Volume 10, Issue 2 (XI)
April - June 2023
ISSN: 2394 – 7780

International Journal of Advance and Innovative Research

Indian Academicians and Researchers Association
www.iaraedu.com

International Conference
on
Emerging Trends in Digital Technologies-2023
(ICETDT- 2023)

21st February 2023

organized by

SVKM's
UPG
USHA PRAVIN GANDHI COLLEGE OF ARTS, SCIENCE AND COMMERCE

SVKM's
Usha Pravin Gandhi College of
Arts, Science and Commerce (NAAC: A+ Grade)

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Publication Partner

Indian Academicians and Researcher's Association

International Conference on Emerging Trends in Digital Technologies-2023

(ICETDT-2023)

Advisory Committee

Dr. Anju Kapoor
Principal, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Maharashtra

Dr. Shastri L. Nimmagadda
Reseracher, Curtin University, Perth WA, Austalia

Dr. Sangeeta Chakrabarty
Associate Professor, S.S. Dempo College of Commerce and Economics, Goa

Ms. Sangeeta Nimkar
Technical Specialist, United Postal Services, New Jersey, USA

Dr. Pallavi Jamsandekar
Director, Bharati Vidhyapeeth (Deemed to be University), Sangli, Maharashtra

Dr. Abhijit N. Banubakode
Pricipal, MET Institute of Computer Science, Mumbai

Dr. Chhaya S. Gosavi
Associate Professor, MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra

Dr. Neel Mani
Associate Professor, Amity Institute of Information Technology, Noida, UP

Dr. Sukhada
Assistant Professor, IIT BHU, Varanasi

Dr. Snehalata Shirude
Assistant Professor, NMU, Jalgaon, Maharashtra

Convener

Dr. Manisha Divate
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Co- Convener

Dr. Neelam Naik
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Organizing Committee

Ms. Smruti Nanavaty
Vice Principal, IQAC Coordinator, UPG College of Arts, Science and Commerce

Dr. Swapnali Lotlikar
Assistant Professor, UPG College of Arts, Science and Commerce

Mr. Prashant Chaudhary
Assistant Professor, UPG College of Arts, Science and Commerce

Ms. Sunita Gupta
Assistant Professor, UPG College of Arts, Science and Commerce

Mr. Rajesh Maurya
Assistant Professor, UPG College of Arts, Science and Commerce

Ms. Neha Vora
Assistant Professor, UPG College of Arts, Science and Commerce

President's Message

Amrish R. Patel

President,

Shree Vile Parle Kelvani Mandal

Vile Parle, Mumbai

Higher Education is an imperative milestone for learners in current times. Indian higher education system is maturing towards its presence in Global Higher Education space. This calls for reformative policy initiatives from stakeholders in curricula, pedagogy, use of technology, partnerships, governance and funding. Encompassing this vision, Usha Pravin Gandhi College has a learner centered paradigm of education where the student is placed in a competitive learning environment of the 21st century to foster excellence, equity and quality.

The academic staff at the college constantly commit themselves towards the growth of students to create the desired intellectual, economic and social value.

With firm faith in the saying, "*Vidhyadhanam Sarvadhanam pradhanam"* – knowledge is the only real wealth in this world, I welcome the students of SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce. I am confident that efforts to excel in the field of higher education through innovative practices, immersed and engaged learning and the inculcation of moral and social values in the learners will continue. May you make the best of these openings to shape your careers and future.

Wishing everyone at Usha Pravin Gandhi College of Arts, Science and Commerce all success this new Academic year 2022-23.

Amrish R. Patel

Principal's Message

Dr. Anju Kapoor

Principal,

Usha Pravin Gandhi College

of Art, Science and Commerce (NAAC-A+)

Vile Parle, Mumbai

I am delighted to once again host the International conference on "Emerging trends in Digital Technologies-2023 (ICETDT-2023) on 21st February at SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, that has received an accreditation of A+ by NAAC in October 2022. The College comprehends that technological change has the potential to create shared prosperity and smart solutions to the world's biggest challenges. At the same time, its exponential pace overwhelms existing institutions, leaving us as individuals exposed to uncontrolled risks. Some of these risks include geopolitical crises, rising polarization and a looming climate crisis, topics of concern that have predominated several discussion forums in the past decade.

There is an urgent need for leaders to seize the opportunity to direct technology towards positive ends. As the late French cultural theorist Paul Virilio observed, when things work in new ways they also break in new ways. When we invented the ship, we also invented the shipwreck. The same holds true for new technologies – and the stakes are high.

This should not make us afraid of technological progress, but it should make us humble and mindful of progress being hard-earned and easily lost. Advancing technological change and rolling out new technologies at scale to meet the world's most pressing challenges is the great imperative of our day and age. If we meet that challenge, the outcome will be more rewarding, prosperous and resilient economies and societies.

The theme of the conference is very relevant in light of the present scenario especially keeping in mind that attention has to be directed to the millennials. The teachers have to be innovative presenters who display extensive research skills that will inspire their students to take ownership while they mitigate the potential dangers of these emerging trends in the new age digital world. I applaud all the organizers and researchers who are associated with this conference and wish them great success.

Dr. Anju Kapoor

International Journal of Advance and Innovative Research

Volume 10, Issue 2 (XI): April - June 2023

CERTIFICATE OF INDEXING (SJIF 2022)

This certificate is awarded to

International Journal of Advance & Innovative Research
(ISSN: 2394-7780)

The Journal has been positively evaluated in the SJIF Journals Master List evaluation process
SJIF 2018 = 7.46

SJIF (A division of InnoSpace)

SJIFactor Project Manager
International Advisory Services
INNOSPACE INTERNATIONAL

SJIFactor Project

International Journal of Advance and Innovative Research

Volume 10, Issue 2 (XI): April - June 2023

CONTENTS

Research Papers

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

WEB SERVICES INTEGRATION – E-LEARNING APPS COLLABORATION AND APPLICATIONS

K. Jayashree[1] and V. Gopi Swaminathan[2]
[1]Assistant Professor, Tagore Arts & Science College, Puducherry UT,
[2]Scientist 'F' & Additional State Informatics Officer, National Informatics center, Puducherry UT, India

ABSTRACT
The e-learning has evolved from computer-based to internet-based learning where the learners get instantaneous access to many resources through various systems as well as system-based instructions. The database operating these systems provide information by connecting through various external systems which are enabled through web services in the background. These databases contain multiple functionalities of information which the student has to learn, together with a number of preference parameters which enables the system to be individualized according to the preferences of each student. With this paradigm shift, two-way communication between the learner and the computer are automatically enabled. It helps whether the student achieved their learning objectives on a satisfactory level. If not, then the processes can be repeated until the student has achieved their desired learning goals. Additionally, educational institutions use computer-managed learning systems for storing and retrieving information about the students, materials, evaluation and exchange of these information among others. To enable such an online learning with multiple functionalities from multiple sources, the systems should be able to communicate in a reliable and secured mechanism. The web services help the e-learning system to connect and share the resources in a secured and standardized manner. The web services help to communicate between multiple sources without the need of sharing the IT infrastructure as well as any sensitive data. Web service, simply invoked remotely using a programmatic interface across the network. The major advantage is web services allow standardized industry protocol for the communication. The standardization helps in the quality as well as reduction of costs. The web application develops User Interface meant for humans to read and understand whereas the web services are meant for computers to read. In order to provide secured web services for web applications, it is essential to address the security issues viz., confidentiality, authentication and network security. In the current scenario, particularly in the post-pandemic era, the learning is more dependent electronically conducted over the net. The students can access their learning resources at any time and from any location. The E-learning concept has been extended in online courses. Web services provide instantaneous access to many external resources. E-learning is also considered an economical and efficient alternative to traditional classrooms and education since it eliminates geographical limits. The paper explains on how the web service integration helps e-Learning applications to collaborate and provide various facilities to students

Keywords: e-learning, web services, portal, xml, multiple data sources

INTRODUCTION
Web services are basically services which are used to accept the request from one data source and gets responses from other data sources digitally. The significant feature of web service is it is independent of the language, platform and location. Extensible Markup Language (XML) are used by the emerging trends of applications whose modules can reside in a single server or servers span around the globe. There is a paradigm shift in distributed computing. Initially, the client server networks dominated and now are giving way to a new kind of network built around the Web using open protocols and a new data representation language called XML With this trend, the organizations focus on the data and its role in providing relationships instead of networks and transport using the internet.

Web provides information through browsers for retrieving and displaying information through HTTP protocol and the file format known as HTML. In mid-1990 when XML emerged, the web started the services as distributed enterprise computing. The major framework like .NET is one such XML-based web solutions and services that helps to build web enabled applications linking various databases from multiple sources. It has changed the development of building enterprise software. The other major companies like Sun and its Java also use XML and Web Services.

E-Learning requires a reliable mechanism of rendering the data which are distributed in different sources for providing necessary training materials to the students as well as enabling any intermediaries for analysis and evaluation. Web Services are able to deliver such mechanisms in an automated way which are auto-enabled and auto-configured.

ISSN 2394 - 7780

XML – Breaking Barriers in Distributed Computing

XML, the Extensible Markup Language, helps to define the data in plain text instead of proprietary binary representation. The simplicity of XML is the definition of sets of rules and guidelines in simple texts. The language is based on the insertion of tags to help describe data. But, the language is actually more than just tags. XML has helped in linking the software architecture in a loosely coupled open Web space to build distributed computing. As it is independent of programming language, operating system or protocols, the data description is adopted easily among the heterogeneous systems. This led to data exchange without the constraints imposed by tightly coupled transport dependent architectures.

SOAP, the Simple Object Access Protocol, used by XML with a set of XML tags for moving XML data around the Web using standard Web protocols. This feature was earlier carried out by client-server computing over a decade. Due to the extensible used of messaging servers and software, the web services are emerged for moving data freely across the web.

- **Web Services**

Web services anybody to create electronic services that can be used by anyone else, thus enabling electronic anybody-to-anybody communications, coordination, and interaction. This feature is essentially required for e-learning. In e-learning, the content management will play a critical role in organizing the course to keep the resources conversational as well as professional. The text, image, and other multimedia blocks require to focus on the content for interaction as well as plug the task based online scenarios. It helps the e-learning content manager to easily assemble the component building blocks to the next level. Webservices help to interoperate from different sources to complete the functional module of content management. Web services provide necessary mechanisms to modules to ensure the communication and also use it. Necessary restrictions are also imposed for secured interaction with the industry standards like XML HTTP etc. E-learning systems can be developed with multiple data sources where SOAP defines an envelope to carry XML data for easy remote procedure calls.

Vulnerability Assessment of Web services

Web based E-learning systems use web services that can be described, published, located, and invoked over the Web. At the same time, the threat of XML-based attacks has significantly increased. Nevertheless, Single sign-on systems using cloud-based e-mail accounts are also attackable and many e-learning systems uses cloud-based email accounts. Therefore, cybercrime occurs in various ways and it is essential for implementing web services to undergo vulnerability assessment. Like Web sites, web services also use common technologies with respect to the programming languages of the application. When a e-learning portal use applications and use both web sites and services for data stores and application servers, the front end both typically use a Web server which in turn uses the HTTP to expose the data. As both use similar technological and architectural similarities, the Web services are subjected to many common Web site security threats.

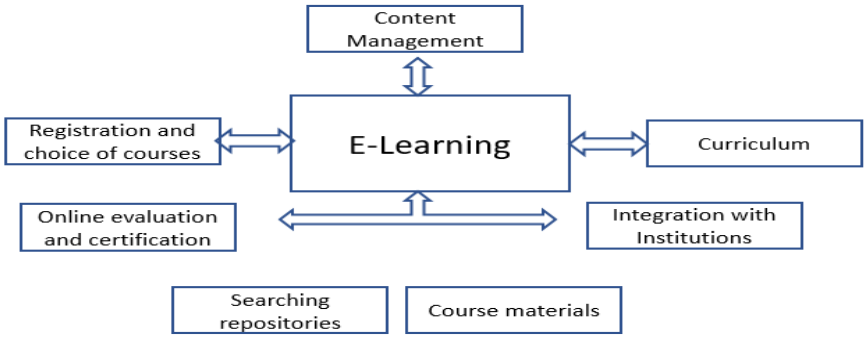


Fig: 1 E-learning framework

The e-learning framework has been represented in the above diagram which contains multiple components. The curriculum, repositories, assessments, personalization may require interoperable access to multiple services but if denial of service attack is targeted, it may affect the entire e-learning system. Therefore, E-learning systems have to be developed with necessary security to mitigate such vulnerabilities.

The database queries which result in large responses, this arises a situation where large payloads are induces. For example, an attacker can submit a query request for a search which is provided to retrieve the available

ISSN 2394 - 7780

items containing a common keyword in the database. If the query interface allows for it, the attacker might send a request to return all the available items in a large data base, which would consume resources and enable a DoS attack against the Web service. There are possibilities for cyber-attacks that can be performed against stateful Web services by abusing the exposed operations to queue a large number of events or store a large number of items in the current session. This is executed by looping around certain operations then requesting the result from other operations in order to generate large responses. Such a vulnerable requests and responses will exhaust the service. Sometimes, web services that allows attachments of files can be used as a vehicle for injecting a virus or malicious content to another system. The attachments that are uploaded to a Web service can then be processed by other applications that have vulnerabilities. If the attachment is an executable, or a package that contains an executable, then it is possible to infect the server that executes the file. Therefore, web services draw attention to the developers for providing necessary security features to mitigate any risks.

Web services – Use Case of e-learning apps

Anywhere, any-time instruction delivered over the internet is called E-learning. It is also called web-based training. This can be extended to a corporate intranet and browser-quipped learners. During the pandemic, one of the imminent challenges posing the challenge is adapting to e-learning. But, E-learning already has proven solutions to offer. The recent advancements in technology have created a revolution in the e-Learning industry. The creation of a new e-Learning experience which was a dream a decade ago has become possible with the introduction of new gadgets, innovative tools for trainers, and cutting-edge equipment. As an example, the virtual reality (VR) utilizes projector environments for creating realistic sounds, images and promotes a user's physical presence in a virtual environment. Another technology, called the Augmented Reality, in which computer generated image can be created by superimposes a computer-generated image on a user's view of the real world, resulting in a composite view, in which real-world sensory input like video, graphics or sound are included. Artificial Intelligence refers to the intelligence displayed by machines, as opposed to natural intelligence displayed by humans and animals. AI can help highlight areas that require improvement and assist students in focusing on areas where they are lagging. Advanced AI models can solve many problems for the users in a more comprehensive way and is having an upper edge over the typical classroom curriculum. Big data gives the learners a fair opportunity to accomplish the best possible outcome by allowing them to pinpoint learning interactions that should be fine-tuned within the e-Learning module to improve their e-Learning courses.

AI-based e-Learning platform has the ability to perform different tasks requiring human intelligence. It creates solutions to human-related problems, like speech recognition, translations involving different languages, decision making, and much more. The auto suggestions for texts are using Artificial Intelligence engine in our mobile devices. Even though the AI-based e-Learning platform has not become a standard learning approach amidst most learning organizations, there is a need for it. The future of e-Learning can be easily influenced by the AI-based e-Learning platform with a positive impact in its development in a diverse way.

In a real-time situation while answering to the questions of the learners the AI-based e-Learning platform plays a vital role. It acts as an able tutor by clearly answering to the leaners questions during the learning process. It acts as a boon to the timid learners by giving them a platform to raise questions on subject matters that are unclear and get immediate solution. By incorporating AI into the learning program, the learners are given the best environment for clarifying their questions at their convenient time and location.

A course content creation is where AI can help to draw out valuable information and convert it into smart content for digital learning. The growth of AI is continuous. An effective e-Learning course can be delivered by combining all the necessary skills. Using an AI system, the human language can be easily and efficiently processed to develop an e-Learning enabling system. The instructors can create content that can be shared to different learners by taking advantage of e-Learning as it is impossible for an instructor to simultaneously cater to the needs of every learner.

e-Learning can be designed in such a way that it can be coupled with multiple sources which provides reliable services available in the network. It can also incorporate the feature of real time data transfers. The solution architect needs to analyze and provide a time plan required for the implementation of the web service and implement necessary milestones for the completion of the integration testing. A parameterized web service may be recommended to cater the multiple types of requests and creating individual web services for each request. Real-time integration and loosely-coupled integration are some of the features of e-learning. The modules of e-Learning which requires AI based analysis are course evaluation and performance analysis. These can be designed as major web services of the e-Learning. Even low-tech integration like feedback, chatbot integration also adopt Web Services and can obtain enormous benefits out of it.

ISSN 2394 - 7780

Webservices has given MOOCs (Massive Open online courses) where different course coordinators, learners, evaluators coming under one platform. The SWAYAM (Study Webs of Active–Learning for Young Aspiring Minds) portal, initiative of Government of India is one such a MOOC online portal for e-learning which provide a front-end user interface for the learners to register for the various courses. It is aimed at access, equity and quality. It is to provide the take the best teaching learning resources to all from the best institutions to all. In order to ensure that best quality content is produced and delivered, different coordinators are identified for each type of courses that include school education, out of school education, undergraduate and postgraduate programs.

Under each course catalog, the learners can have course mode, duration, exam and credit details. The details of the course are provided with necessary access provided through social platforms. Web Services are extensively used for the registration to capture the details and provide the details to the coordinators. The course materials, online classrooms and payment details are provided through web services. A separate mobile app is also provided for accessing the portal features through web services. NPTEL – National Program on Technology Enabled Learning, conducted by IITs are also accessible through SWAYAM. Necessary web services are used for SWAYAM and NPTEL integration. SWAYAM also issued necessary guidelines for developing and integration of online courses to MOOC programs.

CONCLUSION

E-Learning has been synonymous to web learning where learners where multiple, potentially unrelated functionalities and services are incorporated and made available with common interfaces. Portals are becoming single point access to resource offerings, such as study materials, attendance, and online answering, or to faculty offerings, such as schedules of class rooms, evaluation tests, and other useful resources. Such e-learning portals help to avoid spending a lot of time searching through multiple locations, portals aggregate the relevant information at a single, easy-to-remember location. In order to achieve this major objective, the web Services help to build such user-friendly and resourceful education Web portals. Web Services are easy configurable as well as remotely-managed by the designers and content providers, Web portals simply have to either allow or disallow access to each Web Service. Additionally, since Web Services use industry-standard technologies and can be accessed from any computer, they ensure interoperability with the portal as well as with the portal's users.

REFERENCES

[1] Integration of Web Services with E-Learning for Knowledge Society by Jui Pattnayaka , Sabyasachi Pattnaikb * 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016)

[2] Adaptive Content Creation for Personalized e-Learning Using Web Services by K.K. Thyagharajan and Ratnamanjari Nayak

[3] https://www.ugc.ac.in/pdfnews/3885329_MOOCs-Guideline-(Development--Funding).pdf

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

A STUDY ON MARKET BASKET ANALYSIS IN DATA MINING

[1]Aayushi Dedhia and [2]Fatima Shaikh
[1]Department of M.Sc. Big Data Analytics, Jai Hind College (Autonomous), Mumbai
[2]Assistant Professor, Department of IT, Jai Hind College (Autonomous), Mumbai

ABSTRACT
Data analysis plays an important role in the present era as it helps us to understand the patterns by exploring it in meaningful way. Market basket analysis is a data mining technique, also known as association rule learning or affinity analysis which identifies the relationship associated with different data items. This paper aims at analysing some of the different techniques associated with market basket analysis. This paper discusses the data mining technique i.e. association rule mining which may be helpful to study the customer purchasing behaviour and helps in increasing the sales. It discusses various algorithms such as Apriori, FP-growth, Tertius, etc. It presents an overview of the recent work in analytics research for market basket analysis in data mining

Keywords: data mining, market basket analysis, big data.

I. INTRODUCTION

Big Data Analytics is a major research topic in the business world and providing a good Customer Relationship Management is a major requirement nowadays. Handling Big Data according to Velocity, Volume and Variety is a major issue. There are many challenges in using Big Data in Customer Relationship Management (CRM) [1].

CRM 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. 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 [1].

1. Data Mining

Nowadays, data mining is widely used in several aspects of science such as manufacturing, marketing, CRM, retail trade etc. Data mining is the process of analysing large amounts of information to identify trends and patterns [2]. Data mining can be used for everything from determining what customers are interested in and what they want to buy, to fraud detection and spam filtering. Data mining programs analyse data for patterns and connections based on information requested or provided by users.

Artificial intelligence, neural network, statistical techniques, pattern recognition, clustering and classification approaches are areas included in the data mining. With increasing data mining popularity, most researchers apply data mining techniques to extract information from data sets [2].

The data mining process is divided into five steps. The data is collected and stored and managed into a data warehouse. Further the data is accessed by business analysts, management teams, and IT professionals and they decide how to organize it. Later, the data is sorted by the application software based on the user's results and finally it is presented in a format that is easy to share such as a graph, diagram or table [1].

1.1 Techniques under Data Mining

Various major DT techniques, like affiliation, rule classification, clustering, prediction as well as measurement trends, etc., have been residential and applied to DT projects recently and are used in database exploration.

- Association rules, also known as market basket analysis which evaluates trends and identify interesting ties inside results [3].
- Classification: This technique is used to obtain important relevant information about data and metadata. It is the role of searching for a pattern or function and differentiate data groups to determine where an entity is placed whose class mark is unknown [3].
- Clustering: which is similar to classification but it is a method of classifying a collection of specific or abstract items of related artifacts. It identifies similarities between objects and groups those items based on what makes them different from other items [3].

ISSN 2394 - 7780

- Prediction: the classification predicts categorical names, prediction models. This is to estimate the numerical data values lost or inaccessible instead of label class [3].

2. Market Basket Analysis

A data mining technique used to uncover purchasing patterns in any retail environment is known as market basket analysis. Market basket analysis in data mining is to analyse the products which have been bought together. This is a technique that can thoroughly examine a customer's purchases in a supermarket. This concept identifies patterns in items that customers frequently purchase. This analysis helps drive deals, offers, and sales by businesses, and data mining techniques help accomplish this analytical task.

2.1 Types of market basket analysis

Market Basket Analysis techniques can be categorized based on how the available data is utilized:

- **Descriptive Market Basket Analysis:** Descriptive market basket analysis is also known as unsupervised learning. It is one of the most commonly used approach which derives insights only from historical data. The analysis here does not make predictions, it just uses statistical techniques to assess the associations between products.
- **Predictive Market Basket Analysis:** This type uses supervised learning models like classification and regression. For example, buying an extended warranty is more likely to follow the purchase of an iPhone. Essentially, it considers items purchased in a sequence to determine cross-selling. It is not as widely used as a descriptive MBA but it is still a very valuable tool for marketers.
- **Differential Market Basket Analysis:** This type of analysis is a beneficial tool for useful analysis. It compares purchase history between stores, between seasons, between two time periods, between different days of the week, etc., to understand interesting patterns in consumer behaviour. For example, it may help determine why some users prefer to buy the same product from Amazon and Flipkart at the same price. The answer can simply be that the Amazon reseller may have more warehouses and may deliver faster.

2.2 Algorithms used in market basket analysis

There are many algorithms which can be used for analysis.

Market basket analysis works primarily with Association Rule {If} -> {Then}.

IF stands for Antecedent: The antecedent is the item found in the data. Then stands for frequently. A consequent is an item found in combination with the antecedent.

With the help of these, the retailers can predict customer behavioural patterns. From this, the retailers can create offers and specific combinations that customers are likely to purchase. This would increase the company's sales and revenue.

2.2.1 Apriori Algorithm

Apriori algorithm also known as frequent pattern mining is a well-known algorithm proposed by R. Agrawal to discover the association rules that represent purchase patterns in a supermarket dataset [1]. It is a powerful calculation utilized for mining the itemsets and use of affiliation rules on the conditional data sets [4].

Components of Apriori algorithm

a. Support

Support is for the standard popularity of a product. The Support for item A in scientific terms is the ratio of transactions containing A to the total transaction number [4].

$$\text{Support(A)} = \frac{\text{(Transactions involving A)}}{\text{(Total transaction)}}$$

The closer the support value is to 1, the better. Means the item appears frequently in transactions.

b. Confidence

Confidence refers to the likelihood that a customer purchases both item A and item B. The total of transactions involving item A and item B is distributed by the number of transactions involving item B [4].

$$\text{Confidence (A => B)} = \frac{\text{(Transactions involving both A and B)}}{\text{(Transactions involving only A)}}$$

ISSN 2394 - 7780

c. Lift

So, the likelihood a consumer purchase both item A and item B together is the Lift value times higher than a possibility when buying alone [4].

- Lift (A => B) = 1 means that there is no correlation within the item set.
- Lift (A => B) > 1 means that there is a positive correlation within the item set, i.e., products in the item set, A, and B, are more likely to be bought together.
- Lift (A => B) < 1 means that there is a negative correlation within the item set, i.e., products in item set, A, and B, are unlikely to be bought together.

2.2.2 FP-Growth Algorithm

FP growth represents frequent pattern growth, a scalable technique for mining common patterns in a database [6]. It is an algorithm of association rule technique that can be used to determine the set of data that most appears often in a data set [7]. The Algorithm only searches the database twice. It is a depth first search algorithm combined with direct counting using recursive strategy of pattern growth, it need not generate candidate sets, and instead, the transaction database is compressed into a tree structure that stores only the frequent items [3]. The algorithm works in two steps:

1) Construct FP tree

a. The database is scanned and one item set is found.

b. The items are arranged in order of decreasing support.

c. The database is rescanned and the FP Tree is created.

2) Search Common Item Sets Using FP-Tree

a. Common item sets are searched recursively with common suffixes ending with items having lower the support first.

The FP-Growth algorithm is more preferred since it is faster than other algorithms as it uses the original database's comprised representation.

2.2.3 Tertius Algorithm

Tertius algorithm finds the rule according to the confirmation measures using the first order logic representation. It includes various option like class index, classification, confirmation threshold, confirmation values, frequency threshold, missing values, negation, noise Threshold, number literals, report literals, values output etc. [6]. The disadvantages of this algorithm is that it takes long time for some larger tests and the execution time is relatively long which is largely dependent on the number of literals in the rules.

II. LITERATURE REVIEW

The following section provides an overview of the MBA research with the current state in the respective domain using different data mining algorithms.

Cicekli et al. (2021) implemented the market basket analysis through a case study of a conventional analysis by incorporating demographic variables along with the purchase transactions. The data examined for analysis was taken from Adepo Sanal Market-an e-retail market in Turkey and it consisted of purchase records as well as delivery location, gender and age group. The products taken into consideration were groceries, beverages, cleaning supplies and household items. The data set consisted of 3163 purchase records. The primary motive to integrate those attributes was the opportunity to extract patterns that connect purchased products with demographic variables. The results presented different tables wherein the first table represented the basket as column and the product as row. The second table presented the count of customers which were grouped by their gender and age. The third table presented the distribution of customers grouped by their location. The rule mining technique was used to discover numerous item patterns that was related to the most popular items in the basket data. The results showed that the study observed cross-category purchase patterns. Later, among the association rules discovered in the study, interesting results were chosen based on the lift and confidence measures and were presented in the tables [5].

Yuvamathi et al. (2018) implemented and provided comparison between the three association rule algorithms namely Apriori algorithm, FP-Growth algorithm and Tertius. It provides a brief introduction about the association rule mining for finding frequent patterns and correlations between the items in the dataset. The authors used the supermarket dataset wherein association rule mining algorithms were applied for analysis. The results showed that the FP-growth produced best rules and also took less time to execute the data. The second

ISSN 2394 - 7780

best algorithm observed by the authors was the Apriori algorithm which produced best rules but took some more seconds than FP-growth algorithm. The third best algorithm observed by the authors was Tertius followed by Apriori which took too more seconds [6].

Suharjo et al. (2020) implemented Market Basket Analysis with the Double Association rule method to cover the desired outcome. The data examined for analysis in this study is taken from Niki Laris Supermarket. The various steps taken in this study included data collection, data cleaning, data integration, data selection, data transformation, using the FP-growth approach, etc. Later, the product layout recommendations and product bundling recommendations were applied and evaluated using the questionnaire method. The data used in the study counted to 40,940 transactions and 3,339 items which were divided into 6 categories namely herbs, processed food, snack, processed drink, instant drink and medicines and vitamins. Further, the first FP-growth approach was applied to observe the interrelationships between product categories. The results of this study were the recommendations for product layout and product bundling recommendations. These recommendations were implemented to enhance customer loyalty which would increase the retail revenue. The study concluded by mentioning that 71.1% of respondents agreed with the proposed implementation [7]

Sudirman et al. (2021) proposed a model to reveal changes in consumer buying behaviour in the three periods of the months, namely the first 10 days of the month, the second 10 days and the last ten days of the month. This analysis was done through association rule technique. The dataset of a hypermarket was used in the study for analysis. Due a large number of transactions, the results showed low support at all three time periods. The results of the association rule showed changes in the consumer buying behaviour in the three-period tested. It was seen that in the first 10 days, there was association between instant noodles and eggs. In the second 10 days, still the association between instant noodles and eggs was present but a bit simpler. Also, in this period, a high association was observed between cooking oil and broiler chickens. Lastly, in the last 10 days of the month, only one association was observed namely one variant of instant noodles and chicken eggs with another variant of instant noodles. It was also observed that budget played a major role for the consumers while they purchased in the three-period. The first 10 days, the consumers had a flexible budget and in the last 10 days, the consumers had a lower budget [8].

Kurniawan et al. (2018) implemented the market analysis application at BC UIN Malang Supermarket with 1553 transaction data collected. In additions of the stored transaction receipts of 890753 up to 891319 transaction receipts. In association rule mode, the implemented algorithm in the making of market basket analysis application was the Apriori algorithm. The further step was recording minimum support and minimum values of confidence. Algorithm was used to develop the frequent item set using 1-item first, then value support of every item would be later counted. Item whose support value was above the minimum support value was selected as 1-itemset high frequency pattern and as 2-itemset candidates. By that of 1-itemset, the development of frequent item set into 2-itemset which would then the value of confidence be calculated next was recursively brought off. The results showed that the development and implementation of market basket analysis application worked well with the association rule method using the Apriori algorithm. With an average confidence value of 46.69% and a support value of 1.78%, the generated rule set was 30 rules [9].

Sivabharathi et al. (2020) proposed a model for finding repeated item sets along with apriori algorithm and analyse whether the products were sold more in the morning or in the evening. The dataset examined for analysis in the study was provided by UCI Machine Repository which included channel, region, fresh, frozen, grocery, detergents, delicatessen and session. The data set was provided to the data mining tool Tanagra for analysis and Apriori algorithm was used to find repeated item set from the dataset. Further, Receiver operating characteristic (ROC) curve was drawn to show the results. The ROC analysis is a graphical approach for analysing the performance of a classifier. The curve was created by plotting the true positive rate against the false positive rate at various threshold settings. The results showed that the products were sold more in the morning rather than in the evening [10]

Schonrost et al. (2017) proposed a model of the association rule which was helpful for reducing the number of rules. The study made an analysis of the market baskets of purchases on transaction data from supermarkets according to the Mining Association rules to better understand customer buying behaviour and to discuss the applicability of the method. A supermarket was selected and its transactions were stored in a relational database where SQL was performed to obtain the necessary information such as the year, month, weekday, day period, the time of the purchase, type of products, their quantity and payment method, etc. The results of the study showed that the purchasing increased in the morning, then in the afternoon and later at night. It was analysed that a total of 2,398,050 products was purchased with 13,114 different products in 184 different categories. Further, The Apriori algorithm was used to generate the association rules and find all frequent k-item sets

ISSN 2394 - 7780

contained in a database. In short, the association rule mining technique provided an accurate summary of how the items were related [11].

Kaur et al. (2016) proposed an change modeling which is used to understand the dynamics of data generation process by examining the relevant changes that have taken place in discovered patterns. The algorithm was helpful in examining customer behaviour and helped in increasing the sales. The paper discussed various techniques for data mining emphasising more on the association rule mining. It mentioned that the existing algorithms worked on the static data but they proposed periodic mining which it worked on the dynamic data and performed periodic mining. The dataset of a bakery as used for analysis. A score table was made with Association rules along rows and their attributes in columns with their scores. Later, outlier detection test was performed at the threshold value of 20 which divides the rule into 2 parts, first being the upper association rule meaning rules which were above threshold and 2nd being the lower association rule also known as outliers meaning the association rules which were below threshold [12].

Said et al. (2012) proposed a new scheme for extracting association rules from transactional records and implementing it in a case study to validate the efficiency of the proposed scheme. The proposed scheme comprises of business understanding, nonordfp approach, apriori rule generation approach, model building, post-processing of association rules and results interpretation. Anonymous data was analysed of a retail supermarket over 6 months consisting of the recorded transaction made by someone holding one of the loyalty cards. Each card consisted of a code that identified its owner, personal information such as gender, date of birth, occupation, education, etc. The card helped to analyse the buying behaviour of the owner like quantity of products purchased, what is purchased, whether they followed any promotions, etc. The objective in the case study was to track the buying patterns. The authors considered the measures such as support, confidence and correlation for validating a set of association rules. The run time and the memory consumption were calculated simply by adding the mining time and memory to the generating of the association rules time and memory. Later, the scheme was compared with FP-growth and Apriori approach with the collected data. The results showed that the scheme was more efficient in the case of time consumption and memory consumption in comparison to FP-growth and Apriori approach [13].

Kurnia et al. (2019) implemented the data mining market basket analysis for knowing the sales pattern at a particular restaurant using algorithms. In today's world, the development of the food and beverage industry has grown to a great extent. 150 transaction dataset was selected to determine the sales patterns including various types of food and beverage items available in the menu at the restaurant. The model determined a high degree of accuracy that hold a strong relationship between the items with a minimum support of 4% and a minimum confidence of 60%. The application produces item association rules or menu combinations in consumer purchasing patterns for promotion strategies that are right on target and accurate compared to using promotional strategies manually [14].

III. CONCLUSION

Market Basket Analysis is an unsupervised machine learning technique that helps find patterns in transactional data. This is a very powerful tool for analysing consumer buying behaviour. Market Basket Analysis is a conceptual framework that originated in the field of marketing and has recently been successfully used in fields such as bioinformatics, nuclear science, immunology and geophysics. One of the reasons for the increasing acceptance of MBAs in science is that the use of inductive approaches to theory building allows researchers to assess the existence of relevant rules. By summing up the whole, recommendation systems can efficiently influence marketing and sales research that can be used for strategic business decisions.

IV. REFERENCES

[1] Perera, W.K.R., Dalini, K.A., & Kulawansa, T. (2019). Review of Big Data Analytics for Customer Relationship Management. IEEE. 10.1109/ICITR.2018.8736131

[2] Heydary, M. and Yousefli, A. (2017), "A new optimization model for market basket analysis with allocation considerations: A genetic algorithm solution approach", Management & Marketing. Challenges for the Knowledge Society, Vol. 12, No. 1, pp. 1-11. DOI: 10.1515/mmcks-2017-0001

[3] Saxena, A., & Rajpoot, V. (2021). A Comparative Analysis of Association Rule Mining Algorithms. IOP Conf. Series: Materials Science and Engineering, 1099. 10.1088/1757-899X/1099/1/012032

[4] Sinha, A. (2021). Implying Association Rule Mining and Market Basket Analysis for Knowing Consumer Behavior and Buying Pattern in Lockdown - A Data Mining Approach.

ISSN 2394 - 7780

[5] Cicekli, U. G., & Kabasakal, I. (2021). Market Basket Analysis of Basket Data with Demographics: A Case Study in E-Retailing. Alphanumeric journal, 9(1). 10.17093/ alphanumeric. 752505

[6] Yuvamathi, N., & Porkodi, R. (2018). A Study and Analysis of Association Rule Mining Algorithms In Data Mining. International Journal of Scientific Research in Science and Technology (IJSRST), 4(2), 283-289.

[7] Suharjo, R. A., & Wibowo, A. (2020). Customer Relationship Management in Retail Using Double Association Rule. International Journal of Emerging Trends in Engineering Research, 8(5), 1620-1625. https://doi.org/10.30534/ijeter/2020/23852020

[8] Sudirman, I. D., Bahri, R. S., Utama, I. D., & Ratnapuri, C. I. (2021). Using Association Rule to Analyze Hypermarket Customer Purchase Patterns. Proceedings of the Second Asia Pacific International Conference on Industrial Engineering and Operations Management, 12-23.

[9] Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S., & Hariadi, M. (2018). Market Basket Analysis to Identify Customer Behaviors by Way of Transaction Data. Knowledge Engineering and Data Science (KEDS), 1(1), 20-25. https:// doi.org/ 10.17977/ um017 v1i12018p20-25

[10] Sivabharathi, G., & Chitra, D.K. (2020). Data Mining Techniques to perform Market Analysis. Journal of Emerging Technologies and Innovative Research (JETIR), 7(3), 13-17.

[11] Schonrost, G. B., Paes, V. C., Balestrassi, P. P., Paiva, A. P., & Campos, P. (2017). Data Mining Association Rules Applied to Supermarket Transactional Data Modeling: a case study in Brazil. International Joint Conference.

[12] Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia Computer Science, 85, 78-85. 10.1016/j.procs.2016.05.180

[13] Said, A. M., & Dominic, P.D.D. (2012). A new scheme for extracting association rules: market basket analysis case study. International Journal Business Innovation and Research, 6(1), 28-46. 10.1504/IJBIR.2012.044256

[14] Yusuf, K., Isharianto, Y., Giap, Y. C., Hermawan, A., & Riki. (2019). Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm. IOP Conf. Series: Journal of Physics: Conf. Series, 1175. 10.1088/1742-6596/1175/1/012047

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

PHYSICAL ACTIVITY RELATED STRESS PREDICTION USING SMART WEARABLES

[1]Dr. Swapnali Lotlikar and [2]Ms. Alicia Cotta
[1]Assistant Professor and [2]MSc IT Student, Usha Pravin Gandhi College of Arts, Science and Commerce Mumbai

ABSTRACT
The word stress has grown prevalent in today's culture as a result of its catastrophic impacts on a big number of people all over the world. Stress must be managed at all times because it is the root cause of many significant health disorders. It is important to build effective mechanisms for detecting and managing stress. Wearables for measuring health are becoming more popular among individuals as a result of the rise of a variety of gadgets presently available from various manufacturers. Heart rate variance (HRV), skin temperature, galvanic skin response (GSR) and sleep pattern are a few examples of the many types of data available. Smart wearables are equipped with all of the necessary sensors for collecting data from the aforementioned signals. A few of the sub-fields that smart health care falls under include embedded systems, cloud computing, machine learning, big data, and artificial intelligence. The paper focuses on wearable technologies and using the data collected to predict stress in individuals using Machine Learning algorithms like Random Forest, Gradient Boosting and KNN. The finding of this research could be useful in guiding future research on automated stress prediction systems.

Keywords: Stress Prediction, GSR, HRV, Random Forest, Gradient Boosting, KNN

I. INTRODUCTION
Stress is the psychological response that the brain causes when it encounters a situation that puts the host in risk. Following the psychological response, the body goes through a physiological reaction that includes an increase in insulin levels, giving the body a little more energy to deal with the situation, an increase in heart rate, even more oxygen getting produced to the brain to help it make faster decisions, and a subsequent shortening of the response time. Nowadays, nearly everyone is under some form of stress. Stress poses a significant health risk once it becomes severe, which is why it is an important topic to explore. Deadlines, emotions, work stress, and other variables all contribute to stress.

Stress is sometimes a necessary evil since it can help you perform better in high-pressure circumstances like interviews, exams, and presentations. However, because everything has a limit, stress at a certain level begins to have a harmful effect on the body's normal functioning, such as high blood pressure, sleep deprivation, and so on. A few examples of the signals utilized include changes in sleep patterns, skin conductance, skin temperature, and pulse rate variation. These are examined to see which one provides the best classification accuracy when employing different machine learning algorithms.

In current times, the growth of wearable and mobile technologies is radically changing health care, as it permits people to self-monitor and manage their health practices, usually without the involvement of any health care professional. Our daily lives are flooded with health information that was not known earlier. Modern fitness technology gadgets and firms have attempted to capitalize on this data during the previous decade in order to find a wealth of useful data that, when properly used, has the ability to change the way we view health, particularly in the aftermath of the COVID-19 pandemic [**2**]. The identification of mental stress in people relies heavily on human physiology-based research. Studies on identifying stress through facial expressions have also been conducted. We need to see a doctor and be assessed to determine whether or not we are stressed, but this does not always appear to be possible, especially in the contemporary epidemic environment. In reality, in this digital age, where everyone has a smart phone, there are few new ways to use technology to monitor your stress levels automatically. Wearable gadgets that assess stress levels based on physical activity are emerging.

The following is a breakdown of the paper's structure. In Section II, we go over some of the related research that has been done in the literature. Section III provides some information about the wearable devices. The stress detection methods are described briefly in Section IV. Section V discussed the Machine Learning Algorithms. Data Source information is provided in Section VI, which includes a description of the dataset. Section VII discusses the approach used. The methodology of the study is described in Section VIII. Section IX includes the result of the study. Section X includes conclusion. Limitations are explained in Section XI and Section XII show the references used.

ISSN 2394 - 7780

II. LITERATURE REVIEW

Recent years have seen a variety of experiments and studies undertaken as a result of the evolution of medical science and machine learning, leading to the release of relevant publications.

The most comparable study was carried out by David Liu and Mark Ulrich as part of a machine learning course at Stanford University. They used the heart rate variability and ECG methodology to retrieve ECG and variability in heart rate data features from the dataset and used the Physionet driving database to classify stressful conditions. They were able to achieve an F1 score of 0.7855 using a linear SVM, they do not, however, make it clear how data management and parameter adjustment should be done in order to reproduce their findings. [**3**]

As stress becomes a global issue, there are a plethora of studies being conducted on the subject. Pandey et al. have used the relationship between heart rate and mental stress in their studies. Machine learning techniques such as linear regression and SVM are used for classification tasks, which is effective in stress prediction. The Internet of Things is utilized to alert the person when they are in a dangerous situation. They divide stress into six categories: rest, warm-up, exercise, endurance, danger, and maximum [**4**]. They achieved an accuracy rate of 68%, which can be increased further. The heart rate alone is insufficient to determine a person's stress. Facial expression must also be considered in order to assess their level of tension.

On Kaggle, Sriramprakash.S, O. V. Ramana Murthy and Prasanna Vadana. D suggested a strategy that was tested on the SWELL-KW dataset. They took into account physiological signals from ECG (Electrocardiogram), galvanic skin reaction sensors, and other sources. Data classification was done using the SVM and KNN (K- Nearest Neighbor) Machine Learning algorithm. They proposed a method for identifying dominating features and combining it with overlapping techniques to improve outcomes. They find that temporal and frequency domain features such as galvanic skin response (GSR), heart rate variability (HRV) and heart rate (HR) are sufficient for stress prediction [**5**]. The classification accuracy of the proposed solution is 66.52 percent when KNN is used, and 72.82 percent when SVM is used.

A smart watch and a custom-built GSR wrist sensor were utilized in the study [**6**] to recognize stress in college students. This study is context-based because diverse activities are employed to produce stress, and the data is collected using a heart rate and GSR sensor to determine stress.

The authors of the study [**7**] developed a context-based stress detector to distinguish between psychological stress and real-life acts that cause identical physiological reactions, such as exercise and eating meals. Context aids in boosting the accuracy or precision of stress detection by allowing one to discern between reactions and identify stress more accurately.

III. SMART WEARABLES

According to Market Research Future's new study, the worldwide wearable devices marketplace is expected to reach USD 27,255.6 million by 2023, with a staggering 23 percent CAGR (MRFR) [**8**].

Apple and Fitbit devices include a highly accurate optical heart rate sensor [**9**][**10**] that provides heart rate measurements in beats per minute. However, because they smooth the data to more accurately represent heartbeats per minute, these sensors do not indicate RR intervals. [**11**] The absence of RR samples results in a potential loss of crucial information about an individual's health since increased variability in heart rate shows that the heart is responding appropriately to variations in associated physiological signals like breathing. [**12**] Fortunately, there are monitors on the market now that can offer HRV measurements with a high degree of accuracy. For example, one wearable device that precisely measures RR peaks is the Microsoft Band 2. By recording the observations of 49 students taking a memory test on a computer, the Band 2's capability to measure RR peaks was compared with that of an ECG device in a study done by Chudy et al. [**13**]. The results demonstrate that the two devices had satisfactory consistency.

IV. STRESS DETECTION METHODS

1. Heart Rate Variation (HRV)

It is the difference between the intervals between successive heartbeats [**5**]. Because stress causes a direct rise in the flow of blood, which in turn causes an increase in heart rate, it is the most effective and widely used tool for detecting stress. Consequently , one of the most used approaches for identifying stress is change in heart rate. The average heartbeat is between 60 and 100 beats per minute; anything above that is deemed abnormal. Stress is commonly detected by a sudden elevation in heart rate. Commercial wearables include sensors that can accurately measure heart rate. Based on HRV, two categories are created from the data using machine learning

ISSN 2394 - 7780

techniques (stressed or normal). Because stress creates a smaller time interval between each pulse, a rise in HRV is a great way to detect stress.

2. Skin Temperature

It can fluctuate according to a variety of conditions, including stress. An increase in blood flow rate causes a rise in skin temperature. As blood flow is regulated by the nervous system, stressful situations put a large amount of additional load on the nervous system. The skin temperature's mean, max. value, min. value, and standard deviation are used to determine stress [**14**].

3. Sleep Pattern

Two types of sleep include REM sleep and Non - REM sleep. This stage of sleep, where REM stands for rapid eye movement, is sometimes referred to as the dream phase because dreams occur most frequently during this time. The eyes move and exhibit fast movement, and the fluctuation of brain waves, results in dreams. Non-REM can be divided into three stages. Eyes are closed in the first phase, but waking up is simple. In the second phase, the heartbeat slows down, this is also referred to as light sleep, resulting in a reduction of temperature in the body, indicating that the body is prepared for sleep. The third stage, also known as deep sleep, normally accounts for 30 percent of the total combined sleep time. Amid deep sleep, the body is completely at ease, allowing it to repair damaged tissues while also allowing the immunological and digestive systems to work. The average human needs a duration of deep sleep ranging from 2 to 2.30 hours per night, which declines with adulthood. A rapid shift in sleep pattern, particularly a decrease in deep sleep, may, nevertheless, be an indication of stress. So, while a drop in deep sleep quantity over time might be used to infer that the individual is under stress, this strategy is not a 100 percent correct because there could be other explanations besides stress for decline in the duration of deep sleep. Deep sleep is split into two sections: class 1, that lasts for more than two hours, and class 2, which is shorter than two hours. To determine the class to which an individual sleep period belongs, SVM may be used, which in turn can help determining the person's sleep quality [**15**]. Sleep quality is determined by how frequently a person alternates between the phases of sleep; fewer transitions indicate greater sleep efficiency. Deep sleep is can be impacted by how frequently a person switches positions while sleeping or the intensity of arm motions [**16**].

4. Galvanic Skin Response

Electrodermal activity (EDA), which is also commonly known as the Galvanic Skin Reaction, is a component that includes skin conductance (GSR). There are two elements to EDA. The primary is skin conductance level (SCL), which is a static form of conductance that does not vary dramatically, and the secondary is Skin Conductance Response (SCR), that is an activity based conductance which changes in reaction to some event [**14**]. It is designed to measure perspiration which can be a marker for stress. When the body is under stress, it produces a lot of sweat, which causes the skin's electrical characteristics to change. Perspiration influences skin conductance because sweat lowers skin resistance, which elevates skin conductance. As a result, when a person is anxious or engaged in strenuous physical activity, skin conductance varies. A slight current can be used to measure skin conductance between two electrodes in order to assess skin resistance. Because stress is typically linked to an event, the SCL component can be used to establish a standard while the SCR component can be utilised to identify stress. The data can be used to forecast if the subject is worried, with a '1' for stressed and a '0' for not stressed. [**17**].

V. MACHINE LEARNING ALGORITHMS

1. Supervised Learning

In this case, the machine is trained with labelled data, which implies that the values of both x that is the independent variable and y that is the dependent variable are already established. Where x is the input variable and y is the output variable, and y = f (x); Our objective is to define the function f (x) that acts as a map to output the value of y for a specified value of x. Because the output value is already established in the data set, it can be used to verify the output. Through multiple iteration, the mapping function can be improved and will help in the prediction of unforeseen data. In regression and classification problems, supervised learning is applied.

1.1 Random Forest

Random forest is a technique that can be used for regression as well as classification. It comprises of a trees in a large number (decision trees) that are employed as an ensemble, as the name implies. Every choice provides a prediction, and the result belongs to the voting or prediction class with the largest amount of members. The anticipated result will be more accurate if there are more decision trees.

ISSN 2394 - 7780

1.2 Gradient Boosting

In order to reduce overall prediction error, Gradient Boosting combines the best next model with past models. The purpose of defining the intended outcomes for the subsequent model is to reduce error. The objective is set for each case based on the impact of changing the prediction on the prediction error overall: a minor shift in a case prediction returns in a big reduction in error, the next target output value of the case is high. If the new model's predictions are similar to its goals, the error will be lowered. If no change in the error is measured after a minor change in the prediction for a case, then consequent target outcome value of the case is zero. Error cannot be reduced by means of modification to this prediction.

1.3 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a machine learning technique that falls under the supervised category. It stands for K Nearest Neighbour. It is premised on the idea that comparable items occur in close proximity to one another, which is also true for data. It can be used to identify the class to which a data point belongs in both regression and classification. The Euclidean distance or straight line of the data point is determined using its neighbours nearest to it. Using this distance as the radius, the data point's K nearest neighbours are then identified. The class to which the data point belongs is determined by how many of its neighbours are Euclidean distances away from it.

2. Unsupervised Learning

Unsupervised learning involves unlabelled data as opposed to supervised learning, which uses labelled data. Unsupervised learning is unpredictable since the model must identify the dependent variable on its own, and because the outcome is not visible right away, the computer must predict. Complex processing tasks can be handled more effectively by unsupervised learning algorithms compared to supervised learning. Clustering and association are two types of unsupervised learning tasks.

VI. DATA SOURCE

The original data comes from Healey's PhD thesis project at MIT [**1**], and it consists of body measures obtained on a various youths while driving stressing environments, e.g. rush hour, red lights, and highways as well as a relaxation period to generate a non-stressed baseline reading. Physionet has made the dataset available for free. The dataset is separated into 18.dat files along with 18.hea files with supporting meta data in a physionet-specific format. The data contain signals for the ECG, EMG, GSR readings of foot and hand, HR, and ventilation. All results are float values with a sampling frequency rate of 15.5 samples per second.. The information is kept in Physionet WaveForm DataBase (WFDB) format, and every record is composed from two files, one with an extension of .hea and the other with an extension of .dat. The data is read using the WFDB command rdsamp from the Physionets tools called WFDB native terminal installation, then consolidated and stored as.txt files with column headings, units used for measurement, and timing (in seconds) for every row containing the data samples. The data is placed in a Pandas dataframe after the header names are manually sanitised. A sampling time is assigned to each file, which begins at zero and ends when the sampling session concludes. The time interval is increased based on the latest time interval in the previous file to convert the data into a single continuous time-series.

VII. APPROACH

The goal of the research is to see how effective and accurate Random Forest-based, Gradient Boosting and KNN based models based on data could be in stress prediction. Random forest, Gradient Boosting and KNN have been chosen as a methods for accomplishing this task since it can be easily predicted utilising data from individuals. Although the Random Forest is volatile, it is convenient to utilise a machine learning method that has been set up to produce consistently good results in the absence of hyper-parameter adjustment. The Gradient Boosting Algorithm is typically used to reduce bias error. It is used in regression as well as classification problems. By switching from continuous values to discrete bins, which reduces memory use, training becomes more efficient and faster. Instead of using a level-wise technique, which would yield much simpler trees, it uses a leaf-wise split strategy, which results in considerably more complicated trees and is the key to reaching higher accuracy. Compatible with large data sets and supports parallel learning. KNN produces extremely accurate predictions and can compete with the most accurate models. Consequently, the KNN algorithm can be applied to applications that demand high accuracy but don't need a human-readable model.

The dataset is pre-processed after data collection to recover missing values using the mean approach. Missing values in machine learning algorithms can generate a lot of other issues. As a result, missing values in a dataset must be identified and numerical values substituted. Replacement by mean and replacement by median are two methods for addressing missing values in a dataset. There were many missing values in the data, so

ISSN 2394 - 7780

mean/median replacement techniques were used to fill in the gaps After that, the dataset was subjected to feature scaling to ensure that no attribute intimidates other attributes.

Efficiency measures will be used to compare this study to previous work in order to determine whether the suggested model based on random forest regression, gradient boosting and KNN is efficient and viable.

VIII. METHODOLOGY

1. Data Preprocessing

Preprocessing data is an essential process in creating a machine learning model. The obtained data is frequently uncontrolled, with out-of-range values, missing values, and so on. Such information may cause the experiment's outcome to be biased. Integration, transformation, cleaning and reduction of data are the fundamental phases in every data pre-processing phase.

1.1 Data Cleaning

First ensure that the labels of the target feature ‘stress’ are int value of either 1 (stressed) or 0 (normal). Then manually clean the rows of data, and replace the missing values with mean values.

1.2 Feature Selection and Elimination

Since retrieval of information from wearable devices aimed at the consumer market is not convenient, the galvanic skin response elements that were used for labelling the data are also eliminated, along with the data of ECG and EMG. Since most of the readings in the ultra low frequency (ULF) segment are zero and the small time intervals are likely to blame, it is also disregarded. Additionally, the data from the very low-frequency (VLF) band is disregarded as research by Hjortskov et al. **[19]** shows that the extremely low - frequency band confirms to be an untrustworthy indicator in observations just below 5 minutes and the data points existing in the dataset are inaccurate.

While research by Meredith et al. **[20]** claims that identification of respiration is feasible from analysis of waveform of PPG data, without having a strap or band around the chest, however, the dataset captured respiration with a strap around the chest.

The current feature set consists of heart rate, respiration, handGSR, footGSR, interval in seconds, AVNN, RR intervals, SDNN, pNN50, TP, RMSSD, LFHF, LF, HF. With a total of 14 features and 4129 samples, the RR intervals and heart rate are the mean values from the 30 - second time needed to obtain the features of heart rate variability.

The methods used for Feature Selection include tree based model using Random Forest Classifier and embedded method LASSO. Figure 1 shows a flowchart illustrating the implemented feature selection process.

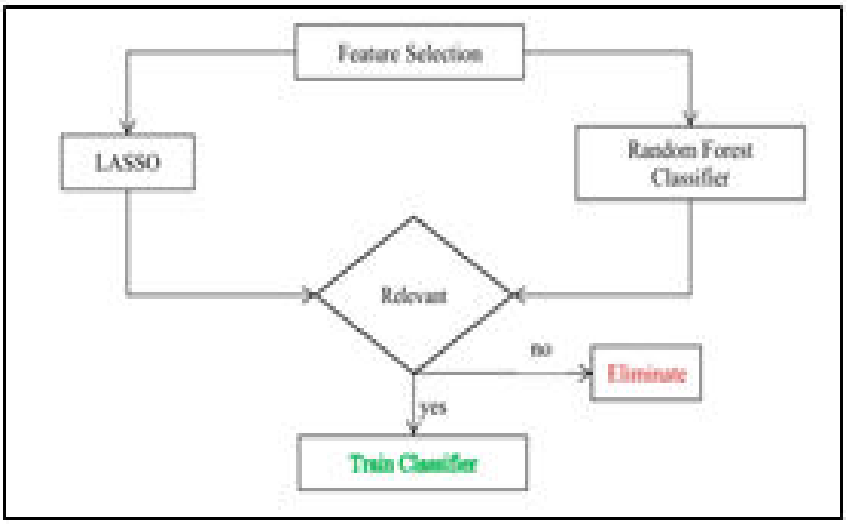


Figure 1: Flow chart of feature selection process

IX. RESULTS

The features used are HR, Resp, footGSR and handGSR and stress as target with 0 - normal and 1 - stressed.

The importance of all the features in the data set is calculated using the feature importance from the Sklearn python module feature importance along with Random Forest Classifier.

ISSN 2394 - 7780

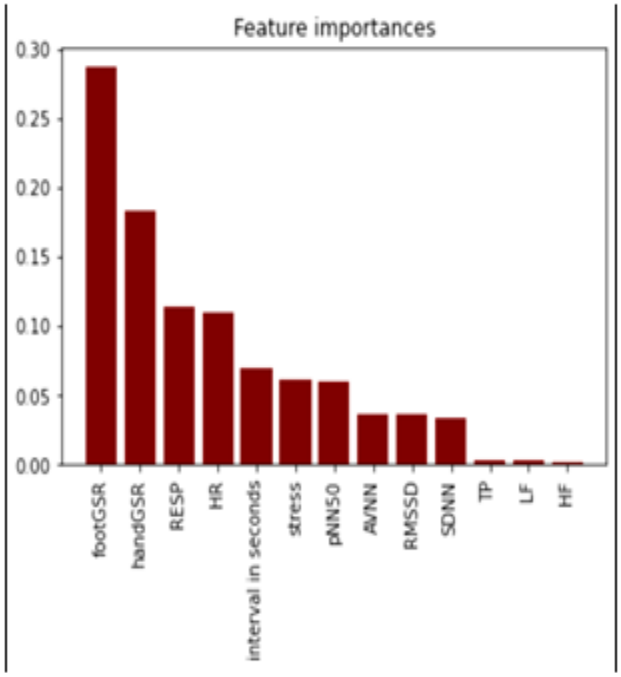


Figure 2: Feature Importance

Figure 2 depicts the estimated feature significance values in a visual manner. The features footGSR, handGSR, RESP and HR were selected

Table 1: Model comparison

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.94	0.96	0.93	0.94
Gradient Boosting	0.93	0.95	0.92	0.93
KNN	0.91	0.93	0.88	0.90

ROC curves of the three models were calculated. The AUC scores have been obtained by the models during stress prediction of the test dataset from the reconfigured dataset. Features included: footGSR, handGSR, RESP and HR.

The Random forest model has an AUC score of 0.95 which means that the model will be able to classify stressed or normal correctly 95% of the times compared to Gradient Boosting which has an AUC score of 0.94 and KNN which has an AUC score of 0.91

The Random Forest model showed a 94.8% accuracy and a precision of 0.96 which means that the model predicts correctly 96% of the times with an F1-score of 0.94 in comparison to Gradient Boosting model with a 93.7% accuracy and KNN model with 91.1% accuracy. Although Gradient Boosting and KNN have high accuracy, Random Forest model is the optimal choice for prediction of stress

X. CONCLUSION

In this research, an experiment has been conducted to identify the suitable algorithm for prediction of stress. There was no solid evidence to define one algorithm as the optimum prediction technique.

As a result, a set of three algorithms was chosen, which included Random Forests (RF), Gradient Boosting and KNN. The dataset was used to train the chosen algorithms to assess the accuracy of machine learning models. The trained algorithms were evaluated using performance measures. After analysing the data, it was discovered that Random Forest had a higher forecast accuracy than both Gradient Boosting and KNN. The proposed stress-predictive system based on a Random Forest Classifier model has achieved 94.8 percent accuracy which is the finding of the research.

XI. LIMITATIONS

Most studies such as Ouwekerk et al. [**18**], that include GSR sensors use custom built wrist devices because they are hardly available in commercial wearables. As a consequence, GSR is a signal that might be regarded the immediate response of stress. T he application of GSR should be minimised. The temperature of your skin isn't the ideal option to measure stress. Because sensors are not optimally operational during the summer, stress

ISSN 2394 - 7780

can be detected as they cannot detect minute changes in heated surroundings. Work should be more focused on recognising stress in a dynamic setting. Additionally, a method for differentiating between physiological and psychological stress should exist. Contextual factors can aid in the detection of stress in the real world (unrestricted), as it is capable of recognising the difference between psychological and physiological stress

Detection of stress based on context, on the other hand, does not have a proven track record. To develop a solid framework, more study is required.

The movement of the body or arm activity is taken into account for sleep pattern change, whereas the lower body and legs will be overlooked. The study's shortcoming is that wearables are capable of detection of upper-body movement or movement in the limbs, that leads to reduction in the accuracy of stress detection

XII. REFERENCES

- Healey, J. A. (2000). Wearable and automotive systems for affect recognition from physiology (Doctoral dissertation, Massachusetts Institute of Technology).
- Greiwe, J., & Nyenhuis, S. M. (2020). Wearable technology and how this can be implemented into clinical practice. Current Allergy and Asthma Reports, 20, 1-10
- Liu, D., & Ulrich, M. (2014). Listen to your heart: stress prediction using consumer heart rate sensors. Online]. Retrieved from the Internet.
- Pandey, P. S. (2017, July). Machine learning and IoT for prediction and detection of stress. In 2017 17th International Conference on Computational Science and Its Applications (ICCSA) (pp. 1-5). IEEE.
- Sriramprakash, S., Prasanna, V. D., & Murthy, O. R. (2017). Stress detection in working people. Procedia computer science, 115, 359-366
- Egilmez, B., Poyraz, E., Zhou, W., Memik, G., Dinda, P., & Alshurafa, N. (2017, March). UStress: Understanding college student subjective stress using wrist-based passive sensing. In 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) (pp. 673-678). IEEE.
- Gjoreski, M., Luštrek, M., Gams, M., & Gjoreski, H. (2017). Monitoring stress with a wrist device using context. Journal of biomedical informatics, 73, 159-170.
- Ruchika Sharma. (2020, October 27). How Big Data and Analytics Reshape the Wearable Device Market. Health Tech Advisor. https://healthtechadvisor.com/how-big-data-and-analytics-reshape-the-wearable-device-market
- Apple, 2017. Apple Watch Series 2 - Technical Specifications. [Online]Available at: https://support.apple.com/kb/SP746?locale=en_GB]
- [FitBit, n.d. Fitbit. [Online] Available at: https://www.fitbit.com/uk/purepulse
- EliteHRV, n.d. Compatible Devices. [Online] Available at: https://elitehrv.com/compatible-devices
- Sisson, M., 2014. Mark's Daily Apple. [Online] Available at: http:// www. marksdailyapple. com/ have-you-checked-your-heart-rate-variability-lately/
- Chudy, N. S. (2017). Testing of wrist-worn-fitness-tracking devices during cognitive stress: A validation study
- Y. S. Can, B. Arnrich, and C. Ersoy, " St ress det ect ion in daily life scenarios using smart phones and wearable sensors: A survey," J. Biomed. Inform., vol. 92, no. August 2018, p. 103139, 2019
- K. A. Kaplan, P. P. Hardas, S. Redline, and J. M. Zeitzer, " Correlat es of sleep qualit y in midlife and beyond: a machine learning analysis," Sleep Med., vol. 34, pp. 162–167, 2017.
- A. Muaremi, A. Bexhet i, F. Gravenhorst , B. Arnrich, and G. T rost er, "Monit oring t he impact of st ress on t he sleep pat t erns of pilgrims using wearable sensors," 2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014, pp. 185–188, 2014.
- M. Zubair, C. Yoon, H. Kim, J. Kim, and J. Kim, " Smart wearable band for st ress det ect ion," 2015 5th Int. Conf. IT Converg. Secur. ICITCS 2015 - Proc., pp. 1–4, 2015.
- Ouwerkerk, M., Dandine, P., Bolio, D., Kocielnik, R., Mercurio, J., Huijgen, H., & Westerink, J. (2013, November). Wireless multi sensor bracelet with discreet feedback. In Proceedings of the 4th Conference on Wireless Health (pp. 1-8).

ISSN 2394 - 7780

- Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. European journal of applied physiology, 92(1-2), 84–89. https://doi.org/10.1007/s00421-004-1055-z
- Meredith, D. J., Clifton, D., Charlton, P., Brooks, J., Pugh, C. W., & Tarassenko, L. (2012). Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. Journal of medical engineering & technology, 36(1), 1–7. https:// doi.org/ 10.3109/ 03091902.2011.638965

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

BIG DATA IN INTRUSION DETECTION SYSTEM: A SYSTEMATIC LITERATURE REVIEW

Jinali Gosar[1] and Fatima Shaikh[2]
[1]Department of M.Sc Big Data Analytics Jai hind College (Autonomous) Mumbai
[2]Assistant Professor Department of IT, Jai hind College (Autonomous) Mumbai

ABSTRACT
In today's world Intrusion Detection System (IDS) has become the main object for monitoring network security. It serves as one of the essential tools for enhancing network security since it can identify threats or anomalous accesses. Information regarding the computers and networks used in defense systems has always been a target for hackers. As a result, deadly attacks are continuously a hazard to these networks. Networks constantly deal with an enormous amount of data. Therefore, it is required to construct an intrusion detection system using big data approaches in order to detect such intrusions. for an accuracy and efficient data analysis process. ML is widely used in lieu of human analysts to construct signatures or categories enormous volumes of knowledge. This article presents an overview of the recent work on big data analytics research for intrusion detection.

Keywords: Big data, intrusion detection system (IDS), machine learning (ML).

I. INTRODUCTION

With over 5.10 billion people using the Internet in 2022 ,since then, 2.5 quintillion bytes of data have been processed per day. Big data can gather huge amounts of automated information to visualize and draw insights which makes it feasible to predict attacks.IDS has become a big data problem because of the tremendous volume and complexity of data to unveil increasing cyber-attacks. Providing protection and privacy for Big Data is one of the most difficult challenges faced by the developers of security management systems, especially with the widespread use of the Internet networks and the quick expansion of data produced from numerous sources give hostile intruders additional space to launch attacks. It also enabled existing attacks to be differentiated by gaining new features. This situation has caused material and moral loss to a considerable number of institutions, organizations, and companies [1]. An increase in technology unfortunately has brought along many risks as well as numerous opportunities. This situation has increased the amount of available data exponentially rather than linearly, so the storage of data has become a separate problem and has increased the costs. Hence, the issue of network security, which allows the protection of integrity, data privacy and accessibility, has taken a critical position and gained a great importance. [1]

With the wide advancement within the web and online platforms, network security conditions have additionally come back ineluctable. Colorful pitfalls associated with network security will be seen presently, like computer code bugs and intrusions. These bugs often occur because of the huge practicality and enormous size of the computer code or the software system. Still, the firewall area unit placed in between two or an additional computer dedicated to segregating these networks is grounded on deciding rules or programs. However, these firewalls are not enough to be secured from similar styles of attacks. The script wherever intrusion discovery systems play an important half in stopping the cyber-attacks and dissecting the safety issues at the time of comparable intrusions so that these things can be dived within the future.[2]

1. Big Data

Big data refers to complex datasets which are to be dealt with by data-processing software. It is a huge collection of data that is growing exponentially with time via various sources. Big data is a combination of unstructured, structured, and semi-structured data which is collected by various organizations for mining information and can be used for predictive modeling, machine learning algorithms, and other advanced analytics applications. Organizations storing and processing big data have a common data-management architecture for their systems, which, when combined with tools, supports the big analytics uses. We usually work in MB (Excel, Word) or maximum GB(code, movies), but data in peta-bytes size is called big data. As per the latest reports, the total damage from cyber-attacks has reached $6 trillion. A cyber attack occurs every 9 seconds. On daily basis 30,000 websites are hacked. In 2021 there were 22billion breached records.

2. Intrusion Detection System (IDS)

An intrusion detection system is a device or software operation that monitors a network or systems for vicious exertion or policy violations. Any intrusion exertion or violation is generally reported moreover to a director or collected centrally using a security information and event operation system. IDS are a monitoring system that will detect suspicious conditioning and generates cautions when they're detected. Grounded upon these cautions, a security operations center (SOC) critic or incident responder can probe the issue and take the

ISSN 2394 - 7780

applicable conduct to remediate the trouble. An IDS is a key element to a truly successful security solution. Whenever a network intrusion gets detected the reported information contains the targeted address type of attack and the IP source address of the intrusion is detected.

2.1 Types of IDS

IDS can be categorized based on the information sources used as input to discover unusual behavior. Regarding data sources, there are generally two types of IDS technologies, namely Host-basedIDS (HIDS) and Network-based IDS (NIDS). Computer security technologies HIDS and NIDS are deployed to safeguard technologies against spyware, viruses, and other destructive file types.

