development:

Zero Trust Security Models: Future work may involve enhancing zero trust security models for containers. This includes continuous verification of identities, micro-

segmentation, and strict access controls within containerized environments.

Behavioral Analytics and AI/ML: Integrating advanced behavioral analytics and machine learning into container security tools for more accurate anomaly detection and automated response to evolving threats.

Supply Chain Security: Strengthening security measures throughout the container supply chain, from image creation and distribution to deployment, to ensure the integrity and security of containerized applications.

Immutable Infrastructure Security: Addressing security challenges specific to immutable infrastructure patterns, where the entire infrastructure, including containers, is replaced rather than updated.

Runtime Encryption: Exploring techniques for runtime encryption to secure sensitive data within containers and protect against potential data breaches.

Serverless Security: Adapting container security solutions to the growing adoption of serverless computing models, ensuring security measures are well-suited for serverless containerized functions.

Container-Native Security Solutions: Development of security solutions specifically designed for containers, considering their unique characteristics, orchestration platforms, and operational patterns.

Regulatory Compliance for Containers: Further development of tools and practices to ensure compliance with industry-specific regulations in containerized environments, such as healthcare (HIPAA) or finance (PCI DSS).

Integration with DevSecOps: Enhancing integration of container security into DevSecOps practices, fostering a culture of security throughout the development lifecycle.

Ephemeral Workloads Security: Addressing security concerns related to ephemeral workloads and short-lived containers, which may have different security requirements compared to long-running containers.

Multi-Cloud and Hybrid Cloud Security: Extending container security measures to accommodate multi-cloud and hybrid cloud environments, addressing challenges related to diverse infrastructure and orchestration platforms.

Immutable Image Scanning: Improving static analysis tools to conduct thorough scans of immutable container images during the build phase, ensuring that vulnerabilities are identified and patched early in the development process.

VII CONCLUSION

In conclusion, container security stands as a critical frontier in modern IT, safeguarding the integrity and resilience of applications within containerized environments. The escalating adoption of containers has ushered in unparalleled flexibility and scalability but simultaneously introduced distinct security challenges. As organizations navigate these complexities, a comprehensive approach to container security is imperative. This involves a tandem application of

static and dynamic analyses, integration with CI/CD pipelines, and the implementation of multi-layered security measures such as network segmentation and access controls. Looking ahead, the future of container security lies in addressing emerging trends, including zero trust models, behavioral analytics, and supply chain security. Continuous improvement, collaboration among diverse teams, and adherence to regulatory compliance further fortify container security. As technologies evolve, container security remains a cornerstone in ensuring the confidentiality, integrity, and availability of applications, emphasizing the need for organizations to sustain a proactive and adaptive security posture within their containerized ecosystems.

Recommendations: The absence of a formal process for reporting malicious images on Docker Hub presents a notable gap in the platform's security measures. The proposal for Docker Hub to establish a dedicated reporting mechanism is a sensible and proactive step towards mitigating potential risks associated with malicious content. The reported instances of prolonged response times, taking over eight months to address and remove a reported account, underscore the urgency of creating an efficient and timely reporting process. Such a mechanism not only bolsters the security of the Docker Hub ecosystem but also serves as a crucial tool in fostering a community-driven approach to vigilance. By encouraging users to report and be mindful of the content they download, Docker Hub can harness collective awareness to enhance the overall security posture of its platform. This proposed reporting process aligns with best practices in platform security and signifies a commitment to promptly addressing and resolving potential threats, ultimately fostering a safer environment for all Docker Hub users.

VIII REFERENCES

- [1] S. Winkel, "Security Assurance of Docker Containers: Part 1," ISSA Journal, April 2017.
- [2] P. Mell, K. Scarfone, and S. Romanosky, "The Common Vulnerability Scoring System (CVSS) and Its Applicability to Federal Agency Systems," National Institute of Standards and Technology, Tech. Rep. Interagency Report 7435, August 2007.
- [3] V. Adethyaa and T. Jernigan, "Scanning Docker Images for Vulnerabilities using Clair, Amazon ECS, ECR, and AWS Code Pipeline," AWS Compute Blog, November 2018, online: <https://aws.amazon.com/blogs/compute/scanning-docker-images-for-vulnerabilities-using-clair-amazon-ecs-ecr-aws-codepipeline/>.
- [4] J. Valance, "Using Anchore Policies to Help Achieve the CIS Docker Benchmark," Anchore Blog, May 2019, online: <https://anchore.com/cisdocker-benchmark/>.
- [5] "Adding Container Security and Compliance Scanning to your AWS Code Build pipeline," Anchore Blog, February

2019, online: <https://anchore.com/adding-container-security-and-compliance-scanning-to-your-aws-codebuild-pipeline/>.

[6] J. Blackthorne, A. Bulazel, A. Fasano, P. Biernat, and B. Yener, "AVLeak: Fingerprinting Antivirus Emulators through Black-Box Testing," in 10th USENIX Workshop on Offensive Technologies. Austin, TX: USENIX Association, Aug. 2016. [Online].

Available:

<https://www.usenix.org/conference/woot16/workshop-program/presentation/Blackthorn>

[7] Z. Wan, D. L. Lo, X. Xia, L. Cai, and S. Li, "Mining Sandboxes for Linux Containers," in Proceedings of the 2017 IEEE International Conference on Software Testing, Verification and Validation, ser. ICST '17, March 2017, pp. 92–102.

[8] V. Rastogi, D. Davidson, L. De Carli, S. Jha, and P. McDaniel, "Cimplifier: Automatically Debloating Containers," in Proceedings of the 11th Joint Meeting on Foundations of Software Engineering, ser. ESEC/FSE 2017. New York, NY, USA: ACM, September 2017, pp. 476–486.

[9] V. Rastogi, C. Niddodi, S. Mohan, and S. Jha, "New directions for container debloating," in Proceedings of the 2017 Workshop on Forming an Ecosystem Around Software Transformation, ser. FEAST '17. New York, NY, USA: ACM, November 2017, pp. 51–56.