1. Host- based IDS (HIDS): HIDS inspect data that originates from the host system and audit sources, such as operating system, window server logs, firewalls logs, application system audits, or database logs. HIDS can detect insider attacks that do not involve network traffic. [3]
2. Network-based IDS (NIDS): NIDS monitors the network traffic that is extracted from a network through packet capture, Net-Flow, and other network data sources. Network-based IDS can be used to monitor many computers that are joined to a network.[3]

2.2 Detection Methods of IDS:

IDS can be classified on the basis of the methods used to identify intrusions namely Signature-based detection method and Anomaly-based detection method

1. Signature-based detection : Refers to detecting network attacks by searching for specific data patterns, this term originates from anti-virus applications, which refer to these detected patterns as signatures.[4]
2. Anomaly-based detection : Primarily introduced to detect unknown attacks or zero-day attacks, this is partly due to the rapid development of malware. Machine learning techniques are trained to create a model and then compare the new behavior to this model.[4]

IDS, however, faces multiple challenges, such as a high rate of false positives and negatives. When an attack is missed by the IDS, a false negative rate occurs. It happens when an attack-related behaviour is recognized as legitimate by the IDS. Since IT experts are unable to predict when an attack will occur, this is the most perilous situation. A false alarm is a false positive. It happens when an activity is flagged by the IDS as an attack despite having approved behavior. Usually not much damage is caused to the network byfalse positives. Common methodologies for merging intelligence with IDS and thus allow ease of detection for all kinds of intrusion attacks and therefore safeguarding the system from all sorts of threats are done with the help of deep learning and machine learning models. Selecting the right dataset is key to building an efficient machine-learning model for intrusion detection.

II. Big Data Approach in IDS:

Whether there is a significant framework that aggregates actions from diverse sources or a primary source of events, it is evident that intrusion detection is confronting a big data challenge. Big data handlers, or parts that manage large data, must be used to address this issue [2].This section provides a quick insight of howbig data handler technologies are used:

The study of big data is an appropriate method for APT identification. An APT detection system might be capable to manage highly unstructured data in arbitrary forms gathered by numerous types of sensors (such as firewalls, IDS, syslog, netflow, and DNS) over prolonged periods of time by using a MapReduce implementation. Additionally, compared to conventional SQL-based data systems, the MapReduce mechanism for significantly parallel processing enables for the implementation of considerably more advanced detection algorithms. Improving current infrastructure like SIEM and data loss prevention (DLP) more sophisticated and safer is the fundamental technique, which enables only the most critical cyber attacks (like APTs) to be diagnosed and segregated. The metrics parameters are tailored, and the data for the analytics is received from both internal and external sources (such as web and mobile activity and ad hoc). In some Big Data analytics platforms, monitoring for malicious threats can be done "in a Google-like approach," which means that organizations can define their own search parameters. Big Data tools like Hadoop and a network monitoring programme called PacketPig to build an NIDS that can manage big data network streams [5]. When used with Hadoop, PacketPig is capable of deep packet inspection, deep network analysis, and even full packet capture. Hadoop is open source framework for working with big data. The biggest strength of Hadoop is its scalability .It suits in long-term event due to processing and fault tolerance provided. Hadoop works on master-slave fashion. HDFS (hadoop distributed file system) provides the parallelism for the Map Reduce which is used for computation on large commodity servers . Apache Spark is yet another the most popular tool which uses faster

ISSN 2394 - 7780

query processing and in-memory caching for quick analytic queries against data of any scale. For ML algorithms Spark has demonstrated excellent outcomes, surpassing Hadoop, MapReduce by up to 100 times.[2][5]

Furthermore, ML algorithm works more accurately in detecting the attacks for huge amount of data under less time .Using ML to extract usable information from big and multidimensional data appears to be the only alternative because handling such massive amounts of data exceeds the capability of individuals [6]. ML algorithms are classified into two major groups:

1) **Supervised ML:** It is a type in which the output is predicted by the machines using well-labeled training data that has been used to train the machines [6].Frequently used algorithms include :

I.1 **Support Vector Machine (SVM):** It is used for regression and classification. The algorithm will output the classification of the data into classes using a hyper plane that optimizes the margin across all attack classes after being trained with labeled data. [6]

I.2 **Logistic Regression (LR):** It is used for discrete set of classes. It makes use of cost function which maps predictions to probabilities. [6]

I.3 **Linear Discriminant Analysis (LDA):** It is used for dimensionality reduction and prediction. [6]

I.4 **Random Forest (RF):** It is used for regression and classification. The model's efficiency enhances as the volume of trees increases. [6]

2) **Unsupervised ML:** The unsupervised learning algorithm will seek for latent structures in unlabeled data so as to detect intrusions. Deficiency of training data affects unsupervised learning. [6] It includes:

2.1 **K-Means:** The identification of categories in the collection is the core of this algorithm, and the parameter can be utilized to represent the number of groups. It is used for pattern matching in time series data. [6]

2.2 **Principle Component Analysis (PCA):** It is used for dimensionality reduction. It can be used as an input for supervised ML by providing new set of variables known as principle components [6].

III. Intrusion Detection Datasets

It is essential to assess the proposed method's ability to detect intrusive behavior through evaluation datasets, which are vital in the validation of any IDS approach..Due to the privacy issues the datasets used for the network analysis in commercial are not easily available. However some datasets available publicly and widely used for IDS are KDD Cup99, NSL-KDD, Mawilab,Cicids [7].In this part, current datasets that are used to construct and evaluate IDS in comparison are discussed.

1) DARPA / KDD Cup-99

The most common dataset used for IDS is KDD98 (Knowledge Discovery and Data Mining (KDD)) dataset made by DARPA (Defence Advanced Research Project Agency) in 1998 where a number of attack scenarios were simulated and features were extracted.To simulate a local US Air Force base with restricted personnel, these statistics were gathered utilizing a number of internet-connected computers. We gathered host log files and network traffic. Lincoln Labs created an experimental testbed to collect 2 months' worth of TCP packet dump data for a local area network (LAN), simulating a typical LAN used by the US Air Force. They modeled the LAN as if it were a true Air Force environment, but interlaced it with several simulated intrusions.This dataset is an important contribution in field of IDS but these datasets are out of date as they don't contain records of the recent malware attacks. However, the IDS research community continues to utilize KDD 99dataset as a benchmark and is still currently being used by researchers. [7]

2) CAIDA

This dataset was collected in 2007 which consists of network traffic traces from DDoS (distributed denial of service). This kind of denial-of-service attack focuses on a specific computer or network and tries to disrupt normal traffic by flooding the target with a large number of network packets, preventing legitimate traffic from reaching the target computer. Major disadvantage because the data collected does not include characteristics from the entire network; it is challenging to distinguish between aberrant and typical traffic patterns. [7]

3) NSL-KDD

Developed from the KDD Cup99 dataset, NSL-KDD is a publicly available dataset. On analysis on KDD Cup 99, the results were misleading in evaluation of AIDS and IDS accuracy. KDD cup 99 had a huge amount of duplicated training and testing datasets. To address the issues raised above, NSL-KDD dataset in 2009 was created from the KDD Cup'99 dataset by removing duplicate records.This dataset has led to reliable and

ISSN 2394 - 7780

comparable findings across a range of research projects. The NSL_KDD dataset comprises 22 training intrusion attacks and 41 attributes. [7]

4) ISCX 2012

Real network traffic traces from the HTTP, SMTP, FTP etc protocols were examined in this dataset to detect typical computer behavior. This dataset is built on tagged, realistic network traffic that includes a variety of attack scenarios. [7]

5) MAWILAB

MAWILAB dataset used for packet traces from the MAWI repository to assess intrusion detection techniques.These data was collected by the Fukuda laboratory, and are published in two versions. It contains fields like: anomalyID, srcIP, srcPort, dstIP, dstPort, taxonomy, heuristic, distance, nbDetectors and label. A basic heuristic and a taxonomy of backbone traffic abnormalities based on protocol headers and connection patterns are the two independent anomaly categorization methods used by MAWILab.

6) CICIDS2017

This is the latest dataset consisting both malware attacks and benign behavior such as Web Attack, Botnet, Infiltration, and DDoS etc. Based on timestamps, destination ips, protocols, and attacks, the dataset was created. This dataset was compiled using a full network topology, which includes nodes running Linux, Apple's MAC OS, IOS, Microsoft Windows (including Windows 10, Windows 8, and Windows XP), and switches, routers, and modems with various operating systems.80 network flow features from the recorded network traffic are included in the dataset. [7]

Since machine learning techniques are utilized to treat, the datasets that are used to evaluate, these approaches are crucial for a realistic evaluation

IV. Tools for IDS

A variety of organizational security objectives are addressed by an intrusion detection device now on the market [8]. Some of them are:

1) **Solar Winds Event Manager:** It is classified as a HIDS. As it maintains data gathered by Snort, it can also be thought of as an NIDS. In Solar Winds, network intrusion detection is used to examine traffic data as it traverses through the network. This enables it to automatically carry out remedial action in addition to detecting suspicious activity. Solar Winds Event Manager is a thorough network security tool in general.[8]
2) **Ossec-Hids:** OSSEC (open source security open source software that is free. It utilizes a Client/Server framework and is compatible with most major operating systems. OS logs can be sent by OSSEC to the server for analysis and data storage. [8]
3) **Snort:** It is an open source and light weight software .To describe the traffic from an IP address, Snort employs a customizable rule-based language. It records the packet in human readable form using protocol analysis, content searching, and a number of pre-processors. Numerous infections, efforts to exploit vulnerabilities, port scans, and other suspicious activity are all caught by Snort. Snort is not only an intrusion detector, but it is also a Packet logger and a Packet sniffer.[8]
4) **Kismet:** It serves as an instruction for WIDSs (Wireless Intrusion Detection Systems). WIDS compromises packet payloads and WIDS activities. It will locate the entry place for a burglar.[8]

V. Recent Studies

The following section gives us an overview of IDS analytics research, using different machine learning algorithms and focusing on big data. The studied papers were grouped into three categories: Real-Time Analytics, Graph Models, Frameworks and Context.

1) Real-Time Analytics: This section references ways to extract security-related data from logs in real-time.

Reghunath K (2017) proposed a system on real-time intrusion system. The proposed system is to integrate a high volume of data which takes into consideration monitoring a wide array of heterogeneous security. It aims to satisfy the need for a robust system that is fault-tolerant, both against hardware failures and human mistakes. The architecture of the real-time intrusion system is initially initiated by Lamba architecture which aims to satisfy the hurdle of fault tolerance against human and hardware failures and can observe a vast range of workloads. The system should scale out instead of up and should be linearly scalable. The real-time intrusion system is mainly about predicting and working based on the different types of data logs available in distributed data processing layer. The architecture of real-time IDS consists of Real-time data collection, Real-time data

ISSN 2394 - 7780

processing, and a real-time output layer. The real-time data/log collection consists of system logs, application logs, errors, and weblogs through the flume, kafta, etc. Further, the data gets distributed for batch processing via HDFS for data storage, via spark for distributed processing, and real-time speed processing via spark for data processing and Solr for intrusion dictionary. After the data processing, it results in a batch output layer which contains various end user-specific outputs in historical and periodic forms, and a real-time output layer which contains end user-specific outputs in alerts, charts, and real-time actions. Each time a user interaction occurs, the user interaction logs are analyzed, the output is recorded in the Real-time output layer, and the output is compared with the intrusion dictionary. If the output matches with the Intrusion dictionary entries, then the system will generate alerts and can take actions such as blocking the user from further interactions, blacklisting the user, blacklisting the IP address, etc. When real-time data analytics are built on predetermined algorithms and queries, they perform well. On the batch layer of the system, the prediction of intrusion takes place with the help of K- means clustering algorithms. Once the prediction output is prepared the desired reports are available to the system. In case of any abnormalities, the system automatically checks and categorizes the log severity accordingly and takes the necessary actions regarding it. The main advantages of the proposed system are it is capable of generating real-time alerts for the system and is capable of processing and storing the log data in a fast manner. As a future enhancement, this Intrusion detection system prediction will be taken care of by the Real-time Intrusion detection system for big data.[9]

FatimaE et al.(2021) proposed a model based on long short-term memory (LSTM) and Attention Mechanism the detection model is been proposed for IDS. Network intrusions are unauthorized acts performed out on digital assets connected to a corporate network with the specific goal of compromising systems. For the above detection model reduction algorithms: Principal Components Analysis (PCA), Mutual information Chi-Square and UMAP have been used. The proposed approaches have been implemented using the NSL-KDD dataset for binary and multiclass classification.KDD is a dataset proposed to address some of the intrinsic issues of the KDD'99 dataset. The major attack classes are Denial of Service (DOS), Probing attacks (Probe), Remote to Local (R2L), and User to Root (U2R).UMAP (Uniform Manifold Approximation and Projection) based on learning techniques, is an algorithm for dimension reduction invented by Leland McInnes et al. which is built for calculating exponential probability based on Riemannian geometry. UMAP has no computational restrictions and it mostly conserves the global structure. The Chi-square technique is used for calculating the chi-square statistics between the target variable and all features and examines the relationship between them. The target variable is significant and crucial if the dependency occurs, otherwise, that variable can be discarded. Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. Based on deep learning an effective network attack detection strategy is presented. For better accuracy, multiple parameters were tested. All the models were applied in multiclass classification which shows the achievements of the models and binary classification showing the performances using an effective network attack detection strategy The Attention-PCA model is able to learn detailed features from the dataset. The experimental results show that the proposed attack detection strategy achieves higher performance and it can also reduce the false negative rate. [10]

Suad Mohammed et al. (2018) introduced the Spark-Chi-SVM model for intrusion detection. Big Data techniques are used in IDS to deal with Big Data for an accurate and efficient data analysis process. In this paper, the researchers have introduced the Spark Chi-SVM model for IDS. On Apache Spark Big data platform using a support vector machine we have to build an intrusion detection model and for the feature selection in the model, we have used ChiSqselector.The model uses the KDD-99 dataset for the testing of the model. A comparative study of the Chi-SVM classifier and Chi-Logistic Regression classifier is done. The proposed model was implemented using Apache Spark's machine learning library: Mlib in Scala Programming. On the basis of the above-proposed method, Chi-SVM is said to be the best classifier. The classification process is highly complex and takes a long time in the case of Big Data due to its high dimensionality. Therefore, ChiSqselector is used for selecting the features and to classify the data into normal or attack SVM-With-SGD is used for the model. The results of the experiment showed that the model has high-performance speed, reduces the training time, and is efficient for Big Data. Future research can expand the model to a multi-class model that can distinguish various attacks kinds. [11]

Mirsky et al. (2018) proposed a plug-and-play NIDS which can detect local networks in an efficient manner without any supervision and online. Using ANNs as NIDS the author's main purpose is to overcome the shortcomings of the typical approaches which include supervised learning requiring a strong CPU for testing and training a model and are limited to known attacks, therefore updating NIDS is necessary for newly trained models. Therefore to overcome the above challenges Kitsune aims for online unsupervised learning which is to

ISSN 2394 - 7780

be a lightweight NIDS and an ANN. It works in real time and is designed to be run on network routers. The Kitsune model consists of a feature extractor (e.g., statistics of the channel), a packet capturer and parser (e.g., meta information of the packet), and a feature mapper to merge features into a wider set that the anomaly detector could process. To differentiate between abnormal and normal traffic patterns uses auto-encoders are defined as an ensemble of small ANNs as the core algorithm for anomaly detection. In an online fashion to track the patterns of every network channel, the feature extractor works efficiently. In a hash table damped incremental statistics are maintained which is useful for the above framework to obtain efficiency. To attain the results experiments were made on the IoT network and IP camera surveillance network deploying several attacks. Kitsune was run on a Raspberry PI and desired results were obtained. The framework gets limited if the network is already contaminated.[12]

2) Graph Models: The below section represents network flows and events using graphs to extract features.

Milajerdi et al. (2019) proposed HOLMES, a system for the APT (advanced persistent threat) campaigns for detecting activities. APT is basically an attack campaign in which an intruder, or team of intruders, which establishes an unauthorized enduring on a network in order to mine highly sensitive data. Basically, the system collects the data from hosts for example Windows/Linux and other available resources to build detection signal. At an advanced level, during attack campaigns, the techniques are developed in such way that they resist the correlation between the suspicious information flows. The system is designed in such way that it enables the generation of alerts semantically close to the activity steps of APT attackers. For map audit logs to APT stages usage of MITREs, ATTACK framework2 is done by the authors to bridge the gap between the system call events and high-level kill chain stages.MITRE's Attack framework defines around 200 behavioral patterns, techniques, tactics, and procedures (TTP) which can be captured on audit data. For the analysis interpretation, HOLMES enables the construction of a high-level scenario graph (HSG) which refers to TTP. The HSG delivers a comprehensive assessment of the attacker's activities.The authors evaluated the system using a dataset corresponding to streams of audit data from 7 computers which are Linux and Windows based with injected attacks. They used a previously created system called SLEUTH to do real-time analysis and detection on the audit data streams after publishing them into the Kafka stream processing server. Attacks constituted less than 0.001% of the audit data volume. The author was able to distinguish between the benign activities and the APT campaigns injected into the dataset. The main limitation of HOLMES is that it has assumed that the initial attack must originate outside the trade using means such as remote network access as well as an auditing system and the OS kernel is well grounded. Furthermore, only a limited percentage of audit log sources were utilized for the analysis. [13]

AhlemAbida et al. (2020) has proposed the given paper based on an ML graph algorithm and cloud computing using Microsoft Azure which detects real-time different attacks as early as possible. The given paper has implemented the K2 algorithm which is the graphical method ML to classify attack techniques. Graphs can display multiple types of information as well as can lead to high communication power which is the key feature of machine learning thus resulting in better analysis and allowing people to navigate through data. Therefore the proposed system is evaluated using the MAWILab intrusion detection dataset. MAWILab dataset is used to evaluate intrusion detection methods which consist of packet traces from the MAWI archive in which the dataset is updated daily to include new traffic from applications and anomalies. The proposed system has the following steps: Data Collection, Data Processing and Intrusion Detection based graph in real-time. The best IDS is the one that gives the minimum value of a false positive rate. The proposed system is said to be distributed and scalable due to the performance of Microsoft Azure. We intend to upload our dataset to Azure blob storage using Microsoft Azure as a unified cloud environment. We employ the K2 algorithm, a machine learning system that uses graphics to analyze intrusions. The proposed method showed great results due to the Apache Spark structured streaming engine and graphical ML.MAWILAB dataset is used for the insights of the real-time traffic data .K2 ML graphical machine learning algorithm is implemented in Azure cloud within a distributed environment. Microsoft on the basis Horton-works Data platform has designed HD-Insight which is a data analysis and data processing service. In the paper to analyze our preprocessed data, we have implemented the K2classification algorithm in Spark MLlib in a distributed manner which is a part of Bayesian networks. Its principle is as follows: The K2 algorithm requires a well-defined order of variables. From this order, the K2 algorithm determines the dependency links between these attributes now called variables or nodes of the graph to be built. Spark Structured Streaming starts reading parquet files in a real-time manner, an acyclic-oriented graph resulting from the application of the K2 algorithm. Three measures are associated at each node of the graph: the measure of necessity, possibility, and probability. We transform the initial graph into a junction tree, then infer the observed pieces of evidence in the tree and deduce the evidence of the variable Label. Next step, show prediction results: for each value of the Label variable, three measurements are assigned: a probability

ISSN 2394 - 7780

measure, a possible measure, and a necessary measure, display the outcome in order to choose the Label variable value with the highest informative probability and one or more predefined thresholds. The K2 ML graphical algorithm showed great performance for the data insights. The results of the prediction and the outcome of the decision-making process showed 98.76% of accuracy and the false positive rate was 0.029%. The limitation of the above-proposed system is it was useful for the small cluster data. The above-proposed system was limited to small cluster data .cluster. In the future, we need to use multiple clusters to achieve faster results and combine two or more datasets. [14]

3) The Frameworks: Explores methods for normalization and event aggregation to achieve contextual awareness. Two researches in this final group emphasize ensuring superior frameworks to meet the needs of security analytics in the future.

Sugandh Seth et al. (2021) proposes a novel approach for a time-efficient and smart Intrusion Detection System. The smart IDS is highly time efficient for predicting intrusions and deploying security systems for analyzing dynamic and current traffic trends. By lowering the model's complexity the Hybrid Feature Selection approach is being used without affecting the attack prediction performance to reduce the prediction latency. Using the CIC-IDS dataset a fast gradient boosting framework i.e. Light gradient Boosting Machine is used to build the model. For deep insights into variations in network parameters depth analysis of the network, parameters have been performed during the malicious session. For time-efficient and smart intrusion, the paper proposes a novel approach. The model uses the CIC-IDS 2018 dataset which is the most recent dataset for the intrusion detection system. This dataset is a massive dataset that reflects modern-day attacks and is a mixture of benign and malicious traffic. Using Random forests the feature selection is done. For dimensionality reduction Principal Component analysis (PCA) is used. In this paper machine learning algorithms, namely, Random Forests KNN and Light-GBM are compared for Recall, Accuracy, F-Measure, Sensitivity, Specificity, Prediction Latency, and the model Building Time. The system can be trained and tested using ML algorithms to detect abnormal traffic trends. The analysis was made that using this approach that by reducing the model's complexity prediction latency reduces due to the lesser number of model's input. 65% of attackers use port 80, and compare to malicious sessions benign sessions have a higher rate of flow per bytes/s, Besides, an in-depth analysis of the network parameters of benign malicious sessions was conducted on the CIC-IDS 2018 dataset. Key findings of the analysis were: Port 80 was utilized by 65% of attackers to carry out the attack, the flow duration of benign sessions is higher than that of malicious sessions, and no packet was returned to the attacker from the server in 50% of the hostile sessions. [15]

Hakim Azeroual et al. (2022) primary goal of the proposed method is to present a methodology for developing, testing, and deploying a deep learning (DL) model for the anomaly, abuse, malware, or bot-net detection in order to create or enhance an intrusion detection system (IDS) within the NTMA framework. This takes into account processing speed and reliability issues. The above-said process will be used for cyber-security and to extract conclusions for better results for intrusion detection systems using Big Data techniques for massive data. The framework for IDS is based on Network Traffic Monitoring and Analysis (NTMA) studies.CNN algorithm is implemented for the model using the CSE-CIC-IDS2018 dataset. The testing of data involves data split i.e. train set for training of the model and test sets for analyzing the dataset. The use of ML in IDS results in great efficiency which exceeds up to 92%. After optimizing our deep CNN model using training data and validating on it data, CNN performed exceptionally well on training data and the accuracy was 99%. Model accuracy was down to 83.55% on validation data after 50 iterations, and gave a good accuracy of 92% after 30 iterations. Thus, it can be deduced that the ideal number of iterations for this model to conclude is 30. It is still vital to improve and test this model on actual networks. [16]

VI. CONCLUSION

The intrusion detection cycle still requires security professionals. The idea of a standalone, completely automated solution seems difficult to realize. Deploying a system with numerous IDPS technology types enables more precise and thorough performance. Applications for real-time monitoring demand distributed stream processing. In stream processing systems, high availability, fault tolerance and fail recovery are crucial. In the security domain handling a huge volume of data for the extraction of useful information requires the use of ML and seems best for real-time threat awareness.IDS are a source of big data because they produce enormous amounts of data, most of which is heterogeneous. The IDS should be able to distinguish the various data structures and communication protocols that define big data. Big Data analytics (BDA) can more quickly sort through enormous amounts of data, enhancing the efficiency and effectiveness of heterogeneous systems.

ISSN 2394 - 7780

VII. REFERENCES

[1] Koca, M., Aydin, M. A., Sertbaş, A., & ZAİM, A. H. (2022, January 27). A new distributed anomaly detection approach for log IDS management based ondeep learning. Tübitak Academic Journals.

[2] Luís Dias, Miguel Correia, (2019), Big Data Analytics for Intrusion Detection: An Overview, IGI Global

[3] Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2, 20 (2019)

[4] Alshamy, Reem&Ghurab, Mossa. (2020). A Review of Big Data in Network Intrusion Detection System: Challenges, Approaches, Datasets, and Tools.

[5] Wang, Lidong. 'Big Data in Intrusion Detection Systems and Intrusion Prevention Systems.' Journal of Computer Networks 4.1 (2017): 48-55.

[6] T. Saranya, S. Sridevi, C. Deisy, Tran Duc Chung, M.K.A.Ahamed Khan,Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review,Procedia Computer Science,Volume 171,2020,

[7] Khraisat, Ansam, et al. "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges - Cybersecurity." SpringerOpen, 17 July 2019, cybersecurity. springeropen. com/articles/10.1186/s42400-019-0038-7.

[8] Tiwari, Mohit & Kumar, Raj & Bharti, Akash & Kishan, Jai. (2017). Intrusion Detection System. International Journal of Technical Research and Applications. 5. 2320-8163.

[9] Reghunath K,"REAL-TIME INTRUSION DETECTION SYSTEM FOR BIG DATA",2017, International Journal of Peer to Peer Networks (IJP2P).

[10] FatimaEzzahraLaghrissi, Samira Douzi2, Khadija Douzi1, and BadrHssin, "IDS attention: an efficient algorithm for intrusion detection systems using attention mechanism",2021, Journal of big data

[11] Suad Mohammed Othman1, FadlMutaher Ba Alwi1, Nabeel T. Alsohybe1 and Amal Y. Al Hashid "Intrusion detection model using machine learning algorithm on Big Data environment",2018, Journal of Big Data.

[12] Mirsky, Y., Doitshman, T., Elovici, Y., &Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. In Proceedings of the Network and Distributed System Security Symposium (NDSS) 2018.

[13] Milajerdi, S. M., Gjomemo, R., Eshete, B., Sekar, R., &Venkatakrishnan, V. (2019). Holmes: Realtime apt detection through correlation of suspicious information flows. In Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P).

[14] AhlemAbida, Farah JemilibUniversite de Sousse, ISITCom, 4011, Hammam Sousse, "Intrusion Detection based on Graph oriented Big Data Analytics", 2020, 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.

[15] Sugandh Seth, Gurvinder Singh, and Kuljit Kaur Chahal, " A novel time-efficient learning based approach for smart intrusion detection system",2021, Journal of Big Data.

[16] Hakim Azeroual, ImaneDahaBelghiti, and NaoualBerbiche, "A Framework for implementing an ML or DL model to improve Intrusion Detection Systems (IDS) in the Ntma context, with an example on the dataset (CSE-CIC-IDS2018)", 2022, ITM Web of Conferences 46.

[17] MukeshChoudhary, Dr. Manish Shrimali ,(2019) ,Intrusion Detection on Big Data using Machine Learning:A Review, IJRAR- International Journal of Research and Analytical Reviews

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

USE OF PREDICTIVE ANALYSIS IN FINANCIAL MARKET AND ITS IMPLEMENTATION USING ALGO THERMIC TRADING

[1]Akshat Bhatia and [2]Fatima Shaikh
[1]Department of Computer Science, Jai Hind College (Autonomous), Mumbai, India
[2]Assistant Professor, Department of IT, Jai Hind College (Autonomous), Mumbai, India

ABSTRACT
This literature review paper intends to focus on applying predictive analytics in the stock market domain and suggest the vital use of Algorithmic Trading in it. The stock market is highly dynamic and the investor is bound to face the critical pricing dynamics which define the profit that an investor may envisage/have the potential to earn in the future. These pricing dynamics are affected by numerous factors that are decoded and their effects are further predicted by numerous ML algorithms, some of which are discussed in this paper.

Keywords: Algo Trading, Machine Learning, Stock Market, Predictive Analytics.

I. INTRODUCTION

Stock Market is the main predictor of the well-being of the economy. It helps us in determining the different economic cycles/phases that our country is going through. It has been a major wealth-creator tool for millions of people in our country. The thin line that demarcates between it being a gamble or a factual investment source is the strategies and the tools that help in identifying a pattern and makes the predicting of stocks trading through it possible. As we advent towards the era of Web 3.0, we have already embraced the transformation and the impact that machine learning is having not only in the IT sector but in every sector whether it be Pharmaceutical, Financial, Transportation, etc. Before disclosing the real application of it in the domain, our paper suggests, let's understand the basic interpretation of the jargon used as well as its branches in our field of study

A. Machine Learning

Machine learning (ML) is a research segment dedicated to exploring and developing ways to "learn." i.e how data can be used to have an increment in the performance of specific tasks. It is said to be a sub-segment of artificial intelligence. Machine learning algorithms construct models based on sample data, known as training data, to predict and decide without being particularly programmed to do so Machine learning algorithms are used in a variety of applications, including medical, email filtering, speech recognition, agriculture, and computer vision, where traditional algorithms would be difficult or impossible to design. Despite the emphasis on using computers to predict outcomes, not all machine learning is statistical learning. There are various types of machine learning algorithms that are used for very specific use cases, but three main methods are currently in use.

i) Supervised Learning

Supervised learning is one of the most fundamental types of machine learning. This type trains machine learning algorithms on labeled data. Although adequate data labeling is required for this method to work, supervised learning has the potential to be quite effective in the right circumstances. In supervised learning, a small training dataset is provided to an ML algorithm. This training dataset, a subset of a larger dataset, provides the algorithm with a general understanding of the problem, the solution, and the data points that must be processed. The characteristics of the training dataset are also highly comparable to those of the final dataset, providing the algorithm with the labeled parameters it requires to solve the problem. In essence, the algorithm determines relationships between the supplied parameters to establish causal links between the variables in the dataset. After training, the algorithm understands how the data works and how inputs and outputs are linked. This solution is available for use with a final dataset that has undergone the same training as the training dataset. This means that when trained on new data, supervised machine learning algorithms improve even after they are put to use.

ii) Unsupervised Learning

Unsupervised machine learning has the advantage of handling unlabeled data. Since no human effort is required to make the dataset machine-readable, programs can handle much larger datasets. The labels enable the system to precisely identify any link between any two of her data points when learning under supervision. Contrarily, unsupervised learning produces hidden structures because there are no labels to cope with. These covert structures are credited with giving them their adaptability. Unsupervised learning algorithms can adjust to data

ISSN 2394 - 7780

by dynamically modifying the underlying structure as opposed to a fixed, preset problem. This provides greater opportunities for post-deployment development than supervised learning methods.

iii) Reinforcement Learning

Reinforcement learning draws its primary influence from how people naturally learn from data in their daily lives. It features algorithms that get better on their own by making mistakes and learning from them. Unfavorable outcomes are discouraged or "punished," while positive outcomes are "encouraged" or "improved." Based on the psychological concept of conditioning, reinforcement learning works by placing algorithms in settings with interpreters and reward systems. After each algorithm iteration, the interpreter assesses the output result before providing it to the user. The interpreter rewards the algorithm to improve the solution if the program discovers the right answer. The method should iterate until a better result is found if the results are undesirable. Most of the time, the effectiveness of the results is closely related to the reward system. Use case B is typical for reinforcement learning. The answer is not definitive when determining the shortest path between two spots on his map. Use percentage-based efficacy ratings as an alternative. The algorithm receives more compensation the greater this percentage is. As a result, the program is taught to offer the best answers for the greatest returns.

B. Predictive Analysis

Data mining, predictive modeling, and machine learning are a few statistical methods used in predictive analytics to examine current and historical data and create predictions about events that are yet to occur or are unknown. [1] Predictive models are used in business to find hazards and opportunities by analyzing patterns in historical and transactional data. To enable the assessment of risks or opportunities associated with a certain set of variables and to inform future trading decisions, models represent the interactions between various aspects. [2] The primary functional outcome of these technical methods is predictive analytics, which gives each individual a predictive value (probability) (customer, employee, medical patient, product SKU, vehicle, component, machine, or other organizational units). a function of an organization that has a broad reach.

Let’s now look at some commonly used financial tools that are frequently used in the financial space

C. Technical Analysis

Before taking a position or making an investment, this area entails using various statistics, patterns, trendlines, indicators, etc. to help make wise selections. Following this transaction, technical analysis enables the trader to get ready for unexpected market reversals, enabling them to make money regardless of market trends. Both conventional assets like commodities, foreign exchange, and stock indices as well as digital assets like Bitcoin and other cryptocurrencies can be the subject of technical analysis.

The following are a few of the tools used in analyzing the stock prices

i) Moving Averages

An indicator that displays the average price of an item over a given time period is called a moving average, and it is superimposed on a price chart. On daily, weekly, or even longer periods, moving averages can be either short-term or long-term. Moving averages are frequently used by traders and investors to determine whether a trend in an asset class is changing as well as to identify levels that serve as support or resistance. A death cross or gold cross occurs when the short-term moving average breaks down or crosses above the long-term moving average. These terms are typically given to the subsequent price movement. A gold cross is optimistic and depicts the wealth investors are anticipated to make from the trend after such an event, whereas a death cross is bearish and frequently indicates an impending slump for an asset.

The following are some of its variants that are used in the paper

a. Simple Moving Average

By combining the current prices and dividing the result by the total number of periods in the calculated average, a simple moving average, or SMA, is produced. For instance, you might add up the closing price of a security over a range of periods, then divide that sum by the same range of periods. The short-term average responds swiftly to changes in the price of the underlying security, while the long-term average does so gradually.

b. Exponential Moving Average

The most recent data points are given additional weight and significance by the Moving Average (MA) version known as the Exponential Moving Average (EMA). Similar to an exponential moving average, exponentially weighted moving averages are more responsive to recent price changes than simple moving averages (SMAs).

ISSN 2394 - 7780

c. Weighted Moving Average

The EMA is a variant of WMA that gives more weight to more recent data items than to those from the distant past since they are more pertinent. The sum of all the weights should be one (or 100%). While the weights in the EMA rapidly decrease between data points, the weights in the SMA are evenly distributed. (Rather than every preceding weight being 1.0 smaller than the weight in front of it, there may be a 1.0 difference between the first two period weights, a 1.2 difference between the two periods following those, and so on.)

ii) The On-Balance-Volume Indicator (OBV)

On-Balance-Volume is a technical analysis indicator that tracks the positive and negative volume flow of an asset over time. an indication of pulse that counts both positive and negative flow. Joseph Granville was the one who came up with this indicator. He thinks that if the volume increases dramatically and the stock price does not change considerably, the price will eventually climb, and vice versa if the volume does not change significantly and the stock price does.

iii) The Accumulation/Distribution line (A/D line)

One of the most popular technical analysis indicators for selecting when to buy and sell stocks is accumulation and distribution. Marc Chaikin created this indicator to track the overall money movement into and out of assets. The Cumulative Money Flow Line was the previous name for this indicator.

By identifying whether investors are purchasing (accumulating) or selling (dividing) a specific stock, this indicator seeks to gauge supply and demand. This indicator illustrates how factors such as supply and demand impact prices. Distribution and accumulation might follow price changes in the same direction or on the opposite slope. The indicator indicates that there is buying pressure and that a price reversal is conceivable when the price of the asset is in a downtrend and the cumulative distribution line is in an uptrend. Similar to the previous example, if a security's price is rising but the cumulative distribution line is falling, the indicator shows that there is selling pressure and that the price may be about to reverse.

iv) The Average Directional Index (ADX)

A trend indicator used to gauge the strength and momentum of trends is the Average Directional Index (ADX). Trading in the trend's direction lowers risk and boosts possible earnings.

When determining whether a price trend is strong, the Average Directional Index (ADX) is used. Based on a moving average of price range expansion over time, the ADX calculation is made. 14 bars are the default setting. To distinguish between trended and non-trended conditions, values are crucial. If the ADX value is greater than 25, the trend is strong enough for trend trading. Similarly, the ADX indicates avoiding the trend trading technique if it is below 25.

v) Aroon

A technical analysis indicator called the Aroon Indicator is used to determine whether a security is trending. These are typically used to determine whether trends are likely to change course. This indicator tracks how long it takes a price to reach a high or low over a given period of time. The Aaron Up Line and the Aaron Down Line make up the indication. The strength of an uptrend is measured by the Aaron Up Line, and the strength of a downtrend is measured by the Aaron Down Line. Below is a list of the three stages for spotting uptrend indications.

- First, the Aroon lines will intersect. Aroon-Up generates an upward signal when it crosses over Aroon-Down. This demonstrates that fresh highs are starting to take center stage in the recent chronology relative to new lows.
- Second, the point at which the Aroon lines cross 50.
- Thirdly, when a certain line reaches 100. Aroon up reaches a value of 100 whereas Aroon down stays at a lower value.

If you try to reverse engineer these stages, you'll get a negative signal.

D. Fundamental Analysis

Fundamental analysts look at all variables that can influence a security's value, from macroeconomic ones like the state of the economy and the state of the sector to microeconomic ones like the efficiency of corporate governance. The final objective is to identify a figure that investors may use to assess whether the investment is being undervalued or overpriced by other investors by comparing it to its existing price.

The following are some tools that are used in Fundamental Analysis

ISSN 2394 - 7780

i) Earning Per Share

The profit attributable to each firm share is expressed as EPS.It is calculated by dividing the entire revenue or profit of the company by the number of shares that are currently outstanding. To express it mathematically:

EPS = Corporate Net Income After Tax / Total Shares Outstanding

Higher EPS translates to higher returns for investors because it is a measure of a company's health. EPS might be diluted or made basic. Diluted EPS comprises shares held by the corporation as well as shares that may be granted to investors in the future. Basic EPS includes the total number of shares outstanding. In addition, there are three types of EPS: trailing, current, and forward. His real earnings per share for the most recent fiscal year that has concluded are his "trailing earnings per share." The project's EPS for the current year is the current EPS. An estimate of EPS for the following year is known as forward EPS. To decide which firms to invest in, an investor might compare its EPS with that of another company in the same sector. However, higher EPS might also result in decreased profitability and a return to normalcy of a higher stock price.

ii) Price per Equity

One key instrument in the fundamental stock analysis is the P/E ratio. This displays the price the corporation paid for the stock price. This will show you whether the stock component is worth the money you paid for it. By dividing the stock price by the EPS, one can determine the P/E ratio. The P/E ratio is 10 for a company with a share price of 50 rupees and an EPS of 5. The likelihood of strong gains relative to the stock price is indicated by a low P/E ratio. He has a low price per share in relation to earnings if his P/E is low. This indicates that there is room for future gains and that the stock is undervalued. If the PER is high, the inverse is true.

P/E can be classified as:

1. Trailing P/E, or P/E for the most recent twelve months

2. The forward P/E ratio, or the P/E ratio for the following year

Profits may be affected if the forward PER is greater than the trailing ratio. If the prospective PER is less than the trailing PER, the company's earnings may rise. P/E has a different lucrative value for different investors. How much you are willing to pay for the company's profits are shown by the price-to-earnings ratio. Your driving force can be different from other investors.

iii) Return on Equity

Return on Equity, or RoE, gauges a company's effectiveness in providing a profit for its investors. It is derived by dividing equity capital by net profit after tax. If a company has a capital of Rs. 50 lakhs and revenue of Rs. 5 lakh this year, its ROE is Rs. 5000000 / Rs. 500000, or 10%. A percentage is used to represent ROE. A high ROE indicates the business is effective. This enables the business to increase its profitability without requiring extra funding. However, businesses with fewer assets may also have higher ROE. As a result, not all businesses with high ROE are good candidates for investment. It is useful to contrast the ROE of businesses operating in the same sector. An ROE in the 13–15 range is regarded as favourable.

iv) Price-to-Book (P/B) Ratio

The price-to-book ratio, commonly known as "equity," contrasts the book value and market value of a stock. The book value of each asset is its acquisition cost minus cumulative depreciation. The P/B ratio is derived by dividing the most recent closing price by the most recent quarter's book value per share. It shows what is left of the company after all debts have been paid off and assets have been liquidated. If the P/B ratio is less than 1, the stock is cheap. If the price is greater than 1, the stock is overvalued. The P/B ratio is significant because it shows whether a company's net worth corresponds to the stock's market value. This ratio is more prominent in businesses with substantial cash balances, such as insurance, banking, investing, and financial businesses. The P/B ratio does not benefit businesses with substantial fixed assets or high R&D expenditures.

Before delving into the different machine learning methodologies, the humongous amount of data generated due to the advent of the technological era, comes with the problem of saving and storage. Some Big data storage techniques especially HADOOP are capable of storing vast amounts of unstructured, semi-structured, and structured data that companies retrieve and process in order to predict future stock prices. This stored data can be further put through different Machine learning models in order to deduce reasonable insight that would further help in predicting future prices.[7]

ISSN 2394 - 7780

II. METHODOLOGY

The following are the Machine Learning techniques used in our reviewed papers

A. K-Mean Cluster analysis

It is an unsupervised machine learning methodology that transforms the data from the K-cluster, where each prototype is grouped according to the mean that is the closest apart, into partition methods like the K-mean Partition. These clusters here stand in for the stock's underlying map. The representative feature that accounts for the cluster's shared chart patterns is then selected from each cluster. Up until a cluster is formed with the best underlying stock, clustering attempts are repeated numerous times. The main feature is used to profile each cluster. The top three stocks are then chosen from the cluster of crucial attributes with the highest ratings that best reflect the possibility of the stock price increasing. (Hargreaves, 2019)

B. Logistic Regression

It is a method of supervised machine learning that is applied to classification issues where the explanatory variable is either continuous or discrete and the response variable is binary. The logistic regression is represented as logit (p) =b0+b1X1+b2X2.....bk Xk.logit (p) =log (p/1-p). Here, p is the likelihood that the target response class will succeed, and X is the set of explanatory factors with the explanatory variables' b0, b1, b2, etc. coefficients. (Hargreaves, 2019)

C. Principal Component Analysis(PCA)

PCA is useful for filtering out redundant characteristics and for locating the pertinent features (variables). We employ the dimension reduction technique to condense the excessive number of variables into a small number of elements. For this, we first gather all 24 technical variables, record their initial eigenvalues, and calculate each technical variable's variance in relation to the overall variance. A general rule of thumb is that the estimated variance should be more than 70%, and the factors are filtered out on this basis (along with the requirement that the eigenvalue be greater than 1), respectively. Once the components have been identified, the Cronbach Alpha Coefficient is used to assess their reliability. For it to be reliable, this value must be higher than 0.70. (Hargreaves, 2019)

D. Support Vector Machine(SVM)

The fundamental idea is to transform the input data using a kernel function into a high-dimensional feature space, which then transforms the data into the necessary forms, like a 3D representation. Next, maximize the margin between classes to find the optimum hyperplane.. (Bilal Elmsil, 2021)

E. Random Forest

It works by creating a huge number of de-correlated decision trees, each using a different bootstrapped training sample and a section of the feature space. The results of individual decisions are aggregated using the majority vote rule and are binary (i.e., 1 or 0). Due to its low variance and low bias (caused by the distribution of data among several decision trees and their training on that particular data), random forest often exhibits overfitting (as the data is being distributed at random). The depth of each tree, the number of trees, and the maximum number of features to take into account while determining the best split are the three adjustable factors in the random forest. (Bilal Elmsil, 2021)

F. K Nearest Neighbor

It is an instance-based learning algorithm that is non-parametric. When a new observation, X0, is made, the distance calculation function is used to find the K points (designated No) in the training dataset that is closest to X0. This new instance's class name is then predicted to be the most prevalent class among N0. In order to calculate the majority, either uniform weights (i.e., weighting all N0 points equally) or weights that are inversely proportional to point distance are used (closest neighbor has greater influence). In this study, the k-d tree is employed to implement the Euclidean distance metric. The factors that can be adjusted are the number of neighbors K and the voting method (distance-based or uniform). (Bilal Elmsil, 2021)

G. Ensemble Learning

It is similar to random tree sampling, where there are multiple subsets of the data(here only columns are present), and this subset can be anything (eg SVM, KNN, LOGISTIC REGRESSION). The individual decisions made by each classifier in this set are often merged using both weighted and unweighted voting. Each ensemble must surpass the individual classifier, therefore the classifier must be accurate (better than guessing at random) and varied (make uncorrelated errors). However, in financial forecasting, it is fairly typical to have highly connected forecasts between classifiers. Because of this, the study employs a consistent voting ensemble that only takes into account predictions that have the support of all classifiers. (Bilal Elmsil, 2021)

ISSN 2394 - 7780

H. Adaptive Boosting

With the use of this method, poor learning algorithms can be changed into ones that attain arbitrarily high accuracy. Adaptive boosting puts this idea into practise by sequentially applying a learning algorithm to reweighted versions of the training data. The data that was incorrectly classified during the previous iteration is given additional weight in each boosting round, which causes the classifier to concentrate on cases that were challenging to categorise in the past. The final prediction is then created by adding predictions from the series of all the weak learners using a weighted majority vote. The ideal number of boosting iterations is still a variable that can be adjusted. (Bilal Elmsil, 2021)

I. LASSO AND RIDGE

Ridge and LASSO, two regularised or penalized techniques to logistic regression, are used to improve generalization and prevent overfitting.. Here Ridge helps in determining the best fit line which would also overcome the problem of overfitting and simultaneously reduce the cost function. Formula--$\Sigma(y- \hat{y})^2 + \lambda*(Slope)^2$.Here the $\lambda*(Slope)^2$ tends closer to zeros. LASSO on the other hand not only reduces the cost function but also helps in the selection of the important parameters/feature among the data set, (as the non-imp dataset feature will eventually become zero in this).Formula--$\Sigma(y- \hat{y})^2 + \lambda*|Slope|$. Here the value of λ is determined through the cross-validation method and hence is important. (Bilal Elmsil, 2021)

III. DATASET REVIEW

The literature papers reviewed had the following data sets

i) **Stock prices from the Australian stock market**
ii) Historical prices from the Ghana stock market
iii) News headlines, forum discussions, and tweets regarding the companies in the picture were extracted through different APIs.
iv) Google trends
v) Historical prices of precious metals list in the commodity market of the US

The prediction of future price movement of Stock datasets that were related to the historical prices was done through the Machine Learning techniques mentioned above.

It was noticed that in order to find the trends of price movement in terms of precious metal commodities, the technical analytics tools such as Simple Moving Average, Weighted Moving Average, and Exponentially Weighted Moving Average were capable enough to address long-term price movement with accurate results. Amongst them, SMA was the most successful in order to predict precious metal prices. (Isaac Kofi Nti, 2020)

On the other hand, the deduction of factual information related to the future price movement through datasets based on trends, Tweets, news, etc required a different approach.

- **Data Preprocessing**

The duplication among the data (like the same news being tweeted and also being the headline of the news forum) was removed through the cosine function. Then the data was passed through the process of slicing i.e removing the unwanted characters among the dataset like URLs, special characters, etc. Then the polarity and subjectivity of the data were measured. Polarity-between minus 1(negative sentiment) and plus 1(positive sentiments). Subjectivity-0(highly objective) and 1(highly subjective).All these factors were taken into consideration while the stock price was being calculated. Returns were calculated by the formula—(Closing Price (i+d)/Closing Price (i)) – 1. Here 'i' stands for a given date, and 'd' stands for the target date(of whose returns have to be calculated). If the Returns>0 then it implies a rise in the prices on a subsequent day, or if the returns<0 then it implies a fall.

- **Data Price Trend Retrieval**

For the current procedure, the multi-layer perceptron (MLP) ANN algorithm was used. A network of connected parts called an MLP receives input, acts on it, and then passes it on to the next layer. The MLP studies a function f(.): R^d->R^o where d and o represents the dimension of the input and output dataset respectively. A combination of accuracy matrics was adopted for the model's performance evaluation. Specifically, Sensitivity, Accuracy, Root square mean error, and Mean absolute percentage error.

IV. IMPLEMENTATION

The implementation of the machine learning techniques is clearly evident in the Algothermic and High-frequency trading that one indulges in order to buy the stock with the potential of giving a positive return(earlier

ISSN 2394 - 7780

gauged by the use of different machine learning techniques used on quality datasets of the study). Considering various advantages, AT and HFT help to reduce liquidity volatility. Increasing the effectiveness of price discovery, it assists in reducing the bid-ask spread and lowers the cost of trading (through the presence of high-frequency quotes). This ratio is more prominent in businesses with substantial cash balances, such as insurance, banking, investing, and financial businesses. The P/B ratio does not benefit businesses with substantial fixed assets or high R&D expenditures.

Several Drawbacks-HFT and AT significantly increase trade volumes and price volatility. The majority of transactions are swiftly made and completed by the HFT and AT-backed investors, generating an irregularity in the investor's universe, which negatively affects the typical investors who use traditional methods (which are slow). Aside from this, unethical practices like spoofing and quote stuffing is frequently used in HFT and AT. Spoofing is when someone places some orders and then immediately cancels them, tricking other investors into believing that there has been an increase in volume. Quote Stuffing is similar to spoofing in that it involves placing a large number of buy and sell orders, which increases the quoted and effective spread as well as the short-term volatility and illiquidity. Speaking of certain HFT and AT regulations by the European and US markets, MiFID ‖ enables the regulatory entities to charge higher transaction fees or taxes on businesses with high order cancellation rates. In order to assure possible risk reduction and that their trading operation has adequate capabilities and resilience, it is also necessary for businesses to have an efficient information system and risk management system in place. The US SEC created the Consolidated Audit Trail (CAT) system and the Market Information Data Analytics System (MIDAS) to retain transaction records and guarantee there were no errors by a careful examination of these records. HTF and AT in Emerging markets (mostly composed of retail investors) could impact retail investors in terms of trading fairness; a lot of unfair trade strategies are being implemented by HFT and AT. The lack of effective technology-driven regulators in such markets could also be a reason for such malpractice. Plus due to the high speed of such trades, HFT and AT absorb news affecting the trade in milliseconds thus depriving the retail investors of opportunities and making much slower and ineffective profit fewer trades. (Siyuan Yan , 2022)

There are some actionable recommendations that can be implemented. The robustness of the Regulators in terms of surveillance over AT and HFT should improve technologically. The entities using AT and HFT practices should register themselves with the regulators and disclose their trading algorithms with them. The regulators should closely work with stock exchanges and collect data concerning AT and HFT. They should have the power to temporarily cease algorithmic trading in the event of extreme market irregularities (Siyuan Yan , 2022)

V. RESULTS

Duration-Short Term

Table 1

Panel A- 1 Day Ahead Return		Precision
1	Ridge Regression	0.45209
2	Lasso Regression	0.45489
3	SVM	**0.50044**
4	K-Nearest Neighbour	0.46289
5	Random Forest	0.49405
6	Adaptive Boosting	0.48832

Table 2

Panel B- 5 Day Ahead Return		Precision
1	Ridge Regression	0.50469
2	Lasso Regression	0.50306
3	SVM	**0.51901**
4	K-Nearest Neighbour	0.49342
5	Random Forest	0.51341
6	Adaptive Boosting	0.51613

As we can see in Tables 1 and 2, which are comparing different Machine learning algorithms on the basis of precision, the results indicate that the SVM was the most efficient in terms of the 1-day ahead and 5-day ahead returns [1]

ISSN 2394 - 7780

Duration-Long Term

1-Month Ahead		Returns
1	Logistic Regression	11.65%
2	K-Mean Cluster	8.88%
3	PCA	**23.87%**

$30,000 was invested in the top 3 stocks selected by the above mention algorithms and the following were the results that were found-

1) Logistic Regression- It selected the stocks MIN, TWE, and A2M and at the month's end, the portfolio was worth $33,495 which is a return of 11.65%.
2) K-Mean Cluster- It selected the stocks MIN, TWE, and BAL and at the month's end, the portfolio was worth $32,663 which is a return of 8.88%.
3) PCA- It selected the stocks A2M, IMF, and BAL and at the month's end, the portfolio was worth $37,160 which is a return of 23.87%.

So we can conclude from these results that the PCA algorithm's formulated portfolio is successful in generating the most returns in comparison to others i.e Logistic regression and K-Mean Cluster.

VI. CONCLUSION

In the financial market, there is a different problem that an investor face in the financial stock market. Along with it, there are different factors that each problem has

There are two problems that we tried to address in our paper:-

1) Stock selection Error for Long term portfolio
2) Stock price prediction in the shorter term

These two problems cater to two different types of investor mindset; One who believes in holding the stock as a financial asset and reaping the benefits of the power of cumulation and the other who believe these stocks to be investment tool to earn quick money through shorter-term transactions.

If the investors want to select a stock for its inclusion in his/her stock and have a long-term horizon, our study suggests that Principal Component Analysis (PCA) amongst others will be a suitable fit to gain additional returns. Please note that this analysis also involved Fundamental analysis tools for dataset creation on top of which ML techniques were implemented. [1]

If the investor is looking forth to the stock price prediction, especially for a shorter period, a Support Vector Machine (SVM) is the best suited.[2]

VII. BIBLIOGRAPHY

- Batten, A, J., Lucey, M, B., Mcgroarty, Francis, . . . Urquhart. (2018). Does intraday technical trading Have predictive power in precious metal markets? Journal of International Financial markets, 52, 102 113.
- Bilal Elmsil, B. O. (2021). Predicting Stock market movement using machine learning technique.
- International Journal of Accounting, Finance, Auditing, Management & Economics, 2(3), 390-405.
- Hargreaves, C. A. (2019). an Automated stock investment system Using Machine Learning
- Tecchniques: An application in Australia. World Academy of Science, Engineering and Technology International Journal of Mathematical and Computational Sciences.
- Isaac Kofi Nti, A. F. (2020). Predicting stock market price movements using sentiment analysis evidence From Ghana .sciendo, 25, 33-42.
- Siyuan Yan , X. L. (2022). Algorithmic Trading and Challenges on Retail Investor in Emerging Market.
- Journal of Economics, Finance and Accounting Studies (JEFAS), 4.
- Ummiya Mohammedi1, D. Y. (2018). Analysis of Stock behaviour using Big Data Analytics.
- International Journal of Computer & Mathematical Sciences, 7(2).

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

SMART CONTRACT BASED CASH WITHDRAWAL FROM ATM

Rashmi Pote and Sushil Kulkarni
Department of Computer Science, University of Mumbai, Mumbai, India

ABSTRACT
Mobile banking and internet banking has become part of our lives. With the increase in internet banking from any corner of the world, banks had to focus on safeguarding the credentials of customers in universal banking applications. Most of the customer have started using internet banking and cashless transactions, but we can't escape from not withdrawing cash from ATM. There are three stakeholders in cash withdrawal process, customer, payment gateway and bank. We have national and international payment gateway providing us with their service. They are centralized third party to provide service between customer and bank. When customer initiate a cash withdrawal transaction, the customer credential is shared via international gateways, due to which the risk of data burglary is high. Our goal is to eliminate the centralized third party and use decentralized digital ledger by generating smart contracts to protect customer's credential.

Keywords: cash withdrawal, data breach, blockchain, smart contract, SMBA

I. INTRODUCTION
ATMs (Automated Teller Machines) are an important part of the banking industry, providing convenient access to cash for millions of people around the world. However, as with any financial system, ATMs are also susceptible to security threats such as fraud, skimming, data theft and other types of cybercrime.

A. The Process of Cash Withdrawal
The customer initiates the cash withdrawal transaction by inserting debit card or credit card in ATM. The ATM checks validity of card and it checks that your account is active or not. Customer inserts PIN, ATM verify it and if enough funds are available in the customer's bank account, the cash withdrawal is routed through the payment gateway network. Once the payment gateway verifies the credentials, then it dispenses the requested amount and updates the account balance to reflect the transaction. Payment gateway or the third party plays a very important role in the current cash withdrawal transaction from ATM.

When a customer requests a credit card or debit card from bank, bank send the request to third party and a card is issued to the customer via the third party. The payment gateway is also known as third party, which came into presence to take requests from customers of any bank via a different bank ATM or White label ATM to offer different banking requirements. It acts as a payment connection between the customer and the bank. The payment network processes the transaction by communicating the data securely to the individual bank. It's very cost-effective to route transactions to a third-party system then forming banks own payment processing infrastructure. A customer can now withdraw money from any corner of the world with the help of a third-party payment network. This is the main benefit of including a third-party payment network in banking system.

B. Issue with Conventional Cash Withdrawal
Conventional cash withdrawal process, relies on intermediaries such as payment gateways and payment networks to transmit transaction details to bank servers. They securely communicate the transaction details to the individual bank servers and charge maintenance fees from the banks. Banks then charge the fees to the customer. The conventional banking system relays upon payment gateways, namely Visa, Master Card, American Express, RuPay etc. Payment gateways handle sensitive financial information, and there is constantly a risk that this information could be compromised. This can lead to data theft or fraud, which can cause financial harm to customers. Payment gateways may not be available in all locations, which can make it difficult for customers to withdraw the cash. The dependency on payment gateways can make the cash withdrawal process slower and more complex, as the transaction has to go through multiple parties before it is completed.

C. Blockchain
Blockchain has become well-known due to the implementation of cryptocurrency systems and popularity of bitcoins in 2008. A blockchain is a decentralized, distributed digital ledger that is used to record transactions across a network of computers. It allows multiple parties to reach consensus on the state of a database, without the need for a central authority. Each transaction on a blockchain is recorded as a "block," and these blocks are connected together in a consecutive chain, forming a permanent and unchangeable record. This makes it difficult for any one party to alter or delete transactions, which adds an element of security and trust to the system.

ISSN 2394 - 7780

Blockchains are frequently associated with cryptocurrencies like Bitcoin, but they have many other potential uses as well. For example, they can be used to track the movement of goods through supply chains, verify the authenticity of documents, or facilitate secure voting systems and many more.

Overall, the use of blockchain technology has the potential to revolutionize the way we store and exchange information, as well as the way we conduct business and interact with one another.

The era started with implementation of blockchain 2.0 by Ethereum. It has become more popular because of its decentralized, peer to peer transaction [1], distributed consensus, and anonymity properties. It helps in maintaining secure and decentralized records of transactions which cannot be changed. Blockchain technology works as a distributed transaction ledger [2] shared between nodes on a peer-to-peer network. It shares the digitally stored data among nodes on the network. With increase in technology, the tools used to analyze smart contract code also vary in different aspects [3].

D. Smart Contract

Introduction of Smart contract has extended the abilities and application of blockchain in a great aspect. Smart contracts are the future of security for many significant applications. A smart contract is a self-executing contract with the terms of the agreement between two parties being directly written into piece of code. The code and the contracts contained therein exist on the blockchain network. Smart contracts allow for the automation of contract execution. They are often used in blockchain-based platforms to facilitate, verify, and enforce the negotiation or performance of a contract.

Ethereum is a decentralized, open-source blockchain platform which was specifically designed to run smart contracts. Ethereum transactions are completed by miners and they charge gas fees for the transaction [4]. We can decide how we want our transaction to be committed by choosing the speed of transaction, such as fast, medium and slow. Miners mine the transaction and include them in Ethereum blocks [5]. The mining is done based on urgency of the work, which is decided by the gas price value associated to a transaction. The maximum gas price transaction is then added to Ethereum. Miners have restrictions on how many transactions can be comprised in one single block, it depends on the maximum gas limit per block.

II. LITERATURE REVIEW

The decentralized ledger has created an incredible impact on numerous domains. It has recreated the internet architecture with new perceptions, methods and tools [6]. It has created profound social impacts on computer science, economics, law, business, tourism, commerce, landing systems, health care, supply chain management systems, finance, government voting system, e-voting and several more [7], [8], [9], [10], [11], [12], [13], [14]. Let us know the variation in functionality, benefits and effect on different domains after implementation with blockchain.

The authors [15] have studied diverse work on smart contracts from 2017 -2020 and they classify the work in four categories, namely, cryptography, Access Management, Social Application and Smart Contract Structure. This paper discusses about the current advances in smart contracts and its wide usage [15]. A team of authors have implemented a mixture of blockchain and smart contracts to lending systems. Loan processing can be done via the proposed system and it provides advantages like 1-ease in enhancing and safeguarding the security of lending transactions. 2- Competent lending processing and automation in the transaction process. 3- Enhances the supervision of organizations and helps in standardizing the management behavior [16].

A blockchain based multi-tenant 5G application prototype of smart grid system has been employed by [17]. Authors [18] have proposed a blockchain based 5G network slice broker system. The model aims to reduces service time and it allows the empowering of engineering equipment autonomously and vigorously acquiring the slice needed for more competent operation. In this model the lease of the 5G network with a slice ledger is executed on a network operator which preserves record of each transaction.

IoT and blockchain execution systems also secure the data on different networks. Incorporation of IoT and Blockchain can result in increasing the scalability and preserving a decentralized system with more security. It can be supportive to authenticate and manage the devices competently. The blockchain increases honesty of data, micropayments, sharing services, data monetization, operations on self-sufficient devices and any types of transactions are more protected with blockchain [19]. The use of electric vehicles is increasing in this era. Authors [20] have implemented blockchain technology in mobile charging of electric vehicles. A blockchain-based petaelectronvolt (PEV) charging system for electric vehicles will help to transfer energy directly from one vehicle to another by using peer-to-peer (P2P) mode [21]. After implementation of diverse IoT applications based on smart contracts, few authors have implemented Attribute-Based Access Control (ABAC) framework

ISSN 2394 - 7780

for smart cities. The model implements smart contracts with the help of Ethereum and manages ABAC policies, attributes of subjects, objects and performs access control [22]. A team of authors have implemented Ethereum based Open Vote Network. The model was implemented and tested for forty simultaneous voters on the official Ethereum test network with $0.73 cost per voter [23].

Patient data in healthcare sector is a critically valuable asset. Authors [24] have implemented and tested blockchain and smart contract technology-based patient centered health care system. There is very good coordination is observed between the components of the system, such as doctor, patient, nurse, medical men, insurance man etc. Six different algorithms were implemented for synchronizing data between the components in a patient centered healthcare system.

Till now we have already seen different types of applications using smart contracts, we can now move on with ring signature algorithms. When IoT devices transmit the data on a blockchain network, and a smart contract transaction is performed, hiding the information about who initiated the transaction can be done with the help of the ring signature algorithm. The method hides the uniqueness of the device and its address on the blockchain network. The ring signature algorithm is safeguarding the whole process on the network [25]. A team of researchers [26] have presented a novel supply chain model based which is on blockchain and smart contract for solving the problem of the whole Supply Chain Matrix (SCM) system. The advantages of using smart contracts in SCM are, 1- Reducing the Cost, 2- Advancing the Value, 3- Speeding Turnover, 4- Easy Finance, 5- Stronger Risk Management and 6- Easily combining with High Technology [26].

Smart contract and its ongoing implementations are useful in all domains. A team of researchers have worked in an area of improving privacy and concurrency in payments channel networks [27]. Smart contracts are useful in many different fields because they are able to store data, process inputs, check conditions, and write outputs [28]. They can be implemented on the Ethereum network, which makes it easy to use and execute these contracts. For example, a decentralized application using the Ethereum blockchain was developed to allow users to share everyday objects without the need for a trusted third party. The smart contract enabled tool owners to register and reclaim their objects, while renters were only allowed to rent out the objects they chose [29].

When it comes to the money supply process, it is important to consider which type of database system is the most suitable. A study by researchers analyzed the strengths, weaknesses, opportunities, and threats (SWOT) of both centralized and decentralized ledger systems in the money supply process [30]. The authors institute that centralized ledger systems are important for keeping records of financial transactions, while decentralized ledger systems are likely to become increasingly important in the future of finance and may cause disruptions in financial transactions.

III. DESIGN OF NOVEL SMART CONTRACT BASED CASH WITHDRAWAL FROM ATM

Smart contract has few unique characteristics, such as verifiability, distributed nature and auto-enforcing ability. The distinct feature of smart contracts, has enabled to encode business rules to be executed in a peer-to-peer network. Each node in the network is equal and none has any superior authority without the participation of a trusted central authority. Due to this, smart contracts are probable to transform many traditional businesses. We are proposing a model, which will implement smart contracts in the cash withdrawal process using ATM.

In the existing cash withdrawal process, customers have an ATM card to withdraw money from an ATM. To get a debit card, customers apply to the bank. Banks provide customer with these cards, which are labelled with Visa, Mastercard, Rupey etc. These are a type of payment gateway card, which helps the customer to withdraw money from another bank ATM or a third-party ATM. The system suggested by us will help in sharing the customer data from the ATM to the bank server with the help of a smart contract, which will be executed on a specific time stamped. All credentials of the cash withdrawal are reachable in the public Ethereum blockchain and are implemented as indicated. The whole process avoids a Trusted Third Party, the payment gateway and the role of TTP is done by smart contract. The cash withdrawal terms and conditions are set by the bank as well as the customer. Our aim is to remove the dependency of the payment gateway and make the whole process more transparent to customers as well as banks.

IV. IMPLEMENTATION

Currently, the financial records are centralized [31]. They are created, controlled, and maintained by a central party [32]. Use of a decentralized ledger with the help of smart contracts will help the banks in maintaining the ledgers to reflect their transactions securely. A decentralized ledger represents a specific data structure preserved by numerous parties through a consensus protocol, and each party holds a copy of the ledger [33]. The major change between the centralized and decentralized ledger is that, the decentralized signifies a collection of connected nodes, creating the ledger and storing all data concurrently. The distributed consensus

ISSN 2394 - 7780

protocol is the significant protocol in creation of the decentralized ledger. Consensus protocol guarantees that all nodes on the network agree on a unified transaction ledger without the presence of a third party.

Following are the steps for implementation.

Step 1: - Customer install SMBA on the smartphone. [34]

Step 2: - Customer initiate cash withdrawal, answer the security questions, scan QR code of SMBA on ATM and get cash from the nearest ATM [34].

Step 3: - A consensus protocol is used to generate smart contract for the cash withdrawal transaction.

Step 4: - Implementation of AES algorithm for communicating data securely over network and storing the transaction data on block chain [35].

The smart contract must be written in Solidity, JavaScript is used in combination with Solidity to interact with smart contract and test smart contract. Solidity can be understood and supported by Ethereum Virtual Machine (EVM). EVM is runtime environment that executes smart contract on Ethereum blockchain. The frontend of decentralized application can be created in JavaScript. The web based user interface can interact with smart contract via web3.js.

A consensus protocol avoids a single entity from controlling a blockchain or altering the truth of what should be recorded. We are securing the data with the help of AES algorithm to make it more robust. To encrypt data during transmission AES is used, we first generate a secret key and then use an AES library to encrypt the data using that key. While performing data decryption, the authorized party would need to have access the secret key and use the same AES library to decrypt the data. It's important to note that AES provides a high level of data security.

V. DISCUSSION

Cash withdrawal can be performed by many ways, 1) physically visiting bank, 2) visiting an ATM, 3) customer can create N26 account, which does not charge any fee for cash withdrawal, it does not have any transaction limit on this account. N26 account has an international bank account number by which the customer can do all transactions. Customer can withdraw cash using CASH26. All these methods have its advantages and disadvantages. Our aim is to withdraw physical money from ATM using bank ATM card. The method currently used to withdraw physical cash from ATM uses payment gateway.

The method suggested by us uses SMBA. When a customer performs cash withdrawal from SMBA [34], we have tried to thoroughly authenticate a customer by asking security questions and then customer reaches the nearby ATM, scan the generated QR code and get cash from ATM. The motive behind using smart contract is to replace the centralized payment gateway with smart contract.

In the ATM cash withdrawal process, the payment gateway plays a critical role in enabling the transfer of funds from the customer's bank account to the ATM. The payment gateway acts as an intermediary between the customer's bank account and the ATM. When a cash withdrawal transaction is initiated at an ATM, the payment gateway is responsible for communicating with the customer's bank to verify that the customer has sufficient funds in their account to complete the transaction. The payment gateway performs a number of functions, namely, authentication, authorization, transaction processing, settlement and keeping the customer data secure. There are many challenges that can arise while using payment gateway, such as, compliance with wide range of regulation standards, scalability in handling increase in transactions, reliability and availability, integration of payment gateway with various platforms, cost, interoperability and security. Payment gateway must implement robust security measures to protect against fraud and hacking.

The blockchain technology governs regulatory problems and technical challenges. A smart contract is self-verifying, self-executing and tamper resistant, immutable these are the basic advantages for which we are suggesting smart contract. It will make the cash withdrawal process independent in real time with small cost and offer a greater degree of security. Use of blockchain will generates and stores data records in a distributed system. It maintains a digital ledger of connected blocks of information that represent how data is shared, changed or accessed on its peer-to-peer network. All nodes on the same blockchain system will generate identical blocks, when a connected node is involved in any kind of transaction. This provides us with a safety of keeping backup in distributed manor on multiple nodes on the network. If one nodes data is accessed, changed, shared or otherwise manipulated in any way, a block is generated locally, the information on every node will differ from other node and this way, changes to data can be easily identified. It’s a decentralized approach that allows data equality to be attained by comparing every connected node’s blocks.

ISSN 2394 - 7780

VI. LIMITATIONS

A major limitation of smart contract is, it's difficult to change, it's almost impossible to do so. If there is any error in code while creating the smart contract, then it can be time consuming and expensive to correct. There are limited number of transactions which can be processed in a given time on a blockchain. Auditing smart contract helps us to eliminates the flaws in the system, but may be more expensive. Just using AES may not suffice, AES implementation with other security measures such as secure key management and access control will make the whole system robust.

VII. CONCLUSION

This paper concludes that, cash withdrawal is an important activity for many of us. While using payment gateway cards for cash withdrawal, there is chance of data theft, which rises security concern in centralized ledgers. The use of decentralized ledger is likely to bring numerous distractions in the future for cash withdrawal from ATM, such as removing the payment gateway network. We are suggesting to accept decentralized ledgers by their capacity to address the ever-increasing security risks encountered during cash withdrawal from ATM.

Decentralized ledger is expected to become an extensively used tool in banking sector in near future. Use of smart contracts during cash withdrawal will safeguard the customers credential and automate the process with encryption and keeping it on block chain.

In future all financial transactions would be utilizing the security feature of smart contract. Smart contracts are implemented in distributed ledger, they are a digital alternative to physical contract. It helps in banking industry to securely conduct all the banking transactions.

REFERENCES

[1] S. Nakamoto. (2008). "Bitcoin: A Peer-to-Peer Electronic Cash System," [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[2] G. Wood, ''Ethereum: A secure decentralized generalized transaction ledger,'' Tech. Rep. EIP-150. Accessed: Dec. 18, 2020. [Online]. Available: http:// gavwood. com/ Paper.pdf

[3] M. di Angelo and G. Salzer, ''A survey of tools for analyzing Ethereum smart contracts,'' in Proc. IEEE Int. Conf. Decentralized Appl. Infrastruct., Newark, CA, USA, pp. 69–78, 2019.

[4] Giuseppe Antonio Pierro and Henrique Rocha, "The influence factors on ethereum transaction fees, 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain, pp. 24–31, May 2019.

[5] Zhenzhen Li, Wei Xia, Mingxin Cui, Peipei Fu, Gaopeng Gou and Gang Xiong, "Mining the Characteristics of the Ethereum P2P Network," 2nd ACM International Symposium on Blockchain and Secure Critical Infrastructure, pp. 20–30, October 2020.

[6] Hughes A, Park A, Kietzmann J, et al. (2019) Beyond Bitcoin: What blockchain and distributed ledger technologies mean for firms. Bus Horiz 62: 273–281.

[7] Lacity MC (2018) Addressing key challenges to making enterprise blockchain applications a reality. MIS Q Exec 17: 201–222.

[8] Erceg A, Damoska Sekuloska J, Kelić I (2020) Blockchain in the Tourism Industry—A Review of the Situation in Croatia and Macedonia. Informatics 7: 5.

[9] Joo J, Park J, Han Y (2021) Applications of Blockchain and Smart Contract for Sustainable Tourism Ecosystems, In: V. Suma, N. Bouhmala, & H. Wang (Eds.), Evolutionary Computing and Mobile Sustainable Networks, Singapore: Springer, 773–780.

[10] Rejeb A, Bell L (2019) Potentials of Blockchain for Healthcare: Case of Tunisia. World Sci News 136: 173–193.

[11] Khatoon A (2020) A Blockchain-Based Smart Contract System for Healthcare Management. Electronics 9: 94.

[12] Kamilaris A, Fonts A, Prenafeta-Boldú FX (2019) The rise of blockchain technology in agriculture and food supply chains. Trends Food Sci Technol 91: 640–652

[13] Dhairya Shah, Srikant Kamath, Smeet Ramani, Aruna Gawade, "Ethereum-based Quadratic Funding Of Public Commodities", International Journal Of Information Security Science, Vol. 11, No. 3, pp. 1-12

[14] Latif Anil Buyukbaskin, Isa Sertkaya, "Requirement Analysis of Some Blockchain-based E-voting Schemes", International Journal of Information Security Science, Vol. 9, No. 4, pp. 188-212, 2020

[15] Victor Youdom Kemmoe, William Stone, Jeehyeong Kim, Daeyoung Kim and Junggab Son, "Recent Advances in Smart Contracts: A Technical Overview and State of the Art,", IEEE Access, Volume: 8, pp. 117782 – 117801, June 2020.

[16] Qi Yang, Xiao Zeng, Yu Zhang and Wei Hu, "New Loan System Based on Smart Contract," ACM International Symposium on Blockchain and Secure Critical Infrastructure, Pp. 121–126, July 2019.

[17] Matevz Pustisek, Jan Turk, and Andrej Kos, "Secure Modular Smart Contract Platform for Multi-tenant 5G Applications," IEEE Access, Volume: 8, pp. 150626 – 150646, August 2020.

[18] Backman, S. Yrjola, K. Valtanen, and O. Mammela, "Blockchain network slice broker in 5G: Slice leasing in factory of the future use case," Internet of Things Business Models, Users, and Networks, pp. 1–8, Nov. 2017.

[19] A. Reyna, C. Martin, J. Chen, E. Soler, and M. Diaz, "On blockchain and its integration with IoT. Challenges and opportunities," Future Generation Computer Systems, vol. 88, pp. 173–190, Nov. 2018.

[20] N. H. Kim, S. M. Kang, and C. S. Hong, "Mobile charger billing system using lightweight Blockchain," 19th Asia-Pacific Network Operations and Management Symposium (APNOMS), pp. 374– 377, Sep. 2017.

[21] J. Kang, R. Yu, X. Huang, S. Maharjan, Y. Zhang, and E. Hossain, "Enabling Localized Peer-to-Peer Electricity Trading Among Plug-in Hybrid Electric Vehicles Using Consortium Blockchains," IEEE Transactions on Industrial Informatics, vol. PP, no. 99, pp. 3154 - 3164, 2017.

[22] Yuanyu Zhang, Mirei Yutaka, Masahiro Sasabe, Shoji Kasahara, "Attribute-Based Access Control for Smart Cities: A Smart-Contract-Driven Framework," IEEE Internet of Things Journal, Volume: 8, Issue: 8, pp. 6372-6384, April1 2021.

[23] Patrick McCorry, Siamak F. Shahandashti, and Feng Hao, "A smart contract for boardroom voting with maximum voter privacy," Financial Cryptography and Data Security, Springer International Publishing, pp. 357–375, 2017.

[24] Nghia Duong-Trung, Ha Xuan Son, Hai Trieu Le, Tan Tai Phan, "On Components of a Patient-centered Healthcare System Using Smart Contract," 4th International Conference on Cryptography, Security and Privacy, pp. 31–35, January 2020.

[25] Hongzhi Yang, Linyun Yuan and Shu Wang, "Design of Blockchain Smart Contract Based on Ring Signature," 9th International Conference on Communications and Broadband Networking, pp. 108–114, February 2021.

[26] Haitao Liu, "A Novel Supply Chain Model Based on Smart Contract," 2nd International Electronics Communication Conference, pp. 115–120, July 2020.

[27] Giulio Malavolta, Pedro Moreno-Sanchez, Aniket Kate, Matteo Maffei and Srivatsan Ravi, "Concurrency and Privacy with Payment-Channel Networks," ACM SIGSAC Conference on Computer and Communications Security, pp. 455–471, October 2017.

[28] Vitalik Buterin, "A next-generation smart contract and decentralized application platform," Ethereum white paper, 2014.

[29] Andreas Bogner, Mathieu Chanson and Arne Meeuw, "A Decentralised Sharing App running a Smart Contract on the Ethereum Blockchain", 6th International Conference on the Internet of Things, ACM, pp. 177-178, Nov. 2016.

[30] Abderahman Rejeb, Karim Rejeb and John G. Keogh, "Centralized vs. decentralized ledgers in the money supply process: a SWOT analysis," Quantitative Finance and Economics, pp. 40-66, January 2021.

[31] Rahmadika S, Ramdania DR, Harika M, "Security Analysis on the Decentralized Energy Trading System Using Blockchain Technology," J Online Inf 3, pp. 44–47, 2018

ISSN 2394 - 7780

[32] Morgan JS, "What I Learned Trading Cryptocurrencies While Studying the Law," Univ Miami International and Comparative Law Review, 2017.

[33] Lei Xu, Lin Chen, Zhimin Gao, Xinxin Fan and Taeweon Suh, "DIoTA: Decentralized-Ledger-Based Framework for Data Authenticity Protection in IoT Systems," University of T ersity of Texas Rio Gr exas Rio Grande Valley, Jan 2020.

[34] Rashmi Pote. and Sushil Kulkarni., "Securing cash withdrawal from ATM with the help of Smart Mobile Banking Application," International Conference on Interdisciplinary Research in Technology & Management, pp. 298-301, Feb. 2022.

[35] Hao Lin, Xiaolei Li, Haoyu Gao, Jie Li, Yongsheng Wang, "ISC-MTI: An IPFS and smart contract-based framework for machine learning model training and invocation", Multimedia Tools and Applications, Springer, pp. 40343-40359,2022.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

A STUDY ON SMART CITY AND BIG DATA

Saadiya Patrawala
Department of Information Technology, Jai Hind College, Mumbai, India

ABSTRACT
The idea of the smart city has been prominent for many years. Urbanization-related issues are believed to be tackled by smart cities. Unfavorable outcomes including excessive energy use, crime, and poor mobility are expected to get worse as more people move to cities. Utilizing smart city solutions enables cities to expand sustainably while enhancing the standard of living for citizens. In order to better serve citizens and enhance decision-making processes, information systems are increasingly emphasizing on extracting insights and hidden relations from big data. The produced data from multiple city domains need to be combined, analyzed and visualized in order to derive useful insights. This paper is a review of the basic concepts of smart city and big data and how big data impacts smart cities.

Keywords: Smart City, Big Data, Internet of Things

I. INTRODUCTION
The idea of residing in a smart environment is becoming a reality owing to the radical transformation from the world of traditional desktop computing to one of highly sophisticated computing, as well as the considerable growth in linked gadgets and sensors. At the moment, urban performance depends not only on the physical infrastructure but also on the accessibility and quality of knowledge communication and social infrastructure. The Internet of Things (IoT), which connects daily items and gadgets to network technology, may be the primary facilitator of various smart city applications [10]. The concept of smart cities emerged as a strategy to mitigate the unprecedented challenges of continuous urbanization, increasing population density and at the same time provide better quality of life to the citizens and visitors. A smart city is composed of several well-defined smart components including smart mobility, smart governance, smart manufacturing as well as its applications and services such as smart transportation, smart healthcare, smart farming, smart education and more [9].

In this respect, IoT and big data technologies are seen as key drivers behind the emergence of sustainable smart cities. With the invent of the web, the complete world has gone online, every single thing we do leaves a digital trace. The rate of data growth has significantly accelerated as a result of smart objects becoming online. Thus far, the quantity of digital artifacts connected to the Internet (e.g., sensors, PCs, social networking platforms, wearable devices, smartphones, cameras, games and many more) have according to the Cisco report, exceeded the number of living people in the world. The continuously increasing number of networked devices deployed across urban environments will in turn result in the explosive growth in the amount of the data generated [6]. Additionally, the widespread use of Information and Communication Technology (ICT) applications and digital technologies in different city domains has amplified human-to-human, human-to-machine, machine-to-human, and machine-to-machine interactions that produce vast amounts of data, generally known as big data, which is a combination of large size and complexity that no traditional data processing application software can store or process effectively [9].

Big data is a combination of structured, unstructured and semi structured huge volumes of data that requires new technologies and architectures to make it possible to extract value from it by capturing, storing, classifying, analyzing the historical and real time data to identify the patterns, hidden relations between different variables and other useful information which can be used to make better decisions. R6. The five Vs.: volume, velocity, variety, veracity, and value that describe the existence of big data are being produced at booming rates. Big data offers the possibility for the city to gain insightful knowledge from the substantial amount of data gathered from diverse sources [10].

Data can be stored in three major places in smart cities: 1) Cloud Storage: For analytical purposes, smart cities need a plethora of data. Cloud data systems eliminate duplicate data, encrypt data transmission, and use solid-state drives in their data centers. Payment choices for cloud-based solutions are typically more flexible than those for on-premise data centers. 2) Edge Computing: Processing data near to the source is made possible by edge computing. Instead of streaming data to a remote storage facility and subsequently to the relevant local authorities, edge computing can be less expensive. In certain places, edge computing functionalities like AI traffic control is already being developed. Intelligent automation is used in AI traffic management to quickly respond to changing conditions while also detecting accidents and traffic congestion. 3) Hybrid Data Storage:

ISSN 2394 - 7780

The benefits of cloud and edge storage are combined in hybrid data storage systems. Cities can now make new judgments based on rich data stores and real-time alerts on situations thanks to hybrid data storage.

The cities are simultaneously creating new platforms for data exchange and integration that dismantle the conventional operational silos and give access to all possible solution suppliers. For this reason, the benefits offered by big data are an important element of many smart city strategies

A. **Smart City**

Despite the worldwide smart city hype, daily news about new smart city initiatives, cities trying to be the smartest, and governments across the world spending billions in smart cities, there is still the question in people's minds, "But what is a smart city?" Nowadays, many of us use Google to quickly find the answers we need. When someone searches for this specific question, Wikipedia comes up as one of the top results. Wikipedia defines a smart city as "A smart city is an urban area that uses different types of electronic data collection sensors to supply information which is used to manage assets and resources efficiently."

One of the nations that have adopted a national approach for smart cities is India. When responding to this delicate query, the Indian government took a modest, unbiased stance by stating, "The fact is, there is no universally acknowledged definition of a smart city. To various people, it has varied meanings. The conceptualization of Smart City, hence, differs from city to city and country to country, willingness to change and reform, depending on the level of development, aspirations and resources of the city residents."

IoT sensors, connection, and data are the three technological pillars that all smart city projects have in common. Communities are given a strong foundation for innovative and more effective approaches to build more livable cities by linking these three pillars. This is well demonstrated by the following smart city concepts:

1) **Smart Energy:** Conventional energy sources do not power smart cities. Energy assets are decentralized and national grids are upgraded. Cities can remodel existing infrastructure to create tradable assets and income streams by managing their energy assets. Micro-energy stations are created from collections of energy assets that return excess energy to the grid. Smart energy causes significant cost reductions for operating both public and private infrastructure. Cities understand about their energy demand profile owing to smart energy applications. Officials could prioritize reduced consumption; comprehend prevailing loads and daily fluctuations. By avoiding peak hours and consumption periods, techniques like load shifting and demand-side response assist in lowering costs; both the city and energy users achieve substantial savings as a result.
2) **Smart Mobility:** The way city dwellers use their time, enjoy their commutes, and spend their money is altered by smart mobility. It includes smart parking solutions, where sensors maintain track of the parking spaces that are vacant. Drivers are given information about available parking options via smartphone applications or digital signage along the roadways so they can discover and navigate to the optimal parking spot with ease. Applications for booking travel that utilize digital currency and technology that links individuals to travel infrastructure are further examples of smart mobility. It also consists of systems that use predictive analytics to manage public transportation vehicles so that they are available when and where they are needed. There are several types of transportation in smart cities. Urban mobility is improved through systems like integrated multi-modal transportation and intelligent traffic management.
3) **Smart Buildings:** Smart buildings integrate people and systems functionally and dynamically. Different parts of the building often use different security measures, such as integrated access control, network intrusion detection, video surveillance, and fire prevention systems. There are automation systems that, for instance, automatically adjust ventilation and heating to the level of occupancy and make sure that lights are switched off in empty rooms. Additionally, there are intelligent trash cans that include sensors to communicate their level of capacity to the pickup service. Based on real requirements, the collecting routes are automatically optimized.
4) **Smart Healthcare:** By tracking patients, equipment, workers, and many other things in smart cities, the Internet of Things increases access to high-quality healthcare and lowers costs. The potential for smart healthcare is limitless; we have only just begun to scrape the surface. Utilizing technology and smart gadgets, it enhances medical diagnosis and treatment. Smart sensors, for instance, can detect air or water pollution before it poses a concern to the public's health. Additionally, sensors may collect information from health centers and track the spread of diseases.

B. **Big Data**

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." For a number of reasons, the amount of

ISSN 2394 - 7780

data on earth is exponentially increasing. E-commerce sites, stock exchanges, social networking sites, weather stations, and satellites are the main sources of big data. Telecom industry giants like Vodafone and Airtel analyze user trends and publish their plans accordingly; for this purpose, they retain the data of their million subscribers. The following are some characteristics of big data:

1) **Volume:** Volume refers to the 'amount of data', which is growing day by day at a very fast pace. The size of data generated by humans, machines and their interactions on social media itself is massive. A forecast by International Data Corporation (IDC) estimates there will be 41.6 billion Internet of Things (IoT) connected devices by 2025 capable of generating 799.4 zettabytes (ZB) of data.
2) **Velocity:** The speed at which various sources provide data on a daily basis is referred to as velocity. There are 1.96 billion Daily Active Users (Facebook DAU) as of now, with a 6.89% increase year-over-year. This demonstrates how quickly data is created each day and how many people are using social media. You will be able to produce insights and make decisions based on real-time data if you can manage the velocity.
3) **Variety:** The types of data that are being produced by the many sources that make up Big Data vary. It may be structured, semi-structured or unstructured. As a result, several types of data are produced every day. The data nowadays are being generated in the form of images, audios, videos, sensor data, etc. Previously, we used to get the data from databases and excel sheets. Therefore, this diversity of unstructured data makes it difficult to collect, store, mine, and analyze the data.
4) **Veracity:** Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. Data at hand can occasionally become disorganized and may be difficult to trust. Quality and accuracy are challenging to regulate in many kinds of big data, such as Twitter tweets with hash tags, abbreviations, typos, and vernacular language. The lack of accuracy and quality in the data is frequently caused by the volume of data. One in three corporate leaders doesn't have confidence in the data they rely on to make choices because of data uncertainty.
5) **Value:** After talking about Volume, Velocity, Variety, and Veracity, there is one more V that must be considered while examining Big Data, namely Value. Having access to big data is great, but unless we can utilize it to create value, it is pointless. By creating value what is being implied is that, is it increasing the advantages of the organization who are studying big data? Does the organization utilizing big data have a high return on investment? It is meaningless unless using big data increases their profitability.

We anticipate that over time, the term "big" will lose its meaning and that the definition of "data" will naturally be expanded to encompass all sorts of data, such as semi-structured and unstructured, in addition to the traditionally recognized "structured data" [9].

II. BIG DATA ANALYTICS IN SMART CITY

Big data analytics has a role to play across all aspects of smart city operations. It has already proven itself valuable in varied fields when it comes to the development of smart city.

1) **Safety and Security:** Security of residents is a must for smart cities. In order to tackle crime, big data analytics offers alternatives to arming and enlarging the police force. Cities may employ predictive big data analytics to pinpoint the exact crime location and identify which areas are likely to be the centers of crime. Cities can build significantly safer environments by using information like historical and geographic data since the police can be stationed there before the crime ever occurs. The Delhi police and the Indian Space Research Organization (ISRO) have teamed up to develop the Crime Mapping Analytics and Predictive System (CMAPS), an analytical tool that enables the Delhi police to ensure internal security, manage crime, and uphold law and order through data and pattern analysis. They can keep tabs on criminals and individuals with its assistance. The police can benefit from data analytics and possibly achieve greater results in their work.
2) **Better Asset Management:** Asset management is a key use-case for data analytics. Cities spend a lot of money on making their cities smart cities. These expenditures may be made for renovation or remodeling. Big data analysis can make recommendations about what areas require changes and the types of changes are needed. Cities may employ predictive maintenance to lower risks and expenses and to better monitor and manage a variety of civic infrastructure owing to data analytics. The city council can better understand the needs of the people it serves and improve public services by using a system of sensors that gather useful data. We may use the City of Marion in South Australia as a successful illustration. Since December 2017, smart sensors have been installed in playgrounds and parks. The insights from the data gathered by these sensors may be used to improve the public experience overall, park equipment maintenance, and improving

ISSN 2394 - 7780

future park planning, by learning how people utilize municipal resources. Additionally, it is envisaged that using this asset management strategy will contribute to lowering of public liability. The council is informed that something may be amiss and is given the opportunity to act straightaway when the sensor notices that a certain piece of equipment is not being utilized or is being used less frequently.

3) **Sustainable Growth:** A smart city's progress is regularly analyzed, allowing city leaders to get continuous reports regarding required modifications. The main forces behind the growth of sustainability are continuous updates, which give a clear picture of the required developments. The outcomes of development in a smart city are significantly influenced by data. The following are some examples: transitioning to carbon-free cities: Advanced data analytics are crucial to helping cities, utilities, and other partners achieve their ambitious zero-carbon goals by optimizing the flow of resources and energy. Big data analytics is essential for the effective administration of distributed renewable energy, storage, and micro grid-based community energy systems.

III. PROPOSED FRAMEWORKS

A. Gaminess Framework

GAMINESS management system is a smart city management system with 3D city representation. This model was developed using the CityGML open data model and XML-based format for storing and exchanging virtual 3D city models. The extendible international standard for spatial data sharing issued by the Open Geospatial Consortium (OGC) and the ISO TC211 is the basis for the GAMINESS model, which is based on the application schema for the Geography Markup Language version 3.1.1 (GML3).

A conceptual model of Gaminess was designed to explain data with all properties together with smart city components. The IndoorGML standard and IoT standard were added to the CityGML standard in the conceptual model. To describe the spatial attributes gathered from sensors, user-defined types required to be added to packages of classes that previously only carried numerical data. The Gaminess management system's implementation schema was based on JSON. To convert the Json Gaminess management system to Data Frames on Apache Spark, the transformation module was created. RDD schema is part of Data Frames, and RDDs are made up of rows of objects with extra schema details including the type of data in each column, fundamental types, and user-defined types. With the ability to perform a new query over stored data, GAMINESS offers data storage in an RDD structure with all of its elements. A set of already user-defined functions may be used in such SQL queries. As indicated throughout the cluster formulation of the processing network, the system offers volume and value progression. For the purposes of complicated analysis in the context of the smart city for variability and velocity benefits of the Gaminess, Apache Spark offers built-in modules for machine learning components [7].

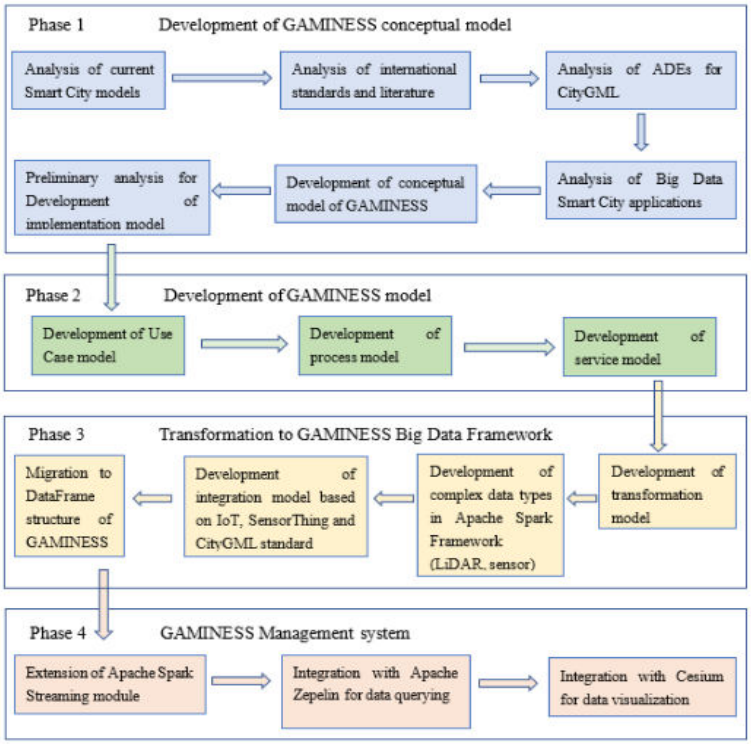


Fig.1 Gaminess management system Architecture

ISSN 2394 - 7780

B. **BDA Embedded Framework**

The proposed smart city framework that embeds BDA comprises of three tiers, namely, data aggregation, data management, and service management. The dataflow and workflow of the proposed architecture is designed bottom-up, starting with the levels of data production and collection, data management, and service management. The foundation of city architecture is made up of smart objects, including sensors, gadgets, and actuators. The last layer creates data and shares them among others that are connected within the network. Aggregated data are sent to the intermediate layer, where they are filtered and analyzed to extract the most useful information from the collected raw data. Intelligent decision-making agents categorize useful facts and provide smart decisions. The service management tier is responsible for formulating and generating service events for smart cities that are based on deduced smart decisions.

By integrating data filtration techniques with BDA, the proposed work's primary objective is to enhance the data processing and decision-making performance of realistic smart city architecture. The proposed work assimilates min-max normalization and data filtration approaches, such as range checking and ambiguity checking, to lessen the quantity of corrupted, noisy and ambiguous data in order to enhance Big Data processing in smart city environments. To identify potentially crucial data, the normalization process determines threshold limits for the minimum desired value Kmin and the maximum desired value Kmax. The initial batch processing of specified genuine datasets yielded these Kmin and Kmax values. Additionally, filtered and normalized data lessen the redundant load that isn't essential on the Hadoop processing cluster. To further speed up the processing activities, the proposed architecture uses a dual-node Hadoop cluster rather than a single-node Hadoop cluster [8].

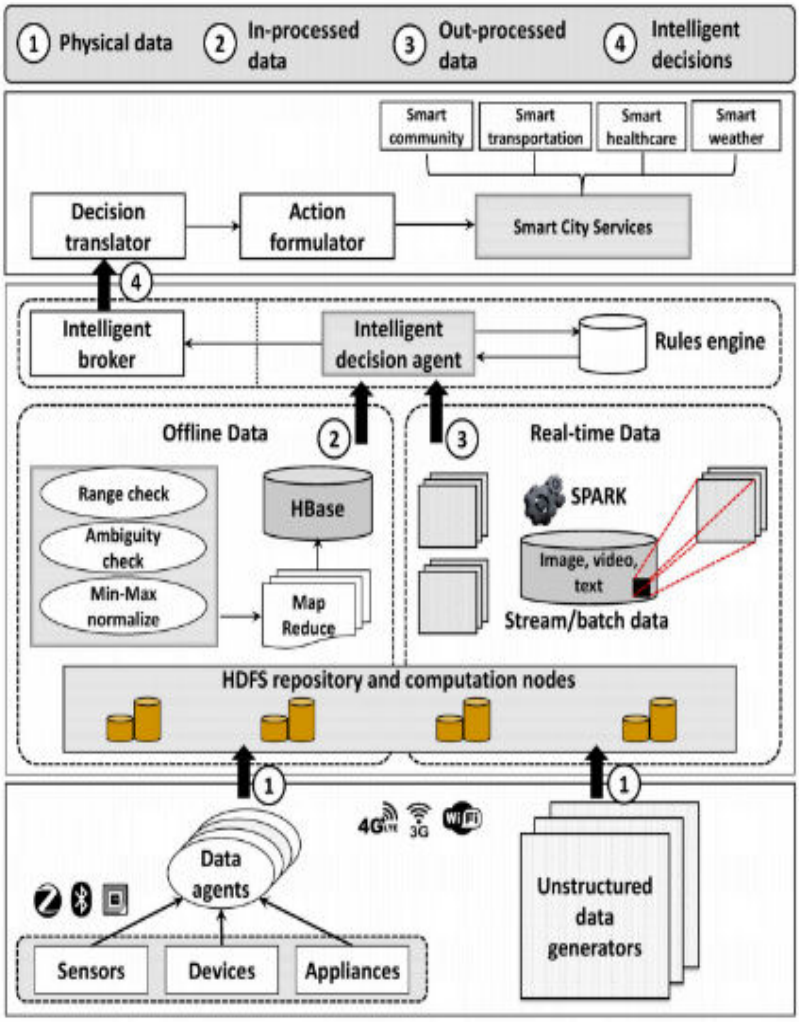


Fig.2 BDA Embedded Architecture

C. **SCDAP Framework**

A "Smart City Data Analytics Panel (SCDAP)" framework is proposed in light of the awareness of the entire perspective of a smart city. The schematic architecture of SCDAP is a 3-layer architecture made up of the platform layer, security layer, and data processing layer.

ISSN 2394 - 7780

Hardware clusters, operating systems, and communication protocols are all part of the horizontally scalable platform layer. Additional computing nodes can be added as required in horizontally scalable platforms. The physical implementation of SCDAP will make clearer the complete functionality of the security layer, for critical analytics, the following security measures should be adhered to on physical design: 1) Critical and sensitive data should be given restricted sign-on access to the framework, 2) Multi levels user authentication, 3) For significant operations, a full audit log should be maintained. Data processing layer is the core data processing engine that offers each and every data processing functionality, from data collection to knowledge extraction. For real-time and historical data analytics, this layer supports both online and batch data processing. Additionally, this layer offers two crucial features that set SCDAP apart: model manager and model aggregation, which allow for the management (i.e., persistence, retrieval, and deletion) and aggregation of extracted data models, respectively [9].

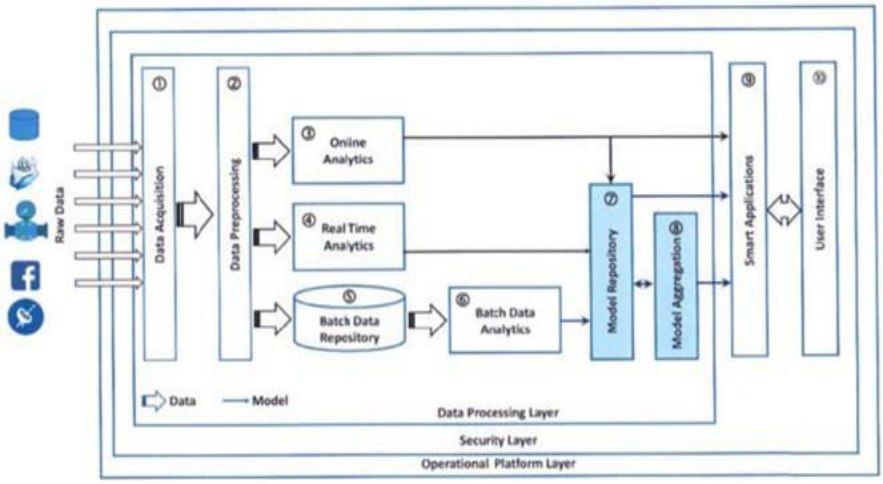


Fig.3 Smart City Data Analytics Panel SCDAP Architecture

IV. RELATED STUDY

This paper provides a general overview of the viewpoints and problems surrounding big data and open data in a smart city, with particular intimation to Trento, Italy. Particular focus is thus given to the outcomes of a working group meeting the municipality conducted on this subject. With regard to big data and open data, the municipality of Trento's existing status is described in this article and some guidelines are offered that will aid the practices moving forward become more smart and responsive to the requirements of the public. In this regard, Trento has already launched a number of projects, particularly in the area of open data, and other efforts are prepared to launch. Given the amount and importance of the research institutes and universities, as well as the fact that local government expenditures in innovation are consistently on the agenda, it is conceived Trento contains the majority, if not all, of the necessary competencies [1].

This paper shows adoption and success of IoT innovations and big data analysis in a smart city. It is seen from its frequent appearance across a variety of industries, including hospitality, oil and gas, military, healthcare, among others, and transportation, to mention a few in accordance to this article. The big data industry is currently well-versed in offering high-quality solutions from fragmented and remote data sets. It is evident from the critique that using big data opens up a wide range of possibilities. The paper also consisted of further opportunities that were available, which led to the creation of smart services that were completed while utilizing all data that was readily available in order to enhance their operations and outcomes. The variety of challenges present in the field and identification of a number of issues that may limit efforts to develop big data services were discussed. Given all the discussion, a list of fundamental requirements due to big data applications in smart cities was proposed [2].

This paper presents findings from an analysis of several use cases of big data in cities throughout the world and four projects with government organizations toward developing smart cities. At least six difficulties must be overcome in order to turn urban data into information for smart cities. The five recommendations made for the gathering, administration, and analysis of urban data will aid in overcoming such difficulties when implementing the reference models in real world applications. This research is distinctive since it analyzed current big data use cases for smart cities using an empirical approach to obtain such findings. Integrating and modifying the knowledge of many experts for the context of using big data for smart cities was one of the hurdles faced in each project proposed by the authors [3].

ISSN 2394 - 7780

In this paper there is discussion of obstacles and challenges that arise during the development of a smart city. There is a framework for the smart city using internet of things which can significantly contribute in the substantial improvement of current models of smart city. The article includes discussion of the components of the framework which is proposed for the city, and analytics for large datasets with their applications in several domains. Because there are fewer cities in India that have made their data publicly available for educational purposes and analysis, the research study's findings have been challenging to visualize and illustrate. New Delhi (DEL) and Jaipur (JAI) are the only cities included in this paper. This framework can be used to guide the growth of other underdeveloped and unfocused cities in order to validate and improve it further [4].

This research focuses on big data applications that are extremely helpful for the implementation of smart cities because they present a richness of potential, but also contain a variety of difficulties and challenges. Furthermore, a smart city service model is proposed, which is helpful for keeping track of the continually created data. These enormous amounts of data are constantly produced by various IoT devices that are used by smart cities. The goal of this article is to present a comprehensive view of the functionality of big data in smart cities. In light of this, revolutionary technologies employed by smart companies and smart cities were discussed [5].

This paper looks at how the growing data-driven smart city is being executed and justified in terms of the development and implementation of its ingenious applied solutions for sustainability. To shed light on this emerging urban phenomenon, a descriptive case study is used as a qualitative research methodology to assess and contrast London and Barcelona as the top data-driven smart cities in Europe. This study demonstrates that these cities have a remarkable development of the applied data-driven technologies, but they somewhat vary in the level of the implementation of such technologies in different city systems and domains in terms of sustainability. They also to some degree differ in terms of their readiness as to the availability and development level of the competences and infrastructure needed to generate, transmit, process, and analyze large amounts of data to derive knowledge that will help them make better decisions and gain deeper insights into how urban operations work, are managed, and are planned in relation to sustainability [6].

This research paper, proposes Gaminess, the biG dAta sMart cIty maNagEment SyStem based upon Apache Spark big data framework. The GAMINESS management system's concept is built on big data modelling ideas, which are very different from those used by conventional databases. The proposed model consists of a sensor component and a geospatial component, which are based on the CityGML and SensorThings standards, respectively. In this study, a technique for reading and structuring data from several sensors into a sensor DataFrame was created. For reading and storing in the sensor RDD, the developed transformation algorithm was utilized to transfigure messages from sensors past JSON messages. Through the application of user-defined functions, varied segmentations, and feature extractions, the suggested solution may simply be expanded with added operations on point clouds and sensor data. The case study's findings demonstrate that Apache Spark outperforms the traditional relational databases. The query's parallel execution is the primary factor in the improved performance [7].

In this research paper a BDA-embedded experimental architecture for smart cities is proposed. The BDA-integrated smart city architecture supports two key functions. First off, it makes it easier to use UBD for smart city planning, design, and maintenance. Furthermore, it requires BDA to handle and process massive amounts of UBD in order to improve the calibre of urban services. Three layers make up the architecture, which is responsible for managing real-time data, aggregating data, and service provisioning. To further improve offline and online data processing tasks, data normalizing and data filtering techniques are incorporated into the proposed work. Analyzing legitimate datasets allows to determine the threshold values needed for urban planning and city operation management. The aforesaid legitimate datasets are used to get performance metrics concerning both offline and online data processing for the proposed dual-node Hadoop cluster. Deep-learning algorithms could be used in decision-making in the future to further enhance this work [8].

V. CONCLUSION

Big data and smart cities are two contemporary ideas that are crucial, thus many people have begun combining them to build smart city applications that will improve resilience, support sustainability, better asset management and improve overall quality of life or livability in the city. This objective of this paper is to provide a concise picture of the role of big data in a smart city. A structured literature review is presented where we explore the definition and concepts of smart city as well as the description and characteristics of big data with straightforward examples. Big data analytics plays a major role in every facet of a smart city. When it pertains to the development of smart cities, it has proved to be useful in a variety of areas. Different proposed frameworks are also delved in our study. The value of smart city technology investment is ultimately realized

ISSN 2394 - 7780

through the use of data to strengthen the quality and efficiency of services, enable real-time operational control, enhance decision-making, and improve engagement with citizens.

REFERENCES

[1] Andrea Molinari, Vincenzo Maltese, Lorenzino Vaccari, Andrea Almi and Eleonora Bassi, "Big Data and Open Data for a Smart City," IEEE-TN Smart Cities White Papers, December 2014.

[2] Arulkumar V, Charlyn Pushpa Latha and Daniel Jr Dasig, "Concept of implementing Big Data in smart city: Applications, Services, Data Security in accordence with Internet of Things and AI," in International Journal of Recent Technology and Engineering(IJRTE), September 2019.

[3] Chiehyeon Lim, Kwang-Jae Kim and Paul P. Maglio, "Smart cities with big data: Reference models, challenges, and considerations," Elsevier, April 2018.

[4] Devesh Kumar Srivastava and Ayush Singh, "Big Data Analytics towards a Framework for a Smart City", Smart Innovation, March 2018.

[5] Shanet Sabu and Minla K.S, "Big Data Analytics Architecture for Smart Cities and Smart Companies," International Journal of Creative Research Thoughts(IJCRT), June 2021.

[6] Simon Elias Bibri and John Krogstie, "The emerging data-driven Smart City and its innovative applied solutions for sustainainability: the cases of London and Barcelona," Energy Informatics, May 2020.

[7] Mladen Amovic, Miro Govedarica, Aleksandra Radulovic and Ivana Jankovic, "Big Data in Smart City: Management Challenges", Applied Sciences, May 2021.

[8] Bhagya Nathali Silva, Murad Khan, Changsu Jung, Jinhun Seo and Diyan Mohammad, "Urbanic Planning and Smart City Decision Management Empowered by Real-Time Data Processing using Big Data Analytics," Sensors, September 2018

[9] Ahmed M. Shahat Osman, "A Novel Big Data Analytics Framework for Smart Cities," Future Generation Computer Systems, February 2019.

[10] Ibrahim Abaker Targio Hashem, Victor Chang, nor Badrul Anuar and Adewole K. S., "the Role of Big Data in Smart City," International Journal of Information Management, October 2016.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

AN EXPLORATION OF FEATURE EXTRACTION AND SEGMENTATION METHODS FOR MEDICAL DISEASE PREDICTION

Gourav Kochar and Dhiraj Khurana
UIET, M. D. University, Rohtak (Haryana)

ABSTRACT
A medical image's ability to be processed depends on the region and retrieved features.The medical photos must be quite specific in order for a segment to identify a tutor, ailment, or type of disease.To detect the disease or the tumor from the disease, an effective segmentation, and feature type are useful.Because of this, some sort of mechanism is needed to enhance the approaches for segmenting and extracting features from the medical image.In this study, feature segmentation and segmentation techniques for medical disorders are explored.The processing of medical images and related challenges have been covered in the study.The paper also looked at a number of image segmentation techniques to enhance medical images.

Keywords: Segmentation, Feature Processing, ROI, Medical Images,

I. INTRODUCTION
Medical disease processing and classification methods are critical application areas that require more accuracy and expert consultation than any other image processing approach. The effectiveness of medical image processing and disease prediction requires a multi-level analysis to detect the disease. Each method or mathematical model is somewhat unique and it cannot be applied to some other image or application. The functional process and responsibilities of medical image processing for disease prediction and detection are divided into four main stages[1][2][3]. These methods and related functional stages are shown in Figure 1. The main objective of medical image processing is to detect, predict or classify the disease or disease type. A classifier-driven framework is applied to detect and predict the disease. This framework contains four stages described in the Figure.

The first stage described in the figure is the image-cleaning stage. The medical images are high-resolution and high-quality images. The existence of minor impurities or disturbances in the medical image can be identified as some tumor or disease during analysis. An effective cleaning method is required within the preprocessing stage to improve the performance and reliability of medical disease prediction methods. The cleaning methods improve the image quality against noise, brightness, and contrast balancing. The cleaning method can repair the image and expose the effective features of the image [4][5].

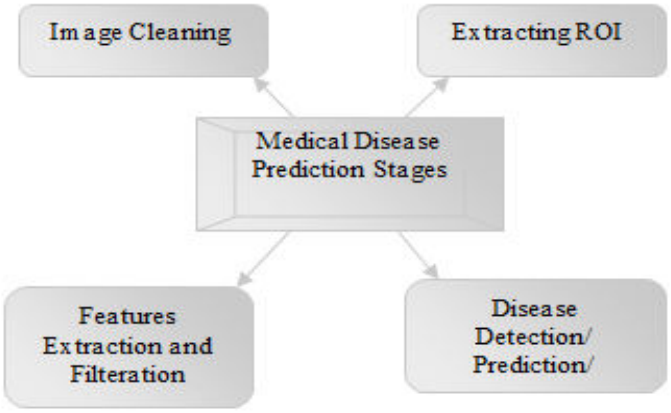


Figure 1: Stages of Medical Disease Prediction

After the filtration and cleaning stage, the next stage is to identify the effective and infected region from the medical image. The segmentation methods are applied in this stage to isolate the normal and infected regions. This stage locates the region of interest (ROI) and present it as the key region. The identification of ROI improves the efficiency and reliability of the disease prediction method. It avoids processing over the complete image which reduces the dimension of the dataset. The segmentation method is also effective to identify the features from the image. There are number of segmentation methods used to identify the effective region or features from the image. These methods include morphological operators, thresholding methods, clustering methods. After extracting the key region, the next work associated with medical disease processing is to extract the features. The features are the key aspects that can be captured from the image or effective region to isolate the normal and diseased person. The features in an image can be extracted based on various factors including

ISSN 2394 - 7780

geometry, color, texture etc. The key categorization of the features are textural features, statistical features, textural features etc. There are number of associated methods and measure to identify and extract these features. Some of these methods includes Principal component analysis, Local Binary Pattern, Local ternary Pattern, Discrete wavelet transformation, discrete cosine transformation etc. After extracting these features, the actual medical image is now represented by these features. Sometimes this features set is quite large and include some of unnecessary information, in such case some filteration methods can be applied to remove the unwanted features. There are number of available feature filters that ranks the features and remove the low ranked features. These methods include infogain, chisquare test, ReliefF filter etc. After this stage, the effective features are obtained, and these features are finally processed by the classifier or the predictor [5][6][7].

In the final stage, the classification or prediction method can be applied to process the features and to isolate the normal and diseased medical images. Various machine learning and deep learning methods can be applied to remove these features and to identify the image accurately. These classifiers are capable to identify the disease criticality or type also. Each of the disease prediction and classification method uses this stage to identify and classify the diseases. The medical disease prediction also depends on the organ type, scan type, image type and disease type. There are number of aspects associated to the medical disease that can be used to detect the disease region and the disease type. Each of the disease have different symptoms, infection and visible feature mark that are reflected in the medical image. The medical image processing is about to detect these variations and infection by applying a series of image processing operations and to identify the disease accurately.

The image format is the main aspect of medical image processing. The image formats define the resolution, and availability of information. The algorithms and disease detection methods are specific to the image type. There are number of specialized medical image formats like CT images, DICOM images, PGM images. Various kinds of scanning and capturing through specialized medical equipment are done to obtain these images. The performance and reliability of the medical image processing depends on the availability of effective information and its visibility. The high-resolution images with better information representation will give better results and accurate detection of disease. In this paper, the discussion is performed on various image segmentation and feature extraction methods that can be used to acquire the effective information or region from medical images.

1.1 Medical Image Segmentation and Feature Processing

Segmentation is an important aspect of medical image processing to acquire the ROI (Region of Interest) from images. In medical disease prediction, the sensitive and critical area is very limited and specific. In such case, the processing on the complete image is not effective in terms of performance and reliability. The segmentation method is applied as an intermediate stage to identify the key region over the image. The segmentation identifies the key area and effective features of an image. There are number of segmentation methods that can be applied on different images under different aspects and criteria. The segmented region can be identified by comparing the normal and abnormal region over the image. The feature-based analysis is performed on normal and abnormal images to locate the sensitive and critical disease region over the image. The variation analysis under different features is conducted to identify the key region over the medical image. There are multiple aspects and methods to identify the key regions. The color, textural or positional analysis is conducted using different methods and measures. The morphological operators, threshold-based methods and the clustering methods are most effective for segmentation. The segmentation is either performed at the pixel level, block level or the region level. In pixel level segmentation, each pixel is analyzed individually for multiple images. The intensity or other feature-based analysis is performed for this segmentation. The accuracy can be high, but the process is slow. In block segmentation, the complete region is divided into the smaller blocks and the statistical and feature based measures are applied on each block to separate the blocks based on their characteristics. The method is effective in terms of performance as well accuracy. The third method is region-basedsegmentation, in which the random areas are identified based on region similarity. Some optimization method with feature analysis is applied over the medical image to identify and separate these regions. The method is effective but does not ensure accuracy. If the region is very specific and smaller, it cannot ensure good results [8][9].

In this paper, a study on various segmentation methods is provided for medical disease prediction. The clustering and intensity-based clustering methods are discussed. In this section, the significance of medical image processing and role of segmentation in medical image processing is discussed. The functional stages of medical disease prediction are also discussed. In section II, some work defined by the earlier researchers on medical disease prediction is discussed. The research contribution in this area for different diseases is provided. In section III, the segmentation methods used effectively based the researchers for medical disease prediction are discussed. In section IV, the conclusion of the work is provided.

ISSN 2394 - 7780

II. RELATED WORK

The medical image processing is one of most popular and significant field of image processing. It is used to identify the chances on any disease instantly by observing the scan report. In this disease prediction process, the segmentation is one of the effective methods used to locate the critical disease region over the image. There are number of methods proposed by the researchers to identify the disease. In this section, some of the major contribution in segmentation methods is provided and discussed. A segmentation work CT and MR image was discussed using wavelet [10] based approach. The author used the multi resolution analysis-based segmentation method. The author applied both region and pixel-based segmentation methods to identify the key region and features of the medical image. Author analyzed the visibility-based features in these methods and performed a statistical analysis with an integrated wavelet approach to locate the disease region. The proposed method improved the segmentation process and improved the performance of disease prediction.

A mathematical [11] filter basedsegmentation work was proposed to optimize the segmentation on CT and MRI scanned images. The method was applied to brain images. In this work, a structured analysis of pathological and anatomical features was conducted to perform the segmentation. The single scale analysis based structural features were acquired and to improve the performance and effectiveness of segmentation process. The structured feature analysis was performed to improve the reliability and to achieve the better outcome for tumor prediction on the MRI medical images. A semantic [12] feature based analysis was performed to identify and separate the regions on medical image. The semantic features were used as the descriptor to perform the content-based analysis. The content driven feature analysis was performed to locate the disease region and to obtain the better result. The conceptual model was defined in this work to optimize the performance of the segmentation method and to gain a higher performance in the real environment. An edge region [13] based segmentation method was defined to extracting the effective inner region. The proposed method improved the behavior and the segmentation process over the medical images. The feature driven analysis was performed to match the normal and the edge area over the image. The statistical comparison on the appearance and intensity was conducted to perform the segmentation.

An adaptive and weighted [14] method was proposed to perform pixel based analysis on medical images. The region priority was performed under local features to performs the segmentation. The method was applied on MRI and CT images. The feature adaptive settlement was performed in the image to generate the effective region. The region-based fusion method was also defined to combine the separated region and to obtain a single region over the image. The proposed method reduced the complications over the image and method and improved the reliability. Another work on weighted features with distance based comparison was performed by Chai et al. [15]. Author defined the block-based analysis under frequency band features. The coefficient analysis was conducted on the extracted region and to separate the normal and the disease regions. The high-performancefeatures of the normal region is compared with distance measure to identify the critical region over the image. Another wavelet [16] based method for medical image segmentation was provided to identify the effective region over the CT image. It was decomposition-based method that performed a weighted evaluation under feature computation to separate the different regions of the image. Huang et al. [8] provided a work on contrast based analysis to perform the segmentation on CT images. The multi-phase analysis was performed to evaluate the block features. The segmentation method with mutual information was analyzed under frequency measure to perform the segmentation. Author defined a multimodal architecture to perform the segmentation on medical images. The frequency feature-based analysis was performed to improve the reliability and performance of the segmentation method.

Another multimodal [17] architecture was proposed to optimize the segmentation on medical images. The method was applied on tomographic images. This multimodal acquired the structural features with anatomic information to divide the image into smaller regions. The inverse problem was applied in this work with parameterized method to perform the segmentation and interpretation. The statistical interpretation method was applied under structural feature evaluation to detect the tumor region over the image. A curvelet [18] based segmentation method for medical region identification and separation was proposed. The wavelet-based approach was integrated with fusion method to separate the critical region over the image. The analysis results show that the method provided an effective outcome to generate an effective region and improved performance. Another wavelet based improved method and coefficient analysis based approach was defined by Yang et al. [19]. Author performed the low frequency band-based analysis method to identify the separated regions. The selection rule was defined to obtain the regional characterization in the region. The selection region was defined with frequency band measure to improve the effectiveness. The analysis results show the clear separation of the region and improved results.

ISSN 2394 - 7780

3. Medical Image Segmentation Methods/Models

In medical disease prediction, the images are of high resolution. The processing of such high-resolution images degrades the performance of the system and it cannot be implemented in real time environment. The segmentation methods are defined in these prediction-based algorithm as an intermediate stage to identify the key and critical region over the image. The segmentation reduces the dimension of the image by performing a deep analysis within the region. The effective segmentation method will identify the main critical region that can be infected region. There are number of mathematical, texture based, statistical and feature adaptive analysis method for improving the behavior of segmentation process [20][21][22]. Some of the common and effective segmentation methods are discussed in this section.

3.1 Intensity Based Approach

It is one of most popular and useful method used in medical and normal images. This method can be applied at block and pixel level. In this method, the region adaptive intensity is performed to separate the pixels of normal and disease region. The minimum intensity based positional analysis can be performed over the block. The pixel averaging on the max or intensity count measures can be applied to perform block level segmentation. The multiple measures and methods can be integrated within this method for improving the performance and reliability. It is one the simplest method but many of researchers used this method with certain chances and amendments. The region intensity-based block analysis feature analysis was performed to perform the segmentation. The algorithm for intensity-based region segmentation for medical image is shown here:

Algorithm 1: Intensity based Medical Image Segmentation

Intensity Segmentation (MImages)

MImages is the set of medical images

{

1. Divide the MImages in the smaller blocks of mxm
2. Read the blocks of each image one by one and perform the following steps
3. Compute the intensity for same block of all MImages
4. Compute the average intensity of all blocks
5. Find the difference between the average intensity and intensity of any ith image
6. Apply the threshold conditions to separate the regions over the images.
7. Move the block in particular segment based on the intensity threshold analysis.

}

3.2 Pyramid based Transformation Method

Another segmentation is a transformation-based method called Pyramid based segmentation. This method uses the transform domain to perform the segmentation. The transformation method was applied to divide the image into low and high pass bands. This band passes are able to divide the image based on the pattern image. The gaussian pyramid-based method on the sequence of medical images and perform the low pass transformation over it. The segmentation method is also applied based on some limit under the higher-level consideration. The spatial frequency-based analysis was applied to perform the segmentation. A fusion feature-based method was also integrated within the medical image to apply the pyramid based transformation. Figure 2 shows the structure of pyramid-based segmentation method.

ISSN 2394 - 7780

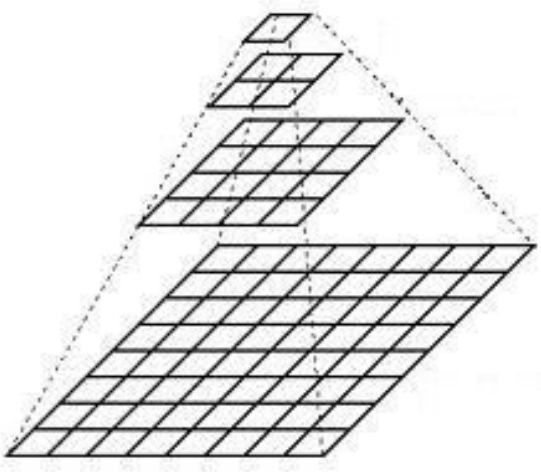


Figure 2: Pyramid based Segmentation

It is one of the most adaptive approaches that perform the task with the specification of three connected functional stages. In the first stage, image decomposition is performed. In this stage, the pyramid adaptive levels are defined, and the segmentation process is performed for each level. The correlation-based analysis is performed at each level to perform the decomposition. The decomposition is performed by using the level-based method. In the second stage, the formation of the pyramid is performed at each level. Each level of this pyramid shows a type or segment of the image. In the final stage, image reformation is performed by collecting the segmentation. The fusion process is applied in the pyramid-based method to regenerate the image. In this formation stage, the merging is performed. After highlighting the regions, the image is finally composed and presented as the segmented image.

3.3 DWT

Discrete Wavelet Transformation (DWT) is a wavelet-based transformation and decomposition method. The method divided the image based on the frequency band. The frequency is the content-basedmeasure and the operational analysis is performed to generate the scales of the frequency. The temporal and scale value analysis are performed over the image for decomposition. The property driven decomposition is performed and the image is divided at row and column level under Low and High frequency segments. In the multi-level decomposition LL, LH, HL, and HH segments are also generated. Figure 3 shows the 1-level decomposition using wavelet transformation. There is number of DWT forms that define the type of scaling and decomposition. Haar is the common wavelet decomposition form. The effectiveness and process of DWT can be controlled under the scaling factor.

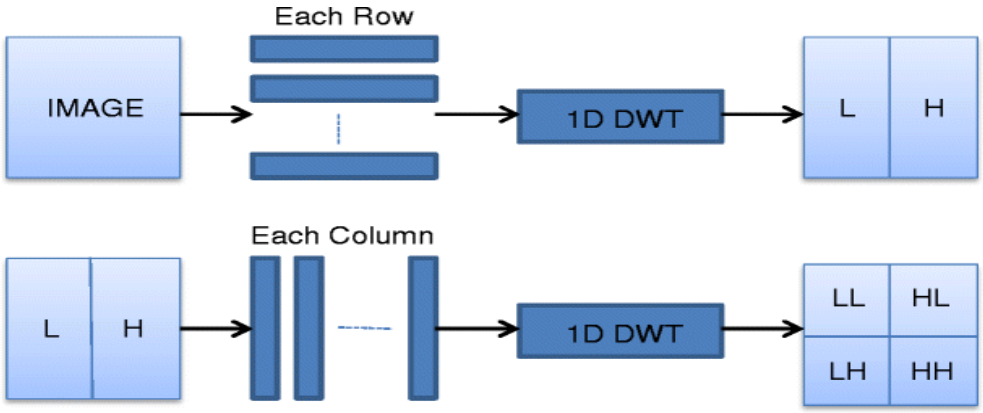


Figure 3: Structure of DWT-based Segmentation

IV. CONCLUSION

The medical images and the disease prediction is these images is one of the major aspects of research. The effectiveness of these methods depends on the effective segmentation and disease region identification. There are a number of segmentation methods proposed by the researchers. In this paper, a study on the significance and effective usage of the segmentation process is defined for medical images and medical disease prediction. The paper has defined various segmentation methods are defined. A detailed study on DWT, pyramid and Intensity-based segmentation methods are defined in this paper.

ISSN 2394 - 7780

REFERENCES

1. Sarvamangala, D. R., and Raghavendra V. Kulkarni. "Convolutional neural networks in medical image understanding: a survey." Evolutionary intelligence 15, no. 1 (2022): 1-22.
2. Giger, Maryellen L. "Machine learning in medical imaging." Journal of the American College of Radiology 15, no. 3 (2018): 512-520.
3. Maier, Andreas, Christopher Syben, Tobias Lasser, and Christian Riess. "A gentle introduction to deep learning in medical image processing." Zeitschrift für Medizinische Physik 29, no. 2 (2019): 86-101.
4. Castiglioni, Isabella, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. "AI applications to medical images: From machine learning to deep learning." Physica Medica 83 (2021): 9-24.
5. Raza, Khalid, and Nripendra K. Singh. "A tour of unsupervised deep learning for medical image analysis." Current Medical Imaging 17, no. 9 (2021): 1059-1077.
6. Aggarwal, Ravi, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis." NPJ digital medicine 4, no. 1 (2021): 1-23.
7. Zhou, S. Kevin, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises." Proceedings of the IEEE 109, no. 5 (2021): 820-838.
8. Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M. and Falk, T.H., 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Information Fusion, 76, pp.355-375.
9. Siddique, Nahian, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. "U-net and its variants for medical image segmentation: A review of theory and applications." Ieee Access 9 (2021): 82031-82057.
10. Sekhar, A. Soma, and MN Giri Prasad. "A novel approach of image fusion on MR and CT images using wavelet transforms." In 2011 3rd International Conference on Electronics Computer Technology, vol. 4, pp. 172-176. IEEE, 2011.
11. Matsopoulos, G. K., S. Marshall, and J. N. H. Brunt. "Multiresolution morphological fusion of MR and CT images of the human brain." IEE Proceedings-Vision, Image and Signal Processing 141, no. 3 (1994): 137-142.
12. Miao, Yumei. "A conceptual model integrating semantic content into CT image fusion." In 2004 Asia-Pacific Radio Science Conference, 2004. Proceedings., pp. 427-430. IEEE, 2004.
13. Jin, Li, Zhou Lulu, Yu Hong, and Liang Hong. "Multi-region Segmentation of CT Images Based on Information Fusion." In 2006 1ST IEEE Conference on Industrial Electronics and Applications, pp. 1-6. IEEE, 2006.
14. Yi, Shen, Jiachen Ma, and Liyong Ma. "An adaptive pixel-weighted image fusion algorithm based on local priority for ct and mri images." In 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, pp. 420-422. IEEE, 2006.
15. Chai, Yong, You He, and Chaolong Ying. "CT and MRI image fusion based on contourlet using a novel rule." In 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 2064-2067. IEEE, 2008.
16. Cheng, Shangli, Junmin He, and Zhongwei Lv. "Medical image of PET/CT weighted fusion based on wavelet transform." In 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 2523-2525. IEEE, 2008.
17. Huang, Xiaoyang, Boliang Wang, Ming Cheng, Shaohui Huang, and Ying Ju. "Image registration and data fusion for different phases of contrast enhanced liver CT data." In 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, pp. 2682-2685. IEEE, 2008.

ISSN 2394 - 7780

18. Hyde, Damon, Eric Miller, Dana Brooks, and Vasilis Ntziachristos. "New techniques for data fusion in multimodal FMT-CT imaging." In 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1597-1600. IEEE, 2008.
19. Ali, F. E., I. M. El-Dokany, A. A. Saad, and F. E. Abd El-Samie. "Fusion of MR and CT images using the curvelet transform." In 2008 National Radio Science Conference, pp. 1-8. IEEE, 2008.
20. Yang, Y. and Huang, S., 2009, June. Fusion of CT & MR Images with a Novel Method Based on Wavelet Transform. In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (pp. 1-4). IEEE.
21. Parmar, Kiran, Rahul K. Kher, and Falgun N. Thakkar. "Analysis of CT and MRI image fusion using wavelet transform." In 2012 International Conference on Communication Systems and Network Technologies, pp. 124-127. IEEE, 2012.
22. Carminati, Maria C., Francesco Maffessanti, Paola Gripari, Gianluca Pontone, Daniele Andreini, Mauro Pepi, and Enrico G. Caiani. "A framework for CT and MR image fusion in cardiac resynchronization therapy." In 2012 Computing in Cardiology, pp. 537-540. IEEE, 2012.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

STOCK PRICE PREDICTION USING MACHINE LEARNING ALGORITHMS

Neelam Naik, Prashant Chaudhary, Melissa D'souza and Isha Manjrekar
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Maharashtra, Mumbai-400056

ABSTRACT
Stock price prediction can be regarded as a critical task and it is highly valued. The successful prediction of stock prices can result in desirable net profit through smart decisions. However, stock market forecasting is a significant challenge due to ambiguously defined assessments and chaotic data. As a result, it becomes difficult for investors to allocate their money in order to gain profit. Multiple methodologies are currently being developed to accurately determine stock market volatility. This paper provides a thorough examination of several research papers proposing methods and approaches for stock price data prediction, such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), Fuzzy approaches and Machine Learning Methods. The papers obtained for literature review are categorized based on stock market prediction parameters, data sets, and various approaches being used to the prediction of stocks. To understand and analyze the stock price, the authors have applied six algorithms on the secondary dataset acquired to find the best fit that gives the highest accuracy and efficiency.

Keywords: Stock prediction, Stock market, stock price, ANN, Machine learning, Random Forest

I. INTRODUCTION
Stocks represent a share of possession in a specific company. The stockholder has a claim on a company's assets in exchange for the money paid to purchase the stock. As an owner, the stockholder shares the wealth and the risk of ownership with other company owners. A company issues stock to raise money to expand and build its operations.

The value of a company's stock influences investors' interpretation of its upcoming potential to earn and increase cash flow. If stockholders are happy and as illustrated by the share value, the organization is prospering. It is said that the broader a business's stock value, the nicer its forecast. A rapidly increasing market capitalization indicates that the top executives of a company are pulling transactions toward revenue growth.

The share market is full of uncertainty and inconsistency. Stock market forecasting predicts the expected activity of the stock price of a capital market. Precise and reliable forecasting of stock prices will lead to far more monetary gain for buyers. It is difficult to forecast the direction of equities because many factors, such as economic growth, interest rates, and politics, make the market volatile and difficult to predict precisely.

II. FACTORS AFFECTING STOCK PRICE
Demand and Supply: A stock's price will fluctuate in response to an imbalance in Demand and availability just like any other commodity. If an organization is flourishing and everybody desires to invest in its stock, it will create scarcity, causing the stock price to rise. The reverse occurs when an excessive number of stocks are available yet no one desires to purchase them. In that case, the stock value will drop.

Political Scenario: If the Government has good policies for development, investors will invest with enthusiasm, but if the government has a weak agenda, then it can cause a decrease in Stock prices. This is among the most significant factors that influence the Indian market.

Government Policies: Government policies have a substantial impact on the capital market's share prices. If investors believe that government policies are favorable, share prices in that industry and sector tend to rise. Investors may lose faith if unfavorable policies, particularly those involving taxation, are implemented.

Status and Performance of the Company: The share values of companies in a similar business world will continue moving in lockstep. This occurs due to market conditions generally having the same impact on companies in the same industry. Furthermore, if two organizations bid in a similar market, a piece of adverse news for either of them may aid the stock price of the other firm.

Inflation: Inflation causes a rise in commodity prices. This frequently reduces sales growth. Price increases frequently result in increased interest rates. For example, in order to slow inflation, the Bank may raise interest rates. Stock prices may collapse as a result of these changes. Commodity markets, on the contrary side, may fare more favorably in terms of inflation, leading to higher costs.

ISSN 2394 - 7780

Investor Sentiment: The market can move up or down due to investor sentiment, also known as confidence. This can create stock prices to climb or crash. The stock market's overall direction, which affects the benefit of a stock, is as follows: A bull market is a powerful share market wherein share price increases and investor confidence increases. A bear market occurs when stock prices fall and investor confidence declines. This occurs if a market is in a slump, labor shortage, and when the prices keep climbing.

Interest rates: To sustain or spur economic growth, financial institutions may also elevate or reduce interest rates. These are recognized as monetary policy. Rising interest rates can influence the price of a company's debt if it borrows money to broaden and boost its operations.

Dividend Declaration: Dividends are post-tax assets that are paid to shareholders and can be expressed in rupees or percentages. For example, if the face value of the stock is Rs. 10 and the company declares a 50% dividend, it means that shareholders will receive a dividend of Rs. 5 per share.

Turnover: On a specific business day, market turnover is the proportion of market activity that occurred in the economy as a whole or in individual stocks. There are two ways to represent turnover in some trades: traded value in rupees and traded volume.

Natural Calamities Affecting the Stock Market: Natural catastrophes such as floods and earthquakes have an adverse effect on stock market prices. This can happen for a plethora of reasons, including damage to property and other assets. Companies suffer significantly resulting in falling share prices.

Current Events that Impact the Stock Market: News and various current events have a considerable amount of influence on the stock market. Current events that have a significant amount of impact on the stock market involve political unrest, civil war or riots, and terrorist attacks. Each of these events will almost certainly cause stock prices to decline and market volatility to rise.

Exchange Rates Influencing Share Prices: Another factor influencing share value in India is the value of the Indian rupee in comparison with the US dollar or any other foreign currency. A strong rupee signifies that our economic system is booming, which means that stock prices will rise.

Monetary Policy of RBI and Regulatory Policies of SEBI: The Reserve Bank of India (RBI) is the supreme authority in charge of monetary policy in India. The RBI is constantly enhancing its monetary policy. Stock prices are affected due to constant change in Repo & Reverse Repo rates.

Gold Prices and Bonds: Stocks are perceived as high-risk investments, whilst gold and bonds are regarded as safe havens. So as a result, investors prefer to engage in safe assets during a major global recession. Along with that, the prices of gold and bonds rise while stock prices fall.

III. PROBLEM STATEMENT

The aim of the present study is to build a model using secondary data, which will predict the stock price on the basis of the parameters namely open, close, high, low, last, turnover, total trade quantity, difference and value of the stock market price.

IV. LITERATURE REVIEW

This paper [2] examined various stock market manipulation techniques. ANN and the fuzzy-based approach are broadly utilized techniques for efficacious stock market prediction. The main problem that stock market prediction systems face is that most established works can't be identified using historical price data.

In this paper [12], the author suggests a PCA-SVM() embedded model for forecasting the direction of share market indices and specific stock prices. Their findings also confirm the impact of Korean co-movement (or Hong Kong) and US stock markets due to the use of the S&P 500 and exchange rates.

In this paper [1], the author used ten events, five of which were terrorist attacks and five of which were natural disasters. According to the study, terrorism has quite a short-term effect on stock market prices, but investors may react by either purchasing or selling their stocks because some people act out of fear. Due to a volatile market, the study also advises making purchase or sale decisions for short revenue. Long-term investors, on the other hand, do not need to react toward such market sentiments since the effect of this act can last only one to five days. Finally, the paper found no effect of natural disasters such as floods or tsunamis, implying that investors should not be concerned because such events have no effect on the stock market because they are limited to a specific region.

In this paper [8], for IMMT, the author used simplistic machine learning approaches along with deep learning and prediction utilizing long short-term memory neural networks. The three models' average accuracy is 53.2%,

ISSN 2394 - 7780

whilst the A-share stock index is 57%, although both are significantly greater in comparison to the stochastic forecast. Models are highly accurate in both the European and American financial markets.

In this exploratory paper [10], the author has come to the conclusion that the currency value, rate of interest, and return on equity (roe) all have an impact on stock returns. When the Rupiah's rate of exchange against the US Dollar rises, international investors typically sell their stock holdings to avoid the anticipated return losses. Stock returns are adversely affected by interest rates; as interest rate increases, stock returns fall.

The author of this paper [9] conducted a stock price prediction survey using machine learning. The author predicted the market using regression and its variants. Prediction is an essential yet complicated and difficult process in the stock market. Using conventional techniques such as fundamental and technical analysis may not verify prediction accuracy. In this paper, the author's survey of a well-established effective regression approach to predict stock price from market data is very critical.

The author of this paper [11] predicts the stock market using ML techniques such as Radial Basis Function and Support Vector Machine. The paper depicts a programme for predicting economic data that holds a raw data of historical share prices. This dataset is used as a training set. Prediction aims to minimize the ambiguity associated with making investment decisions. Using machine learning techniques, we can forecast the upcoming prices of a firm's stocks with greater accuracy and precision. The researchers' most important contribution was the execution of the LSTM Model to determine stock prices. Other ML surfacing models can be explored to see how accurate they are.