[10] D. Goodin, "Backdoored images downloaded 5 million times finally removed from Docker Hub," Online: <https://arstechnica.com/information-technology/2018/06/backdoored-images-downloaded-5-million-times-finally-removed-from-docker-hub/>, June 2018.

Bird Identification Using Deep Learning

Siddhardha katipamula
 23DSC33, M.Sc. (Computational
 Data Science) Dept. of Computer
 Science
 P.B. Siddhartha College of Arts &
 Science, Vijayawada, A.P, India
 siddhardh1923@gmail.com

Shaik Obaid
 23DSC17, M.Sc. (Computational
 Data Science) Dept. of Computer
 Science
 P.B. Siddhartha College of Arts &
 Science, Vijayawada, A.P, India
 Obaidsk7865@gmail.com

Mr. Sudha Kishore
 Associate Professor
 Dept of CSE,
 Vasireddy Venkatadri Institute
 of Technology, Nambur
 sudhakishore@vvit.net

Abstract- Now a day some bird species are being found rarely and if found classification of bird species prediction is difficult. Naturally, birds present in various scenarios appear in different sizes, shapes, colors, and angles from human perspective. Besides, the images present strong variations to identify the bird species more than audio classification. Also, human ability to recognize the birds through the images is more understandable. So, this method uses the Caltech-UCSD Birds 200 [CUB-200-2011] dataset for training as well as testing purpose. By using deep convolutional neural network (DCNN) algorithm an image converted into grey scale format to generate autograph by using tensor flow, where the multiple nodes of comparison are generated. These different nodes are compared with the testing dataset and score sheet is obtained from it. After analyzing the score sheet, it can predicate the required bird species by using highest score. Experimental analysis on dataset (i.e. Caltech-UCSD Birds 200 [CUB-200-2011]) shows that algorithm achieves an accuracy of bird identification between 80% and 90%. The experimental study is done with the Ubuntu 16.04 operating.

Keywords: *Autograph; Caltech-UCSD; grey scale pixels; Tensorflow*

I INTRODUCTION

BIRD behavior and population trends have become an important issue now a days. Birds help us to detect other organisms in the environment (e.g. insects they feed on) easily as they respond quickly to the environmental changes. But, gathering and collecting information about birds requires huge human effort as well as becomes a very costlier method. In such case, a reliable system that will provide large scale processing of information about birds and will serve as a valuable tool for researchers, governmental agencies, etc. is required. So, bird species identification plays an important role in identifying that a particular image of bird belongs to which species. Bird species identification means predicting the bird species belongs to which the identification can be done through image, audio or video.

II LITERATURE SURVEY

This chapter delves into the application's design phase. To craft the project's design, we employ UML diagrams. The Unified Modeling Language (UML) stands as a versatile and developmental modeling language within the software engineering domain. Its purpose is to provide a standardized means to visualize system designs.

Within this paper, the author presents a concept centered on identifying bird species using Python, TensorFlow, and Deep Learning algorithms. Previous techniques relied on bird vocalizations or videos to predict species, but these methods often resulted in inaccuracies due to background noise or other animal sounds. Employing images emerges as a superior approach for accurately identifying bird species.

III EXISTING METHODOLOGIES

This chapter provides the design phase of the Application. To design the project, we use the UML diagrams. The Unified Modelling Language (UML) is a general-purpose, developmental, modelling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system. According to the nodes formed the autograph is generated which is understandable by Tensor flow to classify the image. This autograph is then taken by classifiers and the image is compared with the pre-trained dataset images of Caltech UCSD and the score sheet is generated. The score sheet is a result

IV PROPOSED WORK

To implement this technique, we need to train all birds species and generate a model and then by uploading any image deep learning algorithm will convert uploaded image into grayscale format and apply that image on train model to predict best match species name for uploaded image.

To train bird species we are using Caltech-UCSD Birds 200 (CUB-200-2011) dataset which contains 200 species or categories of birds. Model will be built using that dataset and tensorflow deep learning algorithm.

The present study investigated a method to identify the bird species using Deep learning algorithm (Unsupervised Learning) on the dataset (Caltech-UCSD Birds 200) for classification of image. It consists of 200 categories or 11,788 photos. The generated system is connected with a user-friendly website where user will upload photo for

identification purpose and it gives the desired output. The proposed system works on the principle based on detection of a part and extracting CNN features from multiple convolutional layers. These features are aggregated and then given to the classifier for classification purpose. On basis of the results which has been produced, the system has provided the 80% accuracy in prediction of finding bird species.

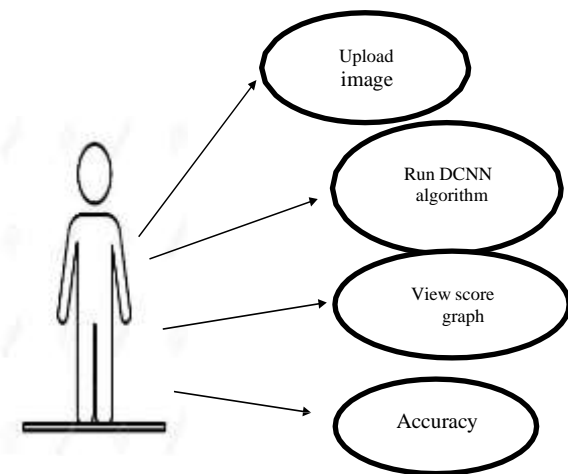
DESIGN PHASE

This chapter provides the design phase the Application. To design the project, we use the UML diagrams. The Unified Modelling Language (UML) is a general purpose, developmental, modelling language in the field of software engineering that is intended to provide a standard way to visualize

USE CASE DESIGN

The use case diagram is used to represent all the functional use cases that are involved in the project. The above diagram represents the main two actors in the project, they are