In this paper [4] the author discusses the uncertainty of the Indian market and points down the main factors that are accountable for the unsteady stock market. The author identifies the following factors as major sources of volatility in the Indian stock market: inflation, political stability, liquidity, GDP growth, the flow of foreign institutional investors and different interest rates, and global-level factors. The research was carried out with the help of a Descriptive Research design. The author asserts that every research model has limitations. There is no Research Model or theory that does not have a single limitation.

This paper [13] overlooks the effects of positive and negative security and political activities on stock demand indicators, using the Beirut Stock Exchange as an example. It highlighted the highly crucial political events and judgements that occurred in Lebanon over the years (2017- 2020). The Lebanese profitable pointers demonstrate that security and political factors throughout Lebanon have quite a significant influence on investors, who tend to be highly susceptible to just about any political news. Investors sell their portfolios in response to any negative news due to Lebanon's elevated threat level and volatile political situation.

The author of this paper [7] states that ML algorithms and Econometric Models are confirmed to be two of the most efficient general methodologies with high accuracy and stability in the stock market prediction domain. The analysis of each model is presented separately, with an overall comparison of ML algorithms and Econometric Models. Overall, a total of four ML algorithms are discussed, which are LR, KNN, SVM, and LSTM. The authors can see the combination of SVM and KNN exhibits an even better prediction capability than the SVM or KNN. From the perspective of Econometric Models, three models are discussed and summarized in total: the ARIMA model, CAPM, and FF Factor Model. In the end, it is concluded, ML algorithms tend to have relatively higher accuracy than Econometric Models with a much longer training time in most cases.

[6] According to the authors, forecasting the share market during the COVID-19 pandemic was difficult as the information wasn’t an arbitrary, stationary, and complex nonlinear system. The medical industry in the health sector, as listed on the IDX, is the focus of this study. The proposed model for forecasting stock prices includes COVID-19 trend indicators and the Indonesian government's response tightness index to COVID-19 as input variables. All conceptual model systems obtain strong precise estimations for the price of stocks prediction, according to the research findings. Standard strategic indicators such as, Root mean square error (RMSE) MSE, and MAPE (Mean Absolute Percentage Error) are utilized to evaluate the models. All deep learning models can make highly precise and reliable predictions, in accordance with the results. As a result, with limited data availability, During COVID-19, the ANN-based model can generate precise projections for stock market projections. During COVID-19, a new technique for acquiring highly relevant stock market predictions was postulated using Google Trends and GUI features.

According to the authors of this research paper [5], stock market prediction can be achieved by applying machine learning techniques on data collected from social media platforms and financial news. Experiments show that the New York and Red Hat stock markets are tricky to estimate, whereas the London and Microsoft

ISSN 2394 - 7780

stock markets are heavily affected by financial news. This article provides a blueprint for forecasting stock market trends in the future that use social media and the news as extrinsic variables. The authors looked at how social media and financial news affected stock forecasting for the next ten days. They discovered that by including sentiment criteria, social media shows greater impact on stock prediction on day 9, whereas financial news has a bigger influence on days 9 and 8.

The research paper [14] contains records pertaining to the stock price dataset of the well-known company Tata Global Beverages Limited. This dataset includes parameters such as a stock's price with opening, closing, high, and low prices, as well as volume traded and turnover on that day.

V. RESEARCH METHODOLOGY

A. Classification and Prediction

Data analysis is typically divided into two types, which are able to extract models depicting various important classes or for predicting future data patterns. The two forms are Classification and Prediction. The importance of classification and prediction is seen while extracting a model that represents the data classes which helps to plot future trends in data. Classification uses prediction models to forecast categorized labels of data. Classification techniques predict categorical class labels, whereas prediction models infer continuous-valued functions. Regression is a widely used methodology for prediction whereas clustering is most commonly used for classification.

B. Types of Algorithms

A feed-forward neural network that gives an array of outcome from a learning algorithm is known as a multilayer perceptron (MLP). MLP is distinguished by multiple layers of input nodes and is represented as a directed graph inside the hidden state and the output. Backpropagation is used by MLP to instruct the network. MLP is a powerful learning method.

Formula: weight = weight + learning_rate * (expected- predicted) * X

J48 is an algorithm inspired by Ross Quinlan that is applied to produce a decision tree. Quinlan's initial ID3 algorithm is extended in C4.5. C4.5's decision trees are used for classification, hence the reason why sometimes it adhered to as a Classification algorithm. It rose to prominence after ranking first in the pre-eminent paper Top 10 Algorithms in Data Mining published by Springer LNCS in 2008.

Random Forest is perhaps a classification method that evaluates the estimation of different decision trees on different subsets of a specific dataset to improve the data's expected accuracy. So instead of relying on a single decision tree, the random forest aggregates the results of all trees and indicates the overall result based on the majority of prediction votes.

Classification via Regression uses the principle of a decision tree algorithm as well as linear regression. This method is divided into two major steps:

1. Build a standard decision tree by maximising the splitting of criteria, parameters, and attributes, in addition to their variations depending on the target/output values. When creating a decision tree, the deviation reduction standard must be calculated.
2. Pruning the decision tree into several sub-trees & filling this with a regression function as required, usually on leaves. The second method was the random committee. Such a strategy is perhaps a designed ensemble mechanism based on a handful of shaky hypotheses.

Logistic regression is a machine learning (ML) classification technique which predicts the prospect of particular groups based on dependent variables. To summarize, the modeling approach computes the results logistically by introducing the feature values (most of the cases have a biased term). The logistic regression output is always between 0 and 1, which is appropriate for a binary classifier. The greater the value, the more likely the existing sample is categorized as class=1 and vice versa.

The Naive Bayes Classifier is a straightforward and effective Classification technique that facilitates the development of fast ML models capable of producing accurate predictions. It is probabilistic, which means it anticipates based on the likelihood of an object. The Naive Bayes Algorithm is widely used for sentiment analysis, spam filtration and article classification.

C. Dataset Used

The authors have collected a secondary dataset containing records of the stock price of Tata Global Beverages Limited [3]. The dataset includes the date-wise price of the company stock along its high, low, opening, and

ISSN 2394 - 7780

closing prices in addition to the volume traded along with the turnover on the day. It's a very effective and excellent dataset for data modeling and various forms of data processing. In the Tata Global Beverages Dataset, the author has made a few necessary changes to make the dataset more efficient for the implementation of models. Two additional columns were inserted in the dataset: Difference and Value. The "Difference" column displays the difference between the "Close" column and the "Open" column. The "Value" column is dependent on the "Difference" column. The columns with a negative value display "Decrease" in stock value and vice versa.

Weka tool provides the statistically computed output of the model processing. It includes a visualization tool for inspecting the given data. The different classification algorithms are applied on the dataset. The authors have compared the outputs of various algorithms and chosen the best one as per the accuracy of prediction.

VI. RESULTS AND DISCUSSION

Authors have reviewed the research paper on the following three categories: [i]Factors, [ii] Datasets, and [iii] Algorithms that affect the Stock market. Every parameter focused on different factors that affected the stock prices for example, as depicted in the review paper [1] the calamities did not bring much of an effect on the stock market as compared to the situation of terrorism where the act of fear led to a collapse in the stock prices. Use of regression, forecasting, and deep learning is also seen here. The authors focused on a secondary dataset that spoke about the Global Tata Beverages Stock Data. The Authors have also applied the following tests on the dataset: Multilayer Perceptron, J48, Random Forest, Classification via Regression, Logistic Regression and Naive Bayes.

After analyzing the models, it is determined that Random Forest Algorithm works well with the dataset as the accuracy is the highest and time taken to implement fits best. Following Random Forest Algorithm, Logistic Regression algorithm is the next best algorithm for the chosen dataset.

Reduced overfitting risks and quicker training periods are features of the Random Forest Algorithm. It also has a very high level of accuracy. The Random Forest method performs well in large databases and delivers exceptionally accurate predictions when used to approximate missing data. It can do all classification and regression tests. It handles missing values and maintains good accuracy even when large amounts of data are missing. Therefore, Random Forest worked the best with the chosen secondary dataset giving the highest accuracy as compared to the other algorithms.

Table I. Algorithms and their prediction accuracy

Sr. No.	Algorithm	Accuracy	Build Time
1	Multilayer Perceptron Model	97.7%	1:63 secs
2	J48 (Decision Tree)	98%	0.06 secs
3	Random Forest	99.9%	0.52 secs
4	Classification via Regression	98.9%	0.26 secs
5	Logistic Regression	99.5%	0.28 secs
6	Naive Bayes	75.43%	0.02 secs

VII. CONCLUSION

This article provides an examination of various classification and prediction algorithms used for stock price forecasting. The purpose of this study is to identify and categorize the distinct parameters, methodologies implemented, datasets used, and existing techniques, by taking reference of several research studies. It is found that the technologies such as PCA-SVM, LSTM, KNN, ARIMA, RBF(Radial Basis Function), forecasting, shallow machine learning, deep learning, and regression are used to predict the stock price. The majority of the current methodologies are not able to determine stock price based historical data as they are influenced by a variety of factors, including market sentiment, government policy beliefs, and other factors.

After implementation of different algorithms on the secondary dataset, the authors conclude that Random Forest works the best in terms of accuracy followed by the Logistic Regression algorithm.

The present study can be improved by considering almost all parameters mentioned in this paper to predict stock price. In the future, the authors want to focus on creating a model that predicts the stock price with a better accuracy and efficient outcome.

ISSN 2394 - 7780

REFERENCES

[1] Azhar, Syed & Ramesh, Dr & Rai, Akshita. (2020). Effects Of Terrorism And Natural Calamities On Stock Market. International Journal of Recent Technology and Engineering (IJRTE). 8. 1082-1086. 10.35940/ijrte.E5909.018520.

[2] Dattatray P. Gandhmal, K. Kumar, "Systematic analysis and review of stock market prediction techniques, Computer Science Review", Volume 34,l 2019, 100190,ISSN 1574-0137.

[3] https://github.com/Matt-Jennings-GitHub/Tata-Global-Forcasting/blob/master/NSE-Tataglobal.csv

[4] Joshi, Mrunal. (2013). Factors Affecting Indian Stock Market. SSRN Electronic Journal. 10.2139/ssrn.2238539.

[5] Khan, Wasiat & Ghazanfar, Mustansar ali & Azam, Muhammad Awais & Karami, Amin & Alyoubi, Khaled & Alfakeeh, Ahmed. (2022). Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing. 13. 10.1007/s12652-020-01839-w.

[6] Melina, Melina & Sambas, Aceng & Firman, Sukono. (2022). Artificial Neural Network-Based Machine Learning Approach to Stock Market Prediction Model on the Indonesia Stock Exchange during the COVID-19. Volume 30, Issue 3: September 2022

[7] Ni, Zhehan & Chen, Weilun. (2022). A Comparative Analysis of the Application of Machine Learning Algorithms and Econometric Models in Stock Market Prediction. BCP Business & Management. 34. 879-890. 10.54691/bcpbm.v34i.3108.

[8] Pang, Xiongwen & Zhou, Yanqiang & Wang, Pan & Lin, Weiwei & Chang, Victor. (2020). an innovative neural network approach for stock market prediction. The Journal of Supercomputing. 76. 10.1007/s11227-017-2228-y.

[9] Sharma, Ashish & Bhuriya, Dinesh & Singh, Upendra. (2017). Survey of stock market prediction using machine learning approach. 506-509. 10.1109/ICECA.2017.8212715.

[10] Trajadi, Evita. (2022). Effect of Exchange Rate, Interest Rate and Return On Asset on Stock Return. Indikator: Jurnal Ilmiah Manajemen dan Bisnis. 6. 118. 10.22441/ indikator.v6i3.15457.

[11] Vanukuru, Kranthi. (2018). Stock Market Prediction Using Machine Learning. 10.13140/ RG.2.2.12300.77448.

[12] Wang, Yanshan. (2014). Stock Price Direction Prediction by Directly Using Prices Data: An Empirical Study on the KOSPI and HSI. Int. J. Bus. Intell. Data Min.. 9. 145-160. 10.1504/IJBIDM.2014.065091.

[13] Youness, Mohamad. (2022). the Impacts of the Political Factors on the Stock Market Index. 13. 168-178. 10.15341/jbe (2155-7950)/03.13.2022/006.

[14] Tata Global Beverages Stocks Data, https://data-flair.training/blogs/download-tata-global-beverages-stocks-data/

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

INVESTIGATING THE SECURITY AND IMPLEMENTATION CHALLENGES OF OAUTH2.0 IN MICROSERVICE

Manisha Divate, Sunita Gupta, Ashish Mondal and Mihir Jalgaonkar
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Maharashtra, Mumbai, 400056

ABSTRACT

The Aim is to make a review study on Authorization platforms available to authenticate web services in the microservices era, a rapid demand for APIs (Application programming interfaces) and microservices in enterprises like banking sectors, social media, mobile computing, web 3 and cloud-based web services.

It constantly keeps storing users' data on large scale, data like (credentials, daily basis activities, payment information, personal information etc).

The major aspect of APIs is used to communicate between the microservices or MSA it is been increasingly applied by MNCs.

This is the way of creating a software in which the services of a single application are decomposed and then deployed and execute separately, the individual services communicate via API.

In this article, we are going to examine the best security combination that goes with Oauth2.0 and we are studying this in spring framework and OAuth2.0 on microservices Architecture API which is built using java maven spring framework, we are doing a literature review.

[12] [R. Yang, W. Cheong Lau, T. Liu] Modern Identity Providers (IdPs) have adopted the OAuth2.0 protocol in large numbers to offer single-sign-on services. Since OAuth2.0 was originally developed to satisfy the need for authorization for third-party websites, a variety of difficulties have been encountered while implementing it to support mobile app authentication.

As far as we are aware, every assault found thus far, including those at Blackhat USA'16 [15], CCS'14 [14], and ACSAC'15 [17], requires interaction with the victim.

Examples of this interaction include using malicious apps or network eavesdropping.

We will discuss the problem regarding OAuth2.0 as it is an authorization protocol and is most used by companies.

Keywords: Microservices and architecture MSA, Monolithic Architecture, APIs, Security threats, Security in Spring Framework, authorization platform like OAuth2.0, OpenID.

I. INTRODUCTION

Here we go with the concept of security in APIs, as we know nowadays web services are moving towards microservices architecture rather than using monolithic architecture.

In the microservices architecture the code is written separately so it makes it easy to deploy and scale without interrupting other services.

Here the code works on the request-response mechanism it use to communicate between service to service or service to the database or front-end to back-end.

The communication is done sending a request in form of JSON (JavaScript Object Notation) or XML format, here the risk stands between the two API's or services.

This innovation has brought some new security threats can interrupt the data transferred in between, there are some security measures we can apply, measures like digital signature, encryption-decryption(cryptography), Hashing algorithms (SHA256), SSL certificate (Secure Socket Layer), Authorization platforms (OAuth2), Spring-security, OpenID connect, Zero downtime architecture, Kerberos Architecture, etc.

As we are going to demonstrate using the Spring framework, we are going to cover the concepts of Spring Security and OAuth2.0 in brief.

OAuth2.0 is a protocol to authorize the user to access the web page or service, OAuth2.0 is an Open Authorization Protocol.

ISSN 2394 - 7780

Which allows the user to access resources of the resource owner by enabling client applications on HTTP services like Facebook, Google Etc.

OAuth2.0 is an authorizing framework that enables the application web security to access the resource from the client.

II. The Need for API Arises and why we Should Secure Them

That's a pretty much valid question that arises in the mind of young developers or academic's students, nowadays enterprises believe in individual programmes which are independent of each other.

They might be written in any different language and make them communicate between here the concept of API arises now if one programme is written in java and its needs to communicate with a programme written in python so what should it do, because both are in different language how they should transfer the data.

Here the concept API comes in the picture they communicate via API i.e., request and response, they use JSON or XML format for data transferring.

When it comes to security it's an important concern as increasing innovations are also leading to an increase in cyber-attacks when it comes to API it is worst if we don't secure our APIs cause it can lead to violation of sensitive data, stealing credentials and tempering with them so here OAuth2.0 plays an important role in securing API's, API is very commonly used in web development and as it enables to access the sensitive software functions and data, they are becoming a primary target for attackers.

APIs may have vulnerabilities like broken authentication or authorization, code injection and lack of rate limiting, Organizations must check on these vulnerabilities and resolve these vulnerabilities using security best practices.

Using OAuth2.0 the application can access the user's data without disclosure of the user's credentials to the application.

In the context of microservices, the OAuth2.0 client credential flow supports secure server-to-server communication between client API and server API.

A. The Top Security Threats

1. With broken user authentication, an attacker may take advantage of an incorrectly applied Authentication mechanism.
2. Broken Object-level Authorization, due to the fact that APIs typically make endpoint handling object identifiers available, any methods that need user input to access the data source may result in a level access control problem.
3. Excessive data exposure, developers usually rely on the client end to filter the data before making it available to the user, however this poses a serious security risk because only important data should be sent to the client side.
4. Lack of resource and rate limiting, in many cases, APIs don't limit the quantity or size of the resources that the client or user requests. This can affect the speed of the API server, cause a denial of service (DoS), disclose authentication flaws, and open the door to brute-force attacks.
5. Inadequate default settings, haphazard or insufficient configuration, incorrect HTTP headers, or improper HTTP methods frequently lead to security misconfiguration.
6. Injections flaws including SQL injection or NoSQL injection or command injection

B. Solutions to Prevent the Attacks

1. Use secure communication channels (HTTPS/TLS) to protect against man-in-the-middle attacks.
2. Implement proper token expiration and revocation mechanisms to reduce the impact of leaked tokens.
3. Store sensitive information securely, such as refresh tokens, client secrets, and authorization codes.
4. Use strong cryptography to secure access and refresh tokens, such as JSON Web Tokens (JWT).
5. Verify the authenticity of the authorization server, client, and resource server to prevent impersonation attacks.
6. Validate incoming requests to ensure that they are well-formed and contain valid information.

ISSN 2394 - 7780

7. Limit the scope of access granted to tokens, to reduce the impact of leaked tokens.
8. Monitor logs for suspicious activity and implement proper incident response processes to respond quickly to security incidents.
9. Conduct regular security audits to identify and address vulnerabilities.
10. Keep software and systems up to date with the latest security patches and upgrades.

C. Microservices Architecture vs Monolithic Architecture

Applications developed in monolithic architecture are developed in a single package, the application starts from a modular layer and is further divided into a presentation layer, business logic layer, data access layer, and Application integration layer.

- **Easy to Develop and Deploy:** Monolithic applications are straightforward to develop, test, and deploy. The entire application is built as a single unit, making it easier to manage dependencies and coordinate updates.
- **Simplified Testing:** Monolithic architecture facilitates testing, as all components are integrated into a single package. This makes it easier to automate testing and reduces the risk of compatibility issues.
- **Improved Performance:** Monolithic architecture provides improved performance as it allows all components to communicate with each other without the overhead of inter-process communication.

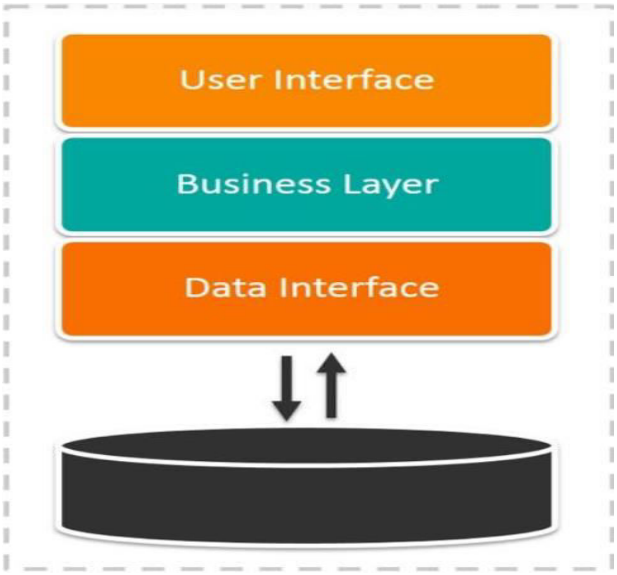


Fig 1.1 Here the above figure 1.1 shows the visual representation of Monolithic Architecture

Monolithic architecture is much more complex and have many cons as compared to pros

- Lack of modularity: Monolithic applications can become very large and complex, making it difficult to make changes or add new features without affecting other parts of the system.
- Scalability challenges: Monolithic architecture can be challenging to scale as the entire application must be deployed on each new server, which can result in increased resource usage and deployment time.
- Difficult to maintain: Monolithic applications can be difficult to maintain as the entire codebase must be managed as a single unit, making it difficult to update or fix individual components.

On the other hand, microservices architecture is comprised of a cluster of compact, autonomous, independent modules that provide variety of services; each service should be capable of being independently implemented by the corresponding business unit; it has a small independent unit that functions as a single application.

- **Scalability:** Each service can be scaled independently, allowing for more efficient resource usage and improved performance.
- **Resilience:** A failure in one service does not impact the entire system, reducing downtime and improving reliability.
- **Flexibility:** Services can be developed and deployed independently, allowing for more rapid and frequent releases.

ISSN 2394 - 7780

- **Improved Maintainability:** Smaller, self-contained services are easier to understand and maintain than a monolithic codebase.
- **Technology diversity:** Services can be developed using different technologies, providing greater flexibility and a better match to specific requirements.

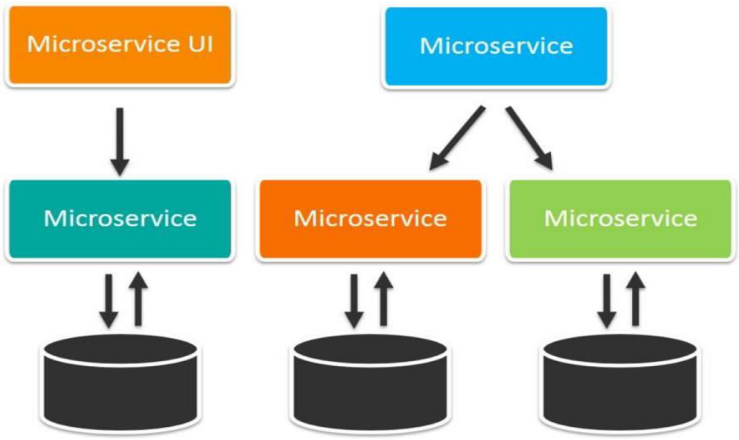


Fig 1.2 the above figure 1.2 represents the flow of MSA.

In Microservice Architecture the complexity is there but its dynamic and easily scalable but there are some cons too.

- Complexity: The distributed nature of microservices increases the complexity of the system, requiring careful management and coordination.
- Increased latency: The need to communicate between services can lead to increased latency and decreased performance.
- Testing and deployment challenges: Testing and deploying a system composed of many small services is more challenging than with a monolithic architecture.
- Cost: Developing and maintaining a microservices architecture can be more expensive than a monolithic architecture due to the increased number of moving parts.
- Inter-service communication: Effective communication between services is crucial for a microservices architecture to function properly and can be a significant challenge.

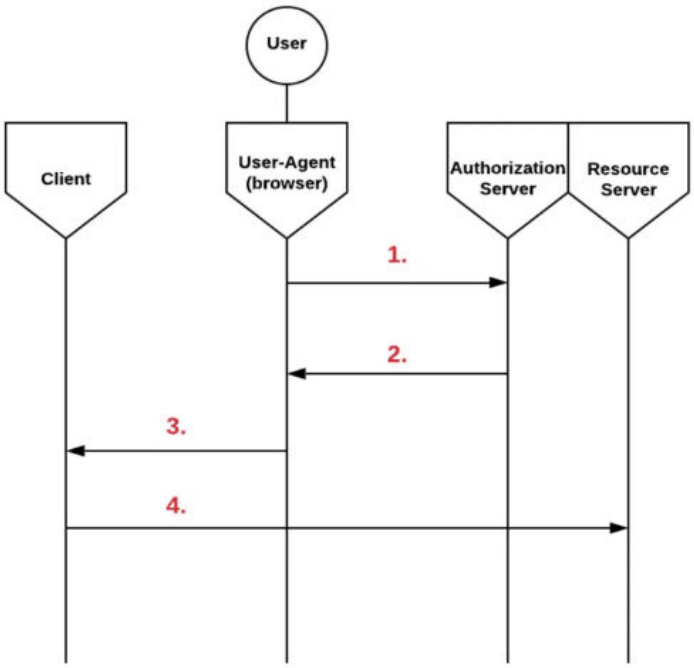


Fig 1.3 the above figure 1.3 represents the diagrammatic illustration of request and response mechanisms on the basis of access tokens

ISSN 2394 - 7780

The major advantage of MSA is every service can be independently scaled in several programming languages, but on the other hand in monolithic architecture, we have to completely develop it in a single programming language so we have to be more consistent while choosing the programming language.

MSA services are very understandable and easy to scale and deploy as it is written independently, altering one service won't affect the other service while in the case of monolithic architecture, it's pretty complicated to scale or deploy one service as it is one single complex code we have to alter or modify carefully

Microservices architecture code base size is comparably less than monolithic code base size, MSA code takes less time to start up's than monolithic code

D. Working of OAuth2.0

OAuth2.0 is open authorization protocol standard for APIs here it allows to client to access the resources of the resource owner.

For example, if you are using a web app XYZ and you need to log in and do so you have options to login with Facebook or Gmail account or Twitter etc here the XYZ app would request for you essential credentials like username and a tempering password which you don't have to worry about it. It can change for every login automatically generated by the resource server.

So here it works as an authorization framework. It secures your profile in the resource server here if you are granting client app (XYZ app) to access your resource server (Gmail account or Facebook or Twitter etc) only for login and nothing more.

Fig 1.4 the above figure 1.2 illustrates the use case of client and Resource communication using OAuth2.0

OAuth 2.0 is an open standard for authorization that uses various algorithms for secure communication. The most commonly used algorithm in OAuth 2.0 is the "Authorization Code Grant" flow.

Which involves the use of an authorization code and a redirect URI to grant access to an API. Other common algorithms include the "Implicit Grant" flow and the "Resource Owner Password Credentials Grant" flow.

Additionally, OAuth 2.0 supports the use of different types of digital signature algorithms, such as RSA and HMAC.

E. Security Measures of Spring Framework

Spring framework is majorly used for creating rest APIs in java language, it provides lots of annotations so the code length and complexity is been maintain, we use thousands of line code by using @, by the use of annotation we don't have to write big-lines of codes anymore.

And as per we are creating APIs in spring, we need some measures to save our data or secure our APIs from cybercriminals so here are some of the libraries' "dependencies" in spring to secure our APIs

<dependency>

<groupId>org.springframework.boot</groupId>

<artifactId>spring-boot-starter-oauth2-client</artifactId></dependency>

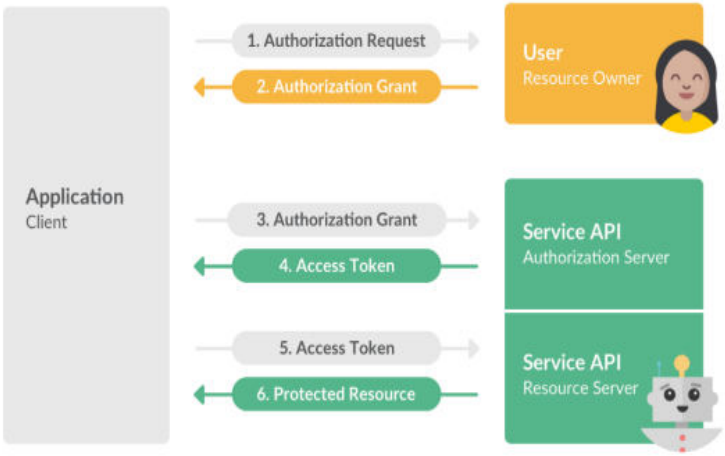


Fig 1.5 the above figure 1.5 represents the illustrations of Access Tokens.

ISSN 2394 - 7780

III. There are Several Drawbacks to the Oauth2.0 Framework

Complexity: OAuth2.0 can be complex to implement and understand, particularly for developers who are new to the protocol.

Security Vulnerabilities: OAuth2.0 is vulnerable to certain types of attacks, such as phishing and spoofing.

Token Storage and Management: Tokens used in OAuth2.0 need to be securely stored and managed, which can be a significant challenge for developers.

Lack of Standardization: There are many different implementations of OAuth2.0, which can lead to compatibility issues and confusion.

Lack of User Experience: OAuth2.0 is not designed to provide a seamless user experience, which can lead to confusion and frustration for users.

Limited to Web-Based Applications: OAuth2.0 is primarily designed for web-based applications and may not be suitable for other types of applications.

In [12] [R. Yang, W. Cheong Lau, T. Liu] paper, they have discovered a previously unexplored flaw that may be remotely exploited by an attacker to take control of the victim's account on a mobile app without the victim's knowledge or consent.

They have shown how these well-known applications are affected by this new vulnerability by looking at the implementations of the Top-200 US and Chinese Android Apps that employ the OAuth2.0-based authentication service offered by three top-tier IdPs.

Our research demonstrates that the various parties need to review their SSO implementations and perform the recommended corrective steps as a result.

A. OAuth2.0 Authorization Vulnerabilities

While surfing on the internet we sometimes cross path with some website which lets us log in without registering using our Gmail account or social media account, there is a chance of they are built into the OAuth2.0 framework.

OAuth2.0 is particularly intriguing to attackers since it is both widely used and intrinsically prone to implementation errors, which can lead to a number of vulnerabilities and allow attackers to steal valuable user information and potentially fully circumvent authentication.

One of the vulnerabilities of OAuth2.0 is characterised by a general lack of built-in security features; security is mostly dependent on the developer's use of the appropriate configuration parameters and the implementation of their own enhanced security mechanisms.

According to [7] [OpenID Foundation], [16] [N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, and C. Mortimore] Online users are increasingly using OpenID and OAuth2.0 together for authentication and authorization.

They are deployed in many web domains, particularly if the applications and user base are massive, like social networking, because of their ease of understanding, use, and robustness.

This also reduces the strain of having to type the password each time for permission and authentication, particularly with handheld devices.

After discussing a straightforward problem situation, it became apparent that the OpenID+OAuth2.0 combination had some issues with authentication.

The difficulty of setting up two protocols independently for authentication and authorisation is one of the two main issues mentioned here, along with issues resulting from the transmission of user credentials over the Internet.

By extending OAuth2.0, both issues are solved. The user can avoid sending their login credentials over the Internet by deploying Kerberos-like authentication.

It is important to note that OAuth2.0 uses some kind of tokens for authorizations, much like Kerberos.

It might be suggested that extending OAuth2.0 to handle authentication eliminates the requirement for OpenID and all associated issues.

ISSN 2394 - 7780

B. Various Protocol for the Security of Microservices

According to [1] [E. Jahjaga 31-10-2020] the best choice depends on the specific needs and constraints of the project.

According to my understanding, most of the Popular choices for authorization frameworks include OAuth2.0, which is widely used for token-based authorization, and JSON Web Tokens (JWT), which are self-contained and can be easily passed in HTTP headers.

Another popular choice is OpenID Connect, which is built on top of OAuth 2.0 and provides additional features such as user authentication.

Ultimately, the best choice will depend on the requirements of your specific microservices architecture and the trade-offs you are willing to make in terms of security, scalability, and ease of development.

C. Wide Popularity of Oauth2.0 Despite of its Drawbacks

OAuth2.0 is widely used despite its drawbacks because it provides a balance between security and ease of use. Some of the potential drawbacks of OAuth 2.0 include:

Complexity: OAuth2.0 can be complex to implement and may require a significant amount of developer resources.

Security: OAuth2.0 is only as secure as the implementation and there have been known vulnerabilities in the past.

Privacy: Some users may be concerned about the amount of personal information that is shared with third-party applications when using OAuth2.0.

More flaws are discussed in 4.1,

However, OAuth2.0 is still considered a good option because it provides a simple and secure way for users to grant access to their resources without having to share their login credentials.

Additionally, OAuth2.0 is flexible and can be used with a variety of different authentication methods, making it easy to integrate into different types of applications and systems.

Another reason why OAuth2.0 is widely used is that it is an open standard, meaning that it is not controlled by any single organization or company, which allows for greater interoperability and flexibility.

Overall, OAuth2.0 is a widely used protocol because it is flexible, simple, and secure, despite its drawbacks.

IV. CONCLUSION

Oauth2.0 is an authorization framework built to make data easily available but it is effective when it is combined with good security measures or else it can be breached and crucial data can be intercepted by the cybercriminal.

The best combination for the Oauth2.0 depends on the specific use case and the requirement of the application.

However, a common combination of Oauth2.0 is to use it with conjunction with OpenID (OIDC) protocol, OIDC is built on top of Oauth2.0 and provides additional features such as Authentication and user information.

[1] [E. Jahjaga 31-10-2020] has concluded that OAuth2.0 was reviewed by many in the software industry to be insufficiently secure by default. Because of this, it was viewed as a framework rather than a protocol in and of itself.

Because of this, the majority of these protocol implementations are made particularly to meet the requirements of the application. The biggest issue with OAuth2.0, according to industry experts, is that there is a fine line between security and usability. Despite these shortcomings, OAuth2.0 is currently one of the most popular implementations among application developers.

It has evolved into a standard for creating new protocols due to its architecture and simplicity of usage.

[12] Understanding the proposition of [R. Yang, W. Cheong Lau, and T. Liu], I have concluded that, another problem with OAuth2.0 is that it can be vulnerable to phishing attacks. Since OAuth2.0 relies on redirecting users to a third-party website to authorize access, attackers can create fake websites that impersonate legitimate sites and steal user credentials.

ISSN 2394 - 7780

A. There are several known security flaws in OAuth 2.0, including

1. Lack of transport layer security (TLS) in some implementations, can allow attackers to intercept and steal access tokens.
2. A lack of binding between the authorization request and the resulting access token can allow attackers to steal access tokens by intercepting and modifying the authorization request.
3. Inadequate validation of redirect URIs, which can allow attackers to steal access tokens by redirecting the user to a malicious site after the authorization grant is issued.
4. A lack of state in the authorization process can allow attackers to steal access tokens by intercepting and modifying the authorization request.
5. Lack of support for proof of possession of the client's secret,

Which can allow attackers to steal access tokens by intercepting and replaying the authorization grant.

OAuth2.0 is not a security mechanism rather it is a protocol, but though ultimately it is an authorization framework.

It provides a way for a user to grant a third-party application access to their resources on an online platform.

It's important to use best practices and libraries while implementing OAuth2.0 to avoid these vulnerabilities.

V. REFERENCES

[1] E. Jahjaga https://knowledgecenter.ubtuni.net/conference/2020/all_events/325/ , 31-10-2020.

[2] OAuth 2.0 authentication vulnerabilities. https://portswigger.net/web-security/oauth

[3] OAuth 2.0-based authentication solution for FPGA-enabled cloud computing https:// dl.acm.org/doi/abs/10.1145/3492323.3495635 2020

[4] J. Jose Diaz Rivera, W. Cheol Song. https:// ieeexplore. ieee.org/ abstract/ document/ 9919940

[5] J. Trammel, Ü. Yalçinalp, A. Kalfas, J. Boag & D. Brotsky. Device Token Protocol for Persistent Authentication Shared across Applications https:// link. springer. com/ chapter/ 10.1007/978-3-642-33427-6_20

[6] OAuth Working Group, Hammer, E. (ed): IETF, The OAuth2.0 Authorization Protocol draft 28 http://tools.ietf.org/html/draft-ietf-oauth-v2-28 2012.

[7] OpenID Foundation, Open ID Connect Protocol Suite, http://openid.net/connect/ 2012

[8] Google, Android Account Manager API, http:// developer. android.com/ reference/ android/ accounts/AccountManager.html 2012.

[9] O. Shurdi, A. Biberaj, I. Tafa and G. Mesi, "OAuth2.0 in Securing APIs" (2020). UBT International Conference. 325. https://knowledgecenter.ubt-uni.net/ conference/ 2020/ all_ events/325 31-10-2020

[10] Attacking and Defending OAuth2.0 https://www.praetorian.com/blog/attacking-and-defending-oauth-2-0-part-1/

[11] Q. Li, J. Kong. https://doi.org/10.1117/12.2656672

[12] R. Yang, W. Cheong Lau, T. Liu, https://oauth.ie.cuhk.edu.hk/static/papers/eurobh16.pdf

[13] "Social login continues strong adoption," [Online]. Available: http:// janrain.com/ blog/ social-login-continues-strong-adoption/ 2014.

[14] E. Y. Chen, Y. Pei, S. Chen, Y. Tian, R. Kotcher, and P. Tague, "OAuth demystified for mobile application developers," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014.

[15] C. Eric, P. Tague, R. Kotcher, S. Chen, Y. Tian, and Y. Pei, "1000 ways to die in mobile OAuth," in Blackhat USA, 2016.

[16] N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, and C. Mortimore, "OpenID Connect core 1.0," The OpenID Foundation, p. S3, 2014.

[17] H. Wang, Y. Zhang, J. Li, H. Liu, W. Yang, B. Li, and D. Gu, "Vulnerability assessment of OAuth implementations in Android applications," in Proceedings of the 31st Annual Computer Security Applications Conference. ACM, 2015.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

REVIEW PAPER ON HEART DISEASE PREDICTION MODEL

Manisha Divate, Smruti Nanavaty, Krupa Panchal and Meet Tank
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce Maharashtra, Mumbai-400056

ABSTRACT
This paper briefly reviews various advancement and recent works that are done using machine learning in prediction of heart disease. Between 1990 to 1924 runner heart disease alone is anticipated to increase by 120% for women and 137% for men in developing countries. The main factors contributing to rise in heart disease are obesity and two type two diabetes some other factors here include age, sex, pulse rate, weight, diabetes, blood pressure, cholesterol rate and many more. Majority of heart related diseases can be reversed if people start focusing on their health which includes physical activities having a balanced and a nutritious diet. From treating cancer, delivery babies, recovering lung damage from Covid-19 virus to dealing with heart attacks Doctors have exceled in developing technology and improve techniques. Advanced techniques and new medical development have changed the face of health care over the centuries medical field has been upgraded to bring together the best and brightest and the society to help those in need. Some of the machine learning techniques and other techniques which are used to predict the heart disease used in this paper are Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbour(KNN), Naïve Bayes Support Vector Machine (SVM)and many more. This review paper provides an inside of existing algorithm and it gives an overall summary of existing work.

Keywords: Machine Learning, Heart Disease Prediction.

LITERATURE REVIEW
Manoj Diwakar el al [28], 2020 provide an overview of working algorithm and provides a description of the current work for diagnosis of heart disease and treating it at initial stage. According to related study algorithm in literature review k nearest neighbour algorithm gives 97.3 percent accuracy in diagnosing patient with heart disease 2005 by Shouman et al. Naive Bayes (NB), Artificial Neural network (ANN), support vector machine are the algorithms mentioned by the author from which the best and accurate results were only

INTRODUCTION
In this modern living lifestyle heart disease has become most common among people due to multiple changes in day-to-day life. Heart disease can be fatal, can risk one's life and can create many other severe situations leading to other diseases. According to 2022 statistics, 70% of people died due to heart attack in India last year between the age group of 30 to 60. Deaths taken place due to heart attack has remain consistently over 25000 in last 4 years and around 28000 and above in the last 3 years. According to WHO survey 17 million people die all around a globe due to cardiovascular disease that is 29.20% among all caused death mostly in developing countries. Machine learning algorithms have been quite efficient in producing accurate results with high level of correctness there by preventing the onset of heart diseases in many patients and reducing the impacts in the ones which are already affected by heart diseases. Massive amount of data is collected all over the world from various medical organisations this data can be exploited using machine learning techniques to gain useful insights in heart disease prediction .Data sometimes can be noisy which are two overwhelming for human minds to complaint and can be easily explode using machine learning techniques this algorithm is useful in prediction of presence and absence of heart related diseases accurately. This data proves to be helpful for medical researchers and also doctors all over the world to recognise the patients which ultimately leads to early detection of heart disease. There is a need of getting rid of this complicated task CVD using advanced techniques in order to discover knowledge of heart disease prediction.

Given by ANN algorithm compared to another algorithm.

Pavithram M el al [29], 2021 gives an overview of predicting heart disease effectively using data mining and artificial intelligence techniques which helps in curing diseases earlier and reduces death rate. According to survey annual number of death suffering from heart diseases in India is forecasted to increase from 2.26 million in 1990 to 4.77 million in 2020. A decision tree classifier technique which is C 4.5 algorithm gives 89% accuracy for predicting heart disease which helps a drastic decrease in mortality rate.

Senthil Kumar Mohan el al [30], 2019 propose a novel method that AIMS at finding significant features by applying machine learning technique resulting in improving accuracy in prediction of cardiovascular disease. Hybrid HRFLM proved to be quite accurate in prediction of heart disease. The various data set with DT, RF,

ISSN 2394 - 7780

LM are applied to find out best classification methods. RF and LM are the best RF error rate for data set 4 is high 20.9 %. LM method for data set 9.1 % method to DT and RF method. HRFLM which also produces high accuracy and less classification error in predicting the heart disease.

Dr S Anitha [32] , 2019 gave an overview of how supervised Machine learning algorithm is used to predict heart disease in this research paper. This paper concentrates on classification model like SVM, Naive Bayes and KNN to predict heart disease and compare their performance. R tool is used for programming purpose for implementing the classification technique. According to this research and clinical data mentioned, Naive Bayes predict the disease with highest accuracy of 86.6% when compared to KNN and SVM.

Monther Tarawneh el al [33], 2019 explains hybrid approach for heart disease prediction using machine learning in this article. According to a specific data set Naive Bayes and SVM technique gives a better performance while others depend on the selected feature.

N Satish Chandra Reddy el al [34], 2019 gives an overview of classification of feature selection approach by machine learning technique for predicting heart disease. The algorithms explained in this research paper are k nearest neighbour algorithm, support vector machine (SVM), random forest (RF), Naive Bayes (NB), Neural network which are data mining and machine learning algorithm. Various feature selection method approach such as wrapper method, filter embedded, ensemble and hybrid method have been applied to study prediction of heart disease. Among 5 classification algorithm used highest accuracy has been observed in random forest with 91.39% to 94.96% and average accuracies respectively.

V V Ramalingam el al[36] , in the year 2018 survey they present various model based on such algorithm and technique analyses of the performance of predicting heart disease. Dimensionality reduction is an important step considered while building any model is generally achieved by two methods Feature Extraction and Feature selection. Alternating decision tree when used with PCA have performed extremely well but the same gave poor results in other cases which cause due to overfitting. Naive Bayes Classifier work computationally very fast and has also performed well. SVM to perform extremely well for most cases.

SP Rajamhoana el al [39], 2018 analysis various research work done on heart disease prediction classification using various machine learning and deep learning techniques to conclude which technique is effective and accurate. Publicly available Cleveland and Statlog heart disease database from UCI repository are used which consists of 303 records and 270 records respectively. The proposed prediction model obtained

83.9 % accuracy. Most of research work uses classification methods such as Association rule mining, Naive Bayes, Decision tree, Artificial Neural network provides better performance of heart disease prediction from used amount of medical data. About technique for eliminating the existing drawbacks and improving the prediction rate thus providing a way for improving the survival rate for well-being of mankind.

Amita Malave el al [41], 2017 overview an service efficient prediction technique to determine and extract the unknown knowledge of heart disease using hybrid combination of k-means clustering algorithms and artificial neural network in this research paper. Algorithms used are k means algorithm and artificial neural network during this research. Hybrid approach here gives higher accuracy rate of 97% of disease detection than earlier proposed method. To perform grouping of various attributes k-means algorithm was used and for predicting back propagation technique neural network was used in this study.

Review on Heart Disease Prediction System

Sr No.	Year	Title	Author	Best Algorithm	Accurac Y
1	**2013**	Disease Prediction in Data Mining Technique -A survey	Vijayarani Mohan el al. [1]	Naïve Bayes	**74.00%**
2	**2014**	Prediction Of Heart Disease Using Classification Algorithm	Hlaudi Daniel Masethe [2]	J48	**99%**
				REPTREE	**99%**
				Simple CART	**99%**
3	**2015**	Heart Disease Prediction Using Data Mining Techniques	Andrea D'Souza[3]	ANN	**79.38%**
4	**2015**	Hear Disease Prediction	Jaymin Patel[4]	J48	**56.76%**

ISSN 2394 - 7780

		using ML and Data Mining Techniques			
5	**2015**	Expert System With Application	Cong Long[5]	Naive Bayes	**85.20%**
				SVM	**81.50%**
				ANN	**81.50%**
				New approach	**87.30%**
6	**2015**	Hear Disease Prediction using Hybrid Genetic Fuzzy Model	T. Santhanam[6]	Stratified k fold technique	**86.00%**
7	**2016**	Efficient Heart Diesase Prediction System	Purushottam[7]	using 10 fold method	**86.70%**
8	**2016**	Data Mining Apriori Algorithm for HDP	Alireza Alinezhad[8]	neural network	**97%**
				clustering algorithm	**85.81%**
9	**2017**	Prediction of HD Using K-Mean and ANN as Hybrid approch to improve accuracy	Kalyani kadam[9]	Naïve Bayes	**88%**
				KNN	**93.00%**
				Hybrid	**97%**
10	**2018**	A Review on Heart disease prediction using Machine Learning and Data Analysis Approach	Marimuthu Muthuvel el al [10]	Support Vector Machine	**85%**
11	**2018**	Heart Disease Prediction using ML Techniques- A survey	V V Ramalingam[11]	Naïve Bayes	**88.49%**
				SVM	**92.10%**
				KNN	**87.50%**
				Decision Tree	**78.46%**
				Random Forest	**97.70%**
				Ensemble Model	**93%**
12	**2019**	Classification and feature slelection approcches by ML-HDP	N. Satish Chandra reddy[12]	KNN	**85.50%**
				SVM	**82.77%**
				Random Forest	**90.76%**
				Naïve Bayes	**86.97%**
				neural network	**85.50%**
				Avg Accuracy	**86.30%**
13	**2019**	Neural Network based intelligent system for HDP	Vikas Boddu[13]	Decision Tree	**83.60%**
				Logistic Regression	**85.20%**
				Naïve Bayes	**90.20%**
				Random Forest	**85.20%**
				SVM	**76.57%**

ISSN 2394 - 7780

				Genralized linear model	**85.20%**
				Gradient boosted tree	**85.20%**
				Deep Lerarning	**88.50%**
				MLPNN-proposed algo.	**94.00%**
14	**2019**	Effective HDP using Hybrid ML Techniques	Senthilkumar Mohan[14]	Naïve Bayes	**75.80%**
				Genralized linear model	**85.10%**
				Logistic Regression	**82.90%**
				Deep Lerarning	**87.40%**
				Decision Tree	**85.00%**
				Random Forest	**86.10%**
				Gradient boosted tree	**78.30%**
				SVM	**86.10%**
				VOTE	**87.41%**
				HRFLM(proposed)	**88.40%**
15	**2019**	Data Mining Apriori Algorithm for Heart Diseasae Prediction	Alireza Alinezhad & Mirupouya Mirmazaffari[15]	Apriori	**97.40%**
16	**2019**	Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique	R Venkatesh & C Balasubramanian & M Kaliappan[16]	Naïve Bayes	**97.12%**
17	**2019**	Heart Disease Prediction Using Data Mining Techiniques	DR. S. Anitha[17]	KNN	**76.67%**
				Naïve bayes	**86.60%**
				SVM	**77.70%**
18	**2019**	Intelligent Diagnosis of Cardiac Disease Prediction using Machine Learning	Ravindhar NV[18]	Logistic Regression	**81.86%**
				Naive Bayes	**61.46%**
				Fuzzy KNN	**87.33%**
				K-Means Clustering	**43.24%**
				BP-Neural Network	**98.20%**
19	**2020**	Prediction of Coronary	Rami Mustafa A	Naïve	**73%**

ISSN 2394 - 7780

		Heart Disease using Machine Learning: An Experimental Analysis	Mohammad[19]	Bayes	
				Support Vector Machine	**71%**
				Decision Tree	**67.50%**
20	**2020**	Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifier	Komal Kumar Napa[20]	Random Forest	**85.71%**
21	**2020**	Heart disease prediction using machine learning	Rishabh Magar [21]	Logistic Regression	**82.89%**
				SVM	**81.57%**
				Naive Bayes	**80.43%**
				Decision Tree	**80.43%**
22	**2020**	Heart disease prediction using machine learning	Apurb Rajdhan[22]	Logistic Regression	**85.25%**
				Decision Tree	**81.97%**
				Random Forest	**90.16%**
				Naive Bayes	**85.25%**
23	**2020**	Heart disease prediction using machine learning techniques	Devansh Shah[23]	Naive Bayes	**88.157%**
				KNN	**90.789%**
				Random Forest	**86.84%**
				Decision Tree	**80.263%**
24	**2020**	Heart Disease prediction using Machine Learning	N. Saranya[24]	Random Forest	**100%**
				KNN	**91.36%**
				Logistic Regression	**87.65%**
				Ensemble model without Logistic Regression	**98.77%**
25	**2021**	Heart Disease prediction using machine learning algorithms	Harshit Jindal[25]	KNN	**88.52%**
				Logistic Regression	**88.5%**
				KNN & LR based model	**87.5%**
26	**2021**	Prediction of Heart Disease using Machine Learning Algorithm	Aadar Pandita[26]	Logistic Regression	**84.38%**
				KNN	**89.06%**
				SVM	**87.50%**
				Naive Bayes	**85.94%**
				Random Forest	**87.50%**

ISSN 2394 - 7780

27	**2021**	Machine learning algorithms for predicting coronary artery disease : efforts toward an open-source solution	Aravind Akella[27]	Generalized linear model	**87.64%**
				Decision Tree	**79.78%**
				Random Forest	**87.64%**
				SVM	**86.52%**
				Neural Network	**93.03%**
				KNN	**84.27%**

CONCLUSION

Majority of researchers mentioned in this review have used the Cleveland Heart Disease Dataset available from the UCI repository which indicates 76 attributes and 303 instances, where 14 attributes are used due to missing values in the data set. Literature survey mentioned here gains marginal for heart disease patients which is achieved by the predictive model. Therefore, there is a need to increase models for more combinations and complexity to find accurate predicting model for early onset of heart disease prediction. From the comparative study Random Forest gave highest accuracy . Many other algorithms such as Naïve Bayes, KNN, J48, Logistic Regression, Clustering algorithms and neural networking also gained success. There is a need for improvement to increase scalability and accuracy for prediction system, hence there is need for more research work to be done in future.

REFERENCES

1. https://www.researchgate.net/publication/329799991_ Disease_Prediction_in_Data_Mining_Technique_-_A_Survey
2. https://scholar.google.co.in/scholar?q=Prediction+Of+ Heart+Disease+Using+Classification+Algorithm&hl=en &as_sdt=0&as_vis=1&oi=scholart
3. https://scholar.google.co.in/scholar?q=heart+disease+p rediction+using+data+mining+techniques+Andrea+D%27 Souza&hl=en&as_sdt=0&as_vis=1&oi=scholart
4. https://www.researchgate.net/publication/309210947_ Heart_Disease_prediction_using_Machine_learning_and_ Data_Mining_Technique
5. https://www.sciencedirect.com/science/article/abs/pii/S 0957417415004261
6. https://www.researchgate.net/publication/277904519_ Heart_Disease_Prediction_Using_Hybrid_Genetic_Fuzzy _Model
7. https://www.sciencedirect.com/science/article/pii/S187 705091630638X
8. https://www.semanticscholar.org/paper/Data-Mining- Apriori-Algorithm-for-Heart-Disease-Mirmozaffari-Alinezhad/1f6b4f71530b9ea29a43f4e16b2c48d9f90f2719
9. https://www.researchgate.net/publication/319486202_P REDICTION_OF_HEART_DISEASE_USING_K-MEANS_and_ARTIFICIAL_NEURAL_NETWORK_as_ HYBRID_APPROACH_to_IMPROVE_ACCURACY
10. https://www.researchgate.net/publication/327722009_ A_Review_on_Heart_Disease_Prediction_using_Machine _Learning_and_Data_Analytics_Approach
11. https://www.researchgate.net/publication/325116774_ Heart_disease_prediction_using_machine_learning_techni ques_A_survey
12. https://ijic.utm.my/index.php/ijic/article/view/210
13. https://www.semanticscholar.org/paper/Effective- Heart-Disease-Prediction-Using-Hybrid-Mohan-Thirumalai/2bc3644ce4de7fce5812c1455e056649a47c1b bf
14. https://www.researchgate.net/publication/317952833_ Data_Mining_Apriori_Algorithm_for_Heart_Disease_Pre diction

15. https://www.researchgate.net/publication/334261584_ Development_of_Big_Data_Predictive_Analytics_Model _for_Disease_Prediction_using_Machine_learning_Techn ique
16. https://hal.science/hal-02196156/
17. Chromeextension://efaidnbmnnnibpcajpcglclefindmka j/https://www.ijitee.org/wpcontent/uploads/papers/v8i11/J 97650881019.pdf
18. https://dl.acm.org/doi/abs/10.1145/3342999.3343015
19. https://ieeexplore.ieee.org/document/9074183
20. https://www.jetir.org/view?paper=JETIR2006301
21. https://www.researchgate.net/publication/341870785_ Heart_Disease_Prediction_using_Machine_Learning
22. https://www.researchgate.net/publication/346484346_ Heart_Disease_Prediction_using_Machine_Learning_Tec hniques
23. https://www.ijrte.org/portfolio-item/F9780038620/
24. https://iopscience.iop.org/article/10.1088/1757- 899X/1022/1/012072
25. https://www.researchgate.net/publication/352453917_ Prediction_of_Heart_Disease_using_Machine_Learning_ Algorithms
26. https://www.future-science.com/doi/10.2144/fsoa- 2020-0206
27. https://www.researchgate.net/publication/344821374_ Latest_trends_on_heart_disease_prediction_using_machin e_learning_and_image_fusion
28. https://scholar.google.co.in/scholar?q=Effective+Hear t+Disease+Prediction+Systems+Using+Data+Mining+Te chniques&hl=en&as_sdt=0&as_vis=1&oi=scholart
29. https://www.researchgate.net/publication/333888974_ Effective_Heart_Disease_Prediction_Using_Hybrid_Mac hine_Learning_Techniques
30. https://www.researchgate.net/publication/365152872_ An_Efficient_Heart_Disease_Prediction_System_based_o n_Supervised_Machine_Learning_Methods
31. https://www.researchgate.net/publication/330880739_ Hybrid_Approach_for_Heart_Disease_Prediction_Using_ Data_Mining_Techniques
32. https://ijic.utm.my/index.php/ijic/article/view/210
33. https://www.researchgate.net/publication/325116774_ Heart_disease_prediction_using_machine_learning_techni ques_A_survey
34. https://ieeexplore.ieee.org/abstract/document/8431153
35. https://www.researchgate.net/publication/323277733_ A_hybrid_approach_for_Heart_Disease_Prediction_using _Artificial_Neural_Network_and_K-means
36. https://www.researchgate.net/publication/314666453_ Analysis_of_data_mining_techniques_for_heart_disease_ prediction

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

SECURITY IN MACHINE LEARNING

Manisha Divate and Keshav Sharma
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle - (W), Maharashtra, Mumbai- 400 056

ABSTRACT
In this paper, we present a broad dataset of vulnerabilities and exploits, approaches to fighting these pitfalls, and the numerous open questions in this area. The dataset consists of a wide range of security-related datasets, which are analyzed with the help of machine literacy and his intelligent anthropological exploration. This paper presents about Machine Learning, Importance of Machine Learning, detail a security in Machine Learning. Cyber-attack talks about training set poisoning with Backdoor Training Set. attack on security system, methods of securing the data. Emerging Threats in cyber security, Social media, Cloud Computing that is of tone-service on request and Ubiquitous network access and as such Other emerging areas of concerns. The main Part of the Paper is the part of Methods for security with the help of Clustering, Decision Trees and Bayesian Network. We will discuss about Security Data Sets of KDD Cup 1999 Dataset, HTTP CSIC 2010 Dataset and UNSW-NB15 Dataset.

Keywords: Security, Clustering, Vulnerabilities, Training and Training Set.

I. INTRODUCTION

The artificial intelligence (AI) and machine literacy fields of computer wisdom focus on Using data and algorithms to mimic the way humans learn and incrementally ameliorate the delicacy of prognostications. It enables software programmes to improve their propensity to anticipate outcomes without having been, it's especially designed to do so. To prognosticate new affair values, machine literacy algorithms use literal data as input.

Data Security in necessary at the same time this technology is being used. This reduces the threat of attacks on IT systems and data breaches. Use security measures to help unauthorized access to sensitive data. Avoid interruptions similar as denial of service attacks. Cover computer networks and systems from exploitation by outlander.

The most crucial component of machine learning is training. Consider, Carefully the features and hyper parameters to use, People make decision, not machines. The most crucial Step in machine Learning is data cleaning.

Machine Literacy in security is always learning to dissect data and find patterns. This allows us to more describe malware in translated business, identify bigwig pitfalls, prognosticate where the" bad areas" are online and reduce the number of people suds the internet. Improve security and protect data in the cloud and Uncover suspicious user behaviour.

II. LITERATURE REVIEW

A) Cybersecurity Attack

Malware was created simply to reveal security vulnerabilities or, in some cases, to demonstrate specialized prowess. The primary purpose of malware today is to steal confidential information from others for the benefit of individuals, finance or company [1, 2].

The cyber trouble geography requires nonstop shadowing and relating millions of external and internal data points across an association's structure and druggies. It is just impossible to manage this data volume with a small group of individuals [3]. It excels at recognizing patterns in large amounts of data and predicting hazards at machine speed. By automating analysis, Cyber brigades can snappily uncover pitfalls and insulate cases that bear further mortal disquisition.

1) Training Set Poisoning.

Malicious manipulation of training sets intended to mislead machine learning model predictions is known as the poisoning attack [4]. Exploration shows that a small, precisely drafted dependent training data set can dramatically degrade the performance of machine literacy models. A summary of attacks is displayed in Figure 1. This article breaks down the poisoning work according to whether it is intended for neural network (NN) models[5].

ISSN 2394 - 7780

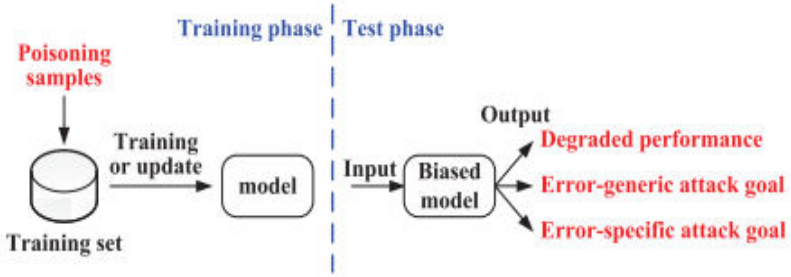


Figure 1: Overview of Poisoning attacks.

a) Targeting SCMs:
Biggio etal.[6] propose poisoning attacks against support vector machines (SVMs). This attack injects set training data to increase the test error of the SVM classifier. This system can produce an optimized expression of the poisoning data and convert it to a kernel[6], but requires thorough knowledge of the algorithm and training data.

b) Target Clustering Algorithm:
The clustering algorithm is extensively used in data analysis and security operations similar as B. Market Segmentation, Website Bracket, Malware Detection[7]. still, the clustering process itself can be demoralized by a sophisticated bushwhacker.

2) Backdoor in the Training Set.
As machine literacy systems grow in size, so do the training data conditions, and interpreters need to automate and outsource training data conservation to achieve state- of- the- art performance[8]. Lack of dependable mortal oversight of data collection processes exposes associations to security vulnerabilities. You can manipulate the training data to control and reduce the downstream of the learned model. The purpose of this work is to totally classify and bandy a broad dataset of vulnerabilities and exploits, approaches to fighting these pitfalls, and the numerous open questions in this area. It not only describes colourful poisoning and backdoor trouble models and the connections between them, but also develops a unified taxonomy for them.

Lately, experimenters have shown that bushwhackers can produce retired backdoors in training data or retrained models. Figure 2 shows an overview of backdoor attacks. The backdoor doesn't affect the normal functioning of the model, but when certain detector conditions do, the backdoor case is inaptly classified by the model as targeted by the attack. similar backdoor attacks are stealthy due to the cryptic nature of deep literacy models.

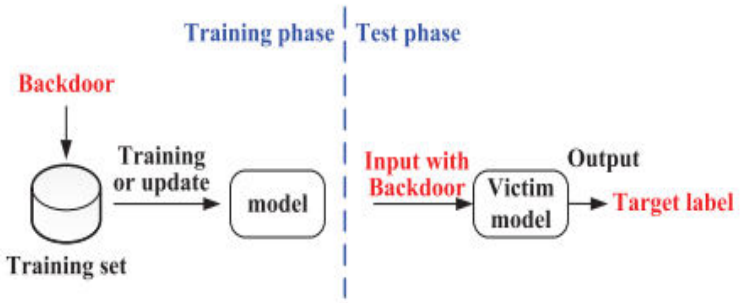


Figure 2: Overview of backdoor attacks.

3) Backdoor Attack
Ward etal [9]. proposes a virulently trained network called BadNet. BadNet can make the model bear worse when certain inputs arrive. The demonstrates such an attack on an early skin cancer discovery system without taking the bushwhacker to have any knowledge of the system or the training process. still, bushwhackers can directly manipulate model parameters to fit backdoors. This supposition is delicate to satisfy in practice[10].

Chen et al. propose a backdoor attack on deep learning models by using data poisoning. Specifically, poisoning samples are injected into the training dataset so as to implant a backdoor. Their attack can work under a weak attack model, which means it doesn’t require knowledge about the model and the training set [11].

B) Emerging Threats in Cyber Security
Substantially, cyber culprits that are modified are malware. The autographs that exploit the vulnerability are new technology. Technologies just explore finding unique selling points for new technologies and malware injection loopholes. We select two of his new technological advances as follows:

ISSN 2394 - 7780

1) Social Media
Social media, like Facebook and Twitter, have endured explosive growth lately. Social networks are veritably popular and are becoming the preferred means of communication for youngest people. The Koobface worm report for IT security and sequestration company Sophos revealed an astounding increase in attacks against druggies on social media websites[12]. In particular, its spread via social media spots in 2009, is best known as an illustration of malware taking advantage of the proliferation of social media spots. The Koobface botnet leverages its zombie magazine to automate the creation of new social media accounts to befriend unknowing druggies and spam seductive links that reflect malware. Victims of social engineering attacks have their social media accounts turned into spammers for the victim's musketeers, turning their computers into zombies.

2) Cloud Computing
Cloud Computing also known as Pall Computing has become an essential tool for businesses, a useful way to store and partake data. pall computing trends are common in areas similar as operation software and structure, business processes, and systems structure. pall computing is moving from a disruptor to a must- have IT strategy. This explosive growth and abandonment of computing are the result of companies prioritizing strictness, dexterity, and new sources of competitive advantage. It offers unique features that differ from traditional styles. Features of pall computing include:

A) Tone- Service on Request.
On- demand tone- service refers to a service handled by pall computing providers that enables the provision of pall coffers on- demand whenever they're demanded. In on- demand tone- service, druggies access this service through an online press. Druggies can assign themselves fresh coffers similar as storehouse or processing automatic feeding without mortal intervention [13].

On- demand tone- service provisioning is a core point of utmost, pall services where druggies can gauge the needed structure significantly without dismembering machine operations. proprietor. These coffers are those that are handled without mortal commerce.

The advantage of on- demand tone- services is that druggies can incontinently emplace or remove coffers to meet their requirements without the need for mortal interposers[14]. Compared to a service that may bear director support, on- demand tone- service services don't have staying times or the threat of miscommunication.

B) Ubiquitous Network Access.
Universal Access represents the broad availability of a pall service. By establishing universal access to Pall services, different impulses, security technologies, transport protocols, and interfaces can be supported. Enabling this position of access frequently requires a pall service armature that's acclimatized to the specific requirements of different consumers of pall services [15]. Therefore, ubiquitous computing relies on the aggregation and operation of contextual data, transparent and intuitive access points, and flexible payment systems.

Ubiquitous computing, also known as Pervasive computing, is the growing trend of embedding computing power (generally in the form of microprocessors) into everyday objects to help them communicate efficiently and perform useful tasks in a way that minimizes the need for end-stoner commerce with a computer similar to a computer[16]. Common computing bias are connected to the network and are always available.

Popular computer operations are designed to be used by consumers and help people do their jobs. A terrain in which bias, ubiquitous, is able to perform some form of calculation can be considered a pervasive computing terrain. Because common computer systems can collect, process, and transmit data, they're adaptable to the environment and operation of the data[16]. They can also collect, process and communicate data, they can acclimatize to the environment and operation of the data.

3) Other Emerging Areas of Concern.
Machine Literacy is the rearmost technology to make swells in information security, and for good reason. Support for complex algorithms that" learn" and evolve is inestimable to mortal judges, allowing them to concentrate on further critical politic combat and bolstering security systems nearly. Bulletproof. In the midst of frequent and structural changes in information security, machine literacy plays a decreasingly important part and will continue to do so in the times to come.

The machine literacy assiduity is changing veritably snappily late technologies and scientific exploration determine how new products and services are erected. As 2022 draws to a close, everyone from machine literacy masterminds to incipience authors is looking for the most promising trends for the coming time. Pall structure is decreasingly powered by machine literacy to give enhanced security structure. Machine literacy can

ISSN 2394 - 7780

learn from data without unequivocal programming[17]. It allows information to be prognosticated and reused more directly. As machine literacy becomes advanced, it can make better data-driven recommendations without mortal intervention.

Pall security is a set of programs, controls, processes, and technologies that work together to cover pall-grounded operations, data, and structure. These security measures are configured to cover this data, support nonsupervisory compliance, cover client sequestration, and establish authentication rules for each stoner and device. Likewise, these rules can be configured and managed in one place[18]. Machine literacy will continue to be a promising and-fleet evolving field with numerous instigation inventions. Big language models, multimodal machine literacy, mills, TinyML, and low-law and zero-law results are technologies that will gain immense significance in the near future. Gartner predicts that in these times, machine literacy will indeed access more areas of business, adding effectiveness and safety at work.

C) Methods for Security in Machine Learning.

1) Clustering

Clustering is a set of styles that alter the patterns of advanced, unlabelled data. It's unsupervised pattern recognition fashion that groups data grounded on similarity[20]. Normal and abnormal network activity is generated by a viscosity-based clustering scheme called Simple Log File Clustering Engine (SLCT). Two clustering schemes are used. The first is detecting normal and attack scripts. The alternate is to use other schemes to determine normal business in a supervised manner[19]. One of the main benefits of clustering for attack discovery are that system directors can learn from content data without having to explicitly consider different types of attacks.

2) Decision Trees

A decision tree is a tree shape where companies are represented by leaves and branches are the representation for the purposeful main connection to the bone companies. A decision tree has its two bumps, a decision knot and a splint knot. Decision bumps are used to make opinions and have multiple branches, while splint bumps are the result of those opinions and contain no further branches[20]. A decision or test is made grounded on the characteristics of a particular data set.

3) Bayesian Network

Bayesian networks are an extensively used class of probabilistic graphical models. They correspond of two corridors known as structures and parameters. Structures and parameters. It's a compact, flexible, and easy- to-understand representation of the general probability distribution. Directed acyclic graphs can also be used for knowledge discovery by creating hamstrung connections between variables. Bayesian networks are generally learned from data.

III. METHODOLOGY

A) Security Data Sets

The data is presently being prepared by colourful exploration groups for their own analysis and for making it available in the community depository. In This section, we bandy being security- related datasets with the help of machine literacy and his intelligent anthropological exploration.

1) KDD Cup 1999 Dataset

The KDD Cup is a regular data mining and knowledge discovery competition hosted by the ACM Special Interest Group on Knowledge Discovery and Data Mining, the leading assiduity association for data miners. Annual libraries of records, attendants, and winners are available at all times. The KDD Cup 1999 is a database containing standard, auditable records containing colourful simulated intrusions in military network surroundings[21].

2) HTTP CSIC 2010 Dataset

CSIC created the original record of HTTP/1.1 packets, including web operation penetration test packets[21]. It has two markers (normal and abnormal). The original purpose of these reformatted CSIC datasets was pre-classification point selection and case selection analysis.

The CSIC 2010 HTTP protocol contains thousands of automatically generated web requests. It can be used to test protection systems against web attacks and was developed by the CSIC Information Security Institute(Spanish Research National Council).

ISSN 2394 - 7780

3) UNSW-NB15 Dataset

The UNSW- NB15 computer network security dataset was published in 2015 (Moustafa & Slay, 2015). This data set consists of realistic, current normal and abnormal (also known as attacks) network applications. These records were collected by the IXIA business creator using three virtual waiters[21]. Two waiters are configured to distribute normal network business, and a third garcon is configured to prevent abnormal network business which record of a network attack. It includes 9 different attacks, including DoS, worms, backdoors and fuzzers[23]. The dataset contains raw network packets. The training set has 341 records and the test set has 332 records with different types of attack and normal.

IV. CONCLUSION

In this paper, we present a broad dataset of vulnerabilities and exploits, approaches to fighting these pitfalls, and the numerous open questions in this area. The main Part of the Paper is the part of Methods for security with the help of Clustering, Decision Trees and Bayesian Network. The field of Artificial Intelligence (AI) Computing and Machine Learning concentrates on using data and algorithms to mimic how humans learn and gradationally ameliorate the delicacy of prognostications. To anticipate new values, systems studying algorithms use ancient statistics as input. Research shows that a small fraction of carefully constructed poisoning training data can dramatically degrade the performance of machine learning models. You can manipulate the training data to control and reduce the downstream of the learned model. Analogous backdoor attacks take a place surreptitiously due to the cryptic nature of the perceptively patterns. In this section, we attack security-applicable datasets using machine knowledge and its intelligent anthropological disquisition. It can be used to test protection systems against web attacks and was developed by the CSIC Information Security Institute.

REFERENCE

[1]. E.E. Schultz Where have the worms and viruses gone? New trends in malware Comput. Fraud Secur., 2006 (7) (2006), pp. 4-8.

[2]. U. Bayer, I. Habibi, D. Balzarotti, E. Kirda, C. Kruegel, A view on current malware behaviours, in: USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), April 2009.

[3]. J. Cano, "Cyberattacks-The Instability of Security and Control Knowledge", ISACA Journal, vol. 5, pp. 1-5, 2016.

[4] B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S. Lau, S. Rao, N. Taft, and J. D. Tygar, ‘‘ANTIDOTE: Understanding and defending against poisoning of anomaly detectors,’’ in Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf. (IMC), Nov. 2009, pp. 1–14.

[5] P. Li, Q. Liu, W. Zhao, D. Wang, and S. Wang, ‘‘Chronic poisoning against machine learning based IDSs using edge pattern detection,’’ in Proc. IEEE Int. Conf. Commun. (ICC), May 2018, pp. 1–7.

[6] B. Biggio, B. Nelson, and P. Laskov, ‘‘Poisoning attacks against support vector machines,’’ in Proc. 29th Int. Conf. Mach. Learn., Jun. 2012, pp. 1467–1474.

[7] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli, ‘‘Is data clustering in adversarial settings secure?’’ in Proc. ACM Workshop Artif. Intell. Secur. AISec, Nov. 2013, pp. 87–98.

[8] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, ‘‘How to backdoor federated learning,’’ 2018, arXiv:1807.00459. [Online].Available: http://arxiv.org/abs/1807.00459

[9] T. Gu, B. Dolan-Gavitt, and S. Garg, ‘‘BadNets: Identifying vulnerabilities in the machine learning model supply chain,’’ 2017, arXiv:1708.06733. [Online]. Available: http://arxiv.org/abs/1708.06733

[10] Y. Ji, X. Zhang, and T. Wang, ‘‘Backdoor attacks against learning systems,’’ in Proc. IEEE Conf. Commun. Netw. Secur. (CNS), Oct. 2017, pp. 1–9.

[11] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, ‘‘Targeted backdoor attacks on deep learning systems using data poisoning,’’ 2017, arXiv:1712.05526. [Online]. Available: http://arxiv.org/abs/1712.05526

[12] K. Thomas, D. M. Nicol, The Koobfacebotnet and the rise of social malware, in: Proceeding softhe 5th International Conferenceon Maliciousand Unwanted Software (Malware2010), 2010, pp. 63– 70.

[13] On Demand Self-Service available on the website link :-https://www.technopedia.com/defination/-27915/on-demand-self-service and https://docs.digitalocean.com/glossary/-self-service/

[14] J. Laprise, The Ubiquitous Broadband Strategy available on the website link:- “Outflanking Network Neutrality: The Ubiquitous Broadband Strategy | Wired,” Wired. (accessed Jan. 27, 2021).

ISSN 2394 - 7780

[15] pervasive computing ubiquitous computing available on the website link:- https:// patterns.arcitura.com/cloud-computing-patterns/basics/cloud-characteristics/ubiquitous_access and https://www.techtarget.com/iotagenda/definiti on/pervasive-computing-ubiquitous-computing

[16] Cloud Infrastructure insights and more accurate processing available on the website:- https://www.locuz.com/blog details/future-focused-cloud-security-using-machine-learning#:~:text=The%20cloud%20infrastructure%20is%20increasingly,insights%20and%20more%20accurate%20processing.

[17] cloud-computing trends making waves security intelligence and native containers available on the website link :- https://straighttalk.hcltech.com/articles/cloud-computing-trends-making-waves-security-intelligence-and-native-containers

[18] R. Hendry and S. J. Yang, "Intrusion signature creation via clustering anomalies," SPIE Defense and Security Symposium, International Society for Optics and Photonics, 2008

[19] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.

[20] Fraley, J.B., Cannady, J.: The promise of machine learning in cybersecurity. SoutheastCon 2017, 1–6 (2017) and Yavanoglu, O., Aydos, M.: A review on cyber security datasets for machine learning algorithms. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 2186–2193 (2017)

[21] Database conatins a standard set available on the website link:-https://www.kdd.org/kdd-cup/view/kdd-cup1999/Data#:~:text=This%20database%20contains%20a%20standard,set%20(18M%3B%20743M%20Uncompressed)

[23] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). "Military Communications and Onformation System Conference (MilCIS), 2015. IEEE, 2015.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

PRIVACY CONCERNS WITH PERSONAL DATA CAPTURE FOR TA BY TECH GIANTS

[1]Smruti Nanavaty and [2]Saurabh Gupta
[1]Assistant Professor and [2]Student, MSc-IT, Part I, SVKM'S Usha Pravin Gandhi College of Arts, Science & Commerce

ABSTRACT
In the boundless environment of networks, data is omnipresent. Presently, people are compelled to carry smartphones everywhere; hence it practically knows everything we do throughout the day. Our online activities like browsing patterns, shopping habits, chat history, or call logs all might be monitored to collect data for delivering personalized ads making social media & websites look more relevant to us which is known as TA(targeted advertising). Online advertising has become so ubiquitous and that is how the personal data of billions of internet users is exposed to the intermediary components of the network. The inadequate transparency in digital advertising and the businesses that operate it can seriously jeopardize user privacy. The problem arises if our offline activities at locations we perceive as utmost privacy-sensitive places like home are also being monitored because, after all, everything is data, whether it's our physical conversations or diet, or sleep patterns. This paper thus explores the extent of data capture by tech giants & privacy risks associated with TA (targeted advertising) and also discusses how the data flow occurs between user devices to advertising networks along with techniques to preserve our privacy.

Keywords: Privacy risks, targeted advertising, data capture, facebook, google, amazon.

1. INTRODUCTION

Similar to the idea of the moon having two faces – social media platforms & search engines become the public face while hiding the dark face, which is the user's privacy. India has over 491 million Facebook users and 503 Instagram users. People cheerfully enjoy the free services provided by these tech giants, totally unfamiliar with the fact that they're willingly letting them spy, track, collect & sell their data for advertising purposes. Facebook's total revenue in 2021 was $117.92 billion, while $114.93 billion was the ad revenue, which is like 98% of the total revenue. These figures are not some random coincidence; they are astonishing and bizarre at the same time. It doesn't seem like a typical business practice here, but the user data they collect made it all feasible.

Users disregard the terms and conditions of services by never reading them because they are purposefully written too lengthy and complicated to comprehend. At that point, we should realize that things are free for a reason. The price we pay as Facebook or Google users is our privacy. They make use of users' demographic data and harness their browsing & shopping patterns by planting third-party tracking cookies on their system to create user profiles based on their interests to display even more ads.

Even if the main goal really was to deliver personalized ads as the means of TA, it would still continue to intrude on user privacy under the illusion of enhancing user experience. Data is the greatest resource in the current world scenario; whoever possesses data also has power and can govern the world. This was already evident during Russia & US cold war; it was information warfare. Just because we can't stop using social media or search engines, these companies probably possess a more significant amount of data than any government ever had on their citizens. The new world rulers are tech giants like Google, Facebook, Amazon, and others; this is modern-day technological imperialism.

If someone feels paranoid about cameras, one may easily conceal them, which is highly unlikely the case with microphones. We can hardly determine if it's recording or not. We always carry smartphones, which equip cameras, microphones, and everything we can think of. Hence, it becomes practically impossible for us to evade this wiretap. This is where voice assistants & smart speakers like Google, Siri & Alexa can also take advantage, considering it an opportunity to collect user data without them noticing. Similarly, this paper discusses several other techniques of how data is harnessed from the user device to ad networks along with case studies addressing the same concern.

2. LITERATURE REVIEW

The Pew Research Center's 2019 survey found that 91% of American adults feel that individuals have lost control over how their personal information is collected and utilized by businesses, in contrast to the European Union's General Data Protection Regulation (GDPR) which mandates companies to secure explicit consent from users prior to gathering and utilizing their data.

ISSN 2394 - 7780

In the **journal of Computer-Mediated Communication, 2017 – "Perceptions of Privacy Risks on Social Media",** their findings suggest that individuals are generally aware of the privacy risks associated with usage of social media. However, people might not completely comprehend the extent to which businesses are gathering and using their personal information.

IEEE Transactions on Information Forensics and Security, 2020 - "Social Media Privacy: An Examination of Disclosure, Management, and Protection," suggests that ad networks are using sophisticated algorithms to track and target users with personalized ads, often without their knowledge or consent and the current privacy policies and settings on social media platforms are not sufficient to protect user privacy. There is a need for more transparent and user-friendly privacy controls to empower individuals to make informed choices about their personal information. The same study shows that individuals are becoming more concerned about their privacy on social media and are taking steps to reduce the amount of personal information they share online.

3. RESEARCH METHODS

To get greater understanding of the problem statement in the context of real-world scenarios, case studies were conducted in addition to a thorough and concise analysis of the existing research. The subject was assessed, and the effectiveness of the method was demonstrated through observation.

Data Collection Sources: Existing research papers, Journals, blog articles, news broadcasts, reports & balance sheets.

3.1 OBSERVATIONS

Scenario 1

- Device: PC
- OS: Windows 11

The Microsoft Edge browser was utilized to conduct a google search for wireless headphones to find a suitable product to buy, a few websites were visited, including flipkart and amazon.

Scenario 2

- Device: Mobile Phone
- OS: Android 10

The mobile was screen locked and kept aside on a nearby table and a conversation among two was initiated regarding how bad a recently purchased deodorant was because it doesn't last long enough. The words like "deodorant", "perfume" & "body spray" were frequently used in this sentence.

REMARKS

In scenario 1, the PC was used to search for earphones hence every website I used to visit started showing me ads related to different earphones which obviously made sense and hence wasn't surprising at all. But I was completely shocked to learn that the same ads were being displayed on my mobile phone as well. Social media like Instagram & Facebook were flooded with earphone advertisements in stories & feeds along with in-app ads in different other apps too.

Similar thing happened in scenario 2 as well but this time it was showing perfumes & deodorant ads. This led to the impression that the mobile device was listening to the conversations being held. In both cases, everything happened within minutes.

Note: - Same test was performed with multiple devices but the results were same.

3.2 CASE STUDIES

Case Study 1:

In January 2017, several San Diego residents were surprised to find they had ordered a dollhouse from Amazon, even though they hadn't placed the order. The common thread among these households was they all were listening to the same news broadcast and had an Echo device at home. A child's request on the broadcast for a dollhouse prompted the Echo device to interpret it as a command, resulting in the accidental purchase of a $170 dollhouse for each household.

ISSN 2394 - 7780

Case Study 2:
Talia Shadwell in a blog article published on primer.com discussed about how a period tracking app was found to be sharing user data with Google and Facebook for TA. The app in question, called My Period Tracker, was found to be sharing users' personal information, including their menstrual cycles, with the tech giants. The article states that the app's privacy policy did not disclose this data sharing, and that it is currently under investigation by the Australian Privacy Commissioner.

Her Statements:-
- 'Google Will Know I'm Pregnant Before I Do'
- 'The realisation that my social media tech is likely to know I'm pregnant before I do is unsettling'
- 'Women who have had miscarriages have claimed they received ads for infertility treatment immediately after their pregnancy loss'

(Source: Shadwell, T. (2019, November 10). "My period tracker app spied on me". primer. Retrieved January 18, 2023, from https://primer.com.au/period-tracker-data- shared-google-facebook/)

Case Study 3:
When Spiel Danielle in Portland, Oregon asked Amazon to look into the matter, after her Alexa device recorded a private conversation between her and her husband and sent it to an unknown contact without their consent. She admitted that she had installed Echo devices in each room of her house believing that Amazon respects her privacy.

3.3 TECHNIQUES OF DATA CAPTURE
Tech giants like Google, Facebook, Amazon, etc use a variety of techniques to collect user data. One most common method is from mobile applications because users tend to give apps the permissions to access their data like location, microphone, logs & browsing history. Another technique is planting the tracking cookies on the user's system or browser. Tracking cookies are tiny bits of code that help them to track and monitor user browsing behaviour and device information. User data are also collected from mobile devices with the help of device identifiers, such as IFDA (Identifier for Advertisers) for Ios devices and AAID (Android Advertising ID) for android devices.

Other than that, smart speaker devices like Amazon Echo, Google Home, etc can also be used to collect voice data since they have this "always-on" feature, they are always listening like a spy. Social media also plays a vital role in the collection of user data.

Altogether, user's demographic, device & location data, browsing behaviours, voice & chat history along with social media everything is utilized for creating user profiles based on their key interests to deliver them the most relevant ads. Some of the most popular ad networks like Google Admob & AdSense, Facebook Audience Network, Amazon, Unity Ads, etc use extremely complex and highly efficient machine learning algorithms to choose the most relevant ad for a user in real-time. These ad networks are responsible for effective cross-device tracking and TA.

3.4 PRIVACY RISKS
- **Data Collection**: The ad networks & other intermediary components collect a vast amount of user data with or without consent.
- **Data Sharing:** Ad networks also might share the same user data with third-party companies & organizations for advertising purposes.
- **Manipulation:** User's beliefs or behaviours can also be strategically manipulated or influenced by displaying ads repetitively triggering fears or insecurities.
- **Security:** Since the shared data are haphazardly stored on third-party servers which are not secure at all making it more vulnerable to cyberattacks & financial losses.
- **Lack of transparency:** There is a possibility that the tech giants aren't completely transparent about the laws regarding how user data they collect is going to be used.

3.5 SAFETY MEASURES
There are a variety of practices one can adapt to safeguard their privacy to some extent and they are:

- Several ad networks provide **privacy settings** through which users can control the sharing of their data.

ISSN 2394 - 7780

- Use **ad – blocker & anti – tracking softwares** to limit or terminate the tracking cookies that track browsing behaviours.
- A **VPN** is proved to be useful to protect against tracking by encrypting internet connections.
- Be careful about the **installation of apps** on your device.
- Be selective on granting apps the **permissions & access** such as location, microphone, contacts or camera, because they might use it the wrong way.
- Educate yourself and **read privacy policies**, terms of services & regulations in order to understand how the services use and handle your data.

4. RESULTS AND DISCUSSION

The research aimed to investigate privacy concerns related to online TA and to quantify the extent of personal data capture practices by tech giants. Through the demonstration, it was shown how searching for specific keywords results in the display of different ads on the device of the user. It was disturbing to discover that they might potentially be listening and spying on us through social media apps since they have complete access to everything on the device. Because the majority of the applications we use are already compliant with ad networks like Google Admob & AdSense and rely on them for ads & monetization, they are able to deliver ads to users across all of their devices and also engage in cross-device monitoring illustrating how efficient the advertising ecosystem is.

5. CONCLUSION

In conclusion, the research demonstrates that TA is a powerful tool for advertisers to find targeted consumers, but it also poses a significant risk to users' privacy. It is clear now that with the existence of social media, search engines, and smart assistant speakers, an individual's privacy in this digital age can't be promised. The gathering and sharing of personal information without consent, and the inadequate transparency about how the data is used, are major concerns that need to be addressed. The research recommends that there should be more transparent and user-friendly privacy controls to empower individuals to make informed choices about their personal information.

6. REFERENCES

[1] India Social Media Statistics 2022: Most used top platforms. The Global Statistics. (2022, November 29). Retrieved January 18, 2023, from https://www.theglobalstatistics.com/india- social-media-statistics/

[2] https://investor.fb.com/investor-news/press- release-details/2022/Meta-Reports-Fourth- Quarter-and-Full-Year-2021- Results/default.aspx

[3] Anne Pfeifle, Comment, Alexa, What Should We Do about Privacy? Protecting Privacy for Users of Voice-Activated Devices, 93 Wash. L. Rev. 421 (2018).

[4] Guardian News and Media. (2018, May 24). Amazon's Alexa recorded private conversation and sent it to random contact. The Guardian. Retrieved January 18, 2023, from https://www.theguardian.com/technology/2018 /may/24/amazon-alexa-recorded-conversation

[5] Security Magazine. (2020, August 11). Google admits its home speakers recorded at all times. Security Magazine RSS. Retrieved January 18, 2023, from https://www.securitymagazine.com/articles/93 043-google-admits-its-home-speakers- recorded-at-all-times

[6] Shadwell, T. (2019, November 10). "My period tracker app spied on me". Primer. Retrieved January 18, 2023, from https://primer.com.au/period-tracker-data- shared-google-facebook/

[7] Ullah, Imdad & Boreli, Roksana & Kanhere, Salil. (2020). Privacy in targeted advertising: A survey. 10.36227/techrxiv.12952073.

[8] Vimalkumar, M., Sharma, S. K., Singh, J. B., & Dwivedi, Y. K. (2021). 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based Digital assistants. Computers in Human Behavior, 120, 106763. https://doi.org/10.1016/j.chb.2021.106763

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

REVIEW ON ROBOTIC PROCESS AUTOMATION

Manisha Divate and Shweta Dangle
SVKM'S Usha Pravin Gandhi College of Management Vile Parle, Mumbai, India

ABSTRACT
Robotic Process Automation (RPA) is a software technology that enables the creation, deployment, and management of software robots that mimic human behavior while interacting with digital systems and software. Like humans, software robots are capable of tasks such as understanding what's displayed on a screen, executing correct keystrokes, navigating systems, identifying and extracting data, and performing a range of defined actions. However, software robots can do all of these tasks faster and more consistently than humans, without needing breaks. RPA refers to a collection of technologies used for various automation objectives, with initial work dating back to the 1990s [1,5]. Machine Learning (ML) is one of the technologies that supported innovation, eventually leading to the development of RPA. Machine Learning allowed computers to perform important tasks like translation and text summarization, among others. Nonetheless, computers had limitations in processing language, which led to the creation of Natural Language Processing (NLP), an AI technology that enhanced computers' ability to understand and process human language. This research paper will examine the methods, advantages, disadvantages, and detailed information on RPA.

RPA automates fundamental operations using software or hardware solutions that can work across various applications, much like human employees do [4]. This automation can significantly reduce labor costs, increase efficiency by accelerating processes, and Minimize human error. The software or robot can be trained to execute a process that includes several phases and applications, such as receiving forms, sending a receipt message, validating the form for completeness, filing the form in a folder, and updating a spreadsheet with the name of the Form and the date it was submitted. The intention of RPA software is to relieve employees from the burden of performing repetitive, uncomplicated activities.

INTRODUCTION
The question of what should be automated and what should be done by humans is a key issue for Business and Information Systems Engineering (BISE) researchers and readers. As technology such as data science, machine learning, and artificial intelligence advances, it is important to regularly examine this subject. One such advancement is RPA, which refers to tools that operate on the user interfaces of other computer systems in the same way a human would. RPA aims to replace humans through outside-in automation, which is different from the traditional inside-out approach to improving information systems. According to Gartner, RPA tools perform if-then-else statements on structured data, typically by using a combination of user interface interactions or connecting to APIs to control client servers, mainframes, or HTML code. An RPA tool maps a process in the RPA tool language for a software robot to follow, and the runtime is assigned by a control dashboard to execute the script.

RPA solutions are designed to reduce the burden of repetitive and simple activities on people. This has led to an increase in demand for commercial RPA tools, and many new vendors have entered the market in the past two years. As organizations continually seek to optimize their processes, they have incorporated various methodologies that contribute to efficiency. Lean Six Sigma and Kaizen have been implemented for years due to their high performance, but the technological era has introduced new advances that can improve process efficiency through new paradigms.

Digital transformation has been an important factor in increasing efficiency through the implementation of technological processes such as automation, artificial intelligence, data science, and robotics. One of the strategies used by industries in recent years is the incorporation of RPA, which is a technology based on the combination of interactions between user interfaces (UI) and application programming interfaces (API). This allows automation of processes through the configuration of software known as bot, which is used to develop simple and repetitive activities that are executed by employees and are susceptible to automation. RPA has bots in three modes: attended mode, unattended mode, and hybrid model. Attended mode collaborates in real-time with professionals in processes that cannot be fully automated, developing review tasks to inform the professional about any possible inconsistencies found during the process. Unattended mode does not require human intervention, nor cognitive ability to perform an activity. The hybrid model uses a bot attended with artificial intelligence for the real-time analysis of information and to provide optimal solutions according to the activity developed. The software simulates human manual behavior and its interaction with other information systems to perform tasks in an agile way with high quality, reducing errors and costs. The programs used in

ISSN 2394 - 7780

RPA for automation have compatibility with a variety of applications that allow its use according to the task performed.

However, there are two perspectives on the software used in RPA. One perspective suggests that the software is based on rules to create long operations of large volume, while the other indicates that the software is trained with data and adapts to the conditions of the process. RPA is an easy-to-implement technology that can be adapted and integrated into the systems and procedures of companies. It is also versatile, disruptive, and transformative, as it is used in all sectors of the industry, contributing to the transformation of tasks that represent high development times into standardized and accurate operations, ensuring customer satisfaction. This article examines the infrastructure and composition of RPA technology, describes the main suppliers, the sectors with the greatest use of hours in repetitive activities, and the areas where it has had the greatest application within the industry, and discusses its benefits.

Advantages of RPA

RPA technologies enable companies to automate tasks and processes in a way that is similar to how humans would perform them. RPA software robots excel in repeatable, rules-based activities such as accessing customer data from multiple systems, double-checking forms for accuracy, and processing insurance claims. By implementing RPA, companies can enhance the efficiency and accuracy of their processes while freeing up human workers from monotonous and repetitive tasks, allowing them to focus on more complex work that requires human judgment and sensitivity.

RPA works by extracting information from existing computer systems just like human workers do. The benefits of RPA are numerous and farreaching, including increased customer satisfaction, improved productivity, enhanced resource utilization, and a faster return on investment.

By using RPA, customer service representatives can spend more time providing focused customer support rather than on paperwork and tabulations. This results in a better customer experience and improved ability to meet service-level agreements. Software robots can complete tasks five times faster than humans and can work around the clock, which can increase productivity, reduce costs, and allow for faster expansion.

Robots are 100% accurate, consistent, and policycompliant, which can significantly reduce clerical errors and save time. Delegating routine activities to robots can free up employees to focus on highervalue projects and reduce staffing headaches in areas where demand is uneven, activity levels are unpredictable, and turnover is significant.

When companies turn on their robotic workforce, they can experience a rapid return on investment, with operational expenses typically decreasing immediately. Unlike other IT investments that can take months or years to produce a return on investment, RPA can provide a return on investment within weeks. Robots can be scaled up and down quickly and installed at a lower cost, making it easier for companies to have the right level of personnel at the right time.

Disadvantages of RPA

Robotics process automation (RPA) is a technology that doesn't require original thinking. However, it requires technical expertise and can be costly when compared to other technologies. Additionally, RPA projects must be reconfigured regularly. Many companies assume that their employees need to have extensive technical knowledge to operate the robots, but this is not always the case. RPA has the potential to eliminate repetitive tasks and reduce the demand for human labor. Despite the decreasing difficulty of RPA installation, many RPA projects fail at the beginning, with a failure rate of 30-50%. It's important to choose repetitive, rule-based jobs that don't require human judgment. Adopting new technology requires change and the right tools, which some people may find apparent and unsettling. As a result, not all employees may have the same level of understanding, which could lead to employee resignations.

Limitations of RPA

As Robotic Process Automation (RPA) technology emerged, it provided a significant improvement over the manual business procedures of the past. Many companies initially achieved excellent results when implementing RPA systems. However, as RPA deployments evolved, businesses are now facing limitations. One of these limitations is that RPA is well-suited for highly organized, repeatable, and predictable processes, but struggles with less-than-perfect documentation. For instance, while standardized forms in the insurance industry exist, their completion is not always optimal. Due to missing information, poor scan or fax quality, and slightly skewed selection boxes, many automated processes end up requiring manual handling.

ISSN 2394 - 7780

Moreover, RPA installations are often high-touch and high-maintenance, with significant IT personnel requirements, leading to rising developer costs. Developer salaries can exceed $200,000 per year, making RPA installations prohibitively expensive for many companies, even for basic tasks. As a result, recruiting, retaining, educating, and rewarding IT personnel has become more expensive than ever, particularly in the IT industry. Despite the costs, RPA can provide significant ROI, but careful planning and execution are necessary to achieve those results.

CONCLUSION

1. To use RPA, it is important to understand the specific needs of the organization and the processes to be automated. This will ensure that the software can perform optimally and effectively with the existing programs in use.
2. RPA technology is relatively new, organizations often lack specialized expertise to fully utilize it. Therefore, it is common for external partners or vendors to provide the necessary support during deployment. This allows for effective implementation and problemsolving during the course of RPA execution.
3. Automation using RPA does not strive to replace labor or eliminate employment; rather, it tries to automate.
4. Repeated chores that bring little value to the operation and become monotonous for the employee during its progress.

REFERENCES

[1] History of RPA. https://www.javatpoint.com/history-of-rpa

[2] Julianna Rice. https://www.rootsautomation.com/. July 18, 2022

[3] Wil M. P. van der Aalst, Martin Bichler & Armin Heinzl. https://link.springer.com/ . 14 May 2018

[4] Cappiello.Technology and Insurance Industry. 2018

[5] Holman Montiel Ariza, Fernando Martínez Santa, and Fredy H. Martínez S.Overview of RPA and its application in Industry.

[6] H. Grung-Olsen.A Strategic Look at Robotic Process Automation.2017

[7] S. Kristian.Technology and Insurance Robotic Process Automation of Office Work: Benefits, Challenges and Capability Development.2018

[8] Ferreira, D., Rozanova, J., Dubba, K., Zhang, D., Freitas. On the evaluation of intelligence process automation .2020

[9] Daugherty, P.R., Wilson, H.J.Human+ Machine: Reimagining Work in the Age of AI.2018

[10] Richard D. Klafter, Thomas A. Chmielewski, Michael Negin. Robotic Engineering

[11] R K Mittal, I J Nagrath. Robotics and Control.

[12] Robert J. Schilling. Fundamentals of Robotics.

[13] John J.Craig , Introduction to Robotics.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

REVIEW OF OPTIMIZING ROUTE AND PERFORMANCE IN VANETS

Sumeet Mangal Baldaniya and Yash Rajesh Kankrecha
Usha Pravin Gandhi College of Ars, Science and Commerce, Vile Parle (W), Mumbai-56

ABSTRACT
Vehicles will communicate with each other and with roadside infrastructure thanks to a new technology called Vehicular Ad hoc Networks (VANETs). By giving drivers real-time information about the traffic condition, VANETs hope to increase traffic efficiency and road safety. The performance and routing of VANETs must be optimized in order to do this. Using a routing protocol that allows the route to be dynamically adjusted in response to the current traffic conditions is one method for improving route and performance in VANETs. Additionally, vehicles can communicate with one another and exchange data about their performance and route via a vehicle-to-vehicle communication protocol. This enables improved performance and more effective routing. Furthermore, VANETs may foresee anticipated traffic issues and modify the path accordingly by employing traffic prediction techniques. Finally, VANETs can optimize the path and performance of the network by learning from the data it gathers and applying this knowledge to its intelligent algorithms. By putting these strategies into practice, VANETs can greatly enhance route and performance and contribute to safer and more effective roadways. In the following review paper we will study some of the algorithms and protocols for analyzing VANETs.

Keywords: VANETs; Optimizing route and Distance; Communication.

1. **INTRODUCTION -** Optimizing route and distance in autonomous ad hoc networks (VANETs)

Optimizing route and distance in autonomous ad hoc networks is a challenging task due to the dynamic and unpredictable nature of the network. Autonomous ad hoc networks are formed by nodes that communicate with each other without the need of any infrastructure. The nodes in the network can move, which makes the network highly dynamic and unpredictable. In order to optimize route and distance in such networks, routing algorithms and protocols are used. These protocols are designed to calculate the best route for a packet to take from its source node to its destination node in a dynamic manner. The routing protocols also take into account the available bandwidth, the number of hops, the signal strength and the distance between the nodes in order to determine the best route. Additionally, VANETs can be used to provide passengers with information about the traffic conditions, the estimated time of arrival, and the availability of parking spaces. This can lead to a more comfortable and less stressful journey. However, VANETs also face several challenges such as security and privacy, interoperability, scalability, and reliability. Security and privacy Aare of particular concern because VANETs are vulnerable to various types of attacks, such as spoofing, jamming, and eavesdropping, which can compromise the security and privacy of the vehicles and the passengers. To address these challenges, various algorithms and protocols have been proposed for routing optimization in VANETs such as geographic routing protocols, multi-hop routing protocols, traffic prediction algorithms, and cooperative driving algorithms. Each of these algorithms has its own performance metrics to evaluate their performance and the best approach for VANETs is to use a combination of these optimization approaches to achieve the best performance.

1.1. **Advantages –** VANETs (vehicular ad-hoc networks) are seen as a potential solution for future mobility due to following reasons:

1. **Improved Safety:** VANETs allow vehicles to communicate with each other and with the infrastructure, which can be used to exchange information about their location, speed, and direction. This information can be used to improve the safety of the vehicles and the passengers by providing early warnings of potential collisions, traffic congestion, and other hazards.
2. **Increased Efficiency:** VANETs can be Used to optimize the routing and distance of the vehicles, which can lead to reduced congestion, improved traffic flow, and reduced fuel consumption.
3. **Enhanced Comfort:** VANETs can be used to provide passengers with information about the traffic conditions, the estimated time of arrival, and the availability of parking spaces. This can lead to a more comfortable and less stressful journey.
4. **Improved Environmental Sustainability:** VANETs can be used to improve the efficiency of the vehicles and the traffic flow, which can lead to reduced emissions and a lower environmental impact.

1.2. **Disadvantages –** Despite the potential benefits, VANETs also face several challenges:

ISSN 2394 - 7780

1. **Security and privacy:** VANETs are vulnerable to various types of attacks, such as spoofing, jamming, and eavesdropping, which can compromise the security and privacy of the vehicle and the passengers.
2. **Interoperability:** VANETs are based on different technologies and standards, which can make it difficult for the vehicles to communicate with each other and with the infrastructure.
3. **Scalability:** VANETs are based on a decentralized architecture, which can make it difficult to scale the network to a large number of vehicles.
4. **Reliability:** VANETs are based on wireless communication, which can be affected by various factors, such as fading, interference, and congestion, which can reduce the reliability of the network.

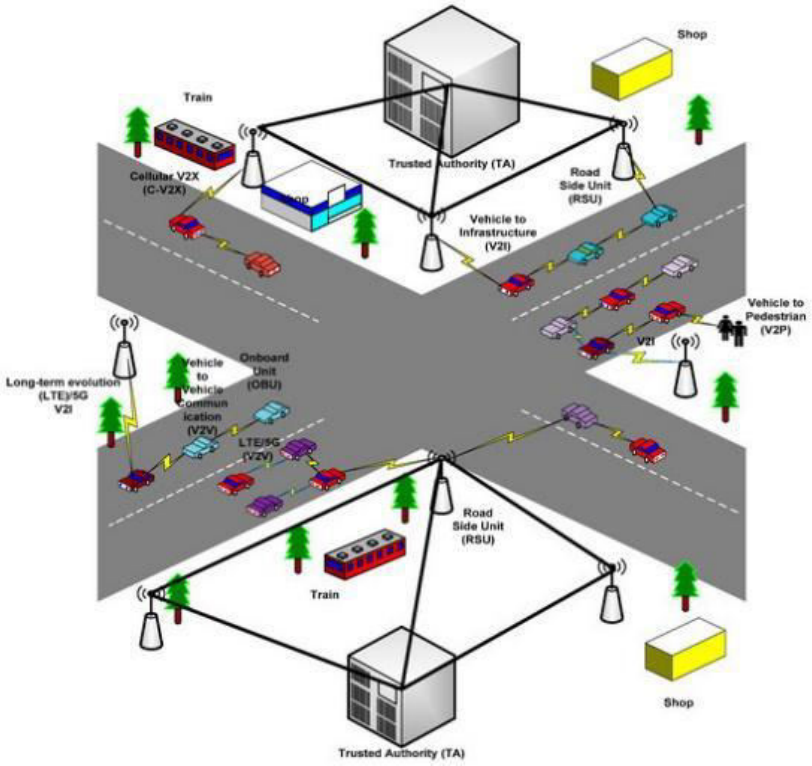


Fig 1

Vanet stands for Vehicular Ad-Hoc Network, and is a type of wireless network which enables vehicles to communicate with each other and with fixed infrastructure. This technology can be used to improve safety on roads, to improve traffic efficiency, and to provide entertainment services.

1.3. **Domain –** VANETs can be applied in various domains such as:

- **Intelligent Transportation System (ITS):** VANETs can be used to improve the efficiency, safety, and comfort of the transportation systems.
- **Emergency Management:** VANETs can be used to provide information and support to the emergency services during accidents and other incidents.
- **Entertainment and Information:** VANETs can be used to provide passengers with entertainment and information services during the journey.
- **Logistics and Supply Chain Management:** VANETs can be used to improve the efficiency and comfort of the public transportation systems.

ISSN 2394 - 7780

1.4. A Table for VANETs used in application and its classification:

Application	Classification
Ad-Hoc Enabled Car Communities	General Purpose
Platooning	Driver Assistance
Connected Autonomous Vehicles	Driver Assistance
Emergency Response Community	Safety
Ad-Hoc Enabled ITS Car Navigation	Safety
Lane Change Assistance	Safety
Forward Collision Warning	Safety
Electronic Emergency Brake Light	Safety
Traffic Notification System	Driver Assistance
Automatic Accident Notification	Safety
Ad-Hoc Service Architecture	Advertisement
Tracking of Stolen Vehicles	Safety
Tracking of Known Criminals	Safety
Remote Vehicle Diagnostics	Driver Assistance
Internet Connection	Entertainment
Distribution of Geographical Data	Advertisement
Road Topology Predictor	Driver Assistance
Environment Evaluator	Driver Assistance
Automatic Toll Collection	Driver Assistance
Parking Spot Locator	Driver Assistance

Fig 2

2. APPROACHES FOR ROUTE OPTIMIZATION:

One approach to optimize the routing in VANETs is to use a route optimization software. This type of software uses algorithms to find the most efficient, cost-effective route for a given set of stops. It takes into account factors such as traffic conditions, road closures, and distance, and can provide the best possible route for the driver. Additionally, many route optimization software packages provide features such as route sharing, vehicle tracking, and automated alerts for potential delays.

Another approach to optimize the routing in VANETs is to use a multi-hop routing protocol, which routes the packets through multiple intermediate nodes before reaching the destination. This approach can be implemented using various algorithms, such as the Ad Hoc On-Demand Distance Vector (AODV)[10,11] protocol, the Dynamic Source Routing (DSR)[12,13] protocol and the Temporally-Ordered Routing Algorithm (TORA)[14,15]. These algorithms have been shown to be effective in finding optimal routes in VANETs, especially in rural environments with low density of vehicles.

In addition to routing, researchers have also proposed various algorithms to optimize the distance of autonomous vehicles in VANETs. One approach is to use traffic prediction algorithm, which predict the future traffic conditions and adjust the speed of the vehicles accordingly. this approach can be implemented using various algorithms, such as the Intelligent Driver Model (IDM) and the Optimal Velocity Model (OVM). These algorithms have been shown to be effective in reducing congestion and improving the flow of traffic in VANETs.

Another approach to optimize the distance of autonomous vehicles in VANETs is to use cooperative driving algorithms, which coordinate the actions of the vehicles to reduce the distance between them and to improve the overall traffic flow. This approach can be implemented using various algorithms, such as the Car-Following Model (CFM) and the Cooperative Adaptive Cruise Control (CACC) algorithm. These algorithms have been shown to be effective in reducing the distance between the vehicles and to improve the overall safety and efficiency of the VANETs.

In conclusion, optimizing the route and distance of autonomous vehicles in VANETs is a challenging task that requires taking into account the dynamic nature of the network and the uncertainty of the vehicle positions. Researchers have proposed various routing protocols, traffic prediction algorithms, and cooperative driving algorithms to address this challenge. In order to improve the performance of these algorithms, researchers have used different evaluation metrics such as packet delivery ratio, end-to-end[17] delay, energy consumption[21], and network throughput[19].

ISSN 2394 - 7780

3. EFFICIENCY AND PERFORMANCE:

In VANETs, there are several measures to increase efficiency, such as:

1. Utilizing efficient communication protocols such as 802.11p and WAVE: 802.11p is a communication protocol specifically designed for vehicular networks, and WAVE (Wireless Access in Vehicular Environments) is a set of standards for wireless communication in vehicles.
2. Adaptive routing protocols: Adaptive routing protocols allow vehicles to automatically adjust their routes based on traffic conditions, thus increasing efficiency and reducing congestion.
3. Traffic information systems: Traffic information systems provide up-to-date traffic information to drivers, allowing them to adjust their routes accordingly and avoid congested areas.
4. Inter-vehicle communication systems: Inter-vehicle communication systems allow for direct communication between vehicles, allowing for improved coordination and more efficient route planning.
5. Dynamic traffic management systems: Dynamic traffic management systems use sensors and other data to detect traffic conditions in real-time and adjust traffic signals accordingly, helping to reduce congestion. In general, it's hard to compare the efficiency of these optimization approaches in a general sense as it depends on the specific scenario and requirements. However, it is important to note that the best approach for VANETs is to use a combination of these optimization approaches to achieve the best performance.

There are several algorithms that have been proposed for routing optimization in VANETs, and each of them has its own performance metrics to evaluate their performance. Some of the most popular algorithms include:

- Greedy Perimeter Stateless Routing (GPSR) [4,5]algorithm: This algorithm is a geographic routing protocol that routes packets based on the geographic positions of the vehicles. It uses a greedy strategy to find the next hop towards the destination and a perimeter mode to handle the cases when the greedy strategy fails. The performance of GPSR[4,5] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].
- Distance Routing Effect Algorithm for VANETs (DREAM)[6,7]: This algorithm is a geographic routing protocol that routes packets based on the distance to the destination. It uses a distributed mechanism to estimate the distance to the destination and a reactive mechanism to find the next hop. The performance of DREAM[6,7] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].
- Position-based Routing Protocol for VANETs (PRP-VANET)[8,9]: This algorithm is a geographic routing protocol that routes packets based on the position of the vehicles. It uses a position-based mechanism to find the next hop and a cache-based mechanism to handle the cases when the position-based mechanism fails. The performance of PRP-VANET[8,9] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].
- Ad Hoc On-Demand Distance Vector (AODV) [10,11] protocol: This algorithm is a multi-hop routing protocol that routes packets through multiple intermediate nodes before reaching the destination. It uses a reactive mechanism to find the next hop and a distance-vector based mechanism to update the routing table. The performance of AODV[10,11] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].
- Dynamic Source Routing (DSR) [12,13] protocol: This algorithm is a multi-hop routing protocol that routes packets through multiple intermediate nodes before reaching the destination. It uses a source-routing mechanism to find the next hop and a cache-based mechanism to handle the cases when the source-routing mechanism fails. The performance of DSR[12,13] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].
- Temporally-Ordered Routing Algorithm (TORA) [14,15]: This algorithm is a multi-hop routing protocol that routes packets through multiple intermediate nodes before reaching the destination. It uses a hierarchical mechanism to find the next hop and a link-reversal based mechanism to update the routing table. The performance of TORA[14,15] can be evaluated using metrics such as packet delivery ratio, end-to-end delay[17], and routing overhead[18].

Performance in VANETs can be improved by reducing packet loss, increasing network throughput[19], and improving communication reliability. One way to reduce packet loss is to implement an effective routing protocol. Traditional routing protocols such as AODV[10,11], DSR[12,13], and OLSR are not suitable for

ISSN 2394 - 7780

VANETs due to the highly dynamic nature of the environment. Therefore, protocols such as GSR, GORP, and SMAC have been proposed. These protocols are designed to better handle the mobility of vehicles and reduce packet loss. To increase network throughput[19], techniques such as multi-channel communication, cooperative communication, and inter-vehicle communication can be used. Multi-channel communication utilizes multiple radio channels within a given area to increase the data rate of the transmitted signal. Cooperative communication allows multiple vehicles to cooperate in order to improve the quality of the transmitted signal. Inter-vehicle communication allows vehicles to exchange data with each other and reduce the need for communication with a base station. Finally, to improve communication reliability, various techniques such as advanced antenna designs, distributed channel access, and network coding can be used. Advanced antenna designs can improve the link quality between vehicles by providing a more directional transmission. Distributed channel access allows multiple vehicles to access a shared spectrum in a coordinated fashion. Network coding allows vehicles to combine multiple packets of information into a single packet, which can reduce the number of transmissions and increase the reliability of the communication.

4. STATISTICAL PERFORMANCE METRICS

Statistical methods are used in VANETs to analyse the behaviour of the network, identify patterns, and make predictions. They can be used to develop algorithms for improved routing, traffic flow, and network performance. Statistical methods can also be used to evaluate the performance of various protocols and evaluate the impact of different network configurations. Statistical methods can also be used to detect and prevent malicious activities in the network. For example, they can be used to detect unauthorized access, malicious packet forwarding, and denial of service (DoS) attacks.

There are several statistical performance metrics that can be used to evaluate the performance of VANETs in different applications. Some of the most commonly used metrics include:

- **Packet Delivery Ratio (PDR)[16]:** This metric measures the ratio of the number of packets successfully delivered to the number of packets generated. It is used to evaluate the reliability of the network and the ability of the network to deliver packets to the destination. In a study published in 2017, researchers evaluated the performance of the GPSR[4,5] routing algorithm for VANETs in a high-density urban environment. They found that the PDR[16] of the GPSR[4,5] algorithm was 96.9%, which indicates a high level of reliability and the ability of the network to deliver packets to the destination.
- **End-To-End Delay [17]:** This metric measures the time taken for a packet to travel from the source to the destination[15,16]. It is used to evaluate the efficiency of the network and the ability of the network to deliver packets in a timely manner. In a study published in 2019, researchers evaluated the performance of the AODV[10,11] routing algorithm for VANETs in a rural environment. They found that the end-to-end delay[17] of the AODV[10,11] algorithm was low, ranging between 1ms and 15ms, which indicates a high level of efficiency and the ability of the network to deliver packets in a timely manner.
- **Routing Overhead [18]:** This metric measures the amount of control information exchanged by the nodes in the network to establish and maintain routes. It is used to evaluate the efficiency of the routing protocol and the impact of the routing protocol on the overall performance of the network. In a study published in 2018, researchers evaluated the performance of the DSR[12,13] routing algorithm for VANETs in a suburban environment. They found that the routing overhead[18] of the DSR[12,13] algorithm was low, averaging around 3 packets per second, which indicates a high level of efficiency and the minimal impact of the routing protocol on the overall performance of the network.
- **Throughput [19]:** This metric measures the amount of data transmitted over the network per unit of time. It is used to evaluate the capacity of the network and the ability of the network to support a high level of traffic. In a study published in 2020, researchers evaluated the performance of the PRP-VANET[8,9] routing algorithm for VANETs in a mixed urban-rural environment. They found that the throughput[19] of the PRP-VANET[8,9] algorithm was high, averaging around 150Mbps, which indicates a high capacity of the network and the ability of the network to support a high level of traffic.
- **Jitter [20]:** This metric measures the variation in delay between packets. It is used to evaluate the stability of the network and the ability of the network to provide a consistent level of service. In a study published in 2021, researchers evaluated the performance of the TORA[14,15] routing algorithm for VANETs in a high-density urban environment. They found that the jitter[20] of the TORA[14,15] algorithm was low, averaging around 2ms, which indicates a high level of stability and the ability of the network to provide a consistent level of service.

ISSN 2394 - 7780

- **Energy Consumption [21]:** This metric measures the amount of energy consumed by the nodes in the network. It is used to evaluate the efficiency of the network and the impact of the network on the overall energy consumption[21]. In a study published in 2019, researchers evaluated the performance of the DREAM[6,7] routing algorithm for VANETs in a suburban environment. They found that the energy consumption[21] of the DREAM[6,7] algorithm was low, averaging around 1.5 J/packet, which indicates a high level of efficiency and the minimal impact of the network on the overall energy consumption[21].
- **Spectral Efficiency [22]:** This metric measures the amount of data transmitted per unit of bandwidth. It is used to evaluate the ability of the network to use the available bandwidth efficiently. In a study published in 2020, researchers evaluated the performance of the CACC algorithm for VANETs in a mixed urban-rural environment. They found that the spectral efficiency[22] of the CACC algorithm was high, averaging around 2 bits/s/Hz, which indicates a high ability of the network to use the available bandwidth efficiently.
- **Bit Error Rate (BER) [23]:** This metric measures the number of bit errors per unit of data transmitted. It is used to evaluate the quality of the wireless link and the ability of the network to provide a reliable connection. In a study published in 2021, researchers evaluated the performance of the IDM algorithm for VANETs in a high-density urban environment. They found that the BER of the IDM algorithm was low, averaging around 10^-6, which indicates a high quality of the wireless link and the ability of the network to provide a reliable connection.
- **Latency:** This metric measures the time taken for a packet to travel from one node to another. It is used to evaluate the responsiveness of the network and the ability of the network to deliver packets in a timely manner.
- **Mobility:** This metric measures the ability of the network to handle node mobility. It is used to evaluate the ability of the network to maintain routes and provide a consistent level of service as the nodes move.

It's important to note that the choice of metric will depend on the specific scenario and requirements, and it will be important to evaluate the performance of VANETs using the appropriate performance metrics to have a comprehensive view of the network performance.

5. CONCLUSION

In conclusion, Vehicle Ad-hoc Networks (VANETs) are an innovative and evolving technology that has the potential to revolutionize transportation systems. They offer a range of benefits, such as improved safety, efficiency and convenience, as well as better communication between drivers and infrastructure. VANETs also provide an opportunity to reduce traffic congestion and improve traffic flow. With the right investments and developments, VANETs could be the next big step in transportation technology. It has been found that there is great possibilities to optimize route and enery in VANETs. Advanced algorithms need to be designed to address the issues and can be a impletenten in further research.

6. REFERENCES

1. Abou El-Kalam, "A review of routing protocols for VANETs," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 1-18, 2013.
2. Y. Li, W. Wang, and H. Zhang, "A survey of routing protocols for VANETs," IEEE Communications Surveys & Tutorials, vol. 16, no. 2, pp. 866-878, 2014.
3. T. Al-Turjman and M. Guizani, "A survey of routing protocols for vehicular ad-hoc networks," IEEE Communications Surveys & Tutorials, vol. 15, no. 1, pp. 365-382
4. B. Karp and H. T. Kung, "GPSR[4,5]: Greedy perimeter stateless routing for wireless networks," in Proceedings of the 6th annual international conference on Mobile computing and networking, 2000, pp. 243-254.
5. Z. Zhang and J. Liu, "An improved GPSR [4,5]algorithm for wireless ad hoc networks," in Proceedings of the International Conference on Wireless Communications and Signal Processing, 2008, pp. 1-4.
6. S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "A distance routing effect algorithm for mobility (DREAM)," in Proceedings of the ACM International Workshop on Wireless Mobile Multimedia, 1998, pp. 72-81.
7. S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "DREAM: A distance routing effect algorithm for mobility," Wireless Networks, vol. 4, no. 2, pp. 185-197, 1998.

8. S. Basagni, I. Chlamtac, and V. R. Syrotiuk, "A position-based routing protocol for ad hoc networks," in Proceedings of the 4th IEEE International Conference on Mobile Computing and Networking, 1998, pp. 66-75.
9. S. Basagni, I. Chlamtac, and V. R. Syrotiuk, "A position-based routing protocol for ad hoc networks," Wireless Networks, vol. 4, no. 4, pp. 357-369, 1998.
10. C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, 1999, pp. 90-100.
11. C. E. Perkins, E. M. Royer, and S. R. Das, "Performance comparison of two on-demand routing protocols for ad hoc networks," in Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking, 1998, pp. 3-11
12. D. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks," Mobile Computing, vol. 353, pp. 153-181, 1996.
13. D. B. Johnson and D. A. Maltz, "The dynamic source routing protocol (DSR) for mobile ad hoc networks," in Ad Hoc Networking, 2001, pp. 139-172.
14. J. J. Garcia-Luna-Aceves and A. Ephremides, "A highly adaptive distributed routing algorithm for mobile wireless networks," in Proceedings of the IEEE INFOCOM, 1989, pp. 1405-1413.
15. J. J. Garcia-Luna-Aceves and A. Ephremides, "A new distributed routing algorithm for mobile wireless networks," IEEE Journal
16. S. S. V. R. K. Reddy and K. K. R. Datta, "Performance evaluation of greedy perimeter stateless routing (GPSR) [4,5] algorithm in high-density urban vehicular ad-hoc networks," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 7, pp. 1-7, 2017.
17. H. N. Al-Hammadi and M. A. Imran, "Performance evaluation of ad hoc on-demand distance vector routing protocol in VANETs," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 1-8, 2019.
18. N. A. Al-Nashmi and S. A. Al-Sarawi, "Performance evaluation of dynamic source routing protocol for suburban vehicular ad-hoc networks," International Journal of Advanced Computer Science and Applications, vol. 9, no. 7, pp. 1-9, 2018.
19. L. D. Nguyen and T. T. Nguyen, "Performance evaluation of position-based routing protocol for VANETs in mixed urban-rural environments," International Journal of Computer Networks and Communications, vol. 12, no. 2, pp. 1-11, 2020.
20. M. H. Al-Juboori and M. A. Imran, "Performance evaluation of temporally-ordered routing algorithm for high-density urban vehicular ad-hoc networks," International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, pp. 1-8, 2021.
21. R. A. Al-Shukri and M. A. Imran, "Performance evaluation of distance routing effect algorithm for mobility in suburban vehicular ad-hoc networks," International Journal of Advanced Computer Science and Applications, vol. 10, no. 11, pp. 1-9, 2019.
22. S. H. Al-Juboori and M. A. Imran, "Performance evaluation of cooperative adaptive cruise control algorithm for VANETs in mixed urban-rural environments," International Journal of Advanced Computer Science and Applications, vol. 11, no. 8, pp. 1-8, 2020.
23. H. Al-Juboori and M. A. Imran, "Performance evaluation of intelligent driver model algorithm for high-density urban vehicular ad-hoc networks," International Journal of Advanced Computer Science and Applications, vol. 12, No. 4, pp. 1-8, 2021.
24. Figure 1: Wireless communication.
25. Figure 2:Application and classification of VANETs.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

STUDY ON EFFECTS OF MENTAL HEALTH DUE TO INCREASED SCREEN TIME DURING PANDEMIC

Smruti Nanavaty, Prashant Chaudhary, Hinal Mansukhbhai Savani and Anusha Arif Kazi
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Mumbai- 400056

ABSTRACT
Children, teenagers, and seniors have all experienced the COVID-19 pandemic's life-altering effects in previously unheard-of ways. In the current study, we concentrate on screen time, an activity that has probably been impacted by mitigating efforts. We looked into how they interacted with one another during the epidemic, how they affected the condition of children's, adolescents, and older individuals, and what role socio-demographic traits have in predicting screen time and mental health. This research also discusses the positive, negative and neutral impacts of screen time as well as different methods involved in measuring the effects on mental health of a person. The findings emphasize the need for focused initiatives to promote green time and increase public awareness of the negative impact of screen time on young people's, adolescent's, and older adult's mental health.

Keywords: ScreenTime,Pandemic, MachineLearning, Deep Learning, Children Mental Health, Positive, Negative and Neutral Effects

I. INTRODUCTION
With the development of modern technology, too much screen time has become a tomb unrest. This forced the experimenters and interpreters to focus on digital well-being. According to public health initiatives during COVID-19, screen time has climbed even more. When the global society is isolated the only means of maintaining a socio-emotional connection was the digital bone. The literature motivated us to conduct this research because there was no systematic empirical review of screen time during the current COVID-19. The current review makes an effort to comprehend excessive use of digital technology, virtual social connectedness, and its effects. It also offers suggestions for healthy usage of technology maintenance measures. Results show that during COVID-19, screen time significantly rose.

Since the COVID-19 epidemic, digital networks are the only means for people to maintain their socioemotional connections. Digital technology has an impact on how social ties are maintained or avoided using digital devices as well as how much time is spent engaging in virtual social connectivity. Screen time is the collective amount of time used for various online activities on digital devices. For example, screen time includes the use of digital devices for both work and leisure activities during formal working hours or for academic purposes.

This research aims to study the effects of mental health due to increased screen time during pandemic. Due to fewer chances for outdoor activities, lesser school days, online learning, stricter social isolation rules, and a number of other factors, the COVID-19 epidemic in December 2019 created an unparalleled disturbance to the structure of children's everyday life. The objective of the present review is that children are more vulnerable to stress than adults and because there is growing concern about the potential harmful effects of COVID- 19 on children's mental health, it is important to identify the demographic groups most at risk for mental health problems as well as the risk and protective factors for children's mental health.

II. LITERATURE REVIEW
[Toombs, Elaine et-al] has studied that evidence from Ontario and around the world has linked increasing screen time for children and teenagers with instructions to stay at home, closing of schools and recreation centers due of the epidemic. Any time spent using an electronic device, whether actively (such as online learning or video conferencing) or passively (such as watching television), is referred to as screen time. Public health initiatives related to the pandemic have created difficulties that have probably made it worse for kids and teenagers to use screens. These specifically include switching to a distance-based, online education model, disrupting childcare, and mandating quarantine or physical isolation practices. Family leisure activities, such as athletic activities, play dates, birthday celebrations, summer programmes, and other occasions unrelated to screens, have been curtailed or removed as a result of stay-at-home mandates. When gatherings or activities were shifted to a virtual setting, social distance orders occasionally resulted in more screen time. Additionally, the nature of these directives set off a chain reaction that unintentionally boosted screen time for families. Limiting in-person events has reduced the natural family routines involved with activity preparation and travel time to event places

ISSN 2394 - 7780

[Marciano, Laura et-al] has discussed that by using survey data to carry out an experiment in nature gathered. In the present study, we investigated the long-term effects of the COVID-19 outbreak in Switzerland on teenagers' mental health both ahead of and following the lockdown. We specifically calculated the changes in teenage mental health between Spring 2019 and Autumn 2020, as well as how adolescent mental health was predicted by screen time and activities involving screens over the course of this 1.5-year period. Four key findings emerged from the current investigation. First and foremost, it is not surprising that teenagers' mental health deteriorated over time. Internalizing issues including anxiety and depression as well as inattention, which had medium to large impact sizes, increased in prevalence. From the time before the pandemic to the time after the lockdown, there was a marked decline in the state of adolescents' mental health in Switzerland. The use of social media, messaging apps, smartphones, and the Internet all increased at the same time, as did Internet usage. Even after accounting for additional important indicators, such as mental wellbeing prior to the outbreak of COVID-19, The increase in social media use was remained strongly associated with worse mental health in the post-lockdown period, even after accounting for at least one extended life event. Along with the epidemic, home schooling, and watching screen media for at least one time in life. These results illuminate the effects of the COVID-19 outbreak early stages on teenagers, who seem to be particularly vulnerable and in a fundamental development stage, necessitating greater focus from researchers and policymakers.

Initially, using a variety of conventional machine learning techniques to evaluate their performance while attempting to identify mental depression Second, a software integration pipeline has been interpreted by researchers so they may comprehend how to install and implement machine learning models at the production level. Thirdly, a comprehensive pipeline that used a number of machine learning algorithms and used model evaluation indicators to identify the robust model has been interpreted. The research community will benefit from the explanation of hypothesis testing techniques, including loss optimization metrics, as they work to develop a clinical model for predicting mental stress or depression.

Besides, experts have examined the data and used a variety of machine learning approaches to find insights regarding mental depression. According to our analysis, the four key characteristics of demographics, medical conditions, employment, and alcohol use account for the majority of depression cases. The research community will utilize our experimental results as a standard as it will be simple to identify the components that contribute to the onset of depression. While working on forecasting mental depression, an optimum model that should be taken into account may be determined by paying attention to the classification reports.

III. METHODOLOGIES

Many methodologies used in this study can be divided into four key stages. The data management mechanism will be listed along with the hypothesis and data modelling in the early step. The experimental setting for the predictive model will be built using an experimental setup and model analysis in the second phase. Different conventional classifiers will be used in the architecture part. Model testing will be implemented in the third step when relevant factors have been identified and evaluated. Following that, a technique for comparing the models will be interpreted, along with how the entire software integration is connected. The planned architecture will demonstrate the outcome of predictions on mental pressure and depression in the final stages. The link between common mental illnesses like depression and physical health has been shown. Psychiatrists and psychologists can now use artificial intelligence (AI) tools to help them make decisions based on their patients' prior data, such as their medical records, data, social media usage, etc. Deep learning (DL), one of the most subsequent years of AI technology, has showed impressive performance in a range of real world applications, from computer vision to healthcare. As more data about a person's mental health status become available, researchers have used AI technology and deep learning technologies that are being utilized to enhance our understanding of psychological conditions and have been involved to assist therapists for improved clinical decision-making. Making machine learning algorithms or data sets that can automatically find hidden patterns in data is the aim of machine learning (ML). The latest AI and machine learning innovations, deep learning (DL), aims to build a top mechanism that converts the input's raw features directly into the outputs via a tri network structure that can recognize the data's hidden patterns. Various model designs are employed, including the dense feedforward neural network (DFNN), recursive neural network (RNN), convolutional neural (CNN), and auto encode.

IV. DISCUSSION

Authors have investigated the effects of increased screen usage during the epidemic on mental health. The researchers investigated that how did pandemic influenced people to engage more time in technologies and media. Various methodologies were used by the experts to measure and evaluate the mental health affected due to increased screen time. A method of machine learning and artificial intelligence was used by the experts to understand the patterns of mental health of a being when they were exposed to increased time spent on screens

ISSN 2394 - 7780

during COVID-19. The researchers also studied about the screen time affecting the minds of different age groups of people like; children, adults, older people and also those people who work and need to be socially active.

V. CONCLUSION

Dependence on digital gadgets, which leads to an increase in daily screen time, has a number of negative repercussions on one's physical, mental, or psychological health. Constant use of gadgets, including laptops, smartphones, and televisions, can be harmful to mental health. For instance, it can increase stress and anxiety and lead to a variety of sleep problems in both children and adults. In terms of their general wellbeing and mental health, post-secondary students across North America have also been impacted. Studies show that COVID-19 increases emotions of isolation, lonely, and fatigue as well as mental health problems such depression, anxiety, and PTSD. The pandemic has also led to an increase in drinking and drug use, erratic quality of sleep, and screen time. According to WHO's screen time guidelines there is no screen time for child less than 1 year, maximum half hour for 2 years and no more than one hour for 3-4 year children.

REFERENCES

[1] Toombs, Elaine & Mah, Linda & Short, Kathy & Young, Nancy & Cheng, Chiachen & Zhu, Lynn & Strudwick, Gillian & Korczak, Daphne & Perkhun, Anna & Born, Karen. (2022). Increased-Screen-Time-for-Children-and-Youth-During-the-COVID-19-Pandemic published 20220412. 10.47326/ocsat.2022.03.59.1.0.

[2] Marciano, Laura & Viswanath, Kasisomayajula & Morese, Rosalba & Camerini, Anne-Linda. (2022). Screen time and adolescents' mental health before and after the COVID-19 lockdown in Switzerland: A natural experiment. Frontiers in Psychiatry.13.981881.10.3389/fpsyt.2022.981881.

[3] E. Hossain, A. Alazeb, N. Al Mudawi, S. Almakdi, M. Alshehri et al., "Forecasting mental stress using machine learning algorithms," Computers, Materials & Continua, vol. 72, no.3, pp. 4945–4966, 2022.

[4] Su, Chang & Xu, Zhenxing & Pathak, Jyotishman & Wang, Fei. (2020). Deep learning in mental health outcome research: a scoping review. Translational Psychiatry. 10. 10.1038/s41398-020-0780-3.

[5] Sillcox, Carly. (2022). Implication of COVID-19 on Post-Secondary Students' Mental Health: A Review. McGill Journal of Medicine. 20. 10.26443/mjm.v20i2.922.

[6] Ng, Catalina SM & Ng, Sally. (2022). Impact of the COVID-19 pandemic on children's mental health: A systematic review. Frontiers in Psychiatry. 13. 975936. 10.3389/fpsyt.2022.975936.

[7] Agarwal, Richa & Tripathi, Alka & Khan, Imran. (2022). Effect of increased screen time on eyes during COVID‑19 pandemic. Journal of Family Medicine and Primary Care. Volume 11. 3642-3647. 10.4103/jfmpc.jfmpc_2219_21.

[8] https://www.healio.com/news/psychiatry/20211230/more-screen-time-during-covid19-pandemic-has-negative-effects-on-pediatric-mental-health

[9] Giorgi, Gabriele & Lecca, Luigi & Alessio, Federico & Finstad, Georgia & Bondanini, Giorgia & Lulli, Lucrezia Ginevra & Arcangeli, Giulio & Mucci, Nicola. (2020). COVID-19-Related Mental Health Effects in the Workplace: A Narrative Review. International Journal of Environmental Research and Public Health. 17. 1-22. 10.3390/ijerph17217857.

[10] Menon, Sheila & Bhagat, Vidya. (2021). A Review Study on the impact of COVID-19 on Mental Health in the workplace and on working people. Research Journal of Pharmacy and Technology. 6725-6731. 10.52711/0974-360X.2021.01162.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

CYBER SECURITY: THREATS IN CLOUD COMPUTING

Dr. Neelam Naik and Tushar Varma
Department Of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce Mumbai, Maharashtra, India

ABSTRACT
Cloud computing was born out of legacy systems. Therefore, threats that are applicable to legacy systems are equally applicable to cloud computing. Additionally, new cloud- specific threats are emerging for a variety of reasons, including multi-tenancy, ubiquitous access, and cloud control. Security can be important in cloud environments. It also highlights the various methods that attackers use to inflict damage. To help us reach our goal, we reviewed, the best publications on cybersecurity. Cyberattacks are industry-specific and have proven to vary significantly from industry to industry. Cyberattacks are highlighted by being classified into phishing attacks and distributed denial of service attacks. This work will be of great help to industry and researchers in understanding his different cloud-specific cyberattacks and developing strategies to counter them more effectively. Authors have done a case study of handling cyberattacks in cloud-based computing environment.

Keywords: Security, Threats, Mitigations, Basic Fundamental of cloud.

I. INTRODUCTION
Big data management and analysis are aided by the scalable, on-demand Cloud solutions. It is now extensively utilised owing to its potential to save costs by pooling resources, geographic accessibility, round-the-clock accessibility, and the protection of data loss provided by the existence of multiple copies. The security and privacy concerns, which are distinct from and more numerous than traditional storage methods, are the most pervasive drawbacks of cloud systems. The goal of this study is to examine earlier research on vulnerabilities to cloud systems' cyber security. Additionally, it will locate and assess the main dangers to the environment of cloud systems. Security in cloud systems is a significant problem since it requires a combination of controls, technologies, and policies to safeguard the infrastructure, services, and data. As a result, this combination increases the vulnerabilities. Data on the cloud are outsourced to reputable or illegitimate service providers, jeopardising customer privacy.

II. EASE OF USE

A. Cloud Services in General
Both businesses and individual users choose to outsource their services online rather than keeping their own resources. When technical resources are outsourced, a firm may concentrate on business needs rather than technical concerns, which are handled by information technology (IT) specialists. A web-based paradigm known as cloud computing has arisen and is providing the services on a utility basis to help these consumers. The main objectives of cloud computing are to lower operational costs, boost throughput, and improve availability and dependability. Three different types of cloud services are available to meet the needs of a wide range of consumers. IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service), and SaaS (Container as a Service) are some of these services (CaaS). Haas (Hardware as a Service) and CaaS (Community/Containers as a Service) are other names for IaaS. IaaS provides consumers with computer resources like processors, RAM, storage, etc. as services. Instead of using private services, all of these resources are available via the cloud on a rental basis. The IaaS has the enticing characteristic that customers do not need to alter the infrastructure frequently because it only has to be updated every three years owing to Murphy's Law. Additionally, users are not obligated to update their operating systems or apply the fresh patches that are usually required to close exploited holes. In PaaS, a development environment is made available as a service; in CaaS, a cloud service that enables IT departments and software developers to collaborate Haas (Hardware as a Service) and CaaS (Community/Containers as a Service) are other names for IaaS. IaaS provides consumers with computer resources like processors, RAM, storage, etc. as services. Instead of using private services, all of these resources are available via the cloud on a rental basis. The IaaS has the enticing characteristic that customers do not need to alter the infrastructure frequently because it only has to be updated every three years owing to Murphy's Law. Additionally, users are not obligated to update their operating systems or apply the fresh patches that are usually required to close exploited holes. In PaaS, a development environment is made available as a service; in CaaS, a cloud service that enables IT departments and software developers to collaborate. Private model (also known as on-premises model) refers to cloud resources that are controlled by cloud users, whereas public cloud refers to cloud resources that are controlled by cloud providers (also known as hosted model).

ISSN 2394 - 7780

B. Threats in Cloud Computing

Due to the volume of data exchanged between businesses and cloud service providers, it is possible for private information to be intentionally or unintentionally disclosed to unreliable parties. Malware, insider risks, human mistake, and weak

The majority of data breaches in cloud services are caused by compromised credentials and criminal behaviour. Malicious actors, especially state-sponsored hackers, try to exfiltrate data from target companies' networks in order to make money or for other malicious objectives through taking advantage of flaws in cloud services. In general, corporations find it challenging to prevent illegal access due to the ease with which an employee and her IT systems may access cloud services. The rise of cloud computing and the demise of on- premises data centres, however, have not been held down by the security risks offered by cloud services. In order to lower the danger of illegal data transfers, service outages, and reputational harm, enterprises of all sizes must reconsider their network security policies. Organizations using cloud services are subject to fresh security risks posed by public APIs and authentication. Hackers with experience target cloud systems and obtain access using their skills. Hackers frequently employ built-in features in cloud services to sustain a long- term presence within the network of a target organisation through social engineering, account takeover, lateral movement, and evasion techniques. Their objective is to send private data to the systems they control. There are three steps to this process. We looked at task utilisation trends in the first stage. The use patterns noted in the traces shared several characteristics with one another This has been supported by earlier research. We used clustering algorithms to arrange jobs based on average resource utilisation because of these commonalities. The clustering output is used in the second phase to pinpoint certain categories of virtual machines for the traces we looked into. The third step is the assignment of a corresponding virtual machine type to each task cluster. VM sizing methods should be compared to a fixed VM size baseline scenario. According to experimental findings, the suggested method reduces the number of servers, which lowers the energy use of the data centre

III. CLOUD SERVICES Real-Time QoS and Clouds

It is hard to think of "the cloud" as a single set of business models with a single set of Quality of Service problems since cloud computing is a generic paradigm. Cloud computing problems are inevitably tied to application features and goals to some extent. However, several cloud types, typical stakeholders, and their issues may be identified. There are now three widely accepted classes in the cloud services stack:

Infrastructure as a Service (IaaS) is the supply of "raw" computers (servers, storage, networking, and other devices) for service users to install their own software, typically as virtual machine images.