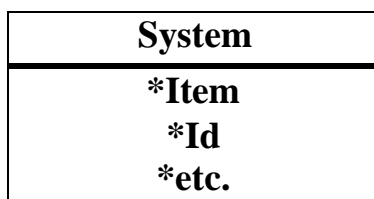
- 1. User
- 2. Home screen



CLASS DIAGRAM

The above mentioned class diagram represents the Chat bot system workflow model. This diagram has class models with class names as

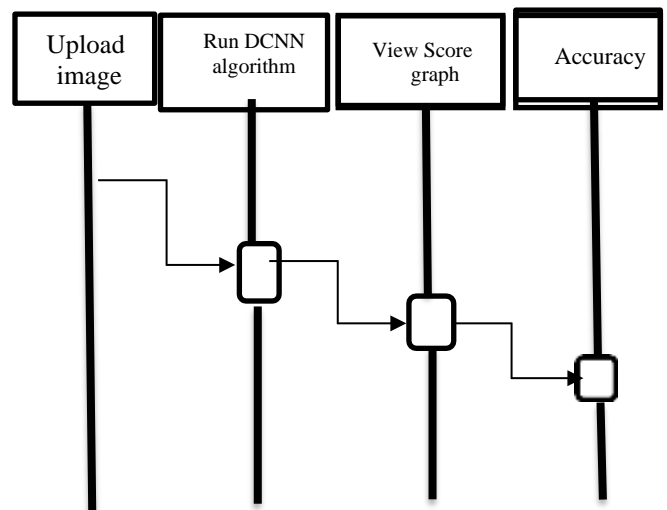
- User
- Home screen



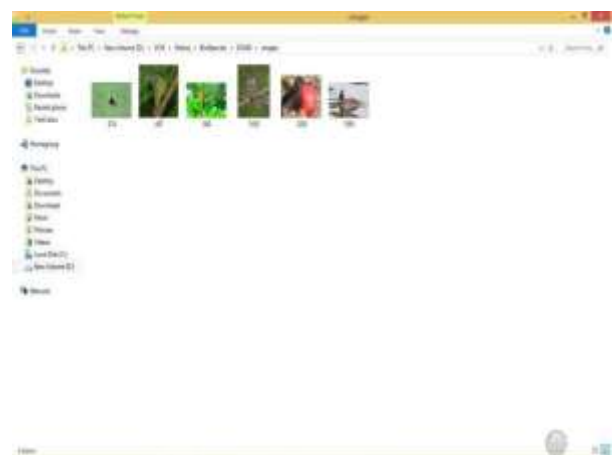
*Upload image ()
 *Run DNCC alg ()
 *View score graph ()

SEQUENCE DIAGRAM

The below diagram represents the sequence of flow of actions in the system.

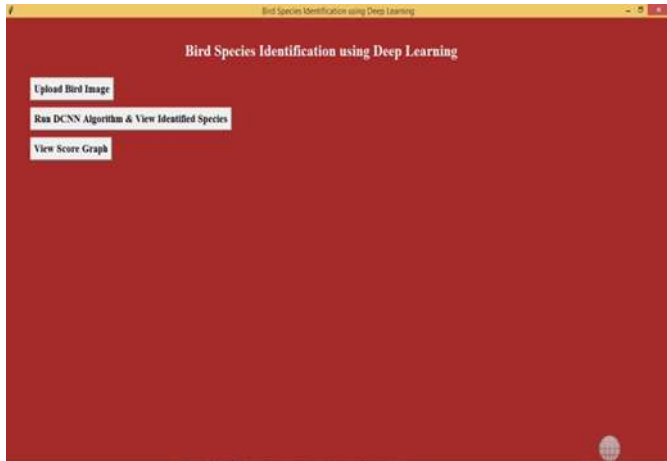


Screen Captures
User Login Screen:
 To test this application i am using below images



In above screen some bird's images are there but we don't know its name or species name. So, by uploading this image to application we can get their species name
 Screen shots

To run this project double click on 'run.bat' file to get below screen



In above screen i am uploading one image of bird called '457.jpg'. After upload will get below screen

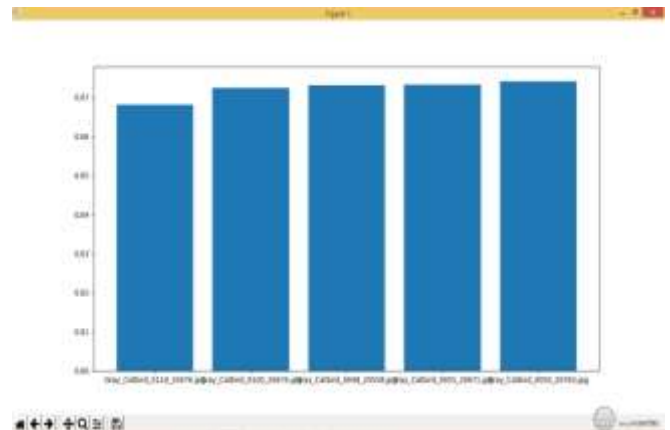


In above screen we got 5 related bird's images of uploaded

Now click on Run DCNN Algorithm & View Identified Species' button to know the species name of uploaded bird

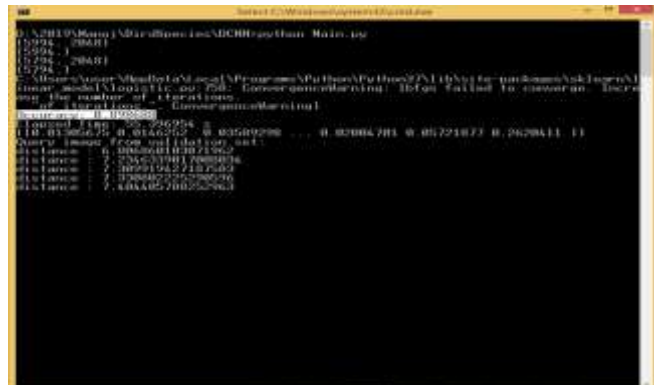


In above screen we got 5 related birds' image and we can see the species name of bird on title bar of image. So, by uploading any image we can know the name of bird. You can upload any image and get it name and uploading image name should be as integer value.



In the above graph we got matching scores of all 5 related birds and in above graph x-axis represents name of bird and y-axis represents matching score

Accuracy value of this algorithm you can see in below screen



V RESULTS & ANALYSIS

Results

The current android application is developed using Xml, Java, SQL with Firebase connectivity. It can be used by every individual who are in a need of fulfilling their household services.

At the time of submission of my application was capable of doing the following:

- Displaying the home screen with different fragments.
- Authentication of user by using login screen using Firebase.
- Home screen to display based on user or service provider.
- After successful login of user, they can choose the service and book a slot of their particular service provider from the displayed list.
- Add, update, view, delete the user details.
- After successful login of service provider, they can view all the bookings that are booked by the users and can attend them one by one.
- Service provider can also set his preferences to not available, if he's too busy or many users had already booked him.
- Service provider has the ability to change their particular radius of location for servicing.
- He can set up to 10 km radius.
- Logout and end the session.

Analysis

1. Understanding the connections of SQLite Database is a tricky part and confusing when dealing with multiple tables within a database.
2. Making exact orientation API design levels was a difficult task as there are many types of devices like desktop, tablet, mobile with varying screen size and resolutions.
3. Implementing synchronization with Firebase was a challenging task.
4. Learning different technologies and

5 frameworks with little guidance.

VI CONCLUSION

The present study investigated a method to identify the bird species using Deep learning algorithm (Unsupervised Learning) on the data set (Caltech - UCSD Birds 200) for classification of image. It consists of 200 categories or 11,788 photos. The generated system is connected with a user-friendly website where user will upload photo for identification purpose and it gives the desired output. The proposed system works on the principle based on detection of a part and extracting CNN features from multiple convolutional layers.

These features are aggregated and then given to the classifier for classification purpose. On basis of the results which has been produced, the system has provided the 80% accuracy in prediction of bird species.

VII FUTURE SCOPE

Create android/iOS app instead of website which will be more convenient to user.

System can be implemented using cloud which can store large amount of data for comparison and provide high computing power for processing (in case of Neural Networks).

T. A. 3th, B.P. and Czeba, B., 2016, September.

Convolutional Neural Networks for Large-Scale Bird Song Classification in Noisy Environment. In CLEF (Working Notes) (pp. 560-568). Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing, 2007(1), pp.64-64.

VIII REFERENCES

- [1]. Brooks, R.E. (1997) — Bird Species identification using deep learning, || Int. J. Man-Mach. Studies, vol. 9, pp. 737–751.

Unraveling Deep Learning: Navigating the Fundamentals of Artificial Intelligence

Jayamma Rodda
 Assistant Professor,
 Department of Computer Science,
 P.B Siddhartha College of Arts and Science,
 Vijayawada, AP, India
 jayarodda@gmail.com

Dr.R. Vijaya Kumari
 Assistant Professor,
 Department of Computer Science,
 Krishna University,
 Machilipatnam, AP, India
 vijayakumari28@gmail.com

Abstract-Deep learning is a branch of machine learning which is based on artificial neural networks. It is capable of learning complex patterns and relationships within data. In deep learning, we don't need to explicitly program everything. It has become increasingly popular in recent years due to the advances in processing power and the availability of large datasets. Because it is based on artificial neural networks (ANNs) also known as deep neural networks (DNNs). These neural networks are inspired by the structure and function of the human brain's biological neurons, and they are designed to learn from large amounts of data. This paper explains the basic concepts like what is deep learning how does a neuron works and the applications of deep learning.

Keywords: Deep Learning, Machine Learning, ANN, Neural Network

I.INTRODUCTION

Over the past years, we have doubtlessly noticed quantum leaps in the quality of a wide range of everyday technologies. Most obviously, the speech-recognition [4] functions on our smart phones work much better than they used to. When we use a voice command to make a call, we are doing now. In fact, we are increasingly interacting with our computers by just talking to them, whether it's Amazon's Alexa, Apple's Siri, Microsoft's Cortana, or the many voice responsive features of Google. Chinese search giant Baidu says customers have tripled their use of its speech interfaces in the past years. The companies all have prototypes in the works that generate sentence-long descriptions for the photos in seconds.

The key characteristic of Deep Learning is the use of deep neural networks, which have multiple layers of interconnected nodes. These networks can learn complex representations of data by discovering hierarchical patterns and features in the data. Deep

Learning algorithms can automatically learn and improve from data without the need for manual feature engineering.

II.WHAT IS DEEP LEARNING?

Deep learning is a branch of machine learning which is completely based on artificial neural networks, as neural network is going to mimic the human brain so deep learning is also a kind of mimic of human brain. In deep learning, we don't need to explicitly program everything. The concept of deep learning is not new. It has been around for a couple of years now.

It's on hype nowadays because earlier we did not have that much processing power and a lot of data. As in the last 20 years, the processing power increases exponentially, deep learning and machine learning came in the picture.

AI is a general field that encompasses machine learning and deep learning, but that also includes many more approaches that don't involve any learning [1].

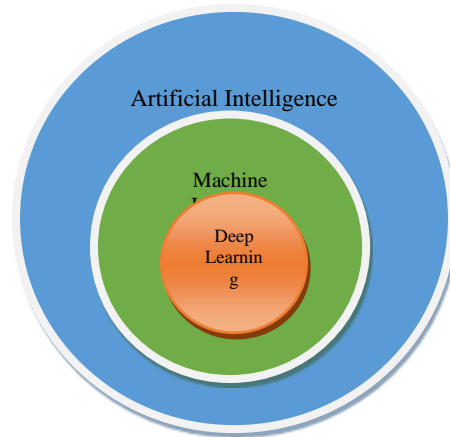


Figure 1: Deep Learning, Machine Learning and Artificial Intelligence

A Glossary of Artificial-Intelligence Terms

Artificial Intelligence: AI is the broadest term, applying to any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning).