Platform as a Service (PaaS): the offering of an environment and development platform that offers services and storage, on the cloud hosted. Software as a service (SaaS) is the delivery of an Internet-based or distributed application as a service. An IaaS serves as the foundation for a PaaS, which employs multi-tenant deployment and development tools. a controlled environment where numerous apps can share resources and user information. Stakeholders in PaaS include: In order to fulfil the expectations of its customers' needs, the PaaS hoster: must supply sufficient resources (usually via an IaaS model), as well as acceptable availability contingencies.

Without extensive domain knowledge of back-end server and front-end client development or website maintenance, the PaaS provider will offer an environment suited for regular developers to construct web apps.

A browser-based development environment, easy deployment to a hosted runtime environment, administration and monitoring tools, and pay as you go pricing are requirements for the PaaS user (developer).

Today, there are a lot of PaaS providers available, including Google App Engine, Microsoft Azure, Salesforce.com Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you. Force.com, Rackspace Sites, Bungee Connect, Engine Yard, Heroku, Intuit, Cloudera, Aptana, Virtual-Global, Long-Jump, AppJet, Wavemaker, and Aprenda. According to our knowledge, none of these PaaS providers provide standardised tools and methods to assist application providers in the administration of QoS assurances for real-time interactive apps hosted by IaaS providers. Of course, one issue is that Service Level Agreements (SLA) from IaaS providers do not provide on-demand QoS flexibility; for example, Amazon EC2 [1].

Platform-as-a-service (PaaS), infrastructure-as-a-service, and software as a service (SaaS) are the other two main elements of cloud computing (IaaS). SaaS is built upon PaaS, which is built off IaaS. SaaS has its own particular development practises and computing infrastructure in addition to its own business strategy. At the

ISSN 2394 - 7780

system level, SaaS is often delivered on a PaaS system like GAE1), EC22), or Azure3), or specific SaaS infrastructure, as opposed to conventional software, which runs on operating systems.

A. PaaS

The majority of PaaS systems are hosted, web-based application development platforms that offer complete or, insome cases, partial environments for creating complete applications online. They take care of everything from managing projects to changing code and troubleshooting. The majority of decisions that affect how the application infrastructure functions in PaaS are made by the system's provider. These decisions include the kind of OS utilised, the APIs, the programming language, and the management tools. The provider's on-demand tools and collaborative development environment are used by users to create their apps. According to David Linthicum, CEO of the software consultant Linthicum Group, "PaaS often delivers a comprehensive set of tools and technologies, from the interface design to process logic, to persistence, to integration." Bungee Labs, Coghead, Etelos, Google, LongJump, Rollbase, and Salesforce.com are a few providers of PaaS. Industry watchers anticipate the entry of additional significant corporations. Before becoming extensively used, PaaS must still overcome a number of important issues. PaaS gives businesses an alternative method for creating applications for sale or internal usage. [2] A virtual platform is created by PaaS for the deployment and development of applications. Through communication with the servers of the PaaS providers, often through a browser, the user creates apps. A developer can express application logic and business process flow using the systems' programming model without having to make use of the underlying physical computer systems and network interfaces. PaaS providers typically attempt to familiarise developers with programming by supporting languages like C, Java, and PHP or by using drag-and-drop business-logic tools that construct code blocks and so minimise development time. The amount of work required

However, some service providers, like Bungee, have created their own languages. According to them, learning the new languages is comparatively simple and simplifies the growth process. [3] Benefits. PaaS advocates assert that by hosting the whole development environment, their technology boosts productivity, enables businesses to produce products more quickly, and lowers software costs. They claim that the method does away with the requirement that application developers set up their own servers for application development, scale their deployment environments, and install and integrate management tools. Additionally, they don't have to deal with OS and server patches, security issues,or storage subsystems. Additionally, they don't have to deal with network interfaces in order to connect Web services or application components. PaaS providers deal with these problems instead. According to Bungee's vice president of marketing, Lyle Ball, developers may just concentrate on producing software. [5] PaaS encourages developer cooperation since it makes it simple to access, edit, and return code because it is handled online. PaaS systems employ a range of methodologies, from the very straightforward—in which users essentially just put together preexisting code blocks—to the complicated. Using the Etelos Development Environment and languages including PHP, Java-Server Pages, C#, and the business's proprietary English Application Scripting Engine, developers may construct apps on the Etelos platform. Salesforce.com Users may connect over 60,000 apps on the PaaS platform to create larger applications. [7] Additionally, they have the option of creating own apps. The platform offers various APIs, a database for usage with apps, and logical ability, according to Gross. Thus, He said, programmers may concentrate on other areas of application development, such the user interface. use the specialised Apex programming language, which concentrates on database modelling. [9] Users may back up data off the platform, but programmes must operate on it. using Google App Engine 30,000 developers are presently testing out the Google App Engine. Scalable environments for development and deployment are provided by the platform. [13] Google's Big-Table database, storage, and the same technologies for access control, security, and Web-services interaction that are used by the company's own applications are also available to developers. The widely used Python scripting language is used by the platform. It also has an authentication mechanism to verify the legitimacy of the participation in the development process of programmers. Organizations can also utilize the system to authenticate the users of the applications they subsequently build. PaaS creates an online virtual environment for application development.

B. IaaS

IaaS containers provide a number of advantages over virtual machines (VMs), including better performance and quicker starting times. The Docker container image must still be pre-configured on the device node. We modified the nova-docker driver, which leverages the hypervisor to retrieve all Docker images on each Docker node, to address this problem and drastically shorten start-up time.

[8] Since both images are already on the device node for this device, there is no delay in deploying the Docker container. True IaaS cloud data centres (CDCs) are rife with heterogeneity, not just between physical servers but also across workloads. Successful modelling is especially challenging due of the variability of IaaS CDC. This paper examined the circumstances in which each Client jobs may use a variety of virtual CPUs. A hierarchical

ISSN 2394 - 7780

probabilistic simulation technique is suggested for performance analysis in IaaS CDC in light of this variability. Discrete event simulations are used to verify the numerical findings from the analytical model that has been proposed for a number of device characteristics. They concentrated especially on his IaaS cloud computing idea when making their contributions. Based on all available operating system features, the host system assigns an IP address to each private and public user. Using Open stack, a strong cloud development platform, we created an IaaS environment. Some devices, including laptops, tablets, and smartphones, could appear to access your preferred operating system (OS) and show the user virtual hardware.

[4] Users of this gadget may observe several operating Client jobs may use a variety of virtual CPUs. In response to this variability, systems using a hierarchical probabilistic simulation technique can operate everywhere without the need for installation or setup. The IaaS model also reduced the cost and time of purchasing and operating hardware. Using OpenStack tools that can track each instance of the client, the findings demonstrate that resources were shared by many users. Using Markov Continuous Time Chain, we presented a monolithic paradigm for IaaS Cloud Data Center (CDC) (CTMC). It has the following characteristics: (1) Active Devices (PMs) and Standby Devices (PMs); (2) PMs assist in switching between Standby PM pools and Active PM Pools; and (3) All work is completed accurately. (4) The job is operating on an idle state if the PM fails on that job. That will go on after PM. The monolithic CTMC model for IaaS cloud performance evaluation demonstrates rigidity and scaling, yet it may be used to test inferred scalable models. The suggested model's state transformation rules are explained in detail, and computation-based metrics such as average reaction time and immediate action probability are included. [6] In order to confirm the correctness of the suggested model, numerical analyses and calculations are made. In this work, a novel virtual machine (VM) placement technique was developed in the semi-programmed backdrop of the anticipated VMs in the IaaS cloud. Before VM placement choices are determined, short-term requests can be made to planned VMs in the semi-scheduled context, and the VM's resource needs are periodically satisfied in real time.

CaaS

Recently, a brand-new category of services known as Containers as a Service (CaaS) was unveiled. The tool Docker, which enables programmers to specify containers for programmes that integrate these two levels, is an illustration of a container management system. CaaS services are often provided on IaaS virtual machines, as indicated in the diagram. Virtual machines and containers, according to CaaS providers like Google and AWS, offer a decent environment for semi-trusted workloads and an additional degree of protection for untrusted applications. You can use dynamic voltage and frequency scaling (DVFS), virtual machine consolidation, or both to lower CaaS power usage. However, these efforts are ineffective unless you size your VMs to more effectively accommodate the deployed containers. wasted. In order to minimise energy waste when running workloads, this white paper addresses the topic of energy efficiency in the context of CaaS. It focuses on determining the most effective sizes for virtual machines that host containers.

[11] Is assuming.

In order to do this, we describe a technique that was created by looking at actual cloud trace logs and accounting for the distribution of cloud workloads, which is crucial for testing and verifying the suggested solution. Only Google's cloud computing services are available to the general public. The first Google log shows the resource utilisation for a series of actions spread out over a period of seven hours. His updated Google Traces, which was launched in 2012, offers more in-depth long-term information. This whitepaper's key contribution is a method for resource allocation that is both effective and efficient. Experimental results show that the proposed approach leads to a reduction in the number of servers, leading to a reduction in data Center energy consumption. (in the form of virtual machines with set CPU and memory entitlements) that closely resembles how deployed containers really use their resources. [12] The goal of this white paper is to quantify the long-term effects of VM size options on data centre power usage. To avoid the impact of prediction errors, we make the assumption that the workload is accurately understood and that there is no method for consolidating virtual machines. There are three steps to this process. We looked at task utilisation trends in the first stage. The use patterns noted in the traces shared several characteristics with one another. Assaults on Hosts: In addition to cloud software called Face Vector, attackers are compelled to download materials that then introduce Trojan horses, malware spoofing, and IP theft attacks.[14]

- **Ransomware Attacks:** It is simple for hackers to target kids by following their online behaviour and demanding money.
- **Email Scams:** By phishing any institute email address or the administration's email, assaults have taken place where rumours of a coronavirus pandemic were propagated, which causes an undetected scenario within the education sector.

ISSN 2394 - 7780

- The effect of business on the promotion of work from home This has been supported by earlier research. We used clustering algorithms to arrange jobs based on average resource utilisation because of these commonalities that closely resembles the real resource utilisation (in the form of virtual machines with specified CPU and memory entitlements). The clustering output is used in the second phase to pinpoint certain categories of virtual machines for the traces we looked into. The third step is the assignment of a corresponding virtual machine type to each task cluster. VM sizing methods should be compared to a fixed VM size baseline scenario.

C. SaaS

Functionality, architecture, usability, vendor reputation, and cost are selection criteria for SaaS. These variables were mostly chosen with our case study of sales force automation in mind (SFA).

Functionality: A functionality factor encompasses characteristics that are sometimes referred to as SFA's functional modules. It contains: I Tracking consumer contacts through contact and activity management. It makes sure that sales efforts are unique. Through opportunity management, opportunities are tracked and managed. across each level of the sales process. It has features like lead generation, lead-to-opportunity conversion, opportunity tracking, and others. (iii) Sales Performance Management facilitates the assignment of territory and quotas to various levels of sales organisations, ranging from regions and districts to specific sales representatives. Dashboards and reports are offered by the Sales Analysis module (iv). [10]

Architecture: The following list of architecture-related factors: The capacity of the product to integrate with other apps is included in the integration attribute. Given that SaaS solutions are hosted off-premises and hence may be seen as being challenging to interface with traditional on-premise systems, the integration feature becomes highly important for SaaS products.

Scalability is the ability of the SaaS product to keep consumers' response times tolerable even during periods of high traffic. The SaaS product's availability to consumers throughout specified time periods is referred to as reliability. Security is seen to be the main worry for SaaS solutions, hence sellers are required to implement monitoring and diagnostic tools. The processing of client data is secured when a vendor has certifications like ISO 27000. Usability

The following are qualities that relate to usability:

User interface attributes include those that are straightforward, simple to use for regularly performed activities, and aesthetically pleasing. The availability of user guides that are simple to follow, eLearning programmes, and context-sensitive assistance is referred to as the help attribute (ii). (iii) Support for mobile devices has grown in importance as the modern sales workforce increasingly relies on portable electronics like PDAs.(iv) Offline assistance is crucial. It means that the SaaS solutions have a technique that enables users to operate on the system when disconnected from the internet and to synchronise once they are back online. vendor standing Two qualities are included in the vendor reputation factor: I The quantity of customers or users reveals the extent of use, which basically identifies whether the product is a recent entrance or an established one. (ii) The vendor's brand value is also significant because, in some cases, consumers may choose a new product from a well-known vendor over one that is offered by a less well-known vendor despite having a large client base. Cost The cost element consists of two components: an annual membership fee and an upfront setup cost. Typically, annual subscriptions cover the cost of the hardware and the support staff, while monthly subscriptions cover the cost of the first consulting, configuration work, etc.

IV. THREATS IN CLOUD COMPUTING Potential Security Risk

Expansion of Security Threats in Remote Learning: Following is a discussion of some of the significant security breaches' root causes.

- **Information Breach:** Any education industry's database is a potential treasure. The university database can be exploited, which can result in serious data breaches. Educational institutions are particularly alluring to cybercriminals looking for high-value targets.
- **Non-Skilled Use:** The education sector is facing a serious challenge as a result of the rapid adoption of cloud computing since most people are unaware of cloud security and lack sufficient training.
- **Psychological Impacts on Cybersecurity:** Colleges and institutions work with students between the ages of 18 and

22. They have a propensity psychologically to flout any law. They can interrupt the system and crack the password by studying ethical hacking and working with intelligence. One of the main assaults in this scenario is

ISSN 2394 - 7780

email phishing. By making a duplicate ID with a name that sounds close to the authority they threaten, and by clicking unintentionally, people become victims of this kind of assault.

- **Attacks Caused by the Dark Web:** Teenagers are naturallycurious and want to experiment with new things. The darkweb and cyber security are both gaining popularity. Manypeople share login information with their officials, whichunintentionally draws them into the realm of cybercrime.
- **Attacks on video tutorial software:** In this case, video tutor or video conferencing apps like Zoom, Cisco WebEx, Google Meet, Blackboard, etc. are required to keep the learning faculties active.
- **Phishing Scams:** By building bogus websites using the names of cloud vendors, resources in the education sector may unintentionally be made public. It is challenging to discover any phishing schemes without knowing the security procedure.
- Even while cloud adoption has already cornered the IT sector, the current health crisis gives cloud services a vitalpush. Enterprises that collaborate with remote workers must move freely on the cloud.
- **Insufficient Authorization And Authentication:** The usage of company resources from a remote workstation is not protected by an adequate industry-specific authentication technique (such as biometrics), which is a common entrypoint for attackers.
- **Connecting Through Home Network:** When connecting to company resources stored on cloud servers, employees utilise their mobile networks, which lack privacy protocols, or untrusted local network services from local service providers. Exposure of Resources: Regardless of classifications, subordinate personnel may currently haveaccess to corporate resources, which poses dangers owingto their unskilled nature.
- Unintentional clicks expose the workforce to security lapses. Lack of Budget for Upgradation: In this case, because of a lack of income and economic development, the budget set aside for this fiscal year was reduced. Radical changes to security regulations are required to accommodate remote working.
- **Social Engineering Attacks:** In contemporary health crisis, the most frequent cyber-attack is a social engineering assault. The weakest place that the attacker may use to deceive us into disclosing sensitive information is the keyboard or visual screen in a cyberattack.
- **Phishing Scams:** In recent days, sharing private work information via email or any other social media platform,such as WhatsApp or Messenger, has resulted in serious phishing attempts. It's a typical practice for scammers to send emails purporting to be someone else.
- An official email; this kind of assault frequently targets employees. Attacks known as "shoulder surfing": When working remotely, it is difficult to guarantee authenticity by looking at the MAC addresses of the workstations orusing any other authentication method from a third party.
- **Spear Phishing Attacks:** Scammers may easily gather interest and preferences from online chat rooms or othersocial media platforms.
- **Scareware Attacks:** All staff must receive the appropriate training and skills before moving to the cloud system. Organizations now operate their businesses without the necessary knowledge or skills.
- By modifying the URL Scammers can obtain the credentials of the intended victim by changing a URL address. The targeted URL will logically come up aheadof the original webpage.
- By clicking on this fraudulent link, authentication information was readily made public. Interest has the advantage of enticing individuals by allowing them to spend time indulging in a passion or activity. The majority of individuals are using social media to follow the newest trend. People use picture, audio, and video editing software by using cloud services. Online video browsing to use a new game platform requires login information, which might be an assault door. Attacks to Crack Passwords: The MAC addresses of the remote workstations are identified using keyboard or mobile device authentication.
- **Cloud Data Center Unable to Be Controlled:** Currently, a cloud service provider (CSP) manages each clouddatacenter, which is another factor to take into account for security breaches.
- **Video Conferencing Attacks:** Businesses deployed various video conferencing programmes to assist distant work, which caused a significant increase in the use of Zoom, Slack, Google Meet, and Cisco WebEx.

ISSN 2394 - 7780

In this remote login viewpoint, it is important to keep in mind some of the most prevalent assaults of cloud security breaches, including Cross-Site Scripting, SQL Injection,Man-in-the-Middle Attack, and Reply Attacks. [15]

Healthcare, banking, and e-commerce perspectives on this epidemic

1. **Electronic banking and money transfers:** On occasion, we might be unable to determine whether any workstations have been faked. When someone transfers money while accessing a financial portal ora website that has been SSH protected, the scenario becomes risky.
2. **Improper Use of Banking Websites:** In these few instances of a health outbreak, most individuals turn to online banking.
3. A separate set of users accessing the e-portal for emergencies can be identified, and they lack considerable security protocol understanding. When users neglect to log out, inefficient usage of the banking portal continues to be active. Recently, this circumstance has become prey for the assailants.
4. **Attack using Coronavirus Safety Apps:** Using contacts to spread malware via Coronavirus Safety Apps is another affected area in this By pandemic. For this spoofing, thousands of coronavirus websites have been developed. In addition, apps running in thecurrent environment, where several recent assaults happened, do not have absolute trust.
5. **Increases in Online Purchase of Vital Goods:** Local online e-commerce portals are profiting from consumers' desire to buy groceries and other essential goods online. By obtaining a customer's cellphone number or email address, hackers can access both their financial information and login credentials.
6. **Using the Dark Web:** In recent days, everyone has been checking daily updates from Facebook, WhatsApp, and other sources due to global concern. Unintentional clicks and a need to stay current with the news propel people to the dark web. Most of the time, clicking on these unauthenticated URLs without meaning to expose individuals. There is no centralised search engine that Google guarantees.
7. **Intruder Attacks:** People download demanding applications (games, movies, personal interest apps) from anonymous URLs on their phones or workstations in order to increase interest by utilising the same authentication information.

V. SOME PREVENTIVE MEASURES FOR CLOUDSERVICES

Cloud services experience serious cyber security data breaches due to the coronavirus epidemic, regardless of the protection measures in place. Some preventative measures can halt the cyber threat's spread.

- A strong password policy and a multi-authentication policy should be implemented to ensure the authentication of workstations or hosts. When logging into a company's cloud services, using external USB devices should be avoided.
- Businesses should provide training programmes regarding preventing security breaches and raising people's awareness of the issue.
- Before clicking, always double-check the email address and URL to prevent social engineering attacks. Unintentional clicks are absolutely forbidden to preventphishing frauds.
- The shared file system (such as Dropbox, Google Drive, etc.) should be used as a channel for internal communication.
- It is strictly forbidden to share files using a free email account or any social media group.
- When using video conferencing, always permit users to connect after the administrator. Avoid screen sharing.
- Before sharing any files on a cloud platform, make a backup of all of them on a hard drive.
- Before downloading any random application, be sure you know where it came from.
- After utilising cloud services, always follow a sign-out routine.
- Update all of your programmes and software often withthe newest security fixes.
- The DMARC, SPF, and DKIM protocols help protect email domains from email spoofing attacks.
- Add anti-phishing scam software on workstations.

ISSN 2394 - 7780

- Before clicking on any link, make sure the spelling and grammar are correct.
- Consider using salutations like "Dear Sir/Madam." Thisunpredictability typically indicates a phishing scam.
- Don't respond to emails that demand action soon or that contain money from prizes or notifications of significant losses (e.g: Lost of ATM card). • Refrain from clicking on or downloading from unidentified websites.
- Be careful while making purchases online. Always checkbefore making a donation to a nonprofit.
- Take note of common salutations like "Dear Sir/Madam." This unpredictability frequently indicates a phishing attack.
- Stay away from emails that demand action instantly andthose that mention receiving a reward or suffering a significant loss (e.g: Lost of ATM card).
- Steer clear of clicking on and downloading from unsecured websites. When paying online, use caution. Before making a donation to a cause, always check.

VI. CASE STUDY OF VARIOUS CLOUD SERVICEPROVIDERS

1. AWS Cloud Platform.

AWS, or Amazon Web Services, is a platform for cloud computing services offered by Amazon that offers users compute, storage, and a variety of carrier options. These software as a service (SaaS), infrastructure as a service (IaaS), and platform as a service (PaaS) options all work together as extendable means to let businesses effectively deploy software. The AWS Computing, The Amazon Elastic Compute Cloud, or E2C, is Amazon's top-rated computing service provider. The carrier is highly versatile and may be significantly less expensive than other options thanks to its interoperability with the great majority of different Amazon web services. To ensure that you don't overpay for your computing needs, E2C can follow your cloud applications and scale your consumption to meet your modern needs with the help of the auto scaling monitor. Amazon Elastic Field Carrier, or ECS, is the name of the container provider used by AWS. To control IP address blocking and gain access to some logs, templates, IAM roles, and security organisations, you may configure such Docker containers as necessary or utilisepreloaded configurations.

AWS offers packaged Kubernetes products in addition to computing capabilities including AWS Beanstalk, AWS Serverless software programme Repository, AWS Batch, AWS Lambda, Elastic Load Balancing, and Amazon Lightsail (EKS).

2. Azure.

Azure is an integrated platform, similar to AWS, that provides computing, storage, development, and database capabilities as SaaS, IaaS, and PaaS. Businesses may set up and manage packages and other offers through the platform using the cloud.

Azure has a network of virtual machines built on an open-source cloud computing platform for computing. As a result, Windows and Linux servers as well as third-party solution providers like SAP and Oracle are compatible with the whole spectrum of computing services, including application deployment and development, pre-deployment testing, scaling, etc. There are hybrid Azure alternatives that combineon-premises private computing with cloud-hosted platforms if your needs differ. The popular Kubernetes-based container service Azure Container Service is known as Azure Kubernetes Service. Microsoft claims that this enables businesses that frequently require various teams to interact online. Additional compute-related technologies and services with built-in Java Spring Cloud compatibility include Azure IaaS, Azure PaaS, Azure Batch, Service Fabric, Azure Functions, and Azure Spring Cloud.

3. Google Cloud Platform.

Despite being officially introduced in 2008, Google has just recently begun to really challenge AWS and Azure. GCP offers IaaS and PaaS, similar to the other two systems, as well as a serverless platform that offers computation, storage, databases, a few networking options, database administration, and IoT. Computing using Google Cloud Platform Although it can perform worse than other systems, especially AWS, which was introduced a few years ago, Google's Compute Engine (GCE) has a number of advantages that can help it catch up to its rivals. For instance, cloud computing skills may deploy code for various Google services, operate and install cloud apps, and expand largely according on the volume of traffic. You also only pay for code distribution that complies with Google policy. Similar to Azure, GCE offers extensive support for Kubernetes-

ISSN 2394 - 7780

based services, including Kubernetes or GKE (like EKS and AKS) and Knative. GPU, Google App Engine, and instant organisations are some of the other GCE capabilities.

- What measures AWS, Azure and GCP have taken to provide cyber attack security for their clients

A. AWS Security.

Because Amazon is the most experienced and established cloud provider, it has historically covered a considerably wider range of issues and solutions. AWS provides API interest monitoring, vulnerability testing with AWS Inspector, risk analysis using protect duty, and information loss prevention in addition to isolation through security organisations (firewalls) and very fine-grained IAM. The easiest problem appears to be related to how long AWS has been used. Before large-scale cloud integration platforms attracted the attention of businesses, a few of the aforementioned safety capabilities were implemented one after the other. AWS is protected. AWS refers to its own solution as a "secrets and methods manager." It also provides a system for certificate storage, however this is mostly used to store secrets and procedures. With a site-to-site connection limit of 10 connections per VPN gateway, AWS VPN enables factor-to-factor and site-to-site options.

B. Azure Security.

While Azure's energy is in a centralised protection device managed from a single listing, AWS' energy is isolated (you need to install different security protocols for each account). However, the possibility for insider attacks is increased by the lack of isolation and manipulation that displays the full enterprise's console and API activities. Businesses also criticise Azure for using uneven compliance documents and less secure setups by default. Key Vault is a service offered by Azure that is used to store confidential information like passwords and encryption keys and also enables certificate storage. With a cap of 30 site-to-site connections per VPN gateway, the Azure VPN Gateway offers point-to-site VPN. Azure offers DDOS protection, clearly. There is a security centre for Azure.

C. Google Cloud Platform Security.

In general, many industry experts view GCP as an effective and simple link between AWS and Azure. While separating projects and using a cosier default setup, Google has taken care to maintain centralised comfortable access. However, compared to AWS, GCP's security features and professional pool are less startling, while the Cloud security Command Center is still capable. .GCP has Google Cloud Armor. GCP Secrets Manager offers the ability to store passwords and certificates and functions similarly to other platforms. Google Cloud VPN does not yet support point-to-point connections; it only supports site-to-site internet VPN connections. A trust and protection centre exists at GCP.

VII. DISSCUSSION

Cloud computing presents a new set of security threats and challenges. By storing data and applications in the cloud, organizations face the risk of data breaches, malicious attacks, and other security threats.

1. **Data Breaches:** Data stored in the cloud is vulnerable to unauthorized access, manipulation, or theft. Hackers can use malicious software to access stored data and applications, or they can use stolen credentials to gain access. Data stored in the cloud is also vulnerable to insider threats, such as employees or contractors who have access to cloud systems.
2. **Malicious Attacks:** Hackers may use malicious code or software to attack cloud systems. They may also use distributed denial of service (DDoS) attacks to disrupt access to cloud services.
3. **Security Flaws:** Cloud providers may not have adequate security measures in place to protect the data and applications stored in the cloud. Insecure cloud configurations and weak passwords can also create security vulnerabilities.
4. **Poor Visibility:** Organizations may not be able to monitor and track all activities in the cloud, making it difficult to detect and prevent security threats.
5. **Compliance Issues:** Organizations may face compliance issues if they do not adhere to industry regulations or data privacy laws.
6. **Malicious Insiders:** Insiders, such as malicious employees or contractors, may have access to data stored in the cloud. These malicious actors can use this access to steal data or cause other damage.
7. **Data Breaches:** Cloud-based systems may be vulnerable to data breaches, which can result in the loss of sensitive data.
8. **Data Loss:** Data stored in the cloud is vulnerable to accidental or intentional data loss, which can have

ISSN 2394 - 7780

serious consequences for an organization.

9. **Insecure Cloud Applications:** Cloud applications may not be secure, and attackers may be able to exploit vulnerabilitiesin them to gain access to data or cause otherdamage.
10. **Denial of Service Attacks:** Cloud-based systems may be vulnerable to denial of service (DoS) attacks, which can prevent users from accessing data or services.
11. **Lack of Visibility:** Without the proper tools, organizations may not be able to monitor and detect security threats in theircloud-based systems.
12. **Unsecured Interfaces and APIs:** Cloud providers mayoffer access to their services via unsecured interfaces orapplication programming interfaces (APIs). If these interfaces and APIs are not secured or properly configured,malicious actors can gain access to data stored in the cloud. Organizations must take steps to protect their data andapplications in the cloud. They should also monitor cloud activities and use cloud security tools todetect and respond to threats.

The various cloud service providers gave us a certain knowledge of how does their security system works on theirplatform and about their secure mechanism and much more.

VIII. CONCLUSION

Cloud computing has become an integral part of businesses, especially in the digital age. But along with the many benefits of cloud computing come the potential for cyber security threats. Cloud computing can increase the attack surface, provide new avenues for malicious actors toexploit, and can leave organizations vulnerable to data breaches, ransomware, and other malicious attacks. Poorlyconfigured security settings can be a major factor in cloudsecurity, as can the use of malicious or vulnerable third- party applications. Organizations must also be aware of thethreat of insider threats and the possible misuse of privileged access. Furthermore, organizations must be aware of the risks posed by data leakage, malware, and other malicious activities that can occur in the cloud. By taking the necessary steps to securely configure and monitor cloud services, organizations can significantlyreduce their exposure to cyber security threats. . Regardingprivacy and legal requirements, security in cloud computing is a crucial concern. The goal of a task force and other organisations is to improve cloud computing security.Working groups provide their suggestions for different countermeasures to serious security concerns as well as their ideas and reports on them. Despite this, several studiesindicate that the hosted approach is safer than the on- premises cloud model. Even Nevertheless, many assaults target hosted models in an effort to find weak points. The two primary techniques used to attack the cloud are DDoSand phishing. Finally, it can be said that phishing and DDoS attacks caused significant financial losses and harm to data privacy since they happened in several vulnerable clouds. A number of Although there are a number of solutions that are there is still a need to strengthen security in both hostedand on-premises cloud to combat various attacks in order torestore user confidence

IX. REFERENCES

[1] Kumari K, Mrunalini M. A Survey on Big Data Security: Issues, Challenges and Techniques. International Journal of System & Software Engineering. 2018;6(2):23-36. Accessed February 13, 2021.

[2] Almaiah MA, Dawahdeh Z, Almomani O, Alsaaidah A, Alkhasawneh A, Khawatreh S. A new hybrid text encryption approach over mobile ad hoc network. International Journal of Electrical and Computer Engineering (IJECE). 2020Dec;10(6):6461-71.

[3] Qadiree, Jahangeer, Neha Prasad, and Pratima Gautam. 2017. “Security and Privacy Approach of Cloud ComputingEnvironment.” International Journal of Advanced Research in Computer Science 8 (7): 648–51. doi:10.26483/ijarcs.v8i7.4355.

[4] Mondal, A., Paul, S., Goswami, R. T., & Nath, S. (2020). Cloud computing security issues & challenges: A Review. In (pp. 1- 5): IEEE.

[5] Adil, M., Khan, R., Almaiah, M. A., Al-Zahrani, M., Zakarya, M., Amjad, M. S., & Ahmed, R. (2020). MAC-AODV based mutual authentication scheme for constraint oriented networks. IEEE Access, 8, 44459-44469.

[6] Almaiah, M. A., & Al-Khasawneh, A. (2020). Investigating the main determinants of mobile cloud computing adoption in university campus. Education and Information Technologies, 25(4), 3087-3107.

[7] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G.M. Voelker, V. Paxson, S. Savage, Spamalytics: an empirical analysis of spam marketing conversion, in: CCS, 2008, pp. 3–14.

ISSN 2394 - 7780

[8] H. Shahriar, M. Zulkernine, Mitigating program security vulnerabilities: Approaches and challenges, ACM Comput. Surv. 44 (3) (2012), Article No. 11.

[9] S. Liu, B. Cheng, Cyberattacks: "Why, what, who and how", in: IT Pro., IEEE Computer Society, May/June 2009.

[10] W. Luo, J. Liu, J. Liu, C. Fan, An analysis of security in social networks, in: Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009, pp. 648– 651.

[11] A. Syed, K. Purushotham and G. Shidaganti, "Cloud Storage Security Risks Practices and Measures: A Review", 2020 IEEE International Conference for Innovation in Technology (INOCON) Innovation in Technology (INOCON) 2020 IEEE International Conference for, pp.14,2020,[online]Available:https://doiorg.sdl.idm.oclc.org/10. 1109/ INOCON50539.2020.9298281.

[12] Z. Balani and H. Varol, Cloud Computing Security Challenges and Threats, IEEE, pp. 1-4, 2020.

[13] F. Ghaffari, H. Gharaee and A. Arabsorkhi, Cloud Security Issues Based on People Process and Technology Model: A Survey, IEEE, pp. 196-202, 2019.

[14] S. Mandal and D. A. Khan, A Study of Security Threats in Cloud: Passive Impact of COVID-19 Pandemic, IEEE, pp. 837-842, 2020.

[15] K. Thomas, D.M. Nicol, The Koobface botnet and the rise of social malware, in: Proceedings of the 5th International Conference on Malicious and Unwanted Software (Malware 2010), 2010, pp. 63–70).

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

A REVIEW ON IMPORTANCE OF CYBERSECURITY IN EDUCATION

Swapnali Lotlikar[1] and Aman Kanojia[2]
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, Maharashtra

ABSTRACT
According to a recent report by the Federal Bureau of Investigation in the United States, India is one of the top five nations in terms of cybercrime victims worldwide (FBI). Although India is in the top five, the United States and the United Kingdom have much more victims of cybercrime globally, according to a report from the FBI's Internet Crime Complaint Centre (IC3). One of the technological infrastructure development sectors with the quickest growth is the internet. Cybersecurity is becoming a prominent and expanding concern for commercial and corporate institutions worldwide. Today's cybercriminals attack educational institutions almost as frequently as they target huge commercial firms and global corporations. Schools and other organizations are struggling. Cybersecurity in education is becoming increasingly important as more schools and universities rely on technology to support teaching and learning. With the rise of online learning, personal data, and sensitive information is at risk. Cybersecurity measures help protect against data breaches, hacking, and other cyber threats, ensuring the safety and privacy of students, teachers, and staff. Additionally, cybersecurity education can also help students develop the skills they need to protect themselves and their future employers from cyberattacks. Overall, the importance of cybersecurity in education cannot be overstated as it is crucial for protecting personal and institutional data and for preparing students for the digital age.

Keywords: Cybersecurity, Information Security, Cyber Education, Threat Intelligence.

INTRODUCTION
The Internet has developed into one of the most essential and practical technology tools in use today. In today's technology age, a large number of people regularly use the Internet for a variety of purposes. You have access to all the information you require. It was the tool that was used the most everywhere [1]. One of the most well-liked and frequently used technologies on the internet is social media. Many people from many cultures and groups in various nations have benefited from poor or non-existent communication, but the unintentional use of social media has come at a pretty hefty cost: an increase. The younger generation is becoming more and more accustomed to social media. These social media encroach on our privacy in numerous ways. You can interact with others [3-7].

All parents and guardians should be concerned about cybercrime targeting children and teenagers. morning. Parents could be unaware that their kids are victims of cybercrime. To lessen the prevalence of cybercrime, this generation must be informed of cybersecurity issues. Social media insults and comments have been used to bully teenagers [8]. Anyone can express their thoughts on social media without any restrictions. Because of this, the majority of cybercrime instances start on well-known sites like Facebook, Twitter, Instagram, and many other media.The majority of cybercrime cases involve harassment, sexual assault, etc. Another negative aspect of the Internet is the use of false identities. Because anyone can make a false identity, the virtual world is less safe. Threats on the internet should be known to parents. Yes, youngsters are capable and skilled phone users. Children are eager to master new skills and explore the virtual world. Everyone is quite interested in the virtual world, often known as the digital world because there is so much to learn about it. Since we don't know the virtual world, there are also pros and cons to be aware of. Internet use by children is increasing rapidly due to recently studied social markets and technological innovations.

Children's entertainment is increasingly moving to internet video, gaming, and music search engines. Songs from cartoons, short films, and animated movies are likely to be viewed by 4 to 5-year-olds on websites like YouTube, Spotify, gaana.com, etc. YouTube videos featuring music, dance, vloggers, YouTube personalities pulling pranks, and much more are frequently viewed by young people above the age of 15. Schools have a crucial role in educating us about online safety in the modern world and relating cybercrime to our everyday lives. [9-15] Cybersecurity education aims to inform people of the dangers of using online communication technologies like social media, chat, online gaming, email, and instant messaging. Numerous research articles exist. [16].

ISSN 2394 - 7780

The Fundamentals of Cyber Security research and education

European Union Cybersecurity Strategy

The European Union's Cybersecurity Strategy is a comprehensive plan that aims to promote freedom and security online for all EU citizens. This strategy includes a range of measures that are both short and long-term and that use a variety of policy tools. It also involves the participation of EU institutions, Member States, and industry.

One of the main objectives of the strategy is to increase national efforts in network and information security (NIS) education and training. This includes introducing NIS training in schools by 2014, providing NIS training and secure software development and personal data protection for computer science students, and offering NIS basic training for government employees. This is to improve the overall performance of the EU.

The strategy also aims to improve the EU's ability to respond to cyber threats by strengthening the EU's cyber incident response capabilities and developing a coordinated approach to cyber crisis management. It also aims to increase the EU's cyber defense capabilities by supporting the development of new technologies and promoting the use of existing ones.

Another objective of the strategy is to enhance the EU's cybercrime-fighting capabilities by improving law enforcement cooperation and strengthening the EU's legal framework for combating cybercrime. The strategy also aims to increase the EU's cyber resilience by promoting the use of secure information and communication technologies and raising awareness of the risks associated with the use of these technologies.

Overall, the EU Cybersecurity Strategy is a comprehensive plan that aims to promote freedom and security online for all EU citizens by improving the EU's ability to respond to cyber threats, increasing national efforts in network and information security education and training, and strengthening the EU's cybercrime-fighting and cyber resilience capabilities. [17].

Digital Agenda for Finland 2011-2020

The Digital Agenda for Finland 2011-2020 is a comprehensive plan that aims to make information resources widely accessible to the general public, so that they can promote innovation and research activities, the development of digital products, services, and markets, the efficiency, impact and transparency of public administration, and encourage participation in decision-making. (MTC, 2010)

One of the main objectives of the Digital Agenda for Finland is to increase the accessibility and use of information resources in the country. This includes making digital services more widely available and improving the functionality and usability of these services. The plan also aims to promote the use of information and communication technology (ICT) in all aspects of society, including education, research, and culture.

Another objective of the Digital Agenda for Finland is to improve the efficiency and transparency of public administration. This includes implementing digital solutions in the public sector to improve the delivery of services to citizens and businesses. It also aims to promote the use of open data to increase transparency and accountability in public administration.

The plan also aims to increase the information society and media literacy of the entire population. This includes providing training and education to citizens and businesses on the use of ICT, and promoting the use of digital tools and services in all aspects of life.

The digital agenda also aims to promote multidisciplinary research on the impact of the information society, and to create a favorable environment for the development of digital products and services.

Overall, the Digital Agenda for Finland 2011-2020 is a comprehensive plan that aims to increase the accessibility and use of information resources, promote the use of ICT in all aspects of society, improve the efficiency and transparency of public administration, increase information society and media literacy, and promote multidisciplinary research on the impact of the information society. [18].

Finland´s Cyber Security Strategy

Finland's Cyber Security Strategy is a comprehensive plan that aims to protect the country's critical infrastructure and citizens from cyber threats. The strategy focuses on three main areas: strengthening the security of critical infrastructure, improving the readiness and response to cyber incidents, and increasing cyber security awareness and education.

One of the key objectives of the strategy is to strengthen the security of critical infrastructure, such as energy, transport, and communication systems. This includes implementing security measures to protect these systems

ISSN 2394 - 7780

from cyber-attacks, as well as increasing the resilience of these systems in case of an attack. The strategy also calls for the creation of a national cyber security center to coordinate the protection of critical infrastructure.

Another objective of the strategy is to improve the readiness and response to cyber incidents. This includes establishing incident response teams to handle cyber incidents and creating a national cyber security exercise program to test the readiness of the country's cyber security capabilities.

The strategy also aims to increase cyber security awareness and education among citizens and businesses. This includes providing training and education on cyber security best practices, as well as promoting the use of cyber security tools and services. The strategy also calls for the creation of a national cyber security awareness campaign to educate the public about the dangers of cyber threats and how to protect themselves.

Overall, Finland's Cyber Security Strategy is a comprehensive plan that aims to protect the country's critical infrastructure and citizens from cyber threats by strengthening the security of critical infrastructure, improving the readiness and response to cyber incidents, and increasing cyber security awareness and education. [19]. (Finland's Cyber Security Strategy, 2013)

The Implementation Programme for Finland's Cyber Security Strategy

The implementation program for Finland's Cyber Security Strategy is a comprehensive plan that outlines the actions and measures that will be taken to achieve the objectives of the strategy. The program is divided into four main areas: strengthening the security of critical infrastructure, improving the readiness and response to cyber incidents, increasing cyber security awareness and education, and promoting international cooperation.

One of the key objectives of the implementation program is to strengthen the security of critical infrastructure, such as energy, transport, and communication systems. This includes implementing security measures to protect these systems from cyber-attacks, as well as increasing the resilience of these systems in case of an attack. The program also calls for the creation of a national cyber security center to coordinate the protection of critical infrastructure.

Another objective of the program is to improve the readiness and response to cyber incidents. This includes establishing incident response teams to handle cyber incidents and creating a national cyber security exercise program to test the readiness of the country's cyber security capabilities.

The program also aims to increase cyber security awareness and education among citizens and businesses. This includes providing training and education on cyber security best practices, as well as promoting the use of cyber security tools and services. The program also calls for the creation of a national cyber security awareness campaign to educate the public about the dangers of cyber threats and how to protect themselves.

Lastly, the program promotes international cooperation to tackle cyber security challenges. This includes working with other countries and international organizations to share information and best practices and to coordinate responses to cyber incidents.

Overall, the implementation program for Finland's Cyber Security Strategy is a comprehensive plan that outlines the actions and measures that will be taken to achieve the objectives of the strategy, by strengthening the security of critical infrastructure, improving the readiness and response to cyber incidents, increasing cyber security awareness and education, and promoting international cooperation. [21].

The 2013 Report of the ICT 2015 Working Group

The 2013 report of the ICT 2015 working group is a document that outlines the key recommendations for the future of the Information and Communication Technologies (ICT) sector in the European Union. The report was produced by the ICT 2015 working group, which was established by the European Commission to provide recommendations on the future of ICT research and innovation in the EU.

One of the key recommendations of the report is to increase investment in ICT research and innovation, with a focus on areas such as big data, cloud computing, and the internet of things. The report also calls for the development of a more cohesive and integrated approach to ICT research and innovation across the EU, with a focus on collaboration between industry, academia, and government.

Another key recommendation of the report is to improve access to and the use of ICT, with a focus on areas such as digital skills, e-inclusion, and e-health. The report also calls for the development of a digital single market, with a focus on areas such as online cross-border trade and the free movement of data.

The report also emphasizes the need for increased international cooperation in the ICT sector, with a focus on areas such as standardization and the development of global ICT standards. The report also calls for the

ISSN 2394 - 7780

development of policies and regulations that promote the use of ICT in areas such as energy, transport, and the environment.

Overall, the 2013 report of the ICT 2015 working group is a document that provides recommendations for the future of the ICT sector in the European Union. It suggests increasing investment in ICT research and innovation, improving access to and the use of ICT, and increasing international cooperation in the sector, in order to ensure that Europe remains competitive in the global ICT market.[22].

INKA Cyber Security Programme

The INKA Cyber Security Programmed is a Finnish national program that aims to enhance the cyber security of Finland. The program was launched in 2018 by the Finnish Ministry of Economic Affairs and Employment and is implemented by the Finnish Transport and Communications Agency Tragicomic.

The main goal of the INKA program is to improve the overall cyber security level of society by promoting the development and use of secure digital services and technologies. The program also aims to increase awareness and knowledge about cyber security among individuals, organizations, and businesses.

The INKA program focuses on several key areas, including the development of secure digital services, the improvement of cyber security in critical infrastructure, and the promotion of cyber security awareness and education. The program also includes measures to support the development of cyber security expertise, such as training and research activities.

The program also includes measures to improve the cyber security of small and medium-sized enterprises (SMEs) and to support the development of a cyber security ecosystem in Finland. The program also aims to increase international cooperation in the field of cyber security.

Overall, the INKA Cyber Security Programmed is a national program that aims to enhance the cyber security of Finland. It focuses on several key areas, including the development of secure digital services, the improvement of cyber security in critical infrastructure, and the promotion of cyber security awareness and education. The program also includes measures to support the development of cyber security expertise, such as training and research activities [23, 24]. (INKA, 2014).

Declarations of Essentials Terms

The technology of the world is growing rapidly and it is not easy to handle the work which is complicated for the old generations tough people need basic information about cyber-attacks which can cause huge losses for people. The following terms can be elaborate on: -

Intelligence- The term Intelligence means the entire process of intellectual exertion, intelligence, counter-intelligence conditioning and secret conduct, and subversive goods. It concludes that in the ultramodern intelligence proposition, the understanding of the notion of intelligence special conditioning of intelligence. literature would be a" systematized exertion or association, at the request and purposeful political forces assess the leading classes of class or state, protect their own interests from opponents and engages. In other conditioning that contributes to the consummation of certain Journal of Information Systems & Operations Management, political pretensions. As we see intelligence or intelligence, under the guidance of operation-operation, attracts, analyses and interprets the information on the other party's intentions (adversary, contender, etc. to prepare the rest of the system to reply, increase Adaptability, save the answer and cover your interests on time and therefore averted an attack or some other type of vicious exertion, conniving the normal functioning and development of the system.

Data Breach: A data breach in education is when sensitive, confidential, or protected information belonging to students, staff, or the institution is accessed or acquired without authorization. This can include personal information, financial data, and academic records. These breaches can happen through various methods such as hacking, phishing, or social engineering and can result in negative consequences for the institution and individuals involved. To prevent data breaches, educational institutions should have strong cybersecurity measures in place, provide employee training, and conduct regular risk assessments to quickly detect and respond to breaches.

Network Security: Network security in the educational sector involves implementing policies, technologies, and procedures to protect computer networks from unauthorized access, use, disclosure, disruption, modification, or destruction. This is crucial for safeguarding sensitive information like student records, financial data, and research data and maintaining the continuity of operations and access to learning resources.

ISSN 2394 - 7780

Firewalls:
A firewall is a device that is used to secure computer networks by controlling incoming and outgoing network traffic. It prevents unauthorized access by monitoring and blocking any suspicious activity.

Virtual Private Networks (VPNs): To secure remote access to the network for students, staff, and faculty, educational institutions use Virtual Private Networks (VPNs). To detect and prevent unauthorized access, intrusion detection and prevention systems (IDPS) are employed. Sensitive information stored on the network or transmitted over it is protected through encryption. Access controls are implemented to regulate access to network resources based on user credentials and permissions. Regular security audits and updates to network security measures are essential to detect vulnerabilities and maintain the overall security of the network.

Two-Factor Authentication: - Two-factor authentication (2FA) is a security measure that requires users to provide two forms of identification to access a system or network. This method adds an extra layer of protection in addition to a traditional password. In the educational sector, 2FA is used to secure access to online resources and systems such as learning management systems, student information systems, and email accounts, to prevent unauthorized access to sensitive information and ensure that only authorized individuals have access. 2FA methods can include a combination of something the user knows, such as a password, and something the user has, such as a security token or a mobile phone. An example is using a password and a one-time code sent to the user's phone. 2FA is considered a best practice in cyber security education and highly effective in preventing unauthorized access to educational resources and networks.

Ransomware: Ransomware is a type of malware that encrypts a victim's files, making them inaccessible. To regain access to the files, the victim must pay a ransom to the attacker who holds the decryption key.

Employee Training: Employee training is the process of equipping employees with the necessary knowledge and skills to perform their tasks efficiently and securely.

Network Segmentation: Network segmentation is the process of breaking up a computer network into smaller parts to limit the impact of a potential security breach. Intrusion detection and prevention systems (IDPS) are security measures that detect and prevent unauthorized access to computer networks.

Remodelling the Education System: Curriculum modernization is the process of updating the curriculum, teaching methods and technology used in education to improve its effectiveness and align it with the needs of the 21st century.

Cybersecurity
Cybersecurity refers to the collection of resources, processes and structure used to protect cyberspace and cyberspace-enabled systems from attacks. With the increasing use of the internet, it's important for people to be aware of potential cyber threats. However, many people, including children and teenagers, may be unaware of these risks and may be affected by them, including through late-night browsing and potential health issues. To address these challenges, it's important to educate people about cybersecurity, especially by incorporating it into children's education. By doing so, India's education system can make progress in this area.

The Need for Cybersecurity Education
Cybersecurity is not just a specialized skill, but a combination of technical and societal expertise. To ensure that individuals are equipped with the knowledge and abilities needed to navigate the digital world safely, it is important to incorporate cybersecurity education in various levels of the education system. This includes basic education to familiarize young people with the basics of cybersecurity and its risks, as well as vocational and higher education that provide more advanced skills and qualifications for specific careers in the field. Universities can focus on the scientific aspects of cybersecurity, while polytechnics can offer practical and hands-on training that meets the demands of the workforce. Additionally, offering a range of degree options such as Bachelor's, Master's, and Doctorate in cybersecurity, as well as polytechnic-specific degrees, can provide opportunities for students to specialize in the field.

Cybersecurity education that spans across various levels of education produces high-level experts for various segments of society, whose skills and knowledge meet the qualifications required for each role. This type of education can include basic degree level education, degree studies, training that prepares for competence-based certification, apprenticeship training, ongoing and supplementary education to enhance and expand a person's professional skills, as well as social and leisure studies that improve general and work-related skills. The efficient growth of public cybersecurity expertise requires identifying the areas of competence and their contents so that each level of education can provide the necessary education. This paper has analyzed the areas of competence in cybersecurity and their contents. Each area of competence encompasses several core

ISSN 2394 - 7780

competencies. Based on the contents of the areas of competence and core competencies, it is possible to determine the necessary courses that aim to achieve the desired mastery.

Challenges of Cyber Security Education

The frequently use social media platforms such as Facebook, Instagram, LinkedIn, YouTube, and Twitter. The abundance of information on these platforms can lead to issues related to privacy and security. It is important for children to learn how to protect themselves and handle potential cyber risks. There are challenges in ensuring that teachers are properly trained and equipped to provide education on critical thinking and safe internet use, as well as guiding students and parents in the use of the internet at home. These challenges include a lack of expertise, funding and resources. Teachers need to have knowledge and expertise in cybersecurity and schools and government departments may need resources and technology to implement cybersecurity education. The rapid advancement of technology creates new risks and poses challenges for educators in keeping students safe. Teachers may struggle to stay current on the latest technology and need access to resources and training. Providing early exposure and education on cybersecurity through workshops and conferences for students at schools is essential to producing a skilled future workforce in cyber defense.

Case Study

Case Study I

One example of a case study in the importance of cybersecurity in education systems is the 2017 cyber-attack on the University of Cambridge. The attack, which was carried out by a group of hackers, targeted the university's administrative and academic systems and resulted in the theft of sensitive personal data belonging to staff and students. The university was forced to shut down many of its systems and services in order to contain the attack, leading to significant disruptions in the institution's daily operations and causing a great deal of distress for the affected individuals. The incident served as a stark reminder of the need for robust cybersecurity measures in education systems, as universities and schools are often targeted by hackers due to the valuable personal and financial data that they possess.

Case Study 1I

The University of California, Berkeley: In 2018, the university experienced a data breach that exposed the personal information of over 80,000 students and staff. The university implemented a number of cybersecurity measures to prevent future breaches, including strengthening network security, increasing employee training on cybersecurity, and implementing two-factor authentication for all accounts.

Case Study 1II

Miami Dade College: In 2017, the college experienced a ransomware attack that affected its entire network, shutting down email, internet, and other services for several days. The college responded by implementing a number of cybersecurity measures, including regular backups, improved network monitoring, and employee training on cybersecurity best practices.

Case Study 1V

In 2016, a security incident at the University of Utah resulted in the compromise of the personal information of a large number of students and staff. To address this, the university put in place various security measures including separating the network, using systems that detect and block unauthorized access and providing education to employees on how to secure their work effectively.

Case Study V

In 2017, the University of Maryland suffered a data security incident which led to the exposure of the personal information of more than 300,000 students and staff. To address this, the university adopted several cybersecurity measures, such as dividing the network into smaller parts, using systems to detect and block unauthorized access, and providing training to employees on the best cybersecurity practices.

Case Study V1

The University of Alabama: In 2018, the university experienced a data breach that exposed the personal information of over 400,000 students and staff. The university responded by implementing a number of cybersecurity measures, including network segmentation, intrusion detection and prevention systems, and employee training on cybersecurity best practices.

RESULTS AND DISCUSSION

In the modern digital environment, education must include cybersecurity. There is a rising need for education systems to become more secure due to the rise in linked devices, a tremendous influx of data, and a growing reliance on technology. To shield academics, staff, and students from internet dangers including hackers,

ISSN 2394 - 7780

malware, phishing, and other cyber-attacks, cybersecurity is crucial. Additionally, it's critical to protect student and staff safety, as well as the confidentiality and security of the school's data. In schools, cybersecurity can aid in preventing unwanted access to the network, systems, and data as well as protecting against data breaches and other cyber-attacks. Aside from ensuring the accessibility and dependability of educational systems, cybersecurity measures can also India. The literature suggests that comprehensive cybersecurity education in schools can be beneficial. Studies show that individuals who received cybersecurity education spend less money or time on cybersecurity workshops or programs. This highlights the importance of schools as central hubs for educating the community on cybersecurity issues. To achieve this, school administrators and teachers must work together to develop and implement cybersecurity programs or events. Additionally, government funding is necessary to cover the costs of organizing these events. Furthermore, cybersecurity training can change people's attitudes and awareness of its importance. A lack of knowledge about cybersecurity is often a result of a lack of understanding and awareness.

CONCLUSION

The significance of cybersecurity education is paramount as technology continues to play a vital role in the field of education. It is essential for educational institutions to adopt effective cybersecurity measures to safeguard sensitive data and maintain the smooth functioning of operations. These measures include but not limited to, network security, staff training, multi-factor authentication, dividing networks, identifying and preventing intrusion, and regular security assessments.

Cybersecurity education is essential for the present and future workforce and society. It is necessary to include cybersecurity education in the education system to prepare students for the digital age. By providing students with knowledge about the risks and threats of cyber-attacks, they will be able to protect themselves and their organizations from cybercrime. The education system must adapt to the needs of the 21st century and equip students with the skills and knowledge to navigate the digital world safely.

In summary, investing in cybersecurity and remodelling the education system is essential for protecting the integrity of educational institutions, the security of student and staff data, and preparing students for the digital age.

REFERENCES

[1] F. Khalid, ―Understanding university students' use of facebook for collaborative learning,‖ International Journal of Information and Education Technology, vol. 7, no. 8, pp. 595-600, August 2017.

[2] F. Annasingh and T. Veli, ―An investigation into risks awareness and e-safety needs of children on the internet, Interactive Technology and Smart Education, vol. 13, no. 2, pp. 147-165, 2016.

[3] L. Muniandy and B. Muniandy, ―The impact of social media in social and political aspects in Malaysia: An overview, International Journal of Humanities and Social Science, vol. 3, no. 11, pp. 71-76, 2013.

[4] V. Ratten, ―A cross-cultural comparison of online behavioral advertising knowledge, online privacy concerns and social networking using the technology acceptance model and social cognitive theory, Journal of Science & Technology Policy Management, vol. 6, no. 1, pp. 25-36, 2015.

[5] M. D. Griffiths and D. Kuss, ―Online addictions, gambling, video gaming and social networking, The Handbook of the Psychology of Communication Technology, Chichester: John Wiley, pp. 384-406, 2015.

[6] L. Mosalanejas, A. Dehghani, and K. Abdolahofard, ―The students' experiences of ethics in online systems: A phenomenological study, Turkish Online Journal of Distance Education, vol. 15, no. 4, pp. 205-216, 2014.

[7] D. Krotidou, N. Teokleous, and A. Zahariadou, ―Exploring parents' and children's awareness on internet threats in relation to internet safety, Campus-Wide Information Systems, vol. 29, no. 3, pp. 133-143, 2012.

[8] R. S. Hamid, Z. Yunos, and M. Ahmad, ―Cyber parenting module development for parents,‖ in Proc. INTED2018 Conference, 5th-7th March 2018, Valencia, Spain, 2018.

[9] F. Khalid et al., ―An investigation of university students' awareness on cyber security, International Journal of Engineering & Technology, vol. 7, pp. 11-14, 2018.

[10] C. S. Kruse et al., ―Cybersecurity in healthcare: A systematic review of modern threats and trends,Technology and Health Care, vol. 25, no. 1, pp.1-10, 2017.

ISSN 2394 - 7780

[11] P. Dong et al., ―A systematic review of studies on cyber physical system security, International Journal of Security and Its Applications, vol. 9, no. 1, pp. 155-164, 2015.

[12] U. Franke and J. Brynielsson, ―Cyber situational awareness — A systematic review of the literature, Computers & Security, vol. 46, pp. 18-31, 2014.

[13] N. H. A. Rahim et al., ―A systematic review of approaches to assessing cybersecurity awareness, Kybernetes, 2015.

[14] D. Mellado et al., ―A systematic review of security requirements engineering, Computer Standards & Interfaces, vol. 32, no. 4, pp. 153-165, 2010.

[15] A. V. Herrera, M. Ron, and C. Rabadão, ―National cyber-security policies oriented to BYOD (bring your own device): Systematic review, in Proc. 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-4, 2017.

[16] F. Mishna et al., ―Interventions to prevent and reduce cyber abuse of youth: A systematic review,‖ Research on Social Work Practice, vol. 21, no. 1, pp. 5-14, 2011

[17] Digile (2014), Strategic Research Agenda for Cyber Trust, 12.6.2014

[18] European Commission (2013), Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace, JOIN(2013) 1 final, Brussels, 7.2.2013

[19] INKA (2014) Innovative Cities programme 2014–2020, cyber security theme, implementation programme v2.1, 1.2.2014

[20] The implementation programme for Finland's Cyber Security Strategy (2014), the Security Committee, 11 March 2014.

[21] Ministry of Education and Culture, MEC (2015), Education System in Finland,

[22] Ministry of Employment and the Economy, MEE (2013), 21 paths to a Frictionless Finland, Report of the ICT 2015 Working Group, 18/2013

[23] Ministry of Transport and Communication, MTC (2010), Productive and innovative Finland – digital agenda for the years 2011-2020.

[24] Finland's Cyber Security Strategy (2013). Government Resolution 24 January 2013.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

A REVIEW ON PHISHING ATTACKS

Swapnali Lotlikar[1], Prashant Chaudhary[2], Prakash Amin[3] and Urvi Rathod[4]
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, Maharashtra

ABSTRACT
Phishing is a type of cyber-attack that involves the use of fraudulent emails, websites, or other digital communication channels to steal sensitive information from an unsuspecting victim. The objective of a phishing attack is to obtain personal information such as usernames, passwords, credit card numbers, and other confidential data from the victim. Phishing attacks are typically carried out by sending out malicious emails disguised as legitimate messages from trusted sources. These emails often contain malicious links or attachments, which when clicked on, will redirect the victim to a malicious website or open an application that will steal the victim's information. Other methods of phishing include social engineering, where attackers use psychological manipulation in order to gain access to a victim's system. This paper will discuss the various types of phishing attacks, the different anti-phishing techniques and how to identify different phishing types. It will also provide recommendations/tools on how individuals and organizations can protect themselves against phishing attacks.

Keywords: phishing, social engineering, cyber-attacks, anti-phishing techniques.

INTRODUCTION
Phishing is a dishonest technique in which the attacker or "phisher" tries to persuade internet users to give up their login credentials or personal information in return for money. [1]. When phishing is compared to fishing, which has a distinct meaning, the attacker uses bait (sending an email with a hyperlink that goes to a phony website embedded) to obtain users' login credentials. The attackers have previously been referred to as "Phreaks" (a Phreak is a person who illegally hacks into phone networks to make free long-distance calls or to tap phone lines), and they are all connected to one another. To connect phishing schemes with phreaks, "ph" has been used in place of "f" [2]. In the last 20 years, phishing has developed into the most dangerous threat, and numerous attacks take place every day. [3, 4]. The first phishing scam was reported on January 2, 1996, by internet service provider American Online (AOL).

The thief generates credit card information at random and then uses that data to build AOL profiles. Later, they ask customers to validate their login details by clicking on a legitimate source by sending them an email utilizing AOL's instant messaging or email system. If the user clicks the hyperlink and enters their login information, the data is delivered directly to the hacker and is immediately sent to him. Therefore, the attackers abuse these credentials for unscrupulous purposes. Every time, a new technique is developed by an attacker to trick a specific user and steal their personal data. The anti-phishing working group (APWG) 2nd quarter of 2021 for the entire month of June. The anti-phishing working group (APWG) 2nd quarter of 2021 documented 2,22,127 unique phishing attacks for the entire month of June 2021. The distinctive phishing sites created by the scammers in 2020–2021 are displayed in Figure 1. Pharming, often known as "phishing without bait," is an improved version of phishing [6, 7].

Pharming, then, is a DNS-based attack where the attacker gains unauthorised access to the DNS and modifies the host file entries to send all users who acquire data from that DNS to the fake website. Since it affects so many people who have experienced DNS poisoning, it is more dangerous and difficult to recognise. More recently, it has been discovered that phishing is carried out by using malware (ransomware) to seize control of access to user information and then blackmailing users into paying a ransom. In 2015, 2,453 ransomware-related complaints totaling $1.6 million were submitted to the Internet Crime Report (ICR).The APWG study (1st quarter 2019) [8] found 1,80,768 distinct phishing websites. 36% of this phishing schemes target software-as-a-service (SaaS) and webmail providers. Various researchers and groups have created a plethora of anti-phishing technologies to protect users against phishing assaults. These anti-phishing solutions primarily function at the user level, with only a few exceptions working on the server. A comprehensive categorization of anti-phishing solutions is offered in our earlier study [9].

Problem Statement
The Internet has dominated the world by dragging half of the world's population exponentially into the Cyberworld. With booming of the internet transaction, cybercrimes are rapidly increasing. Many cyber-attacks are spread via mechanism that exploit weakness of end-user through various forms such as phishing, malware,

ISSN 2394 - 7780

ransomware, and so on. Among all these attacks, phishing reports to be the most deceiving. Our main aim of the paper is to review the different phishing attacks and surveying many of the recently proposed anti phishing techniques, anti-phishing tools and case studies.

LITERATURE REVIEW

Ludl et al, (2007) [10] investigated the effectiveness of phishing and looked for remedies, focusing on two well-known anti-phishing programs. The anti-phishing features built into Firefox 2 (i.e., Google blacklists) and Microsoft's Internet Explorer 7 were investigated for three weeks by automatically testing them against a blacklist of 10,000 bogus URLs maintained by Google and Microsoft. They also examined a large number of phishing sites to see if there were any page characteristics that might be used to identify phishing pages. And how the presence of these elements—links, dubious urls, forms, and input fields—might be essential for visitors to be deceived.

Tanvi Churi et al, [11] offered a sample system for figuring out whether a website is a phishing site or not. In accordance with their report, current phishing protection solutions do not offer code generation techniques that are 100% accurate and only accessible to authenticated users. The suggested system creates a picture, which the visual cryptographic methods then divide into two groups. The user is prompted to match the website with the created picture captcha in order to distinguish the legitimate website from phishing sites when these two portions of the image are combined. The next step involves creating and having a qualified individual authenticate a four-digit code. The tactic can help you spot phishing websites and safeguard credentials from unwanted repercussions.

Simono et al, [12] have briefly discussed the lax security measures present in Android phone password managers, which serve as a breeding ground for phishing attacks. The study identifies a number of architectural weaknesses in password manager implementations that support these attacks.

Rao et al, [13] created a special feature extraction heuristic-based categorization technique. Three categories URL obfuscation features, Third-Party-based features, and Hyperlink-based features—are used to group the obtained characteristics. The proposed method also has a 99.55% accuracy. The drawback of this strategy is that while third-party features are taken into account, website classification is reliant on the speed of third-party services. Broken links feature extraction covers lexical features, URL-based features, network-based features, and domain-based features, and it depends totally on the caliber and volume of training data.

Gupta et al, [14] developed a ground-breaking anti-phishing method that just uses client-side attributes. Since it solely pulls features from URLs and source code and doesn't rely on a third party, the suggested solution is speedy and reliable. According to their analysis, the overall detection accuracy for phishing websites was 99.09%. This article came to the conclusion that this method has limits because it can only identify HTML-coded websites. Non-HTML websites cannot be found using this method.

Dhamija et al, [15] demonstrates that many internet users have difficulty detecting phishing assaults. Even when consumers are given the specific duty of identifying phishing schemes, many of them are unable to discern between a legal website and a faked website. The best phishing site in the survey deceived more than 90% of the subjects. Furthermore, people frequently do not understand certain security indicators, such as the padlock signifying secure transmission, signal the reliability of a website.

Fette et al, [16] introduced PILFER, an email filtering strategy that included ten characteristics, including URL and script-based elements, to identify phishing assaults. By screening phishing emails before they are seen by users, the percentage of fraudulent users can be reduced. Phishers can conceal the URL and utilise tools such as TinyUrl to make it look genuine. Phishers' methods are growing more complex, and they are adding ways to circumvent existing anti-phishing solutions.

Steps in Phishing

A phisher is someone who engages in virus operations. Phishing attempts today generally deploy broad "lures", frightening victims and inciting panic - a frequent example is "we need you to confirm your account credentials or we must shut your account down". A more advanced strategy that is considered to be becoming more popular is context-aware attack: this is a more complex approach since it not only employs threat or inducement but also makes the target think of the communications as anticipated and therefore legitimate.

Phishers often create bogus websites that seem similar to legitimate websites by replicating the HTML code and using the same graphics, content, and sections. Some phishing websites register a domain name that is identical to the real website of a firm or bank. Forms are the most typical approach employed by phishers, such as the

ISSN 2394 - 7780

Internet Banking login page or a form for account verification. Many phishing efforts involve domain spoofing or homographic assaults to persuade victims to provide personal information (Gabrilovich & Gontmakher).

A phisher might target a variety of sensitive information, such as user names and passwords, credit card data, bank account details, and other personal information. According to Gartner research (Gartner Inc, 2004), around 19% of those polled claimed having clicked on a link in a phishing email, and 3% admitted to providing financial or personal information [17].

A typical phishing attack involves the attacker obtaining the victim's authentication information for one website, corrupting it, and using it at another. Given that many computer users reuse passwords—either verbatim or with only minor modifications—this is a significant assault. The lifecycle of a phishing attack can be broken down into:

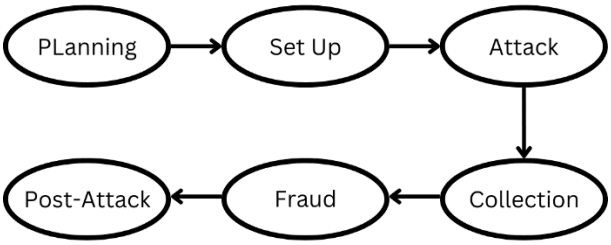


Figure 1: Steps in Phishing

The attack code or message is created by the phisher and sent to the target user. The target site receives a damaging message. The unknowing victim sees the message and takes any action that could lead to a data breach. A trusted and well-known internet interface is then used to ask the user for sensitive information. User discloses private information. Private information is delivered to the phisher using a phishing server. The phisher commits fraud by impersonating the user using private information [18].

There is no single method for preventing all phishing. However, different approaches used at different phases of a phishing campaign can stop it, and correctly used technology can dramatically lower the danger of identity theft.