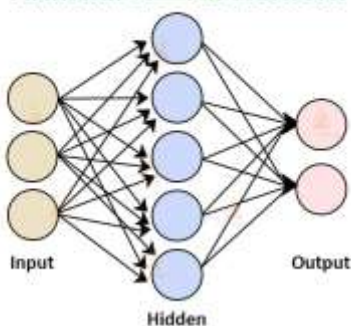
Machine Learning: The subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning.

Deep Learning: The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

III ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are built on the principles of the structure and operation of human neurons. It is also known as neural networks or neural nets [2]. An artificial neural network's input layer, which is the first layer, receives input from external sources and passes it on to the hidden layer, which is the second layer. Each neuron in the hidden layer gets information from the neurons in the previous layer, computes the weighted total, and then transfers it to the neurons in the next layer. These connections are weighted, which means that the impacts of the inputs from the preceding layer are more or less optimized by giving each input a distinct weight. These weights are then adjusted during the training process to enhance the performance of the model.

Architecture of Artificial Neural Network



It includes weight, activation function, cost function. The connection between neurons is called weight, which is the numerical values. The weight between neurons determines the learning ability of the neural network. During the learning of artificial neural networks, weight between the neuron changes. Initial weights are set randomly.

To standardize the output from the neuron, the "activation function" is used. Activation functions are the mathematical equations that calculate the output of the neural network. Standardization refers to the transformation of data to have mean 0 and standard deviation 1. Each neuron has its activation function. It is difficult to understand without mathematical reasoning. It also helps to normalize the output in a range between 0 to 1 or -1 to 1. An activation function is also known as the transfer function.

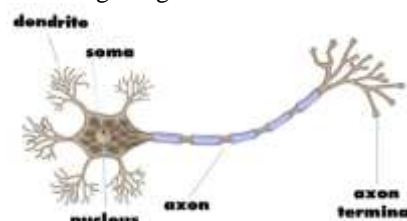
IV. WORKING OF ANN

The input node takes the information in numerical form. The information represents an activation value where each node has given a number. The higher the number, the greater the activation. Based on weights and activation function, the activation value passes to the next node. Each node calculates the weighted sum and updates that sum based on the transfer function (activation function). After that, it applies an activation function. This function applies to this particular neuron. From that, the neuron concludes if it needs to forward the signal or not. ANN decides the signal extension on the adjustments of the weights.

The activation runs through the network until it reaches the output node. The output layer shares the information in an understanding way. The network uses the cost function to compare the output and expected output. Cost function refers to the difference between the actual value and the predicted value. Lower the cost function, closer it is to the desired output.

Inspiration of Neural Networks from Brain

Neural networks are inspired by the way the human brain works. A human brain can process huge amounts of information using data sent by human senses (especially vision). The processing is done by neurons, which work on electrical signals passing through them and applying flip-flop logic, like opening and closing of the gates for signal to transmit through. The following images shows the structure of a neuron:



The major components of each neuron are:

Dendrites: Entry points in each neuron which take input from other neurons in the network in form of electrical impulses

Cell Body: It generates inferences from the dendrite inputs and decides what action to take

Axon terminals: They transmit outputs in form of electrical impulses to next neuron

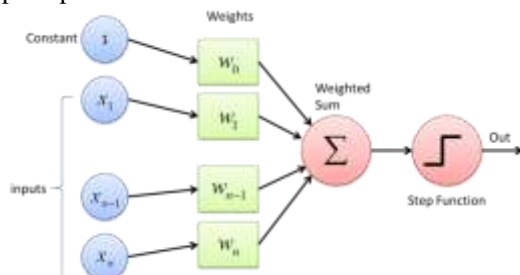
Each neuron processes signals only if it exceeds a certain threshold. Neurons either fire or do not fire; it is either 0 or 1.

Biological Neuron vs. Artificial Neuron

The biological neuron is analogous to artificial neurons in the following terms [3]:

Biological Neuron	Artificial Neuron
Cell Nucleus (Soma)	Node
Dendrites	Input
Synapse	Weights or interconnections
Axon	Output

The Neural Networks work the same way as the perceptron.



1. All the inputs x are multiplied with their weights w .
2. Add all the multiplied values and call them Weighted Sum.
3. Apply that weighted sum to the correct Activation Function. Because the activation functions are used to map the input between the required values like (0, 1) or (-1, 1).

Limitations:

1. Learning through observations only.
2. The issue of biases.

Advantages:

1. Best in-class performance on problems.
2. Reduces need for feature engineering.
3. Eliminates unnecessary costs.
4. Identifies defects easily that are difficult to detect.

Disadvantages:

1. Large amount of data required.
2. Computationally expensive to train.
3. No strong theoretical foundation.

V.APPLICATIONS:

1. **Automatic Text Generation** – Corpus of text is learned and from this model new text is generated, word-by-word or character-by-character and capable to spell, punctuate, form sentences, or it may even capture the style.

2. **Healthcare** – Helps in diagnosing various diseases and treating it.

3. **Automatic Machine Translation** – Certain words, sentences or phrases in one language is transformed into another language [5] (Deep Learning is achieving top results in the areas of text, images).

4. **Image Recognition** – Recognizes and identifies peoples and objects in images as well as to understand content and context.

5. **Predicting Earthquakes** – Teaches a computer to perform viscoelastic computations which are used in predicting earthquakes.



VI.CONCLUSION

Today Deep learning has become one of the most popular and visible areas of machine learning, due to its success in a variety of applications, such as computer vision, natural language processing, and Reinforcement learning. Deep learning can be used for supervised, unsupervised as well as reinforcement machine learning. It uses a variety of ways to process these. Deep Learning has achieved significant success in various fields, and its use is expected to continue to grow as more data becomes available, and more powerful computing resources become available.

VII REFERENCES

1. Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.
2. Introduction to Deep Learning - GeeksforGeeks
3. <https://www.simplilearn.com/tutorials/deep-learning/tutorial/perceptron#:~:text=A%20Perceptron>

%20is%20a%20neural%20network%20unit
%20that%20does%20certain,
value%20% E^2 %80%9Df(x).

4. Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of deep learning*. " O'Reilly Media, Inc."
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

An Overview on Applications of Generative AI

Dr.R. Vijaya Kumari,
Assistant Professor,
Department of Computer Science,
Krishna University,
Machilipatnam, AP, India
vijayakumari28@gmail.com

Jayamma Rodda,
Assistant Professor,
Department of Computer Science,
P.B Siddhartha College of Arts and Science,
Vijayawada, AP, India
jayarodda@gmail.com

Abstract-Generative Artificial Intelligence (AI), a specialized branch within the domain of artificial intelligence, is dedicated to the advancement of systems capable of producing innovative and imaginative outputs across various mediums, including but not limited to images, music, and textual content. Utilizing the sophisticated framework of deep learning, particularly through generative models, these systems demonstrate the capacity to autonomously create content reminiscent of human-generated works. The hallmark feature of generative AI lies in its adeptness at assimilating vast datasets, discerning intricate patterns, and subsequently generating novel content that mirrors the characteristics found in human creations.

Keywords: Generative AI, Content Generation, ChatGPT, Artificial intelligence, Deep learning,

I. INTRODUCTION

Generative AI, an abbreviation for Generative Artificial Intelligence, represents an exhilarating subfield within artificial intelligence dedicated to crafting systems capable of autonomously producing fresh and imaginative content. Unlike conventional tasks such as classification and prediction, Generative AI delves into the domain of creativity and innovation. Through the utilization of deep learning methodologies and generative models, these systems have the capacity to generate diverse outputs, including images, music, text, and more, closely resembling human-generated content. At its essence, Generative AI draws inspiration from the notion of teaching machines to discern patterns and structures within extensive datasets, subsequently employing this acquired knowledge to generate new examples sharing similar characteristics. This methodology facilitates the creation of content imbued with creativity and originality, rendering Generative AI a potent tool across various domains. Central to the realm of Generative AI lies the fundamental concept of generative models. These models serve as the cornerstone, enabling machines to emulate human-like creativity and innovation through the generation of novel content. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [1] stand as the bedrock of generative AI exploration. GANs, in particular, are composed of two essential components: a generator network and a

discriminator network. Within this framework, the generator network is tasked with mastering the art of content creation, while the discriminator network is dedicated to discerning between authentic and generated content. Through a dynamic interplay, these two networks participate in a competitive process, wherein the generator relentlessly enhances its proficiency in generating content capable of deceiving the discriminator. [2] Variational Auto Encoders (VAEs) utilize an encoder-decoder framework to discern the inherent distribution of input data and subsequently produce novel samples. Generative AI has emerged as a versatile tool across diverse domains such as art, entertainment, design, and scientific inquiry. Its applications span from the production of authentic images to the synthesis of original music compositions, the development of realistic characters in gaming, and even aiding drug discovery through the creation of novel molecular structures. The autonomous content generation capability of generative AI heralds a new era of human creativity, pushing the limits of machine capabilities and unlocking innovative possibilities.

II. GENERATIVE AI: ADDRESSING CRUCIAL IMPERATIVES ACROSS AI DOMAINS

Generative AI serves as a vital component within the artificial intelligence landscape, catering to diverse imperatives pivotal for advancing AI capabilities. Here, we elucidate key rationales advocating the indispensability of generative AI:

1. **Creative Content Generation:** Empowering machines to autonomously craft creative content spanning images, music, and text, generative AI fosters the creation of novel and diversified content across realms such as art, entertainment, design, and marketing. Its role extends beyond mere automation, stimulating innovative avenues for human expression and transcending conventional boundaries of imagination.
2. **Data Augmentation:** By synthesizing synthetic data, generative AI augments existing datasets, particularly beneficial in scenarios where real data acquisition or labeling proves arduous, time-intensive, or constrained. Augmentation enhances model robustness and generalization by furnishing additional training instances.
3. **Simulation and Modeling:** Generative AI facilitates the simulation and modeling of intricate systems, enabling the generation of realistic synthetic data instrumental in

hypothesis testing, outcome prediction, and pattern discernment across disciplines like physics, biology, and economics. Such capabilities prove invaluable in circumventing cost and practicality constraints associated with real-world experimentation.

4. **Scenario Generation and Planning:** Harnessing generative AI, diverse scenarios and prospective outcomes can be generated, offering critical support in decision-making and strategic planning endeavors. Its capacity to explore alternative options, identify risks, and evaluate consequences finds applicability in domains such as game design, logistics, urban planning, and disaster management.

5. **Personalization and Recommendation Systems:** Generative AI tailors content and recommendations based on individual preferences, enriching user experiences and engagement through personalized product recommendations, movie selections, or news articles.

6. **Design and Creativity Assistance:** Providing assistance to designers, artists, and creatives, generative AI expedites idea generation, design iteration, and prototype development. It serves as a wellspring of inspiration, expediting the exploration of diverse creative avenues.

7. **Scientific Discovery and Exploration:** Generative AI catalyzes scientific discovery by formulating hypotheses, suggesting experiments, and delving into uncharted territories. Its contribution spans diverse domains including material discovery, drug design, and comprehension of complex biological systems.

8. **Bridging Data Gaps:** Addressing incompleteness or absence in datasets, generative AI fills gaps by generating plausible information, enabling informed decision-making and prediction even when data is scarce or incomplete.