Types of Phishing Attacks

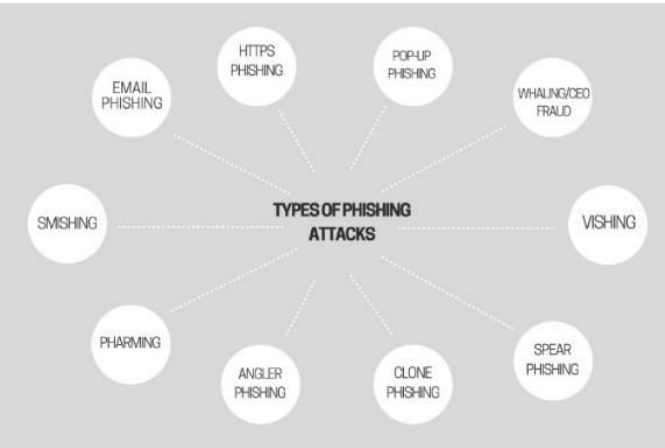


Figure 2: Types of Phishing Attacks

Email Phishing

Email Phishing is a sort of cyberattack in which bogus emails are used to trick users into disclosing sensitive information or clicking dangerous links. Attackers will frequently claim to be from a reputable corporation or organization, utilizing identical language and layout to authentic emails. The emails may pretend to request updated personal information or may include malicious links or files that install malware on the victim's device. People should be wary of unexpected emails and should never disclose important personal information or follow links from unknown sources.

How to identify email phishing:

Look for Spelling and Grammar Mistakes: Legitimate organizations take pride in their communications, so typos and bad grammar are a red flag.

ISSN 2394 - 7780

Check for Generic Greetings: If the email address doesn't include your name, be suspicious. Legitimate organizations will use your name in the greeting.

Inspect Links Carefully: Hover your mouse over links to check the destination URL. It should match the linked text.

Look for a Sense of Urgency: Phishing emails often create a false sense of urgency to get the recipient to take action quickly.

Check for Requests for Personal Information: Legitimate organizations will never ask you for personal information, such as passwords or credit card numbers, via email.

HTTPS Phishing
HTTPS phishing is a sort of attack in which a malicious website masquerading as a genuine site is used to deceive visitors into entering personal information such as usernames, passwords, and credit card information. The malicious website is typically hosted via a secure HTTPS connection, making it difficult to differentiate from a real one. Attackers take advantage of this since users may not be able to notice the malicious website and may submit their personal information into it, resulting in data theft.

How to identify HTTPS phishing:

Look for errors in the website's URL: Make sure there are no spelling or grammar errors in the website's URL.

Check the domain: Hover over the URL and look at the domain. If it doesn't match the name of the website or company, it's likely a phishing site.

Verify the website's authenticity: If the website is asking for personal information, make sure it is from a legitimate source. Check for contact information and customer reviews.

Spear Phishing
Spear phishing is a sort of social engineering assault in which victims are tricked into exposing critical information or providing access to their systems by sending targeted emails with malicious files or links. The emails look to be from a reliable source and are tailored to appear as though they are intended exclusively for the recipient. It is critical to be aware of possible spear phishing efforts and to take protective steps.

How to Identify Spear Phishing:
Watch out for suspicious emails. Spear phishing emails are usually from an unknown sender and usually contain urgent requests or messages that seem to come from someone you know.

Check for grammar and spelling errors. Many times, spear phishing emails contain poor grammar and spelling, which is a sign that the email may not be legitimate.

Look for personalized messages. Spear phishing emails often contain personal information that is tailored to the recipient. If you receive an email that contains your name, address, or other personal information, it could be a sign of a spear phishing attempt.

Whaling/CEO Fraud
CEO fraud, also known as whaling, is a sort of phishing assault in which attackers send emails posing as the CEO or another high-ranking official of a firm. Emails frequently request sensitive information, such as passwords or bank information, and may include dangerous files or links. Victims who fall victim to the attack face identity theft, money loss, and other fraudulent acts.

How to Identify CEO Fraud:
Poor grammar and spelling: Poor grammar and spelling in emails are usually a sign of a scam.

Unusual requests: Be wary of any requests that are out of the ordinary, such as asking for money or confidential information.

Unusual email attachments: Be careful when opening any attachments that come with suspicious emails. They could contain malicious software.

Vishing
Vishing is a sort of telephone fraud that employs social engineering tactics. It is a mix of the term's "voice" and "phishing". It includes attackers phoning victims and posing as representatives of a genuine business or institution, like as a bank, and requesting personal information. Passwords, credit card numbers, and other sensitive information may be requested by the attackers. They often employ social engineering techniques to

ISSN 2394 - 7780

deceive individuals into delivering the needed information, such as generating a feeling of urgency or playing on people's emotions. Vishing may be extremely harmful, resulting in identity theft and financial losses. As a result, it is critical to understand the hazards involved with this form of fraud and to practice safe online practices.

How to identify vishing:

Do not trust caller ID. Scammers can spoof their phone numbers so that it appears to be coming from a trusted source.

Do not provide any personal or financial information over the phone, through email, or in response to a text message.

Be suspicious of emails or text messages that contain urgent requests, threats, or instructions to click on a link or open an attachment.

Smishing
Smishing is a sort of cyber assault in which SMS messages are sent to victims in attempt to steal their personal information or install malware on their devices. Typically, the messages contain a link or phone number that, when clicked or contacted, will take the recipient to a malicious website or an automated speech system. Smishing attempts to dupe victims into revealing personal information or installing harmful software. To defend against smishing attacks, be alert of strange communications, never click on links or phone numbers from unfamiliar sources, and always use up-to-date security software.

How to Identify Smishing:
Look out for text messages that ask for personal information such as bank details, passwords, or credit card numbers.

Be wary of messages that contain urgent language or create a sense of urgency in order to get you to act quickly.

Be cautious of links included in text messages that appear to be from legitimate companies.

Be suspicious of messages that ask you to call a number to update your account information.

Angler Phishing
Angler phishing is a type of cyber-attack in which a hostile hacker manipulates victims into exposing private information such as passwords or financial information. The hacker frequently sends a fraudulent email or link that appears to be from a reputable source. The attacker then obtains the private data and uses it for personal advantage. Angler phishing attempts can be difficult to identify because they frequently look to be legitimate.

How to Identify Angler Phishing:
Be wary of attachments: Phishing emails often contain attachments that are malicious or contain dangerous links. Do not open any attachments unless you are sure they are safe.

Check for threats: Phishing emails often contain urgent threats such as, “Your account will be suspended if you do not take action now.” These are usually attempts to get you to act quickly without thinking.

Look for typos: Legitimate emails rarely contain typos or other mistakes. If you spot any errors, be suspicious that the email may be a phishing attempt.

Pharming
Pharming is a sort of cyber assault in which traffic is diverted from a genuine website to a malicious one. Attackers employ malicious code to reroute users to a bogus website that appears to be the real thing. The bogus website can then capture sensitive information like as usernames, passwords, and credit card numbers. It is also capable of disseminating malware and other dangerous programming. Pharming is a severe danger to internet security that must be addressed.

How to Identify Pharming:
Be aware of any suspicious emails that appear to be from legitimate organizations, such as banks or credit card companies, but may contain malicious links or attachments.

Monitor your bank and credit card accounts for any unauthorized transactions or activity.

Be wary of pop-ups or other advertisements that ask for personal information, as these may be phishing scams.

ISSN 2394 - 7780

Pop-up Phishing
Pop-up phishing is a sort of cyberattack in which a malicious pop-up window is used to collect sensitive information from victims. The window is used by attackers to imitate a genuine website or service, such as a bank or other financial institution, and request visitors to submit personal information or passwords. Pop-up phishing attempts to get sensitive data or credentials that can subsequently be used to gain access to accounts or commit identity theft.

How to identify pop-up phishing:

Check The URL: A phishing popup will usually have an URL that seems suspicious or doesn't match the website you are visiting.

Check The Content: Legitimate popups will never ask you to enter sensitive information. If it is asking for personal information, such as your credit card or banking details, it is likely a phishing popup.

Check the grammar: Phishing popups often contain spelling and grammar mistakes.

Clone Phishing
Clone phishing is a phishing assault in which a hacker copies a legitimate website or email and distributes it to victims. The phony version contains harmful links or malware, or it requests personal information or credentials from users. These malicious links or messages can be sent by email, text messaging, or websites. The idea is for visitors to confuse the false website or email with the actual one, providing the hacker access to important information.

How to Identify Clone Phishing:
Check the sender address: One of the key indicators of a clone phishing attack is an email address that looks similar to a legitimate address, but is slightly different. Be sure to check the sender's email address carefully.

Check the domain name: Clone phishing attacks often come from domains that are similar to those used by legitimate companies or organizations. Carefully inspect the domain name of the sender to ensure that it is not a spoofed version of a legitimate domain.

Anti-Phishing Techniques
In [19], a technique was created that uses an automated list called the Automated Individual White-List (AIWL) to keep track of the user's well-known Login User Interfaces (LUI). The AIWL alerts the user of the potential trap and notifies him/her of the impending assault when the user submits login credentials or sensitive information to a LUI that is not on the white-list.

In [20] the author integrated visual similarity-based techniques with a white list, the authors' solution for defending against phishing attacks was revealed. Computer vision (CV) software includes the Speed up Robust Features (SURF) detector. This detector uses square-shaped filters to extract discriminative key point features. These qualities have been gathered from both reliable and dubious places. The degree of similarity is then determined by comparing the traits that were gathered from the websites. Thus, the degree of similarity helps determine the credibility of the website. Since the real website was being attempted to be imitated, it was deemed damaging if the degree of similarity was high.

In [21], Support Vector Machines (SVM) are being used in a novel way to ascertain whether an email is malicious. The language, design, and structure of the email were all common attributes that the SVM extracted. To assess whether the similarity is accurate, it then compares the information that was retrieved with information already stored in the system. The email is labelled as malicious if the accuracy exceeds a certain threshold.

The study done in [22] used a cutting-edge Natural Language Processing (NLP) technique to determine whether the email was malicious. In this study, they extracted and compared common features using NLP techniques. PhishNet-NLP combined natural language processing techniques with email data, such as the header, links, and body content. PhishSnag used data gathered from emails to identify phishing. To evaluate whether or not the email was phishing, Phish-Sem used statistical analysis and natural language processing (NLP) on the email body.

In [23] A more sophisticated approach to filtering and classification was used. The URLs were examined by the study's authors to rule out any harmful content. They used an automated system to recognise phishing. Pre-filtering and classification were the two divisions. The URL was compared to a black list at the pre-filtering stage using the domain component of the URL. The URL was labelled as malicious and would not move on to

ISSN 2394 - 7780

the Classification Phase if it was on that list. The following stage involved testing the consistency of the domain token location and the URL's randomization (RU). The URL was categorised as malicious or non-malicious based on the results of the classification phase.

In [24] To extract different traits from emails, the authors of used text mining. The emails may be phishing or legitimate in order to better identify the attack. Following a preliminary vectorization of the email, feature selection techniques for classification were applied. Data sets from the publicly available PhishingCorpus and the HamCorpus (legitimate e-mail) of the Spam Assassin project were used to conduct the study (phishing e-mail).

Anti-Phishing Tools
A number of solutions are available to assist defend your business from the sorts of hazards that phishing assaults bring. Knowing what solutions are available and how they may assist safeguard your organization, and consequently your staff and customers, is half the battle.

Avanan
Avanan is a high-end cloud security platform for businesses that offers complete security for cloud-based programmes including Office 365, G Suite, AWS, and Salesforce. Advanced threat protection, data loss prevention, email encryption, and other features are available. IT administrators may easily find and fix any security vulnerabilities because to the thorough information and analytics it offers. Additionally, it provides ongoing monitoring to guarantee the security of all cloud settings.

Barracuda Sentinel
A cloud-based artificial intelligence and machine learning technology called Barracuda Sentinel was created to defend enterprises against complex threats like phishing and malicious URLs. It employs a variety of technologies, including machine learning, behavioural analytics, and natural language processing, to spot malicious conduct and shield businesses against network attacks and data breaches. It offers features for automatic response, proactive alerting, and real-time threat detection.

BrandShield
BrandShield is an automated tool for brand protection and monitoring that assists businesses in safeguarding their brands, trademarks, and domain names against internet threats. It scans the web for possible hazards and threats and notifies brands of any potential difficulties using cutting-edge technology and AI-driven algorithms. To assist defend organizations from online dangers including phishing, counterfeiting, cybersquatting, brand hijacking, and more, its complete platform offers monitoring, enforcement, and reporting capabilities.

Cofense PDR
A complete, cloud-based security system called Cofense PDR (Phishing Defense and Response) enables businesses to quickly identify, assess, and counteract dangerous phishing assaults. It swiftly identifies and prioritises phishing attacks using sophisticated machine learning and automated analysis, and it offers thorough reports and repair advice to make sure your organization's network is safe.

RSA FraudAction
A security system called RSA FraudAction is intended to shield businesses against online fraud, malicious assaults, and other cyberthreats. It makes use of cutting-edge analytics and machine learning to spot suspicious activity and take preventative action to guard against fraud. Additionally, by offering automated fraud detection and prevention capabilities, it aids firms in lowering their risk.

IRONSCALES
A security platform called IronScales combines artificial intelligence and machine learning to find and stop cyberthreats. It offers a thorough system of risk management and response and makes use of an automated system to monitor, identify, and assess the security of companies. This platform assists businesses in defending against harmful attacks to their data, networks, and systems, and it gives security professionals the resources and information they need to keep networks safe.

Mimecast
The cloud-based email security and management platform Mimecast assists businesses in lowering IT expenses, ensuring business continuity, and protecting their data from cyber-attacks. To shield customers from phishing, malware, and dangerous information, it offers services including email archiving, threat prevention, encryption, and secure email gateway. Additionally, it aids businesses in upholding secure email interactions, safeguarding their brands, and conforming to international data privacy standards.

ISSN 2394 - 7780

Microsoft Defender for Office 365
A cloud-based security solution called Microsoft Defender for Office 365 helps shield your Office 365 data from dangers like malware, phishing, and ransomware. It checks emails, documents, and other files for malicious activity using sophisticated threat prevention and machine learning, and it alerts users to any questionable information. Additionally, it offers details on malware trends and aids in data loss prevention.

Case Study

Case No. 1 – Twitter Phishing Case 2020
In 2020, Twitter experienced a phishing attack that impacted a number of high-profile accounts. The attack involved a coordinated effort to gain access to Twitter accounts by sending out phishing links to various accounts. The attackers then used the compromised accounts to send out tweets with links to cryptocurrency scams. Twitter responded to the attack by temporarily disabling features such as tweeting, changing passwords, and verifying emails to prevent further attacks. They also implemented new security measures, such as two-factor authentication, to help protect users from future phishing attacks. They also took action against the accounts that were used to send the phishing links, suspending or deleting them.

Twitter also worked with law enforcement agencies to investigate the attack and identify the perpetrators. In July 2020, the US Department of Justice announced that three individuals had been charged in connection with the Twitter phishing attack. The attack serves as a reminder of the need for users to be vigilant when it comes to their online security. It is important to be aware of phishing attempts and to always take steps to protect yourself, such as using two-factor authentication and avoiding links from unknown sources.

Lessons Learned from the Case
Ensure Your Security Settings Are Up to Date: It's important to ensure that your security settings are up to date to protect yourself from phishing attacks. This includes using strong passwords, two-factor authentication, and regularly updating your software.

Be Wary of Emails: If you receive an email that looks suspicious or contains a hyperlink that you don't recognize, be wary. Do not click on any links or attachments unless you are certain they are from a secure source.

Educate Your Employees: Make sure your employees are aware of the risks of phishing and how to recognize and avoid it. Consider providing them with training on how to spot and protect against phishing attacks.

Case No. 2: Upsher - Smith Laboratories Case 2014
Upsher-Smith Laboratories is a leading manufacturer of generic and branded pharmaceutical products. Based in Maple Grove, Minnesota, the company was founded in 1919 and has grown to become a major player in the global pharmaceutical industry. In 2014, Upsher-Smith Laboratories faced a significant challenge when it was the subject of a class action lawsuit. The lawsuit, which was filed in the United States District Court for the Eastern District of Pennsylvania, alleged that Upsher-Smith Laboratories had engaged in unfair and deceptive practices in the sale of its generic drugs. Specifically, the lawsuit alleged that Upsher-Smith had conspired with other generic drug makers to keep generic drug prices artificially high in violation of antitrust laws.

In response to the lawsuit, Upsher-Smith Laboratories sought to have the case dismissed. In its motion, the company argued that the claims were baseless and that the company had not engaged in any anticompetitive practices. The company also argued that the lawsuit was filed too late, as the claims were barred by the applicable statute of limitations. The court ultimately denied the motion to dismiss, however, and the case proceeded to trial. At trial, the plaintiffs argued that Upsher-Smith Laboratories had conspired with other generic drug makers to keep generic drug prices artificially high. The plaintiffs also argued that Upsher-Smith had violated antitrust laws by entering into agreements with other generic drug makers to divide up the market and limit competition. The jury ultimately determined that Upsher-Smith had violated antitrust laws and held the company liable for damages. The jury awarded the plaintiffs nearly $100 million in damages, including $50 million in punitive damages. The case was ultimately settled out of court for an undisclosed amount.

As a result of the litigation, Upsher-Smith Laboratories was forced to pay out a significant amount of money in damages. Additionally, the company faced a significant reputational risk as a result of the lawsuit, which could have had a negative impact on its ability to attract customers and partners. Upsher-Smith has since implemented several measures to ensure compliance with antitrust laws and to prevent similar issues from arising in the future.

ISSN 2394 - 7780

Lessons Learned from the Case

Don't be afraid to take risks. In order to stay competitive, Upsher-Smith Laboratories had to take risks in order to develop new products and services. They had to invest in new technology and push the boundaries of their existing products in order to stay ahead of their competitors.

Focus on customer needs. Upsher-Smith Laboratories was able to create a successful product by focusing on their customers' needs and understanding what they wanted. By doing this, they were able to create a product that was tailored to the needs of their customers.

Utilize partnerships. Upsher-Smith Laboratories was able to develop their product by partnering with other organizations. This allowed them to combine their resources and create a comprehensive product that could meet the needs of their customers.

Leverage technology. Technology was a key part of Upsher-Smith Laboratories' success. They were able to use technology to create a unique product that was tailored to their customers' needs. This allowed them to differentiate themselves

CONCLUSION

Phishing attacks continue to be a serious issue for both individuals and organisations as well as enterprises. Despite significant improvements in security technology, phishing attempts continue to pose a serious concern. This is as a result of hackers constantly changing their methods and plans of action to avoid discovery.

Even though phishing assaults are challenging to counter, there are procedures that may be taken to lower the likelihood of a successful attack. The first thing people and organisations should do is make sure they have the most recent security technology in place and that it is always current. Firewalls, antivirus programmes, and spam filters are examples of this. Additionally, it is crucial to inform employees about the risks posed by phishing scams, the necessity of being watchful for questionable emails, and other similar issues.

Organizations have to think about utilising two-factor authentication for private accounts like banking, email, and social networking. Due of the additional login information required, hackers won't be able to access the account, lowering the probability of a successful phishing attempt. Organizations should also be aware of the most recent trends and keep a watch out for new phishing schemes.

Finally, if a business or person falls victim to a phishing assault, they should get in touch with their IT provider right once and take action to limit the harm. This can entail doing scans to find any harmful software, monitoring accounts, and resetting passwords.

Overall, phishing assaults continue to pose a serious threat to both individuals and businesses. Even while it is challenging to totally guard against these assaults, adopting the right precautions can greatly lower the likelihood of being a victim.

REFERENCES

[1] Kirda E, Kruegel C. Protecting users against phishing attacks with antiphishing techniques. In annual international computer software and applications conference 2005 (pp. 517-24). IEEE.

[2] Mei Y. Anti-phishing system: detecting phishing e-mail. School of Mathematics and Systems Engineering. 2008.

[3] Yadav S, Bohra B. A review of recent phishing attacks on the internet. In international conference on green computing and internet of things 2015 (pp. 1312-5). IEEE.

[4] IRONSCALES. How modern email phishing attacks have organizations on the hook. 2017.

[5] APWG. APWG phishing trends report 2nd quarter 2021. 2021.

[6] Alfayoumi IS, Barhoom TS. Client â [euro]" Side pharming attacks detection using authoritative domain name servers. International Journal of Computer Applications. 2015; 113(10):26-31.

[7] Ollmann G. The vishing guide. IBM Global Technology Services. 2007:1-16.

[8] Anti-phishing working group. APWG Phishing activity trends report 2nd quarter 2012.

[9] Chanti S, Chithralekha T. Classification of anti-phishing solutions. SN Computer Science. 2020; 1(1):1-8.

[10] Ludl, Christian et al. “On the Effectiveness of Techniques to Detect Phishing Sites”. In: (2007). Ed. by Bernhard M. Hämmerli and Robin Sommer, pp. 20–39.

ISSN 2394 - 7780

[11] T. Churi, P. Sawardekar, A. Pardeshi, and P. Vartak, "A secured methodology for anti-phishing," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018- Janua, pp. 1–4, 2018.

[12] S. Aonzo, A. Merlo, G. Tavella, and Y. Fratantonio, "Phishing attacks on modern android," Proc. ACM Conf. Comput. Commun. Secur., pp. 1788–1801, 2018.

[13] Routhu Srinivasa Rao1 , Alwyn Roshan Pais : Detection of phishing websites using an efficient feature-based machine learning framework :In Springer 2018.

[14] Ankit Kumar Jain, B. B. Gupta : Towards detection of phishing websites on client-side using machine learning based approach :In Springer Science+Business Media, LLC, part of Springer Nature 2017.

[15] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pages 581–590, 2006.

[16] Fette, I., Sadeh, N., Tomasic, A.: Learning to detect phishing emails. In: Proceedings of the 16th international conference on World Wide Web, pp. 649–656 (2007)

[17] Kirk J. (2011). Phishing Tool Constructs New Sites in Two Seconds.

[18] Camp LJ, Goodman S, House CH, Jack WB, Ramer R and Stella M. Chapter 6: Offshoring: Risks and Exposures.

[19] Ye Cao, Weili Han and Yueran Le - Anti-phishing based on automated individual white-list, Proceedings of the 4th ACM workshop on Digital Identity Management, pp. 51-60, October 2008.

[20] Routhu Srinivasa Rao and Syed Taqi Ali - A Com- puter Vision Technique to Detect Phishing Attacks, 5th International Conference on Communication Systems and Network Technologies, IEEE, October 2015.

[21] Madhusudhanan Chandrasekaran, Krishnan Narayanan and Shambhu Upadhyaya - Phishing E- mail Detection based on Structural Properties, IEEE, November 2015.

[22] Rakesh Verma, Narasimha Karpoor, Nabil Hossain and Nirmala Rai - Automatic Phishing Email Detection based on Natural Language Processing Techniques, Research Gate, 2016.

[23] Yi-Shin Chen, Huei-Sin Liu, Yi-Hsuan Yu and Pang-Chieh Wang, Detect Phishing by Checking Content Consistency, IEEE, 2017.

[24] Masoumeh Zareapoor, K.R. Seeja, Text Mining for Phishing E-mail Detection, Intelligent Computing, Communication and Devices: Advances in Intelli- gent Systems and Computing, vol. 308, pp. 65-71, August 2016.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

STUDY OF AUTOMATIC TRAIN STOP SYSTEM IN KAVACH AND IN AMERICAN TRAINS

Sunita Gupta and Khanak Thosani
Department of Information Technology (MSc-IT), SVKM's Usha Pravin Gandhi College of Arts Science and Commerce, Mumbai, Maharashtra

ABSTRACT
An outline of the creation of an automated train stop system for Kavach and Amtrak is provided in this essay. The system is intended to improve the convenience, effectiveness, and safety of train services. The system's components, technological difficulties, and potential advantages are all covered in the paper's discussion of design considerations for the system. The article also looks at the status of technology today, as well as advantages and potential difficulties with the system's implementation. The study concludes with a review of the findings and a prediction about how the system can boost efficiency and safety.

Keywords: component, formatting, style, styling, insert (key words).

I. INTRODUCTION
In cases where the length of the train and subway platform are almost equal, target braking (docking) is employed to stop the train on the platform. One may describe target braking as an automatic method of halting the train at the station. This type of braking is crucial to railway operation since it calls for stopping the train on a platform that is a certain length. The train stopping at the metro station must be very precise and must adhere to the "door to door" braking principle in the event of closed platforms, when the entire platform is shielded from the track by a wall with automatic platform-edge doors. Doorway at the edge of the platform. Target train station braking should enable the train to stop so that it is positioned in relation to the platform so that all of the doors let passengers to exit the train straight onto the platform. From the perspective of the passengers, precise train stopping is equally essential. A passenger cannot be allowed to become trapped in the space between the platform and the train if the train stops in the wrong spot while in the platform. Automatic braking needs to begin at the suitable distance from the stopping point and needs the braking force to be properly controlled in order to ensure stopping at the required location. Goal braking is one of the necessary braking cases (duty).

It boils down to choosing between mechanical (electro-pneumatic) braking and the available electromagnetic braking system, which is based on electric traction motors acting as generators.

II. BACKGROUND

Content	American ATS	Indian Kavach
1. Train Running over Red/Danger Signal.	Yes	Yes
2. Emergency Braking If driver doesn't Respond (This system applies emergency brake when driver is not responding to the given period of time.	Yes	Yes
3. Automatic applying of emergency brakes if drivers exceed to given speed limit.	No	Yes
4. 24/7 SOS Availability.	Yes	Yes
5. Automatic Application of emergency brakes in both the trains before (500meters) in case of head-to-head collision.	No	Yes
6. Automatic Application of emergency brakes in the rear trains before(500meters) in case the front train is in halt.	No	Yes

Figure 1: Overall view of kavach and Amtrak

1. Background KAVACH
Indian Railroads is seeing extremely good gross improvement increase. because 1950, the quantity of passenger kilometres, freight educate kilometres, revenue-producing kilometres (NTKM), gross tonnes per kilometres (GTKM), waggon kilometres, and coach kilometres has been steadily growing. in the years 2012–2013, Indian

ISSN 2394 - 7780

Railroads joined China, Russia, and the USA inside the elite institution of countries that deliver a thousand million tonnes of freight annually. IR is the second one-biggest passenger transporter within the international in terms of passenger kilometres dealt with, at the back of best Chinese language Railroads.. Regardless of its extremely good employer fulfilment, there hasn't been any appreciably multiplied common speed of both passenger and freight trains "so as to greenback this fashion, Indian Railroads has released the rate upgrade plan coupled with the removal of capacity restraints via doubling, tripling, and quadrupling of avenue sections." despite the fact that there were no passenger fatalities in 2019–2020 or 2020–21, Indian Railroads' overall performance at the protection the front is showing a nonstop enhancing fashion. But, with the increase in speed and the doubling, tripling, and quadrupling of the music, it's far hard to maintain this level of protection without offering professional useful resource to loco aviators. Signal Implicit hazards encompass passing at danger, breaking block guidelines, and failing to hold the desired speed by means of loco aviators. on the way to reduce the dangers Indian Railroads has planned to install the current teach manage device on its complete network in a timely manner in reaction to those events and to assist loco aviators in feeling self-assurance driving at excessive speeds. The 265 Km part of the Secunderabad division of South has correctly examined an independently advanced automatic train safety system on this surroundings [1].

Railroad significant. moreover, it has been accepted for in addition deployment of this system in the contiguous portion of Bidar - Prabhani, Manmad - Parbhani, Mudkhed - Secunderabad - Dhone - Guntakal (1200 RKMs), and paintings there may be efficiently progressing. advocated via the effects, a planned selection has been made to have all current train control systems paintings with a locally constructed train Collision Avoidance device (TCAS), that's now known as KAVACH.

2. Background Amtrak’s Acela:

Amtrak finished transforming the Northeast Corridor (NEC) into the first high-speed rail corridor in the country just last year. Although there are many definitions of high-speed rail, in this technical memo it is defined as passenger train operation at speeds more than 125 mph[2]. The installation of the new Acela trainsets is a key tenet of the high-speed NEC route. Amtrak had a simulator created to educate its locomotive engineers on the new Acela trainsets as part of the high-speed rail corridor improvement. The emphasis of the training was on acquainting engineers with the new Acela technology because Amtrak engineers were already familiar with the NEC terrain. Corys Training and Engineering Support Services (Corys T.E.S.S., sometimes known as Corys) created the simulator. along with Bombardier (for the remainder of the technical memorandum) (the original equipment manufacturer, or OEM, for the Acela trainsets).

The U.S. Congress subsequently ordered the FRA Office of Research and Development (OR&D) to assess Amtrak's Acela high-speed rail simulator in order to evaluate whether or not the FRA may use it as a research tool. The FRA has indicated interest in employing the Acela high-speed rail simulator to investigate various human factors or locomotive engineer-centred concerns, including training, alertness and weariness, new technology, and communication.

III. PURPOSE

3. Project Overview of kavach

The Pt. Deendayal Upadhyay - Manipuri - Pradhankhunta (DDU- MPO- PKA) segment of East crucial Railway, which is a part of the Delhi-Howrah HDN (high-Density network) path, desires to be ready with the ATP device "KAVACH" (TCAS) as a part of "project Raftar" to be able to boom the rate to 160 kmph from the contemporary a hundred thirty one of the busiest strains on the Indian Railroads is the Delhi-Howrah course. A crucial component of this Deendayal Upadhyay to Pradhankhunta is a portion of ECR. But, ECR most effective makes up 5 of the Indian Railroads, however it handles 10 in their business. It’ll permit the KAVACH-ready locomotives with loco aviators to perform in this congested direction. The cause of this design is to take use of the enjoy and advantages of in a position driving on a nonstop long stretch of approximately 411 km on the East crucial Railway[1].

4. Purpose of Amtrak

The investigation outlined in this specialized memorandum aims to ascertain whether the FRA can use the Amtrak Acela high speed rail simulator installation to look into issues related to exploration that are focused on people. To adapt the being simulator to the FRA's criteria, it was estimated that some changes would likely need to be made [2]. In light of this, the general design goals are as follows:

- Determine how well the Acela high-speed rail simulator is currently functioning.
- Identify simulator configuration restrictions that impede the capacity to undertake driver performance data collection and mortal-centred locomotive mastermind investigation.

ISSN 2394 - 7780

- To satisfy the FRA's exploration demands and docket, suggest changes to the Acela high-speed rail simulator.

An additional objective of this research is to offer design suggestions for a conventional high-speed rail system, as only a small fraction of all U.S. rail operations are thought to be high-speed. Simulator for a locomotive operating at a lower speed. A typical locomotive simulator would simulate various kinds of freight rail operations in addition to slower-speed passenger rail operations. The bulk of rail operations in the United States today are conventional. The FRA has not stated that it intends to create or support the creation of such a simulator. In the event that the FRA decides to construct or support the creation of a traditional locomotive simulator in the future, it is more interested in utilizing the knowledge obtained from the evaluation of the Acela simulator.

IV. OBJECTIVE

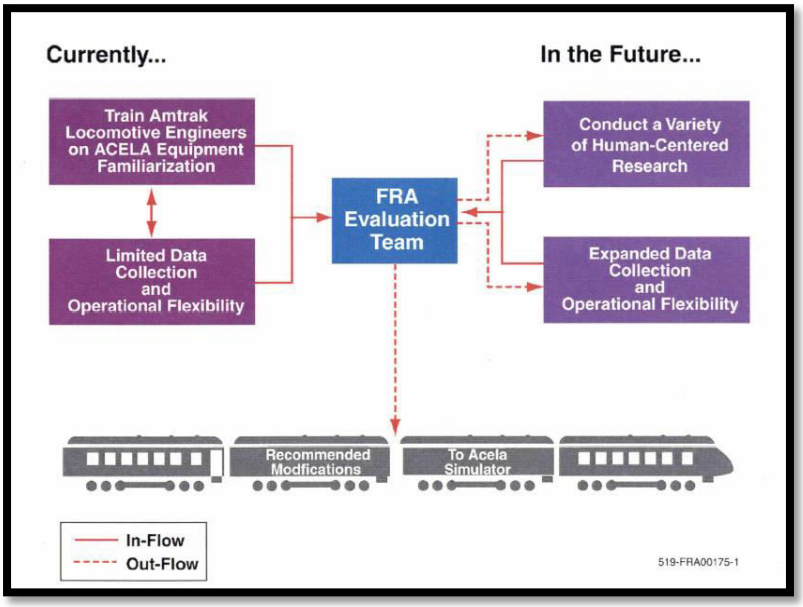
5. Project Objectives of Kavach

Further to leaving guests happy, secure deposit bins and incident-unfastened train operations boom the productivity of pricey everlasting and movable assets like tune, OHE, signalling systems, locomotives, rolling stock, and grave coffers. The Automated protection gadget (ATP) enables a shipping system accomplish the aforementioned dreams..

KAVACH is an automated educate safety machine that was created in-USA in accordance with the vision of "Aatma Nirbhar Bharat" and "Make in India" projects of the Honourable prime Minister. The cost-powerful overall performance improvement of Indian Railroads thru enhancements in both protection and efficacy might advantage KAVACH. cutting-edge train manipulate systems like KAVACH are a key element of Indian Railroads' "VIKAS YATRA" at a time whilst road control structures around the world have historically changed in reaction to outside elements (such injuries, technological breakthroughs, customer pride, and so on.). The KAVACH assignment will substantially make contributions to more secure and extra green teach operation on Indian Railroads.

6. Overall Approach of Amtrak

The overarching strategy adopted in this study focused on the transition between what the Acela simulators can currently perform and what the FRA wants it to be able to achieve in order to meet its research goals. An expert panel was assembled to assess the Acela simulator in its current design and offer suggestions for improving the simulator to meet the objectives of the FRA's research. Information about the FRA's study objectives was gathered using structured interviews with FRA programmed managers. The evaluation team then got together for a multi-day conference to assess the Acela simulator, go over the FRA's research objectives, and offer suggestions on how to modify the Acela simulator to satisfy the FRA's research requirements. Figure 1 shows this general strategy in detail. In Section 2, the technical strategy employed to conduct this research is covered in more depth.



Refigure 2. Overall approach to the evaluation of the Acela high speed rail simulator for FRA research purposes

ISSN 2394 - 7780

V. FEATURES

Features	Kavach	Amtrak
Security	Kavach provides end-to-end security for all users across the network.	Amtrak provides secure access to reservation and ticketing systems for customers.
Reliability	Kavach offers reliable and consistent performance with minimal disruption.	Amtrak provides reliable transportation services with a high level of safety.
Scalability	Kavach can be easily scaled up or down to meet the growing needs of its users.	Amtrak can quickly respond to spikes in customer demand.
Cost	Kavach is a cost-effective solution for businesses.	Amtrak offers competitive fares for customers.
Accessibility	Kavach is accessible from anywhere with an internet connection.	Amtrak provides access to its services through its website and mobile app.

Figure 3: Features of kavach and Amtrak

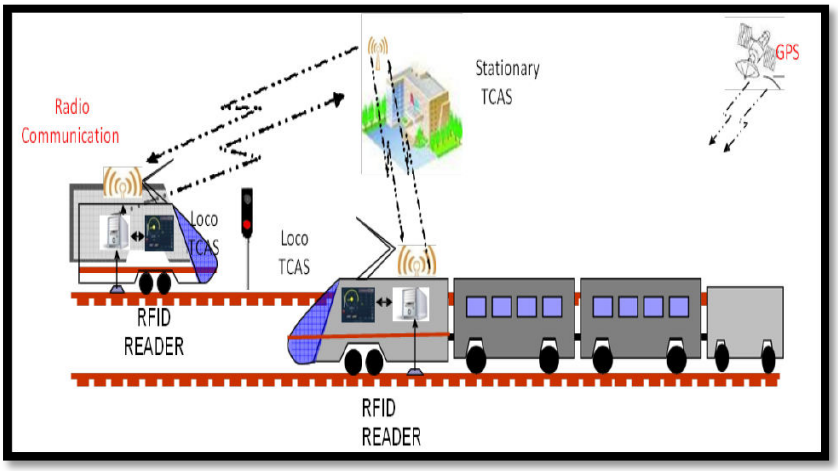
7. Features of KAVACH

Teach Crash Prevention gadget KAVACH, created by way of RDSO, is a country wide automatic teach protection system verified to SIL 4, the highest degree of safety Integrity. The machine warns the Loco Pilot as he methods the signal approximately their implications, ever-gift speed restrictions (PSR), and needs that he renowned the Advising in the event of overspending. The mechanism starts off evolved the boscage motion after a predetermined detention in the occasion that the Loco airman doesn't respond. the following are the device's key characteristics.

- The system pretends to be a signalling gadget and constantly updates the movement Authority.
- Section pace & Loco specific speed presentations component of signal at the DMI within the Locomotive to save you over speed.
- Guide SOS installation from the station and the locomotive in case of extremes.
- Underneath sure predetermined circumstances, the prevention of side collisions, head-on collisions, and avoid-quit collisions.
- Reverse, ahead, and rear movement safety. Automated effervescence upon LC Gates method.

Components of KAVACH

The trackside outfit, which includes a stationary TCAS Unit, and the on-board outfit, make up the train collision avoidance system.



Refigure 3. Components of KAVACH

The desk bound TCAS Unit to Interface with Signalling at Station EI/ RRI/ panel or LC Gate or IB or LSCs of auto Signalling, the Tower and Antennas, and the Radio device shall make up the Trackside Subsystem.

ISSN 2394 - 7780

A. An RFID tag.

B. The on-board subsystem need to include two RFID reader antenna hooked up inside the hot buttress and the loco TCAS important pc.

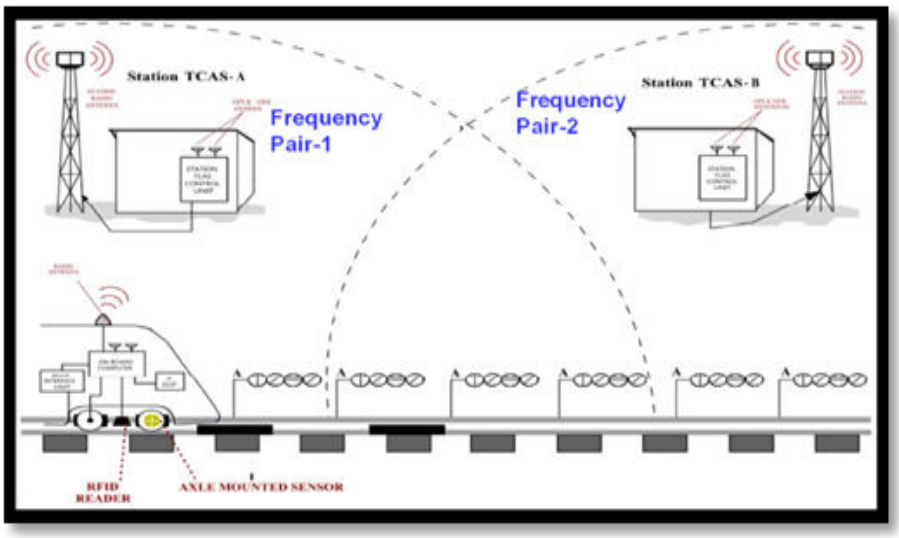
C. A loco TCAS radio unit that includes two radio modems in a heated buttress, every with its own wires and antennae.

D. A driver gadget interface (DMI) for each locomotive hack or a brake interface unit (BIU), as important, for every using motor trainer of an EMU, DMU, MEMU, or DEMU.

The Communication Back –
Bone in order to join the vibrant desk bound Station TCAS/ LC & IB TCAS/Rius gadgets in the computerized Signalling phase as needed, the device will also include a backbone of an optical fibre network.

5. Specifications and Future Developments:
The teach Collision Avoidance device as a way to be transported and established must observe the translation of three.2 OR rearmost found in RDSO Specification No. RDSO/ SPN/196/2012. The RDSO/ SPN/196/2020 interpretation4.0 of the draught specifications is now undergoing finalisation. The plan calls for a fee to be hooked up in the long run on version four. Zero as properly [1].

RADIO TELECOMMUNICATION SUB - SYSTEM:
The KAVACH (TCAS) native ATP gadget is based on continuous "Mama" (movement Authority) updating, which necessitates non-stop radio touch in a "communication obligatory quarter". The mandatory area covers the whole lot of the section for an automated Signalling home. The component in question now has a GSM-R grounded MTRC system, despite the RDSO preferred calling for using a UHF frequency grounded device. Due to the antiquated nature of GSM-R, the region is currently being evolved with a UHF Frequency grounded system in order to subsequently be able to transition to LTE-R once its miles applied on IR.



Refigure 4. RADIO TELECOMMUNICATION SUB - SYSTEM:

TOWER PLANNING:

Additional REQUIREMENT OF TOWERS:
Due to the requirement for steady Mama (motion Authority) updates and the segment's massive use of automated signalling, the best radio signal electricity will play a critical function in how well the layout works. To appropriately check the figures, place, and height of halls required for contemporary UHF grounded communique, it's miles deliberate to have an original software program-based totally simulation of entered sign strength (RSSI test). This FEASIBILITY file also consists of the unique take a look at record as Annexure-I. It’s far supposed to have a 30dB minimal fade perimeter. Later, the chosen contractor gives an intensive evaluation of the design. Additionally, these halls have to be designed to allow for a seamless transition to future KAVACH that is LTE-grounded. The use of MW can lessen the fee of designing and using the structures. And MTRC halls are each recommended. But, care should be required to determine the actual palace height that the TCAS radio outfit can use. [1]

ISSN 2394 - 7780

Refigure 5. TOWER STRUCTURE:

TOWER STRUCTURE:
It is recommended to have a point-particular palace shape in place of a chassis kind or, in some situations, a monopole type, based at the revel in of the South principal Railway and North significant Railway continuous structures. Before IR approves the shape, the contractor can efficaciously design it and get a third birthday party's approval. The contractor desires to be chargeable for upholding IR's moral standards. On the RDSO point, there are a few not unusual designs for the superstructure and substructure of the palace that can be used.

LOCATION OF TOWERS:
With the intention to select the placement and height of the new towers, a complete evaluation of radio communication energy inside the selected KAVACH quarter must be performed for the duration of the primary design phase. Halls for each station are either already gift or are presently being planned for the stations that do not yet have them. New halls will still be needed within the Block phase to preserve the proper UHF signal power inside the preliminary and closing LTE-R designs. Any new halls within the medical component need to be planned next to LSC/LC gates in compliance with all site visitors' policies. Comparable exams will want to be made in order to properly handle this.

SIGNALING INTERFACE:
Similarly to functioning as a signalling gadget, the ATP gadget "KAVACH" (TCAS) uses the equal input for automatic train safety. In the interim, all the stations from the DDU- MPO- PKA Absolute section have electronic interlocking, panel interlocking, and direction relay interlocking. The signalling input from the signalling system can be received in certainly one of two approaches. EI may be accessed without delay thru a SIL four protocol motor, or indirectly through the relay connections of the applicable ECR, NWKR, RWKR, and TPR relays. Considering that OEMs haven't but constructed protocol factories, a characteristic that would permit relays to offer input is being taken into consideration. In a comparable vein, it's far supposed to to begin with use Relay connections of the relevant relays and ultimately transfer to protocol factories as they arrive into existence.

Requirement for Building:
The proposed educate Collision Avoidance machine will interface with the Being Signalling device, in which the add-ons have to be established within the Absolute phase's Being Relay residences. But, there are a few places where the modern-day relay room might not have sufficient room to provide relay racks, TCAS device, electricity force additions, and many others. Therefore, a radical exam might be required, and the scope of the work must also take this into account.

Communication Backbone:
That allows you to join the colourful desk bound Station TCAS with the LC & IB TCAS and LSCs thru RIUs (far flung Interface units) inside the automated Signalling component as wanted, the ATP (computerized teach safety) system will even help the lower back-Bone of an Optical Fibre network. figures of at the least 24 middle OFC lines must be laid, one on every aspect of the music, if you want to join the RIUs LC TCAS to the Station TCAS according with demand. Telecom OFC is jogging a contrast in ECR to maintain music of the total for the section. It's been decided to install 24 fibre OFC lines for the automatic educate protection system inside the section underneath consideration. "RING REDUNDANCY" is used during the phase with area range. The compass of labour may even require that this OFC string be ended within the required stations. Further investigation will be required to determine whether or not the existing OFC's spare optic filaments are iced. It

ISSN 2394 - 7780

will likely be essential to plan ahead for shifting items to the brand new OFC if the same isn't always to be had in a specific phase.

Loco TCAS Planning:
The locomotives operating on the section must be outfitted with locomotive TCAS. Due to the impossibility of equipping all locomotives with Loco TCAS simultaneously, the following precedence has been chosen for prioritising the locomotives:

a. High-Speed Locos.

b. Mail express locomotives.

c. Locos for all other passengers.

d. Goods Locos.

The vacuity of these Locos will also be a factor, thus some precedent must be followed as closely as feasible.

Maintenance Strategy:
On Indian Railroads, KAVACH is a novel technology with little background in South Central Railroads. A fresh 4 times disfigurement liability period and a 15 times life cycle conservation contract are envisaged because this technology is new. Additionally, training by OEMs and training facilities on Indian Railroads is planned.

8. Features of Amtrak's Acela:

A. USA ATS System
Mechanical trip stops systems connected to rapid-fireplace conveyance lines built within the first half of the 20 th century are the structures that high-quality match the definition of an automatic train stop in the united states. The Interstate commerce commission (later the Federal Railroad management) has required ATS in view that 1951 as a minimum protection requirement to allow passenger trains to go faster than a pace restrict of seventy nine mph (127 km/h). The non-supervision requirement refers to a mechanism that requires the locomotive mastermind to reply to the caution inside a predetermined quantity of time before the thickets are automatically implemented each time the teach approaches a restricting wayside signal. Beginning inside the Twenties, the general Railway sign corporation produced the maximum widely used ATS for mainline roads, which consisted of inductive coils positioned without delay beyond the right-hand rail in regard to the direction of tour. The system's full call, Intermittent Inductive automated train prevent, is once in a while abbreviated as just ATS in road operating books to differentiate it from mechanical techniques available on the time. Following the advent of tune encrypted hack alerts inside the Nineteen Thirties, ATS's use as a educate safety medium declined.

B. Intermittent Inductive Automatic Train Stop (IIATS)
A train protection tool utilized in North American mainline roads and speedy-fireplace hearth automobile structures is the intermittent inductive automated teach stop (regularly referred to as IIATS or simply computerized teach stop or ATS) [2]. The ATS device's most frequent use case changed into to inform the street supervisor of an approaching threat and, if necessary, to forestall the educate using a full-provider operation of the thickets if the alert wasn't mentioned. Whilst related to indicators, the shoe might be amplified if the sign displayed the idea "clean." different signal hints would de-energize the shoe and motive the hack to head off. If A penalty boscage operation might be released if the mastermind did not silence the alarm inside 5 to eight seconds; it couldn't be stopped till the teach came to a whole forestall. (1) in contrast to mechanical teach stops or different train stop structures, IIATS turned into no longer commonly used to automatically stop a train if it passed a prevent signal, and in fact, it could not be used for this cause because the shoes have been placed just a few bases from the sign they protected and would not present sufficient retarding distance for the teach to prevent. As locomotives handiest have a detector to detect the shoes on one aspect of the train, on bi-directionally gestured lines, two "footwear" could be required, one for every route of tour. The receivers might also moreover built for simple junking to reduce fees while best a small phase of the road required ATS prepared locomotives or to lessen damage whilst operating in non-geared up houses. Every so often, "inert" inductors are hooked up as an alarm earlier to hurry limits or at system outstations to test the operation of the ATS system. IIATS has been used on numerous mild rail traces in a manner just like mechanical train stops, stopping the teach if it passes an absolute prevent sign. due to the fact mechanical passageways may be harmed by way of or obstruct freight operations and due to the fact light rail motors may additionally forestall much greater quick than mainline avenue trains, it's far helpful whilst light rail crosses tune with those trains. Without incorporating difficult sign overlaps.

ISSN 2394 - 7780

C. ATS

ATS is considered an obsolescent signalling system and is 80 times old. Inside the locomotive or hack auto the ATS sensor is principally a attraction. As the train passes the ATSÂ wayside inductorÂ the detector on the truck detects a glamorous field if there is not a clear or greensignal.Â when this happens thisÂ turns on a buzzer and an eight alternatetimer.Â the mastermind also must push a button to turn off the buzzer. If the button is not pushed within eight seconds the thickets will be applied and the train brought to a stop. The train does not go into exigency retardation. The Â ATS system primarily ensures that the mastermind is paying attention to the signals and isn't incapacitated [2].

VI. CONCLUSION

The Automatic Train Stop System in Kavach and Amtrak is an effective and reliable system that can help improve safety and efficiency on the railways. It can reduce the risk of accidents by alerting the train operator to any potential danger and can also help to increase the speed of the trains, enabling them to reach their destinations faster. The system is also cost efficient, as it requires minimal maintenance and is easy to install. The system is also highly reliable, as it can function even in the most extreme weather conditions. All in all, the Automatic Train Stop System is an excellent solution for both Kavach and Amtrak.

REFERENCES

[1] Indian Railways Feasibility Report for “Provision of Train Collision Avoidance System (KAVACH) along with two 24 Fiber OFC backbone in sections between Pt. Deendayal Upadhyay (DDU) and Pradhankhunta (PKA) of the East Central Railway”

[2] Mr. Stephen Reinach Foster-Miller, Inc.350 Second AvenueWaltham, MA 02154-1196 Prepared for: Dr. Thomas Raslear U.S. Department of Transportation Federal Railroad Administration Office of Research and Development Washington, D.C. 20593.

[3] Advisory Group for Aerospace Research and Development. (1980). Fidelity of Simulation for Pilot Training. Technical Report No. AGARD-AR-159. Neuilly sur Seine, France: North Atlantic Treaty Organization.

[4] Biirki-Cohen, J. Boothe, E., Soja, N., DiSario, R., Go, T. and Longridge, T. (2000). Simulator Fidelity—the Effect of Platform Motion. In Proceedings of the International Conference Flight Simulation—the Next Decade. May, 2000. London, UK: Royal Aeronautical Society. Pp. 23.1-23.7.

[5] Biirki-Cohen, J. Soja, N. and Longridge T. (1998). Simulator Platform Motion—the Need Revisited. The International Journal of Aviation Psychology, Vol. 8 (3), pp. 293-317.

[6] Gamst, F. (1991). Occupational Tasks and Responsibilities of Locomotive Engineer, Conductor, Engine Foreman, Brakeman, Switchman, and Train Baggageman. Unpublished Technical Report No. FCG-BLE-91-6.

[7] Lee, A. and Bussolari, S. (1989). Flight Simulator Platform Motion and Air Transport Pilot Training. Aviation Space and Environmental Medicine. Vol. 60, pp. 136-140.

[8] Prasad, J., Schrage, D., Lewis, W., and Wolfe, D. (1991). Performance and Handling Qualities Criteria for Low Cost Real Time Rotorcraft Simulators - A Methodology Development. In Proceedings of the 4 7th Annual Forum of the American Helicopter Alexandria, VA: American Helicopter Society.

[9] Orlady, H., Hennessy, R., Obermayer, R., Vreuls, D., and Murphy, M. (1988). Using Full- Mission Simulation for Human Factors Research in Air Transport Operations. National Aeronautics and Space Administration Technical Report No. NASA-TM-88330. Moffett Field, CA: NASA Ames Research Center

[10] Rehmann, A., Mitman, R., and Reynolds, M. (1995). A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research. U.S. Department of Transportation Federal Aviation Administration Technical Report No. DOT/FAA/CT-TN95/46. Washington, DC: U.S. Department of Transportation.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

RISE OF CYBERCRIME IN BANKS AFTER COVID 19

Fizza Jatniwala and Raza Ali Kadaya
Usha Pravin Gandhi College Vileparle

ABSTRACT
The centre released a broader economic support package worth Rs. 1.7 lakh crore on March 26, 2020, which includes the food security programme. Direct cash transfers, free gas cylinders for In order to defeat a deadly pandemic like covid-19, the Indian government announced full lockdown in the country from 24 March 2020 and was then extended till3 May 2020 by Indian Government with increasing covid cases number of cyberattacks also increased which as a result affect our Indian Banking System. Banking is the heart of the Indian economy. This article tries to assess the casual effect of pandemic like covid-19 on banks due to cyberattacks. This article shows a very serious impact of lockdowns and cybercrime on bank. In this paper we got an insight of cybercrime being the highest in banking sector and Social Engineering and Phishing attack being the highest in online banking sector.

I. INTRODUCTION
The coronavirus 2019 pandemic is a worldwide lethal disease caused by acute respiratory failure. According to Indian official resources, India had the second-highest number of confirmed cases in the world on 9 January 2023, with 44,681,318 recorded cases, and the third-highest number of COVID-19 deaths, with 530,721 deaths. As in Fig 1.1 we can see rise of covid cases in corresponding years. The World Health Organization forecasts that COVID-19 will cause 4.7 million extra deaths by May 2022,both directly and indirectly.

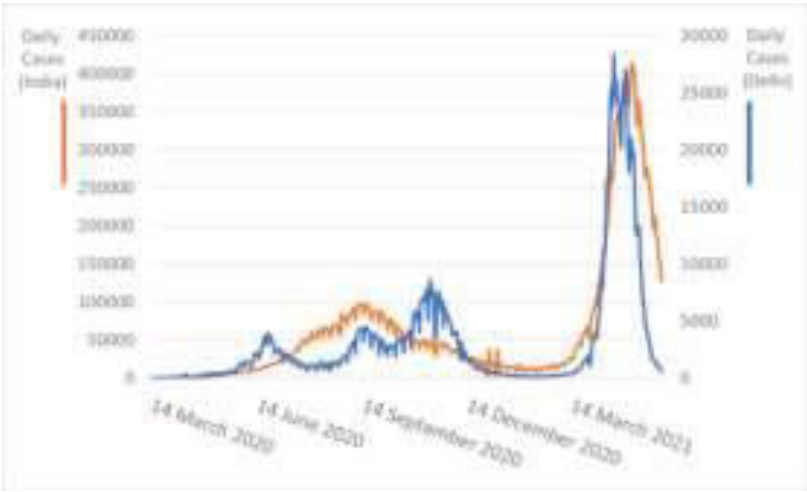


FIG 1.1 Rise of Covid Cases

Three months, and the RBI were also included in this, especially for migratory and daily wage workers. lowering repo rates, bringing in more cash, and enabling banks to impose a three- month loan hold. Then came the RBI. Repo rates will be lowered, liquidity will be added, and banks will be permitted to lower interest rates on all loans for the following three months. Through improved ways and means advances, tax payments were made simpler and states were granted short-term credit. The economic package for overdue pay was announced by the prime minister on May 12 and is worth Rs 20 lakh crore. This included previous government initiatives like the statements made on March 26 by the Finance Minister and RBI. On October 12 and November 12, the government unveiled two or more economic stimulus packages, bringing the total amount of stimulus to Rs 29.87 lakhcrore.

The second wave's economic impact has been less severe so far due to the second wave's constrained social movement restrictions as opposed to the first wave's lockdown measures. Socioeconomic indices like the need for electricity, labour force participation, and rail freight traffic decreased less in the second wave than in the first. The first wave strengthened domestic economic resilience, which is seen throughout the second wave despite the severity of the second wave. The Indian Finance Ministry stated that "Economic activity has learnt to work with covid" in their monthly economic review for April 2021, which was published on May 7, 2021. Poverty has risen since the pandemic's start in India, and livelihoods have been impacted.

A study conducted by Azim Premji University and released in May 2021 found that women who were employed before to the outbreak had a seven-fold higher risk of being fired than males did.

ISSN 2394 - 7780

Furthermore, among those who had previously lost their employment, women had an 11-fold higher likelihood of not going back to work than did men. The union health ministry's admission of two further confirmed cases caused a brief decline in the value of Indian financial markets on March 2, 2020. The greatest decline in Indian financial markets since June 2017 followed WHO's declaration of the pandemic on March 12, 2020. India's stock markets saw their biggest losses ever on March 23, 2020. The NSE NIFTY fell 1150 points (12.98%) and the SENSEX fell 4000 points (13.15%), on March 25, 2020, one day after the prime minister announced a complete 21-day lockdown, the SENSEX registered its greatest gains in over a decade. From October 2020 to April 2021, the domestic stock markets saw a positive surge.

Phishing, internet fraud, and the dissemination of false information were the criminals' immediate choice of the best tactics to profit from the pandemic and the disoriented public. By May 2020, the number of cyberattacks linked to the coronavirus had already risen to over 5,000 each day, from just a few hundred at the start of the pandemic. According to FBI reports, crime hassurged by 300%.

30% of all cyber attacks target healthcare companies and banks. Hacker attacks are increasing by 240% in both areas. Given the crucial services that both organisations offer, this is very troubling. For instance, the Brno University Hospital was allegedly the target of a cyberattack by the Czech Republic. To stop the onslaught, staff members were forced to shut down the whole IT system, postpone essential surgical procedures, and reroute seriously ill patients to other nearby hospitals. Think about the importance of the personal information held by both of these organisations. Such sensitive data loss or theft could have tragic consequences for the victims. According to PwC's Global Economic Crime and Fraud Survey 2022: India Insights, the Covid-19 pandemic's shift to remote working has exposed businesses to a new set of risks, with customer fraud (such as fraud involving mortgages, credit cards, claims, and checks) at the top of the list and reported by 47% of businesses.

The top crimes in India have also changed significantly over the last two years, with 45% of Indian organisations identifying cybercrime as the second-most common fraud. Additionally,34% of Indian businesses that experienced fraud, corruption, or financial or economic crime did not comply with KYC rules, placing them third overall.

Additionally, 95% of these 52% of Indian organisations experienced new types of fraud as a result of the disruption brought on by COVID-19, and 52% of Indian organisations experienced fraud or economic crimes in the past two years.

II OBJECTIVE

The purpose of this paper is to analyse the risks associated with both current and developing technologies while also exploring the procedures that may be used to continuously monitor the state of cyber security and new threats.

- Analyze the effect of cybercrime on the banking industry.
- Plans to research cutting-edge solutions to address the difficulties posed by the cyber threat
- Propose the implementation of various security protocols and standards, liaising with stakeholders and recommending the right use of policy.
- To study cybercrimes and its implications on theBanking Sector.
- Recognize the gravity of the cyberthreats posed to theInternet banking sector.
- To comprehend the causes and effects of cybercrime.
- To assess the degree of security and how it is being usedin the Internet banking industry.
- To evaluate and employ the available preventivemeasures to control fraud..

III RESEARCH METHODOLOGY

To arrive at logical conclusions or solutions to the research question, information/data already available from various sources is gathered and compared. White papers, government records, published research papers, journals, print media, findings from the RBI, NCRB, NTI Aayog, and CERT-IN, statistical data banks, and historical records are the most commonsources..

A secondary method of data collection and analysis is employedin this since a good collection of data is already present in written form. Performing a direct study might not be a practical choice given the subject and time restrictions. finding prior reports scenarios of cybercrime and a list of Historical prevention measures are used to build an anti-cybercrime framework. Morework has gone into adopting a case study approach to evaluate the

ISSN 2394 - 7780

effects of cyber attacks. With a focus on bank fraud cases in India after COVID, the research aims to pinpoint the parts of the banking process that are more susceptible to assault as well as the kinds of cyber-attacks that banks are likely to encounter frequently.

IV Indian Banking System

With about half of the nation's financial assets held by banks, banks are the main sources of credit in India's financial system. Since the 1970s, government-controlled banks have played a crucial role in India's development strategy by providing finance for sectors that the government prioritises, like infrastructure and agriculture (RBI 2005). The Indian government has made an effort to develop a domestic corporate bond market, although it is still quite small and mostly used by major businesses and financial institutions. Despite the fact that non-bank financial companies (NBFCs) have expanded recently as alternative finance intermediaries, banks continue to provide the majority of NBFC funding. Along with financing private and state-owned businesses, banks are a significant source of funding for governments. financing and purchasing of federal and state government debt. India's capital account has generally stayed closed, suggesting that the nation is still more dependent on domestic financing than other developing market nations.

India's banking sector is dominated by government-owned "public sector banks" (PSBs), which hold about 60% of the assets in the commercial banking system. Since the middle of the 2010s, these banks have struggled with non-performing loans (NPLs) and low capital levels. Despite significant failures in recent years, private sector banks have been more prominent over the past 20 years and, on average, have healthier balance sheets with lower NPL levels. Even while foreign banks only make up 7% of the assets in the commercial banking system, they are the ones with the best financial standing. Smaller banks that operate outside of the commercial banking system, such as rural cooperative banks, small finance banks, local area banks, and payment banks, cater to the needs of more specialised borrower groups.

In comparison to many other emerging market countries, India's credit to the non-financial sector is large at over 165 percent of GDP. The impact of banking sector stress on economic growth is amplified by India's high level of debt and reliance on bank loans. Although Australia and India have limited financial ties directly, Australia may be affected by possible flaws in the Indian financial system through trade. Only 0.6% of Australian foreign investment and 0.05% of Australian foreign investment goes to India. A few Australian banks operate a small number of subsidiaries in India. India, on the other hand, took in about 4% of Australia's exports in 2020.

V Digital Banking

E-banking, also referred to as online banking, virtual banking, or internet banking, is something we've all heard about. It is a system that enables online financial transactions including fund transfers, loan and EMI payments, cash deposits, and withdrawals to be made without physically going to a bank's premises. Customers can benefit from services including online banking, SMS banking, ATMs, mobile banking, e-cheques, and debit/credit cards by using e-banking.

Digital banking is another term that is widely used as a synonym for e-banking. Figure 1.2 shows a growth in the number of people using digital banking. Both terms are interchangeable. Digital banking, however, doesn't or hardly rarely use paper money. Although it is still extensively used, paper money. Nowadays, ATMs are a crucial component of the banking system since they enable customers to withdraw money whenever they need it. In a digital economy, there aren't many exchanges of real money.

Cash is typically perceived as being gratuitous. However, utilising cash comes at a significant cost. The cost of cash is examined in a report published in the Harvard Business Review titled "The Countries That Would Benefit the Most From a Cashless World." The price of cash includes the price paid by consumers, businesses, banks, and other institutions, as well as the price paid by tax income and the price of printing money. In a comparison of cash costs worldwide, India ranks among the highest.

The majority of Indian banks have developed their net banking and mobile banking websites to give customers access to practically all banking products online. Today, accessing dependable and useful banking services is frequently done through internet banking.

A digital payment instrument called internet banking, often known as net-banking or online banking, enables a bank or financial institution's customer to undertake commercial or non- commercial transactions online over the internet.

Almost all banking services that were previously only accessible through a local branch can now be accessed online by customers through this provider, including fund transfers, deposits, and online bill payments.

ISSN 2394 - 7780

Any financial foundation or an operational ledger can be accessed by anyone who has enrolled for internet banking at the bank andhas a computer with internet access. Once a customer registers for online financial services, they are no longer need to visit the bank whenever they require financial assistance. It is not only a useful way to do banking, but it is also safe. Access to net financial gateways requires creative user/customer IDs and passwords. Anyone who utilises any financial institution, has a current bank account, or has enrolled for online banking at that institution is eligible to access internet banking. A consumer no longer needs to go to the bank each time they want to utilise a banking provider after signing up for online banking services. It is not only practical, but it is also a trustworthy way to do banking. Websites for online banking are safeguarded by specific User/Customer

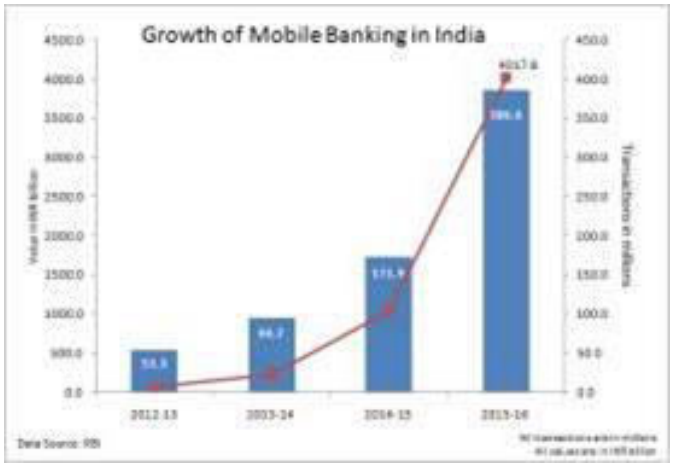


FIG 1.2 Rise of Mobile Banking Users

VI UNIFIED PAYMENT INTERFACE

UPI is a smartphone application with multiple bank accounts that integrates unrestricted financial routing for merchant payments along with other banking capabilities. The value of the 180 crore UPI transactions that were made in September 2020 exceeded Rs 3 lakh crore. Compared to the same quarter last year, UPI payments saw increases of 82% and 992%. In Q2, 19 banks joined the UPI ecosystem, with UPI services covering all banks. As of September 2020, 146 banks' customers began receiving NBCI's PIM App at the same time. With over 180 billion transactions in September 2020, UPI had surpassed value in general by Rs 3 lakh crore. UPI transactions will significantly increase over the holiday season. The main banking system run by a bank is inherently linked to or a component of the internet banking system, and this is in accordance with the branch .This is the conventional manner for clients to access banking services.

VII CYBER CRIME IN DIGITAL BANKING

FIGURE 1.3 proves that cyberattack has mostly affected banking sector. Some of the crimes in internet banking sector are:

Identity theft is a common tactic employed by thieves when dealing with electronic businesses, particularly Internet banking services. This involves using someone else's identification, such as their name, date of birth, and address, to conduct fraud. Cyber thieves can then use the identity theft information they've obtained to create new bank accounts, get credit cards or loans, or apply for government benefits.

Phishing is a method that imposters and cybercriminals use to deceive victims into exposing their personal and financial information. Cyber criminals employ a range of phishing techniques, but one of the most popular ones involves sending phishing emails to Internet banking users by pretending that a legitimate business or organisation is providing online services.

Vishing with speech is a technique used by computer fraudsters to gain financial information and customer information from online banking users by employing a phoney call centre and voice over IP (VOIP) technology. To achieve this, scammers utilise an email system to request that users of internet banking confirm their banking information and other details as part of a security routine check at the phone number provided in the phishing email.

The main danger posed by internet thieves to users' accounts is malware, which they employ to steal their bank information and other sensitive data (viruses, worms, trojans, and other threats). More harmful software is being developed as mobile devices like smartphones and tablet computers proliferate exponentially.Automating online banking fraud: With the use of Automatic Transfer Systems, cybercriminals and computer fraudsters have recently advanced the situation. Using Spy Eye and Zeus malware versions as part of Web inject files, which are

ISSN 2394 - 7780

transcript files with a lot of JavaScript and HTML coding, a new system has been started for automating online banking fraud.

Social engineering is the practise of coercing people into disclosing private information or taking specific behaviours. Social engineering is a social science field that is regularly used by cybercriminals and computer fraudsters to get sensitive information and financial data.

Web applications typically make use of cross-side scripting. As a result, attackers can add client-side scripts to web pages that users are viewing. This is how attackers get around access restrictions.

A "BOTNET" attack refers to a network of personal computers that have been infected with malicious software and are being managed by a group without their owners' knowledge.

Ransomware: This is one of the biggest risks in the online environment. Such malicious software aims to block access to a computer or set of computers unless a specific sum of money is paid.

FIG 1.3 Affect of different attacks on different sectors

VIII Analysis and Statistics:

The amount of loss brought on by cyber crime, which includes frauds involving ATM/Debit card, Credit card, and Internet Banking, increased to Rs 63.40 crore in FY20-21 from Rs 58.61 crore in FY19-20, the government informed parliament on Monday.

The customer will not be held responsible for the financial fraud involving cards and online banking if they notify the bank of an unauthorised electronic transaction within three working days of receiving information about it from the bank. The bank will also credit the customer's account with the amount involved in the transaction.

The customer's maximum liability will also be between Rs 5,000 and Rs 25,000 if they report the unauthorised electronic transaction within four to seven working days; if they report it after seven days, the liability will be calculated in accordance with the bank's board-approved policy.

Aside from that, the RBI has given banks instructions to ensure that complaints in this regard are handled and any consumer liability is established within the 90-day window permitted by the bank's Board-approved policy.

The publication of information security best practises for public officials, the organisation of cyber safety and security events, the distribution of messages on cybercrime via short message service (SMS), radio campaigns, and social media accounts of the Indian Cybercrime Coordination Centre are just a few of the initiatives made to increase awareness of cybercrimes (I4C).