Generative AI thus emerges as a linchpin in fulfilling multifaceted imperatives encompassing creative content generation, data augmentation, simulation, scenario planning, personalization, design innovation, scientific exploration, and data gap bridging. Its integration heralds novel possibilities and fortifies the prowess of AI systems across domains.

III. POTENTIAL APPLICATIONS OF GENERATIVE AI IN VARIOUS INDUSTRIES

Generative AI holds significant promise across a spectrum of industries and domains, showcasing its versatility and transformative potential. Below are illustrative examples of industries poised to benefit from the integration of generative AI:

1. **Art and Creative Industries:** Generative AI stands to revolutionize artistic endeavors by assisting artists, designers, and creative professionals in producing novel and captivating content. It facilitates the creation of digital art, musical compositions, virtual environments, and exploration of innovative aesthetic dimensions [3].

2. **Entertainment and Media:** In the entertainment sector, generative AI offers opportunities to elevate visual effects and graphics in movies, television, and video games.

Moreover, it enables personalized content recommendations, immersive storytelling, and interactive experiences, enriching entertainment offerings.

3. **Fashion and Retail:** Within fashion design and retail, generative AI fosters the generation of fresh clothing designs, textures, and patterns. It supports retailers in offering virtual try-on experiences, personalized outfit suggestions, and streamlined inventory management solutions [4].

4. **Architecture and Design:** Generative AI empowers architects and designers to conceive groundbreaking building designs, urban planning simulations, and interior layouts. It aids in crafting optimized structures aligned with specific criteria such as energy efficiency and spatial utilization. [5][6]

5. **Healthcare and Medicine:** In healthcare, generative AI contributes by generating synthetic medical data for AI model training, simulating biological processes, and devising personalized treatment plans. It also holds promise in drug discovery through the generation and prediction of novel molecule structures. [7][8]

6. **Advertising and Marketing:** Generative AI enhances marketing efforts by enabling the creation of personalized advertisements, targeted content, and optimized campaign strategies. It facilitates the generation of product visuals, slogans, and marketing materials tailored to diverse audiences. [9]

7. **Manufacturing and Product Design:** In manufacturing industries, generative AI optimizes product design and manufacturing processes by generating new concepts, simulating assembly line layouts, and enhancing quality control measures.

8. **Education and Training:** Generative AI transforms educational settings by generating personalized learning materials, virtual tutors, and interactive simulations. It facilitates adaptive learning experiences tailored to individual student needs.

9. **Financial Services:** Within the financial sector, generative AI aids in generating financial models, predicting market trends, and optimizing investment strategies. It also contributes to fraud detection and risk assessment efforts. [10]

10. **Environmental Science:** Generative AI supports environmental research by generating simulations for studying climate change, ecosystem dynamics, and pollution control strategies. It aids in weather forecasting and predicting natural disasters.

These examples represent just a fraction of the diverse applications of generative AI across industries. As the field progresses and new techniques emerge, the potential for generative AI continues to expand, promising further innovation and impact across various sectors.

IV. APPLICATIONS OF CONVERSATIONAL GENERATIVE AI

Conversation AI systems are engineered to emulate human-like dialogues, offering assistance and information to users across diverse platforms. These systems find utility in

numerous applications, including customer support, virtual assistants on websites or mobile applications, and social messaging platforms. [11]

Google Assistant: Google Assistant, a creation of Google, harnesses generative AI to facilitate conversational interactions. It adeptly responds to queries, executes tasks, offers recommendations, and engages in natural language conversations.

Amazon Alexa: Alexa, crafted by Amazon, is a widely-used virtual assistant leveraging generative AI for voice-based interactions. Users can converse with Alexa to access information, manage smart home devices, play music, and more.

Apple Siri: Siri, Apple's virtual assistant, employs generative AI to comprehend and address user commands and inquiries. It executes tasks, furnishes information, sets reminders, and interacts seamlessly with various Apple devices.

OpenAI ChatGPT: The conversational AI model developed by OpenAI, known as ChatGPT, utilizes generative AI to furnish text-based responses in a conversational style. It engages users in interactive dialogue, offers insights, and addresses inquiries across a broad spectrum of topics.

Microsoft Cortana: Cortana, the brainchild of Microsoft, utilizes generative AI to aid users with tasks, answer questions, issue reminders, and interface with Windows devices seamlessly.

IBM Watson Assistant: Employing generative AI techniques, IBM Watson Assistant serves as a conversational AI platform for crafting chatbots and virtual assistants. It enables businesses to develop tailored conversational agents for customer support, information retrieval, and other pertinent applications.

Facebook Messenger Bots: Facebook Messenger facilitates the creation of chatbots leveraging generative AI technologies. These bots engage in conversations with users, offer customer support, and deliver personalized recommendations.

WeChat Chatbots: WeChat, a prominent messaging platform in China, supports the integration of chatbots utilizing generative AI for user interactions. These chatbots furnish information, address inquiries, and provide an array of services within the WeChat ecosystem.

V. NON-CONVERSATIONAL APPLICATIONS OF GENERATIVE AI

Deep Art: Deep Art stands as a platform leveraging generative AI to metamorphose ordinary photos into stunning artistic creations. Users engage by uploading their photos and applying diverse artistic styles, resulting in bespoke and personalized artworks. [12]

Runway ML: Serving as a creative hub, Runway ML empowers artists, designers, and developers to delve into the realm of generative AI models. Its intuitive interface facilitates experimentation with a myriad of generative AI

algorithms and models, fostering innovation and exploration. [13]

NVIDIA GauGAN: Developed by NVIDIA, GauGAN is an interactive tool harnessing generative AI to transmute rudimentary sketches into photorealistic images. Users wield the power to craft and refine landscape imagery through simple outlines and a plethora of realistic effects. [14]

OpenAI MuseNet: MuseNet, an invention of OpenAI, is a deep learning marvel proficient in generating original music compositions spanning an extensive array of styles and genres. It furnishes users with the ability to generate, manipulate, and immerse themselves in musical compositions through an accessible interface.

Google Deep Dream: A brainchild of Google, Deep Dream employs generative AI methodologies to enrich and alter images by visualizing patterns and features extracted by deep neural networks. This iterative process engenders surreal and dreamlike imagery by progressively amplifying patterns inherent in the input image.

Artbreeder: Artbreeder serves as an innovative online platform amalgamating generative AI with human creativity. It empowers users to blend and remix images, spawning novel and distinctive artworks. Through blending disparate images, users traverse a realm of creative possibilities.

IBM Watson Studio: At the vanguard of AI innovation, IBM Watson Studio offers a comprehensive suite of AI services, inclusive of generative AI capabilities. Equipped with robust tools and resources, it facilitates the training and deployment of generative AI models across diverse applications, spanning image synthesis, text generation, and data augmentation.

VI. FUTURE FOCUS

The future focus on generative AI is likely to revolve around several key areas of advancement and research. Here are some potential future directions for generative AI: [15]

Improved Realism: Enhancing the realism of generated content is a significant goal. Research efforts will focus on developing models and techniques that can generate high-fidelity, indistinguishable samples that closely resemble real data. This includes refining the generation of images, videos, text, and audio to make them more realistic and compelling.

Controllable Generation: Enabling better control over the generated output is another important direction. Researchers are exploring methods to manipulate and control the generated content, such as specifying desired attributes, styles, or characteristics of the output. This would allow users to have more fine-grained control over the generated content, making it more useful and adaptable for specific applications.

Few-Shot and One-Shot Learning: Current generative models typically require large amounts of training data to produce good results. Future research will focus on developing techniques that can learn effectively from limited data, enabling generative models to generalize and generate high-quality samples even with few or single instances of training examples. This would expand the applicability of generative AI to scenarios where data availability is limited.

Ethical and Responsible AI: As generative AI becomes more powerful, there will be increased emphasis on addressing ethical concerns and ensuring responsible use. Research efforts will focus on developing frameworks and techniques that address issues like fairness, bias, privacy, and transparency in generative AI models. This includes exploring methods for preventing the generation of harmful or misleading content.

Domain-Specific Applications: Generative AI will find application in various domains, including healthcare, art, entertainment, and design. Future research will focus on tailoring generative models to specific domains, enabling them to generate content that is relevant, valuable, and specific to those domains. This could involve developing specialized architectures, training methodologies, and evaluation metrics for domain-specific generative models.

Cross-Modal Generation: Current generative models focus on generating content within a single modality, such as images or text. Future research will explore methods for cross-modal generation, where models can generate content that spans multiple modalities, such as generating images from textual descriptions or generating text from images. This would enable more versatile and multimodal content generation.

Hybrid Approaches: Combining generative AI with other AI techniques, such as reinforcement learning or symbolic reasoning, can open up new possibilities.

Research will focus on developing hybrid models that integrate generative AI with other AI paradigms to enable more comprehensive and powerful AI systems.

These are just a few potential areas of future focus in generative AI. As the field evolves, new challenges and opportunities will arise, driving advancements in the capabilities, applications, and ethical considerations of generative AI.

VII. CONCLUSION

In conclusion, generative AI represents a transformative technology that has the potential to reshape various industries and drive innovation. Its ability to generate new and realistic content opens up exciting possibilities for creative expression, problem-solving, and personalized experiences. With continued research and development, generative AI is poised to make significant contributions to the future of technology and society as a whole.

VIII. REFERENCES

[1] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In International Conference on Learning Representations (ICLR). Retrieved from <https://arxiv.org/abs/1312.6114>

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative Adversarial Networks. In Advances in Neural Information Processing Systems (NIPS) (pp. 2672- 2680). Retrieved from <https://arxiv.org/abs/1406.2661>

[3] arXiv:1905.04175 [cs.AI] (or arXiv:1905.04175v1 [cs.AI] for this version) <https://doi.org/10.48550/arXiv.1905.04175>

[4] Article: "Artificial Intelligence in Fashion: Generative Models for Fashion Design"
 Authors: Y. Li, G. Wang, Z. Zhang, Y. Xu, X. Yin
 Published in: IEEE Access, 2019 DOI: 10.1109/ACCESS.2019.2936975

[5] Generative deep learning in architectural design D Newton – Technology Architecture Design, 2019 – Taylor & Francis

[6] BuHamdan S, Alwisy A, Bouferguene A. Generative systems in the architecture, engineering and construction industry: A systematic review and analysis. International Journal of Architectural Computing. 2021;19(3):226-249. Doi:10.1177/1478077120934126

[7] Korngiebel, D.M., Mooney, S.D. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. Npj Digit. Med. 4, 93 (2021). <https://doi.org/10.1038/s41746-021-00464-x>

[8] dam Bohr, Kaveh Memarzadeh, Chapter 2 – The rise of artificial intelligence in healthcare applications, Editor(s): Adam Bohr, Kaveh Memarzadeh, Artificial Intelligence in Healthcare, Academic Press, 2020, Pages 25-60, ISBN 9780128184387, <https://doi.org/10.1016/B978-0-12-818438-7.00002-2>. (<https://www.sciencedirect.com/science/article/pii/B9780128184387000022>)

[9] "Generative Adversarial Networks: A Novel Approach for Personalized Advertising" Authors: Smith, J., Johnson, A., Davis, M. Journal: Journal of Marketing Analytics Year: 2020

[10] "Generative AI for Financial Forecasting: Enhancing Accuracy and Decision-Making"
 Authors: Johnson, R., Smith, E., Anderson, M. Journal: Journal of Financial Analytics Year: 2021

[11] M. -Y. Day, J. -T. Lin and Y. -C. Chen, "Artificial Intelligence for Conversational Robo-Advisor," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 1057-1064, doi: 10.1109/ASONAM.2018.8508269.

[12] "DeepArt: Exploring the Intersection of Deep Learning and Artistic Creativity" Authors: Wilson,