Banks must conduct an annual investigation into frauds and report the results, together with the amount that was recovered, to their boards of directors and regional advisory boards. Additionally, banks are required to establish a special committee of the Board for monitoring and following up on fraud cases involving amounts of Rs 1 crore and above (SCBF), which keeps track of the development of the CBI/Police investigation and the status of the accounts' recovery.

ISSN 2394 - 7780

Banks With Highest Number of Fraud Cases

BANK	NO OF FRAUD	AMOUNT
IDBI Bank	4	2,227
Bank ofBaroda	4	2,011
ICICI Bank	4	1,908
Union bank ofIndia	4	1,742
Canara Bank	5	2,658
Indian Bank	7	1,682
State bank ofIndia	8	3,902
Punjab National Bank	10	4,820
Yes Bank	11	3,869
Bank of India	13	3,925

IX POSSIBLE PREVENTIVE MEASURES:

Analyse Cloud Security:
Make sure your cloud infrastructure is up to date by periodically checking its state. Check your cloud security's current state, best practises, and compliance requirements. Cloud systems and infrastructure can be made more secure by using multifactor authentication.

Monitor Cloud Security:
To automate threat identification and protect against potential assaults before they cause an issue, employ a vulnerability management tool.

Establishing Strict Access Management Policies:
Don't give access to independent contractors, part-timers, etc.; only let workers who actually need it have it. By implementing strict access management policies and granting permissions to employees who want them, you can secure your business from inside.

Increasing Awareness Among Employees:
Banks must put in place a comprehensive training programme that staff members can use in the event of a cyberattack..

Disaster Recovery plan:
By having a backup plan, you may reduce any downtime following a disturbance and prevent data loss. Only if you routinely backup your data may this be used.

Encrypt Your Data:
One approach to encrypt data is through cryptography, which guarantees that your most valuable digital assets are always safe.

Cybersecurity Training:
For cybersecurity professionals, cybersecurity training is a requirement to increase their understanding of relevant material,assess their cyber-awareness, and keep them up to date..

X EFFECT OF COVID ON BANK:

The moratorium offered by the Reserve Bank of India (RBI) for debtors affected by COVID-19 benefited clients who made up 40% of all outstanding bank loans as of August 31, 2020. The majority of industries reported fewer outstanding loans subject to the moratorium in August 2020 compared to April 2020, but micro, small, and medium-sized businesses (MSMEs) saw a slight increase, and in August 2020, 78% of MSME clients were using moratoriums, underscoring the pressure this sector was under. PSBs and non-banking financial companies faced the heaviest hit from the upcoming stress, while urban cooperative banks (UCBs) and non-banking financial companies followed, according to the distribution of the moratorium requested for MSME loans (NBFCs). Small financing banks (SFBs) have the largest share of the moratorium granted for outstanding personal loans, followed by UCBs and NBFCs. In April 2020, about two-thirds of all PSB customers and half of all PVB customers chose to postpone making payments (RBI 2020a). As shown in FIGURE 1.4 we can see mostly phishing and social engineering attacks on bank with 60% .As of August 31, 2020, this was reversed, with PSBs having a sizable customer base across categories (primarily individuals) opting out of the moratorium while PVBs had a larger customer base under the moratorium than the other categories of lenders. This was primarily due to a fourfold increase in their MSME customers taking advantage of the benefit (RBI 2020).

ISSN 2394 - 7780

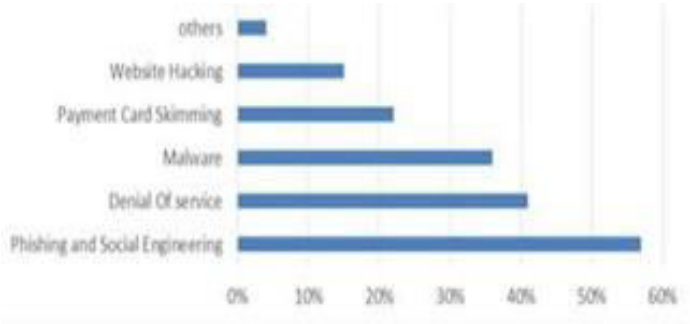


FIG 1.4 Types of attack on banking

Large borowal accounts (exposure of 5 crore and above) accounted for 53.7% of all loans and 79.8% of all non-performing assets as of September 30, 2020. The ratio of restructured standard assets to gross non-performing assets (GNPAs) and the ratio of restructured standard assets to total financed amounts both exhibited a falling trend for PSBs between 2019 and 22. On the other hand, PVBs observed a higher proportion of NPAs in these accounts. The percentage of special mention accounts (SMA-0) dramatically increased in September 2020. This might be the first sign of stress after the ban is lifted on August 31, 2020.SMA-1 and SMA-2 categories of SMAs, however, continued to make up a relatively smaller portion of the overall SMA population (RB1 2020c).

XI CYBERATTACK AFTER COVID:

Indian banks reported 248 successful data breaches by hackers and criminals between June 2018 and March 2022; the government alerted Parliament on August 2, 2022.

11,60,000 cyberattacks were reported by the Indian government in 2022. It is predicted to increase by three times from 2019. India has been a victim of significant cyberattacks, including the phishing attempt that almost led to a fraudulent transaction worth $171 million against the Union Bank of India in 2016.

Union Bank of India was the victim of a hack that involved online banking and caused a large loss. One of the officials clicked on a questionable link in the phishing email, which gave the malware access to the system. Using phoney RBI IDs, the attackers gained access to the system.

Banks are required to improve their IT risk governance framework, which calls for the Chief Information Security Officer to take a proactive role in addition to that of the Board and the Board's IT committee in ensuring that the appropriate standards are being followed.

In just two years, the number of cyberattacks in the nation has tripled, but the funds designated for cybersecurity have been underutilised, with only Rs 98.31 crore of the total Rs 213 crore sanctioned being used.

Government statistics indicates that 3,94,499 cyber security incidents were recorded overall in 2019 by the Indian Computer Emergency Response Team (CERT-In). In 2020, the number climbed to 11,58,208, and in 2021, it rose again to 14,02,809. 6,74,021 cyber security incidents were reported in 2022 up till June.

The cyber attack that took down the systems at the All India Institute of Medical Sciences (AIIMS), Delhi on November 23 is still not fully resolved.

Investigations into the cyberattack on the country's most important installation have been conducted by numerous entities.

The Ministry of Jal Shakti's Twitter account was momentarily hacked by cybercriminals on December 1, marking the second such cyberattack on a government website.

The surge in cyber security events is a global issue because of the unbounded nature of cyberspace, its anonymity, and its rapid expansion.

In its report this year, a parliamentary standing committee noted that there has been a considerable increase in cyber events and cyber security breaches. It is crucial that the nation's capacities and resilience are increased accordingly to deal with impending hazards in cyberspace.

Cybersecurity must continue to be a top priority for the Ministry, and no budget shortage should prevent it from providing a secure ecosystem for the cyberworld. In light of recent challenges in this area, the committee believes the Ministry must coordinate its efforts to create a more secure cyberspace.

ISSN 2394 - 7780

The government is proactively gathering, analysing, and sharingtargeted alerts with organisations across sectors for proactive threat mitigation activities by them, according to the Information Technology Ministry, which is running an automated cyber threatexchange platform.

Regarding the Chief Information Security Officers' (CISOs') primary tasks and responsibilities for compliance, the government has published guidelines for CISOs.

Additionally, before being hosted, all government websites and applications undergo a cyber security audit.

After hosting, there is also a frequent assessment of the websites and applications. Additionally, 97 security auditing organisations have been appointed by the government to support and monitor the application of information security best practises.

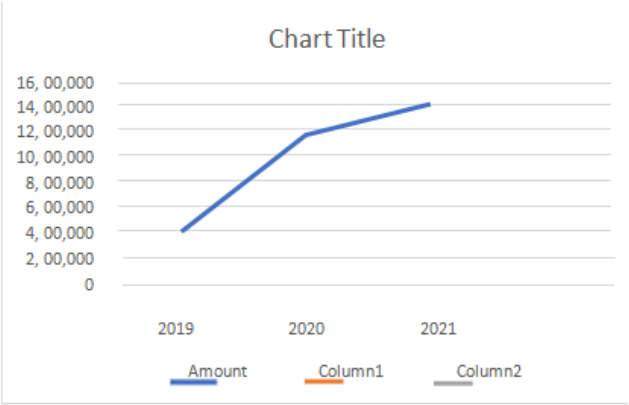


Figure 1.5 The above graph shows increase in number of cyberattacks from 2019-2021.

XII FINDINGS

- Identity theft and hacking have been the main causes of cybercrimes in this industry.
- Since banks hold all of the reserves in the form of cash, banks are frequently the focus of attacks.
- A dramatic shift in household risk and savings is anticipated tooccur in the event of an epidemic like Corona.

The majority of attacks were DDOS attacks. • An epidemic like Corona may boost the demand for borrowing from banks, reducing secured lending like personal loans or credit card

The law enforcement should be highly strict and updated frequently to maintain track of such crimes. Internet banking customers should use strong passwords and distinct username combinations for different sites and accounts.

- Using big data banks, the government should also keep tabs onhow the operating networks are being used.
- Awareness initiatives should be started to educate the public about the current situation and impending threat. • Punishments and penalties must be applied consistently in order to reduce theimpact of these issues and penalise the attackers.

XIV CONCLUSION

The goal of this study is to comprehend and pinpoint the securityconcern when using Internet banking services. The criminals of today attempt to carry out these new crimes utilising computers,

the Internet, and other forms of cyberspace abuse. As a result, cybercrime is growing as a menace. The distinctive and alluring characteristics of cybercrime have begun to progressively exceed those of regular criminality. Cybercriminals find the degree of anonymity, the potential for widespread victimisation, and quick results to be among the most alluring. Due to their ignorance of the most recent assault tactics and established preventative mechanisms, unaware consumers are easily duped. In order to provide quicker and better cybercrime investigation outcomes, itis a step further to engage qualified cyber security professionals.

REFERENCES

1] COVID-19 pandemic in India - Wikipedia

2] PNB: Punjab National Bank reports maximum bank frauds amounting to Rs 4820 cr in first nine months of FY20-21 | India Business News - Times of India (indiatimes.com)

3] 4 Biggest Cyber Security Threats for Indian Banking Sector |Great Learning (mygreatlearning.com)

ISSN 2394 - 7780

4] Impact of COVID-19 on the Indian Banking Sector | Economicand Political Weekly (epw.in)

5] Cyber attacks in India triple in last three years, but securityfunds underutilised : The Tribune India

6] Cybersecurity in Banking Sector: Importance, Threats,Challenges (knowledgehut.com)

7] Machine learning in Banking Risk Management : A LiteratureReview Martin Leo*, Suneel Sharma and K. Maddulety

8] Impact of cyber-attacks on banking institutions in India: A study of safety mechanisms and preventive measures Suman Acharya, Sujata Joshi

9] COVID-19 : Impact on Banking sectorPoonam Sharma, Dr.Neha Mathur

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

WHY INDIA IS NOT ALLOWING CRYPTOCURRENCY?

Dr. Neelam Naik[1] and Mr. Karan Desai[2]
[1]Assistant Professor and [2]Student, Master of Science in Information Technology (MSc-IT), SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, Maharashtra

ABSTRACT
Bitcoin and other cryptocurrencies have evolved from being solely associated with geeks and radicals to being taken into consideration by central banks as a method to adopt digital money. Cryptocurrencies can be completely transferred between digital addresses and are only available in digital form. This is different from both conventional corporeal money, which may be physically held, and conventional electronic money, which functions as a debt claim on a deposit with a reputable financial institution like a private bank. This implies that, despite being open to interpretation, any legal rights related to possessing bitcoins must be distinct. This chapter examines the many ways that money is treated legally and discusses the hazards involved by using examples from everyday life. We draw the conclusion that fraud committed through hacking could potentially hinder the widespread use of cryptocurrencies since the lack of legal remedy against a third party like a bank focuses risk on cryptocurrency holders. Therefore, users should proceed with caution and be aware of the hazards before purchasing cryptocurrencies. It is important to emphasize this warning since many people believe that the technology's cryptography protects them from fraud when, in reality, it does not.

Keywords: Cryptocurrency, Block chain, Challenges in cryptocurrency, Types of cryptocurrency.

INTRODUCTION
According to market demands, the exchange instruments used to facilitate trade transactions have undergone significant development and modification. Money refers to the tools that are used to facilitate the exchange of products. the use of money as a medium of exchange, an accounting unit, and a value storage. In the sense that we all agree to accept it while conducting transactions, money is a medium of exchange. Both employees and merchants consent to accepting payment sin exchange for their goods and services.

Money serves as a straightforward tool for identifying and expressing value in accounting. Money acts as a store of value by enabling us to save the benefits of our work or business in a handy object. From the time of barter to the arrival of cryptocurrencies like Bitcoin and Ethereum, as well as metal and coin money like gold and silver, modern monetary systems and checks, and ending with the most recent advances in world currencies. The introduction of cryptocurrencies has transformed the global payment system on a scale that was unthinkable only a few years ago. A cryptocurrency is a form of digital or virtual money that is secured by encryption. E-cash, a form of anonymous cryptographic electronic money, was created in 1983 by American cryptographer David Chaum. Later, in 1995, he put it into practise using Digicash, a pioneering electronic payment system that used encryption and required user software to withdraw money from a bank and select certain encrypted keys before it could be transmitted to a receiver.

As a result, neither the government nor any other entity, including the issuing bank, could track the digital currency A cryptocurrency's security characteristic makes it challenging to counterfeit. The absence of a central issuing body for a cryptocurrency is one of its distinguishing characteristics. It has zero centralised control.

STATEMENT OF THE PROBLEM
Cryptocurrency is not officially recognized as legal tender in India and its use is currently restricted by the Reserve Bank of India (RBI). The RBI has issued several warnings to the public about the potential risks associated with investing in cryptocurrency and has also instructed banks and financial institutions to refrain from providing services to businesses and individuals dealing in cryptocurrency. .The government has also set up a panel to study the regulatory framework for virtual currencies and the potential use of blockchain technology. The panel was tasked with studying the issues related to virtual currencies and making recommendations for the regulatory framework. The main reasons cited for the restrictions on cryptocurrency in India include concerns about money laundering, financing of illegal activities, and consumer protection. The government also believes that the decentralized nature of cryptocurrency makes it difficult to regulate and monitor. Additionally, the government is also considering launching its own digital currency, which may have further implications for the use of private cryptocurrencies in the country.

LITERATURE REVIEW
A literature review on the reasons for the restrictions on cryptocurrency in India would likely examine a variety of sources, including government reports, academic studies, and news articles. One key source would be the

ISSN 2394 - 7780

Reserve Bank of India's (RBI) statements on the matter. The RBI has issued several warnings to the public about the potential risks associated with investing in cryptocurrency and has also instructed banks and financial institutions to refrain from providing services to businesses and individuals dealing in cryptocurrency. The RBI has cited concerns about money laundering, financing of illegal activities, and consumer protection as some of the reasons for the restrictions. Another important source would be studies and reports from government agencies and committees. For example, the Indian government set up a panel to study the regulatory framework for virtual currencies and the potential use of blockchain technology. The panel was tasked with studying the issues related to virtual currencies and making recommendations for the regulatory framework.

Academic studies on the topic would also be important to include in the literature review. These studies would likely examine the potential benefits and risks associated with the use of cryptocurrency in India and would provide insights into the economic and legal considerations surrounding the issue. Finally, news articles and other media sources would provide a valuable perspective on the current state of cryptocurrency in India and the government's stance on the matter. These sources would give an overview of the current scenario and the reaction of the public and the government towards the cryptocurrency. Overall, a literature review would likely find that the Indian government's restrictions on cryptocurrency are primarily driven by concerns about money laundering, financing of illegal activities, and consumer protection. The decentralized nature of cryptocurrency, which makes it difficult to regulate and monitor, is also a major factor in the government's decision.

Types of Cryptocurrency

The purpose of cryptocurrency is to function as a medium of exchange. Over 1600 cryptocurrencies are currently accessible online, and that figure is rising. Any time a new cryptocurrency is invented. Bitcoin is now the largest blockchain network by market capitalization, followed by Ripple, Ethereum, and Litecoin.

Bitcoin (BTC)

Bitcoin is one of the most well-known currencies and is regarded as the first cryptocurrency. It was Bitcoin was developed in 2009 as an open-source programme that uses block chain technology to peer-to-peer transactions that are transparent. Although these transactions are safe, anyone may observe them.by using the blockchain's algorithm. The transaction is visible to everyone, but only the owner of that Each owner of Bitcoin receives a "private key" that may be used to decode data. There is no central authority, unlike a bank authority in the Bitcoin world. Users of Bitcoin are in charge of transmitting and receiving money, enabling global transactions to be conducted anonymously.

Litecoin(LTC)

Litecoin was introduced in October 2011 as a Bitcoin substitute. Litecoin is a peer-to-peer cryptocurrency and open source software project made available under the MIT/X11 licence, just as other cryptocurrencies.It is entirely decentralised and its generation and transmission are based on an open source encryption technology.Litecoin and Bitcoin are dissimilar in various respects. These digital currencies differ from one another in a key ways:While Bitcoin takes 10 minutes to process a block, the Litecoin network aspires to do it every 2.5 minutes.Due to this, Litecoin may confirm transactions more quickly.There are a total of 21 million Bitcoins and 84 million Litecoins in circulation.

Ethereum (ETH)

The cryptocurrency Ethereum was first put out by researcher and programmer Vitalik Buterin in late 2013. It was first made available in July 2015. It is a platform that is open source and built on blockchain technology. Ethereum blockchain focuses on executing the computer code of any decentralised application in addition to maintaining ownership of digital currency transactions, enabling application developers to utilise it to pay for transaction costs and services on the Ethereum network.

D Ripple (XRP)

Ripple is a network for real-time gross settlement, currency exchange, and remittances developed by the US-based business Ripple Labs Incorporation. The cryptocurrency and digital payment network Ripple was introduced in 2012 and serves as a platform for financial transactions. It is a worldwide settlement network created to make money transfers quick, safe, and affordable. Ripple links to banks and facilitates the conversion of any money, including USD,

Cash Bitcoin

A new kind of digital money called Bitcoin Cash was developed to enhance several aspects of Bitcoin. Block sizes were expanded by Bitcoin Cash to accommodate more transactions. Bitcoin, gold, and EUR. This makes it different from other currencies. Another way that Ripple differentiates from other digital currencies is by emphasising large-scale money transfers rather than one-on-one transactions.

ISSN 2394 - 7780

Dogecoin

Dogecoin is a decentralized, open-source cryptocurrency that was created in 2013. It is based on the popular "Doge" internet meme, featuring a Shiba Inu dog as its logo. Dogecoin is similar to Bitcoin, but it has a faster block time (1 minute compared to Bitcoin's 10 minutes) and a much larger maximum supply (unlimited compared to Bitcoin's 21 million). Dogecoin was originally created as a joke and was not taken seriously by many people. However, over time it has gained a strong community of supporters and has been used for various charitable causes and online tipping.

Dogecoin transactions are confirmed by a network of users who contribute their computing power to solve complex mathematical problems, a process known as mining. Dogecoin uses a scrypt algorithm which make it less energy-intensive than other cryptocurrencies like Bitcoin, which uses a SHA-256 algorithm. In recent years, Dogecoin has seen a significant increase in value, but it is considered a highly volatile asset and its value can fluctuate greatly. It's important to note that it is considered a speculative investment and one should do thorough research before investing in any Cryptocurrency.

G Tether (USDT)

Tether (USDT) is a stable coin that is pegged to the value of the US dollar. It is issued by Tether Limited, a company based in Hong Kong. Tether is built on the Bitcoin blockchain, using the Omni Layer Protocol, which allows for the creation of digital assets that can be stored and transferred in a decentralized manner. Tether is used to facilitate trading in cryptocurrency markets, as it allows traders to move in and out of other cryptocurrencies without having to convert to fiat currencies. This can help to reduce volatility and increase liquidity in the market. However, Tether Limited has faced controversy and scrutiny over the lack of transparency of its reserves and the relationship with its associated exchange Bitfinex

Indian Economy & Crytocurrency

The Indian economy is the fifth- largest in the world by nominal GDP and the third- largest by purchasing power parity. The service sector makes up a majority of the GDP, followed by the industrial and agricultural sectors. The Indian government has traditionally been cautious about the use of cryptocurrencies, with the Reserve Bank of India issuing warnings about the potential risks and a ban on banks dealing with cryptocurrency-related transactions. In March 2020, the Supreme Court of India lifted the ban, allowing banks to again transact with cryptocurrency exchanges and businesses, though the government is still considering a bill that would ban cryptocurrencies altogether.

DISCUSSION

The Indian government's stance on cryptocurrency can be largely attributed to the concerns about money laundering, financing of illegal activities, and consumer protection. The decentralized nature of cryptocurrency, which makes it difficult to regulate and monitor, is also a major factor in the government's decision. The Reserve Bank of India (RBI) has issued several warnings to the public about the potential risks associated with investing in cryptocurrency and has also instructed banks and financial institutions to refrain from providing services to businesses and individuals dealing in cryptocurrency. This has made it difficult for individuals and businesses to access banking services if they are involved in cryptocurrency transactions. Additionally, the government has set up a panel to study the regulatory framework for virtual currencies and the potential use of blockchain technology. The panel has been tasked with studying the issues related to virtual currencies and making recommendations for the regulatory framework. This suggests that the government is still in the process of evaluating the implications of cryptocurrency and how it can be regulated in a way that addresses their concerns. Another important factor that may play a role in the government's stance on cryptocurrency is the potential launch of its own digital currency. The government is reportedly considering launching its own digital currency, which would have implications for the use of private cryptocurrencies in the country. The government may be wary of allowing private cryptocurrencies to operate in parallel with its own digital currency. In summary, the Indian government's restrictions on cryptocurrency are primarily driven by concerns about money laundering, financing of illegal activities, and consumer protection. The decentralized nature of cryptocurrency, which makes it difficult to regulate and monitor, is also a major factor in the government's decision. The government is still evaluating the issue and considering the regulatory framework for virtual currencies and the potential use of blockchain technology.

CONCLUSION

The Indian government's stance on cryptocurrency is primarily driven by concerns about money laundering, financing of illegal activities, and consumer protection. The decentralized nature of cryptocurrency, which makes it difficult to regulate and monitor, is also a major factor in the government's decision. The Reserve Bank of India (RBI) has issued several warnings to the public about the potential risks associated with investing in

ISSN 2394 - 7780

cryptocurrency and has also instructed banks and financial institutions to refrain from providing services to businesses and individuals dealing in cryptocurrency. The government has set up a panel to study the regulatory framework for virtual currencies and the potential use of blockchain technology, which suggests that they are still in the process of evaluating the implications of cryptocurrency and how it can be regulated in a way that addresses their concerns. The potential launch of its own digital currency may also be a factor in the government's stance on private cryptocurrencies. It's worth noting that the Indian government's decision on cryptocurrency is subject to change as the technology and the market evolves. The government is closely monitoring the developments in the cryptocurrency space and may reconsider its stance if the situation changes. Overall, it can be stated that the Indian government's restrictions on cryptocurrency are a precautionary measure to protect the interest of its citizens and economy. As the technology and its use evolve, the government may consider revisiting its stance and create a regulatory framework that balances the interest of its citizens and the technology's potential.

REFERENCES

[1] Bearman, J. (2015, May). The Untold Story of Silk Road, Pt. 1. Retrieved from Wired.com

[2] Bitcoin: A New Global Economy. (2015, August 4). Retrieved July 2016, from BitPay, Inc.

[3] Bovaird, C. (2016, June 24). Bitcoin Rollercoaster Rides Brexit As Ether Price Holds Amid DAO Debacle.

[4] Gerber, R. (2015, January 29). Why Apple Pay And Dollars Are Killing Bitcoin. Retrieved from Forbes Investing

[5] Hileman, G. (2016, January 28). State of Bitcoin and Blockchain 2016: Blockchain Hits Critical Mass

[6] Desjardins, J. (2016, January 5). It‟s Official: Bitcoin was the Top Performing Currency of 2015. Retrieved from The Money Project

[7] Hofman, A. (2014, March 6). The Dawn of the National Currency – An Exploration of Country-Based Cryptocurrencies

[8] Kar, I. (2016, June 30). Everything you need to know about the bitcoin „halving‟ event

[9] Kasiyanto, S. (2016). Bitcoin's potential for going mainstream. Journal Of Payments Strategy & Systems, 10(1), 28-39.

[10] Kelly, B. (2014). The Bitcoin Big Bang : How Alternative Currencies Are About to Change the World

[11] King, R. S. (2013, December 17). By reading this article, you‟re mining bitcoins.

[12] Magro, P. (2016, July 16). What Greece can learn from bitcoin adoption in Latin America.

[13] McMillan, R. (2014, March 3). The Inside Story of Mt. Gox, Bitcoin‟s $460 Million Disaster.

[14] Team, B. (2016, January 20). Understanding Bitcoin's Growth in 2015.

[15] Reuters. (2016, July 6). Two-Year High on Gold Prices Fueled by Brexit-Spooked Investors.

[16] Perez, Y. B. (2015, October 24). European Exchanges React to Bitcoin VAT Exemption.

[17] Price, R. (2016, June 17). Digital currency Ethereum is cratering because of a $50 million hack

[18] Ivaschenko, A.I. (2016). Using Cryptocurrency in the Activities of Ukrainian Small and MediumEnterprises in order to Improve their Investment Attractiveness. Problems of economy, (3), p.267-273.

[19] Angel, J., & McCabe, D. (2015). The Ethics of Payments: Paper, Plastic, or Bitcoin? Journal of Business Ethics,132(3), 603-611.

[20] Murali, J. (2013). A New Coinage: Can Bitcoin, the global online digital currency, be the precursor of a new monetary system? Economic and Political Weekly,48(38), 77-78.

[21] Patterson, J. (2015, August 04). Bitcoin: A New Global Economy. Retrieved from Bitpay

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

A REVIEW OF SOLAR ENERGY: HISTORY, FUTURE,WORKING, BENEFIT AND DRAWBACKS

Prashant Chaudhary, Rahul Chaurasia and Mukesh Chaudhary
MSc IT Department, SVKM's Usha Pravin Gandhi College of Arts,Science and Commerce, Ville Parle(W), Mumbai, Maharashtra, India

ABSTRACT
Renewable energy is eco-friendly in nature, and we get it from different types of renewable resources which will never run out. The focus of this research will be Solar energy which we get directly from the sun. Solar Energy is a clean renewable resource with zero emission. In this research, we will discuss about the working of solar panel, photovoltaic effect, how it can be used to save the nature from pollution and how it brought down the electricity cost. We will go through the past of solar energy consumption, how it became useful in human life, and about the changes it made. Further then we will see, how the usage of solar panel became popular, and we will also discuss about the changes we can see in the coming few years, also positive and negative aspects of solar panel.

Keywords: Renewable energy, Solar energy, Zero emission,photovoltaic effect

I. INTRODUCTION

Energy plays an important part in affecting and shaping our day-to-day life. Basic life requirements such as water and food are also obtained and transported with the help of energy. That is why having good quality and uninterruptible energy is a fundamental need. The demand for environment-friendly renewable energy resources is increasing rapidly because of high fuel prices, energy requirements, pollution, cheapness, sustainability, and greenhouse gases.

Solar Power is one of the cleanest renewable resources of energy with zero emissions. Solar energy is basically a process of catching the light of the sun, which is photons and transferring it to electricity. The process of this conversion is called 'the photovoltaic effect' or PV. Solar energy is one of the energy sources that is expanding the quickest in the globe. The energy generated by the sunlight is transformed into thermal or electrical energy. It is one of the most prevalent and pure forms of renewable energy accessible. Captured solar energy can be used for powering anything from homes, cars, businesses, aircraft, portable power stations, and space programs to small things like calculators, portable power stations, and many more.

Alexandre-Edmond Becquerel, a French scientist, discovers the photovoltaic phenomenon in 1839, which is used to turn sunlight into energy. Later in 1883, American inventor Charles Fritts developed the first rooftop solar system, which produced energy by covering panels with selenium. Albert Einstein, however, contributed to our understanding of the precise mechanism of how light generates energy that humans turn into electricity in 1905. Albert Einstein earned the Nobel Prize in 1921 for his work, which not only altered the way humans see light.

The U.S. government opted to deploy current PV technology in its space program between 1950 and 1960, which led to its introduction at that time. The first solar- powered spacecraft, named Vanguard I, was launched in 1958. Solar panels have since become a crucial component of satellites and other space program vehicles.

Similar to Fritts' solar array from the 19th century, the contemporary solar panel generates power, the only difference is we use silicon solar cells instead of selenium coating on panels because it is more efficient and can produce more electricity to power our lives. However, to use electricity generated through silicon solar cells, direct current (DC) electricity must be converted into alternating current (AC) power. And in order to do this, a device known as an inverter which assists in converting electricity from DC power to AC power is attached to the solar panel. And from this process, 100% clean energy is sent to use for homes, businesses, and many more.

II. PAST PRESENT AND FUTURE

Before 1960, in India, we used to generate electricity from hydro energy and fossil fuels like coal and oil. In 1961, solar energy was discussed for the first time in India. Till 1981, solar energy was not implemented in India. So, we were relying on other renewable resources like Geothermal, Hydro, and Tidal Energies. In 1981, solar energy was implemented for the first time in India in the state of Gujarat. The Department of Non-Conventional Energy Sources (DNES) was Formed in 1982, under the ministry of energy and its main objective was to provide funds to Renewable energy resources (RES). After 1997, the Government of different states started to give subsidies on buying the Solar Panels. It became more affordable as the government started making different policies and implementing new schemes. Till 2005, people started to buy solar panels at a

ISSN 2394 - 7780

large scale but as the cost of the panel was high, many people were not able to afford it even after the government was providing subsidy. Further then the government started to make solar panels at a reasonable price so that many people canafford it.

At present, India is ranked On Number 4th in the world as per report of 2021. Solar power installed capacityhas reached around 62 GW by the end of 2022. On 9th October 2022, Modhera was announced as first solar powered village by the Prime Minister of India which is situated in state of Gujrat. The people living in Modhera are saving nearly 60 to 100 percent on their electricity costs as they are relying only on solar power.

The Biggest solar park of India is located in Bhadla, Rajasthan, it spreads over a total area of 14,000 acres and its total capacity is 2245 MW. To build a solar Project in Rajasthan, the government has spent nearly 10 thousand crore INR and it was built under the Ministry of New & Renewable Energy (MNRE) Scheme. This project was constructed in 4 different Phases. The first 2 phases were created by the Rajasthan Solar power park and the third phase was created by Saurya Urja Company of Rajasthan and the final phases were designed by the Adani Renewable Energy Park and its capacity was 500MW. The Temperature in Bhadla ranges from 46 to 48 degree and this is the main reason why the government decided tobuild this project in Bhadla, Rajasthan as the solar park spread over 14,000 acres and located in Desert area. Solar panels are cleaned by the Robots and monitored by the Humans. India's largest floating solar project is situated inKerala which produces 101MWP. In India, Rajasthan shares the largest solar Capacity of 16.06GW while Gujarat shares 8GW which comes on the second position and Karnataka, Tamil Nadu, Telangana shares 7.8GW, 6.2GW, 4.6GW respectively by the end of 2022. Out of the 10 Largest solar parks constructed in the world, 5 are in India itself as per Institute for Energy Economics and Financial Analysis. In Gujarat alone, out of 3 houses, 2 are running on the Solar rooftop Systems because of the subsidy given by the government. Government decided to cover Eight lakh consumers by 2021-2022 under the rooftop solar scheme Surya Gujarat Yojana. As the Rooftop can be easily managed by the owner of the house and it also brings down the electricity cost used by the household appliance.

Till 2030, the Government is planning to increase the solar capacity by 500GW, and to achieve net zero emission by 2070 as solar energy is eco-friendly. This willhelp in improving the climate and will also increase the job opportunities within the region. The government is also trying to bring down the cost of solar panels so that everyone can afford solar panels and use them at large scale, and this will bring down the usage of the non-renewable resources. Till 2050, every house will be running on solar energy and the government is going to build new projects in different states of India, so that industries will rely on solar energy. Tata has planned to take renewable capacity to 80% by the end of 2030. The government is also trying to bring new schemes to make every house solar powered.

III. SOLAR ENERGY

Solar energy is created by collecting photons from thesun's light and turning them into electricity. We use solar thermal energy, solar architecture, solar heating, and artificial photosynthesis to harness the sun's radiant light and heat. Because it is so widely available, sunlight is a very alluring source of power.

IV. WORKING OF SOLAR ENERGY

The photovoltaic cell within the solar panel absorbs the energy from the sun's rays and converts it to direct current when the sun shines on the solar panel (DC). Then we use a charge controller to prevent the overcharging of the battery. It basically reverses back the extra charges to the solar panel to avoid damage to the battery. After that, we connect the battery to store the current when sunlight is not there like at night or in cloudy weather. And then we use an inverter to convert direct current (DC) to alternative current (AC).

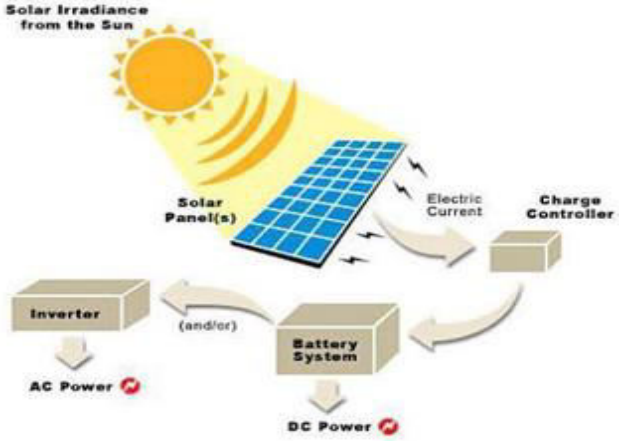


Fig 1: Working of solar energy

ISSN 2394 - 7780

V. BENEFITS

If we talk about renewable resources, while people condemn renewable resources for being too expensive and not so efficient, solar energy has proven them wrong and is very beneficial for the environment and for finance as well. And as a result of the strong demand, technology has advanced and become a reliable source of clean energy. It also has some more benefits, let's discuss them.

A. Solar Power is Truly a Renewable Energy Source

Solar power is one of the best renewable energies as it works on the sun's light. As long as the sun shines, we will be able to use solar energy. Scientists roughly calculate that the sun has at least five billion years of energy left in it. As long as solar panels receive sunlight for a few hours of the day, they will generate sufficient energy for us.

B. Reduces Energy Bill

Generating your own electricity using solar panels will make you free from the government's conventional power plant's electricity bill. In short, we can say that the more energy we will generate, the more money we will save. But exactly how much cost we save depends on the size of our solar project and normal usage.

C. Getting Cheaper Along With Time And Easy To Install

Solar panels are getting cheaper along with developments and improvements. And it can be easily installed on the roof of your house. It is a one-time investment that will save you money for a long period of time. The price depends on a few factors like the quality of your solar panel, your roof's orientation, the size and type of properties, and many more.

D. Very Low Maintenance Cost

This is one good aspect of installing solar panels in your houses. Many owners manage to clean their panels twice a year at most. Since there is no movement in the system, no wear and tear risks are involved. Once you pay the cost of the installment, you can expect very less on repair and maintenance.

E. Reduces Pollution and Impact on the Environment

As we know solar panels generate electricity without greenhouse gasses or any type of water pollution or noise pollution which makes them less pollutant and good for the environment.

F. Improve Energy Independency

Another benefit of being less dependent on the government's conventional power plant and using your own solar panels is that your business or work will not be disturbed or interrupted because of power cuts and other out stages.

G. Safer Than Other

Solar energy is safer than any other conventional power plant whether it is for use or preservation.

There is no harm to any property in the installation of the solar power plant.

H. Technology Development

Technology in solar power plans is increasing day by day. Scientists and Engineers are trying to make it more efficient and easily available for everyone. They are also trying to improve the performance and cost so that it will be more affordable.

VI. Drawbacks

As we saw, installing solar panels is very beneficial, but we cannot ignore the issues and problems with these panels. So, let's discuss significant complaints about solar energy.

A. Cost

Although solar panels are very economically friendly as they do not have any maintenance or charges after installation, but the charges for installation of solar panels are high. This covers

charges for wiring, inverter, batteries, solar panels, and installation. However, technologies are constantly improving, and we can expect better prices in the future.

B. Weather Dependent

As we know, solar panels totally depend on sunlight to effectively collect solar energy. So, in cloudy weather or in the rainy and winter season, the absence of sunlight can affect the energy system.

Although solar energy can still be collected in the rainy and winter season, but the efficiency will drop.

We should also remember that energy cannot be collected at night.

ISSN 2394 - 7780

C. Storing Solar Energy Can Be Expensive

Either you use the energy generated through solar energy immediately, or you need large batteries to store it. These batteries can be charged in the daytime in presence of sunlight and can be used at night-time in absence of sunlight. Using batteries to store energy can be a good solution for using energy all day long but buying batteries can be expensive.

D. Panels Use Lots of Space

The more we collect the sunlight on the panel, the more energy we can produce. But to collect more and more sunlight we need to install panels as much as possible and that can use lots of space.

E. Pollution While Manufacturing

Although the pollution that occurs from solar energy is far less than from any other conventional power plant, but solar energy is somehow related to pollution. The manufacturing of solar photovoltaic systems involves the use of various hazardous and poisonous substances that may have an adverse environmental impact. However, solar energy produces much less pollution than any other alternative energy source.

VII. CONCLUSION

The study reveals that making use of renewable resources is pollution free and best for nature.

Although the government is making an effort to make solar panels cheaper, many people still cannot afford it. The government should make more schemes to make it more affordable. Also, engineers should develop plans to recycle these used panels as it can help in improving the cost and decreasing the pollution that occurs because of manufacturing of these panels. Although there are some drawbacks in solar panels, but still it is better than any other form of energy consumption.

REFERENCES

[1] Swami Prakash Srivastava, Surat Prakash Srivastava, Solar energy and it's future role in Indian economy, ISSN No. 2231-1289, Volume 4 No. 3 (2013).

[2] Mochamad Choifin, Achmad Fathoni Rodli, Anita Kartika Sari, Tri Wahjoedi, Abdul Aziz, A Study of Renewable Energy and Solar Panel Literature Through Bibliometric Positioning for Three Decades, July 2021.

[3] Dinesh Kumar Sharma, Anil Pratap Singh, Versa Verma, A Review of Solar Energy: Potential, Status, Targets and Challenges in Rajasthan, ISSN: 2278- 0181, Vol. 3 Issue 3, March – 2014.

[4] Mohd Rizwan Sirajuddin Shaikh, Santosh B. Waghmare, Suvarna Shankar Labade, Pooja Vittal Fuke, Anil Tekale, A Review Paper on Electricity Generation from Solar Energy, ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887, Volume 5 Issue IX, September 2017.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

MEDICAL EXPERT SYSTEM

Neelam Naik[1] and Chirag ramesh Wala[2]
[1]Assistant Professor and [2]Student, MScIT Usha Pravin Gandhi College of Arts, Science &Commerce

ABSTRACT
Healthy and prompt treatment of diseases is required. If they are not handled in a timely manner, lying can lead to a variety of health issues, some of which may be fatal. Due to a lack of practitioners, medical facilities, and specialists, these issues are getting worse. Studies have attempted to create and develop expert systems that can provide guidance for doctors and patients to make patient diagnosis and treatment recommendations easier in an effort to address these issues. This review paper provides an in-depth investigation at the use of medical expert systems for disease diagnosis. It gives an examination of previous investigations combined with a brief review of medical diagnostic expert systems.

1. INTRODUCTION

1.1 EXPERT SYSTEM

An expert system is a computer software created to solve complicated issues and offer decision- making capabilities similar to those of a human expert. This is accomplished by the system retrieving information from its knowledge base in accordance with user queries, utilising reasoning and inference procedures.

The first expert system (ES), which was the first effective use of artificial intelligence, was established in the year 1970 and is a subset of AI. By drawing on the knowledge that is kept in its knowledge base, it can solve even the most complicated problems like an expert. Like a human expert, the system aids in decision-making for complex issues by using both facts and heuristics. It is thus named because it has in-depth expertise in a certain field and is capable of resolving any complex problem of that particular domain.

1.2 MEDICAL EXPERT SYSTEM

There are many expert systems in the medical field. Any medical expert system's main goal is to diagnose and treat illnesses. Programs and a medical knowledge base are the foundation of a medical expert system. The knowledge gleaned via a medical expert system is comparable to that provided by experts in that field.

Medical the development of a medical expert system depends on the knowledge of specialists. This information is constructed in two stages. During the first phase's formation of individual meetings with doctors and patients, the medical conditions of diseases are documented. The second phase involves the formation of a deposit of rules, each of which has an IF part that contains the symptoms and a THEN part that contains the disease that should be realized.

2. WORKING OF EXPERT SYSTEM

Artificial intelligence and machine learning are used by contemporary expert knowledge systems to mimic the actions and decisions of subject matter experts. Just like humans, these systems can get better over time as they gain more experience.

Expert systems compile knowledge and experience into a knowledge base and combine it with an inference or rules engine, which is a system of rules the software uses to apply the knowledge base to given scenarios.

The knowledge base is accessed by the inference engine via one of two methods:

Forward chaining reads and analyses a collection of data to generate a logical forecast of what will occur next. Making forecasts regarding the direction of the stock market would be an example of forward chaining.

Backward chaining processes and reads a set of facts to reach a logical conclusion about why something happened.

3. DATASET

First, we need some datasets to help us train our model and gain some insights. As a result, we conducted some surveys in the medical area, looked up some data online, and then combined all of that to create a raw dataset. We now possess a dataset.

After gathering the data, we must prepare it because it is raw data and will be used to train our machine learning model. We have prepared that data for machine learning models by using certain Python tools, including NumPy and pandas.

ISSN 2394 - 7780

Now that our data is available, machine learning algorithms can utilise it to forecast some results. We chose the Support Vector Machine, Random Forest Classifier, and Naive Bayes algorithms because our problem falls under the unsupervisedmachine learning technique.

4. LITERATURE REVIEW

The prediction of diseases based on the symptoms displayed by an existing employing machine learning algorithms has been the subject of several discussions: Monto et al. [1] developed a statistical model to prognosticate whether a patient had influenza or not. They included 3744 adult and adolescent cases of influenza who were unvaccinated and had fever and at least two other influenza symptoms. Out of 3744 people, 2470 hadinfluenza that was confirmed by a lab. Based on this information, their model predicted a delicacy of 79.

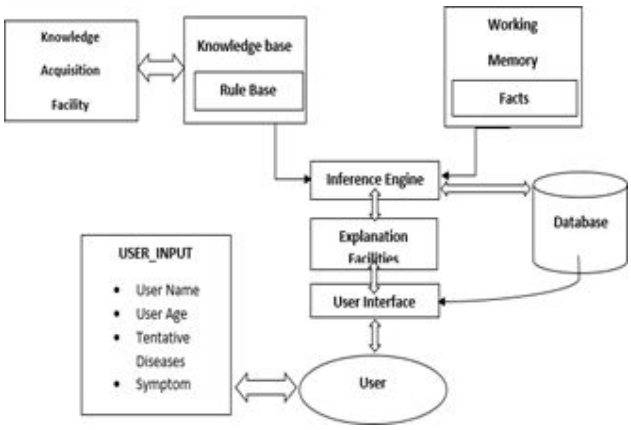
Chen et al. [2] simplified colourful machine learning methods for the accurate prediction of an onset of habitual complaints. The information gathered for the purpose of training was insufficient. A model with idle factors was appliedto get around this.

The LSTM model provided accurate predictions when conditions were extreme, but the DNN model outperformed them both in terms of average performance. A database with data on patients with any type of cardiac problem was employed by Haqet al. [3]. They used the least absolute loss and selection motorist, relief, minimal redundancy andmaximum connection (mRMR), and these three selection algorithms to extract features, which were then cross-verified by the K-fold system. Six different machine learning techniques were used to classify the extracted data based on the presence orabsence of cardiac complaints.

ML algorithms were used by Maniruzza-Man et al.

[4] To categorise the diabetic complaint. Using logistic regression (LR), the problem variables for diabetes complaints were found. The ML-based system's overall delicateness was 90.62%.

5. FRAMEWORK FOR EXPERT SYSTEM



Patients typically go to hospitals to complain about their illnesses, then MES users interview thepatients about their illnesses and look up symptomsin databases. The user delivers the patient the prescription if the patient's symptoms match those in the database. Figure 1 depicts the MES's suggested framework. The different components that come together in this diagram to really achieve a complete rule-based expert system are presented.

The knowledge base is the ESs' intellectual capital because it contains all the information needed to create the rules. The primary source of rules for the ESs is this knowledge. Consultation with general practitioner doctors, Internet medical websites, medical literature, research papers, and journals were the most significant sources of knowledge for the MES. The knowledge base is composed of learning the diseases' symptoms. Rules serve as a representation of the knowledge.

6. DISCUSSION

Different machine learning models were employedto analyse the illness prediction for the input dataset that was made public. For the prediction, 11 different ML models were employed. Six of the 11 models we tested were accurate to 50% or more. we achieved the greatest accuracy for the WeightedKNN model, which was 93.5%, out of all themodels. The weighted KNN was high since the value of K changed in this model, which contributed to the accuracy being high. This number fluctuated based on our dataset; for the training set, it was both tiny and

ISSN 2394 - 7780

huge. It turned out to be the most accurate model when compared to the other ML algorithms because of this variation.We gathered raw data and separated the groups according to gender, age group, and symptoms.

The RUSBoosted Tree, with an accuracy of0.5%, was the least accurate model. 21.8%accuracy was exhibited by the fine tree. The accuracy of the medium tree was 12.3%. The accuracy of the coarse tree was 6.4%. The accuracy of the Gaussian Nave Bayes model was 16.8%. The accuracy of Kernel Nave Bayes was 16.8%. The accuracy of Fine KNN was 80.3%. A medium KNN had a 61.8% accuracy. The accuracy of coarse KNN was 5.3%. Accuracy for the subspaceKNN was 73.2%.

The RUSBoosted Tree, with an accuracy of 0.5%, was the least accurate model. 21.8% accuracy was exhibited by the fine tree. The accuracy of the medium tree was 12.3%. The accuracy of the coarse tree was 6.4%. The accuracy of the GaussianNave Bayes model was 16.8%. The accuracy of Kernel Nave Bayes was 16.8%. The accuracy of Fine KNN was 80.3%. A medium KNN had a 61.8% accuracy. The accuracy of coarse KNN was 5.3%. Accuracy for the subspace KNN was 73.2%.

In an emergency, doctors and other medical personnel are constantly needed. Our prediction technique can be useful in the present COVID-19 situation, when adequate facilities and resources are not available, and it can be utilised to diagnosea disease.

7. CONCLUSION

The described a method for diagnosing a patient's condition based on their symptoms, age, and gender. Weighted KNN model provided utilizing the aforementioned parameters, illness prediction had the best accuracy of 93.5%. Nearly every ML model produced good accuracy values. Some models couldn't forecast the illness and had low accuracy rates since they were dependent on the parameters. We could simply manage the medical resources needed for the therapy once the sicknesswas predicted.

This strategy would aid in reducing the expense involved in treating the sickness and would also speed up the healing process.

8. REFERENCES

[1] Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin, M., 2020. Classification andprediction of diabetes disease usingmachine learning paradigm. Health information science and systems, 8(1), pp.1-14.

[2] Chen, M., Hao, Y., Hwang, K., Wang, L. andWang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, pp.8869-8879.

[3] Haq, A.U., Li, J.P., Memon, M.H., Nazir, S.and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heartdisease using machine learning algorithms. Mobile Information Systems, 2018.

[4] Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., 2020. Disease predictionfrom various symptoms using machine learning. Available at SSRN 3661426.

[5] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

[6] Kaushik, K., Kapoor, D., Varadharajan, V. and Nallusamy, R., 2014. Disease management: clustering–based disease prediction. International Journal of Collaborative Enterprise, 4(1-2), pp.69-82.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023

ISSN 2394 - 7780

LAND-USE AND LAND COVER CLASSIFICATION USING GEO- SPATIAL ANALYSIS FOR MUMBAI REGION

Maurya Rajesh Kumar[1] and Kaprawan Sandhya[2]

Department of Information Technology, Usha Pravin Gandhi College of Arts, Science and Commerce Vile Parle(w), Mumbai, 40056

ABSTRACT

The field of Land Use and Land Cover (LULC) research is a crucial area of study that aims to understand how human activities are impacting the natural environment. This research has become increasingly important as it helps to understand the impacts of human activities on biodiversity, ecosystem services, and climate change. LULC research involves the usage of remote sensing and Geographic Information Systems (GIS) to map and analyze land use and land cover patterns over time. The primary goal of this research is to identify areas where conservation and land management efforts are needed, and to understand how human activities are affecting the natural environment. One of the key areas of focus in LULC research is the study of land use change, which is the transformation of land from one use to another. Other key areas include land cover change and the identification of areas important for conservation and land management. The paper varous methods used to classify land use and land cover. The results of these methods were compared and evaluated, and it was found that the spectral angle method was the most accurate and efficient for land cover classification. The paper also suggests areas for future research in the field of LULC, including the improvement of classification methods, assessment of the impacts of land use change, and development of new indicators for land use and land cover patterns. These findings are crucial for the preservation of the natural environment and the provision of essential ecosystem services.

Keywords: LULC, Spatial Analysis, Geo-visualization, GIS

I. INTRODUCTION

Geospatial analysis plays a crucial role in the gathering, display, and manipulation of various types of data including imagery, GPS data, satellite photography, and historical data. This data can either be explicitly described in terms of geographic coordinates or implicitly defined, such as in the case of a street address, postal code, or forest stand identifier. The integration of these data sources into geographic models helps in a more comprehensive and accurate analysis [1]. Urbanization is a rapidly growing phenomenon and geospatial technology plays a decisive role in capturing, analyzing, and managing the effects of urbanization on our cities. Consideration and wise use of land are critical for the advancement of human culture and thus, the study of land use is an essential aspect to consider. The evolution of land use and land cover (LULC) is a crucial subject, as the amount of land used for each purpose is constantly evolving. Land use is influenced by various factors, such as physical, economic, and social factors, and thus, information on LULC is vital for all types of natural resource management and action plans [2]. Notably, the application of geospatial analysis and technology in the study of land use and land cover is critical for understanding and managing the impacts of human activities on the environment. The information gathered from these studies helps in formulating land use policies, conservation and management strategies, and sustainable development plans that balance economic, social and environmental considerations.

Land cover refers to the physical land types present on the Earth's surface, including water bodies, vegetation, roads, rural areas, urban areas, and many others. Land use/land cover change (LULCC) refers to the transformation of the terrestrial surface caused by human activities [3]. The difference between land use and land cover is that land use helps to determine how individuals utilize the parcel of land for variety of socioeconomic objectives, whereas land cover indicates the physical composition on the land surfaces. This include features such as forest region or wetlands [4]. The change in the extent of land cover and use around the globe is commonly referred to as LULC change. Several factors influence LULC changes, including ecological conditions, altitudes, geological structures, slope, and the technological, socioeconomic, and institutional setup, which also influences land use patterns [5]. To classify and understand LULC, various techniques and approaches have been developed, including manual classification, numerical and digital classification, hybrid approach, and other methods of classification.

In the manual classification method, remotely detected data is organized to characterize land use when the investigator is aware with the study region. This method takes advantage of the ability to interpret satellite images. In numerical methods for satellite image classification, two types of approaches have advanced and

ISSN 2394 - 7780

remained as the primary possibilities. These approaches differ in the assumptions made about the information in the scene to be classified. In supervised classification, information on all cover types is expected from prior knowledge to be included in the classified scene. Studies have found that manual classification can be effective in areas where the investigator has knowledge of the study region, but numerical and digital methods can provide more consistent and accurate results in complex or unfamiliar regions. Hybrid approaches, combining both manual and numerical methods, have also been found to produce promising results in LULC classification [6].

Overall, LULC research plays a crucial role in understanding how human activities are impacting the natural environment and identifies areas where conservation and land management efforts are needed. Further research is needed to advance our understanding of LULC and the factors that influence its change, as well as to develop more effective techniques for LULC classification.

The study highlights the effectiveness of the Hybrid Approach in LULC classification. This approach combines the benefits of both digital classification and manual classification methods to produce a more accurate and refined land cover maps. By using a computerized method for an initial classification and then relying on manual techniques to correct errors, the Hybrid Approach can quickly produce a reasonably good classification while also allowing for refinement of classes that were not accurately marked. One of the widely used LULC classification methods is the Normalized Difference Vegetation Index (NDVI), which is one of several methods for determining vegetation cover from remotely detected data. Other methods include Spectral Mixture Analysis Modified Soil Adjusted Vegetation Index (MSAVI), Normalized Difference Built Index (NDBI), Soil- Adjusted Vegetation Index (SAVI) and Normalized Difference Water Index (NDWI) [6]. These methods vary in terms of their complexity and accuracy. It is important to note that different approaches and techniques are used to best suit the data and the desired output. By combining the benefits of multiple techniques, the Hybrid Approach can produce a more accurate and detailed map of LULC, providing valuable information for natural resource management and planning.

II. LITERATURE SURVEY

Land cover change detection is a crucial aspect of geospatial analysis that assesses the transformation in land cover over time. The process involves evaluating the differences between two or more land cover maps and determining which changes have occurred in the designated area. There are several methods for change detection, including aerial difference computation, post classification comparison, image rationing, image regression and image differencing. One of the widely used methods for post classification change detection is the Normalized Difference

Vegetation Index (NDVI). NDVI is a remote sensing technique that provides information about the presence and health of vegetation by using the difference between the near- infrared and red bands of an image [6]. The NDVI is calculated on a per-pixel basis using the following formula:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

where, NIR is used to refer to the near-infrared band value and RED is used to represents the red band value for the corresponding cell. The NDVI assumes the value in the range from +1 to -1, where a value close to +1 highlights a denser and greener vegetation, while a value close to 0 indicates a less green or dry vegetation.

A. LULC Study in in Current Context

Studies related to LULC for spatial analysis earth coverage have been ongoing for decades, with various researchers investigating different aspects of the field. Some key studies in the past include:

(a) **Landsat Program:** The Landsat program, initiated by the U.S. government in 1972, has provided a continuous record of high-resolution imagery of the Earth's surface, allowing for long-term monitoring of LULC changes.

(b) **The European Space Agency's GlobCover Project:** The GlobCover project, launched in 2005, used satellite data to produce a detailed map of the world's land cover, providing valuable information on the extent and distribution of different land cover types.

(c) **The National Land Cover Database (NLCD):** The NLCD, produced by the U.S. Geological Survey, is a comprehensive, multi-temporal land cover database of the United States, which has been widely used by researchers, policymakers, and other stakeholders to monitor and analyze LULC change.

ISSN 2394 - 7780

(d) **Urbanization and Land Use Change Studies:** Urbanization and its effects on LULC have been the focus of numerous studies in recent years, with researchers investigating the causes, patterns, and consequences of urbanization on land use and land cover.

These studies, along with many others, have significantly contributed to our understanding of LULC and its impact on the environment, economy, and society. They have also provided a foundation for ongoing research and practical applications of LULC data in decision-making and policy formulation.

B. LULC study in India

There have been several studies on Land utilization and LULC changes in India. Here are a few examples:

(a) "Land Use and Land Cover Changes in India: A Review of Literature" by Kalaivani and Raman [7] - This study provides an overview of the land use and land cover changes in India over the past few decades and identifies the major drivers of these changes.

(b) "Assessment of Land Use and Land Cover Changes in Urban and Peri-urban Areas of Hyderabad, India" [8] - This study analyzed the land use and land cover changes in the urban and peri-urban areas of Hyderabad, India, using satellite imagery and GIS techniques.

(c) "Land Use/Land Cover Changes and Its Impacts on Environment in the Himalayan Region of India" [9] - This study analyzed the land use and land cover changes in the Himalayan region of India and their impacts on the environment.

(d) "Urban Sprawl and Land Use Change in the Delhi Metropolitan Region, India" by Chatterjee et al. [10] - This study analyzed the patterns of urban sprawl and land use change in the Delhi Metropolitan Region, India, using satellite imagery and GIS techniques.

(e) These studies highlight the importance of monitoring and analyzing land use and land cover changes in India and the need to address the impacts of these changes on the environment and society.

NDVI plays a crucial role in the field of Land analysis and LULC study. It helps to detect changes in land cover and monitor the vitality and growth of vegetation in a particular region over a period of time. This information is vital for the sustainable management of natural resources, as well as for observing the impacts of human activities on the environment. The accurate and timely monitoring of vegetation through NDVI is critical for informed decision- making in environmental management. Additionally, there are various data formats used for LULC classification, including some of the commonly used ones that are listed below.

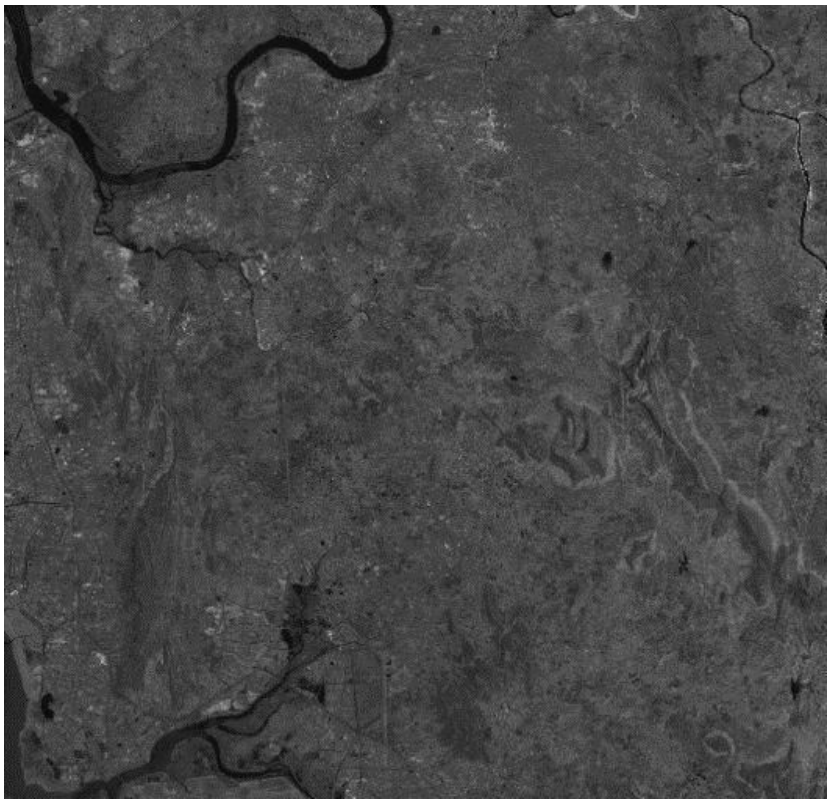
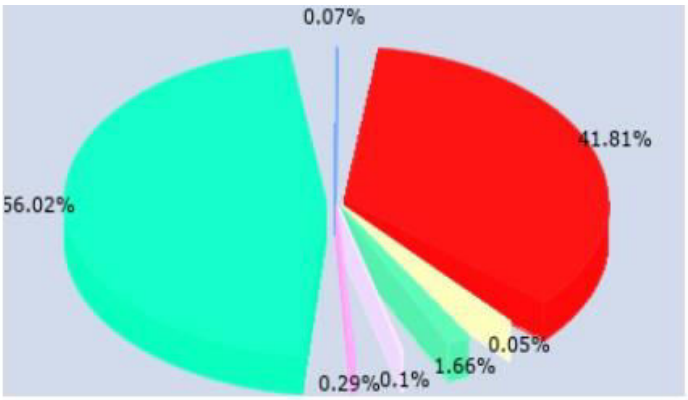


Fig. 1 TIF image from LISS III sensor from Bhuvan Portal

ISSN 2394 - 7780



LULC Class	Area (Sq.Km)	LULC Class	Area (Sq.Km)
Builtup,Urban	65.64	Agriculture,Fallow	0.08
Forest,Swamp/ Mangroves	2.6	Barren/unculturable/ Wastelands, Sandy area	0.16
Barren/unculturable/ Wastelands, Barren rocky	0.46	Wetlands/Water Bodies, CoastalWetland	87.95
Wetlands/Water Bodies, Reservoir/Lakes/Ponds	0.11		

Fig. 2 LULC information for Mumbai city (2015-16) for Total geographic area of 157 sq. Km

Fig. 3. False Colour Composite

Spatial Analysis of earth surface and LULC classification is most widely used application of remote sensing. This section discusses the research work which has already been developed and implemented for LULC change detection in various geographical areas. It also focuses on different methods for extracting information from satellite imagery and various classification methods for detecting LULC changes. The observations of various previous research indicate that LULC modifications are strongly related to a wide range of environmental changes, such as the release of greenhouse gases, increase in temperature, loss of biodiversity, and depletion of water supplies [4].

Hassan et al. [3] 2016 has conducted LULC change detection evaluation of Islamabad, Pakistan from 1992 to 2012. In his research, he has accomplished the change detection by using two satellite imagery first is Landsat 5 multispectral 30 m resolution imagery for 1992 and second is SPOT 5 multispectral 10 m resolution imagery for 2012 from United States Geological Survey (USGS) and SUPARCO. For the change detection overall LULC was divided into different classes of Built-up area, Agriculture, Forest, Water and Barren area. ERDAS 2011 software was used for geometric correction of satellite images and ArcGIS 10 was used for post-classification change detection analysis. Supervised classification method is used for preprocessing the images by applying maximum likelihood algorithm on images in this work. During post-classification change detection ArcGIS 10 was used for processing and visualization.

Another technique called Overlay procedure was used for obtaining the changes in LULC during specific time period. Classification accuracy was measured using Stratified Random Method and Non-Parametric Kappa test-classification accuracy. This method achieved 89% accuracy for overall map of 1992 and 2012 using confusion matrix and

ISSN 2394 - 7780

0.89 kappa statistics for LULC of two periods. This was reasonably good overall accuracy for change detection and analysis. During the study period they observed increase in Urban, Agriculture and Water areas. The study highlighted that there is a significant decrese by 19.5% for Forest area, Barren land region and drinking water reservoir areas in the city during the study period. After analysis the major driving force of urban land expansion were Population growth and economic development. [3]

Md. Inzamul Haque et al. [6] 2017 has conducted Land cover change detection spatio-temporal study on Tanguar Haor, Sunamganj, Bangladesh from 1980 to 2010. In his research, he has accomplished change detection of Land Cover using Primary Data and secondary data from Landsat satellite for 02-02-1980, 28-01-1989, 29-01-2001 and 30-01-2010 with spectral resolution of 60 m for Landsat 3 – MSS and 30 m for rest all. Satellite data was acquired from United States Geological Survey (USGS). For change detection, land cover was divided into different classes as Deep water, settlement, vegetation and shallow water. ENVI

5.1 software was used for change detection statistics. The methods such as CVA, NDWI, NDVI was utilized for change detection for pre-classification purpose. For post- classification, maximum likelihood technique and Image differencing (Transition probability matrix) were used [6].

During the study period major increasing change were observed in Settlements (140%), Shallow water body (33%), 71% of deep-water body change its state mostly to the shallow water and Decreasing trend were seen in Deep water (49%) from 1989 to 2010, vegetation cover (52%) from 1980. Overall change influence approximately 40% of total landscape within 1980 to 2010. Thematic change statistics maps, Bayesian probability function, from-to change matrices were computed for evaluating change detection of LULC of 1980 to 2010. The major resason behind LULC changes in the region of Tanguar Haor [2] was found to be anthropogenic influence and has been highlighted in the study.

Prabuddh Kumar Mishra et al. [5] 2019 has conducted LULC change detection in Rani Khola watershed of Sikkim, Himalaya from 1988 to 2017. In this research, the researcher has used pre-processing for TOA to BOA reflectance in Sentinel 2A/MSI and Landsat 5/TM satellite data with 10 to 30 m resolution. Supervised Classification was used on ASTER DEM. Maximum Likelihood Classification method was used for LULC classification.

Accuracy assessment was done using kappa accuracy assessment technique and it was observed that the kappa coefficient was 0.8117and 85.67% was overall accuracy. After evaluation increase in Water body by (0.11%) and Built-up area by (2.13%) and decrease in Open Forest by (13.98%), Barren land by (1.82%) and Agriculture land by (2.83%) was observed. Major cause for the change during study year was conversion of open forest areas into dense forest and paddy fields [5].

Bhanage et al. [4] 2021 has conducted Prediction of LULC in Mumbai and its surrounding region using Multilayer Perceptron Neural Network based Markov Chain Model. For their research purpose they have used satellite data from Landsat TM, ETM+, SRTM DEM. For LULC classification MLPNN based MCM was used. From their research work following comparison can be made: The prediction model used (MLPNN) has an accuracy of 81.69%, and there are three types of accuracy (user, producer, and overall) and a nonparametric kappa coefficient index (k) used to evaluate the model. Since valdation is important in such work, different variations of kappa index were claculated and used. The included the kappa for no information (kno), stratum level location kappa (klocation strata), kappa standard (kstandard), and kappa corresponnding to grid cell- level location (klocation).

The overall accuracy of the 1992, 2002, and 2011

LULC maps [5] was 87.56%, 85.57%, and 89.05%, respectively, with a kappa coefficient higher than 0.80. For the simulated LULC of 2011, the value of each kappa coefficient was higher than 0.75. Urban, Forest, and Coastal feature classes increased by 405.84 km2 (43.97%), 58.27

km2 (9.73%), and 15.57 km2 (8.00%) respectively by 2050.Water decreased by 1.14%. By 2050, the Urban class will cover the area of 1328.77 km2 which is 46.87%. Agricultural/Sparsely Vegetated land, Coastal feature, Forest and Water will occupy the 17.75%, 7.41%, 23.16%, and 4.79% respectively.

III. ANALYSIS OF LAND CLASSIFICATION METHODS

User accuracy, on the other hand, measures the extent to which the LULC map is correct from the perspective of the map viewer. It is calculated as the ratio of the total number of correct classifications for a particular class to the total number of classifications. This accuracy is crucial from the perspective of the end-user, who needs to trust the information presented in the map for decision making [10]. Finally, the study mentions overall

ISSN 2394 - 7780

accuracy, which is the ratio of the number of test samples classified correctly to the total number of test samples. This measure provides an overall view of the accuracy of the LULC map and can be used to compare different LULC maps and methods. Producer accuracy, user accuracy, and overall accuracy are important measures to evaluate the accuracy of LULC classification, and they provide different perspectives on the accuracy of the LULC map.

Reference	Study Area	Dataset	Method	Land cover with Max. accuracy	Overall Accurac	Kappa Coefficie
Prabuddh et al. [5] 2019	Rani Khola watershed of Sikkim Himalaya	Sentinel 2A - MSI and Landsat 5 - TM (ASTER DEM)	Maximum likelihood Classification	Dense Forest, Open Forest	85.67%	0.8117
Md. Inzamul et al. [6] 2017	Tanguar Haor, Sunamganj, Bangladesh	Landsat 3 - MSS, Landsat 4 - TM, Landsat 7 - ETM, Landsat 5 - TM	Maximum likelihood Classification	Shallow water, Deep water, vegetation	89.13%	0.8543
Bhanage et al. [4] 2021	Mumbai and its surrounding region, India	Landsat TM, ETM+, SRTM DEM	MLPNN (multiple perceptron neural network) based MCM (Markov chain model)	Water, Forest, Agriculture	89.05%	0.85
Hassan et al. [3] 2016	Islamabad, Pakistan	Landsat 5 and SPOT 5	Maximum likelihood Classification	-	89%	0.89
Saurabh et al. [9] 2020	Haridwar district, Uttrakhand	Landsat 5 - TM, Landsat 7 - ETM+, Landsat 5 - TM, Landsat 8 - OLI	Maximum likelihood Classification	Water bodies, Orchards, Watershed	93%	0.91
Khyat et al [20] 2021	surrounding Nirma University, Ahmedabad, Gujarat, India	Sentinel 2B - Near-Infrared Band multispectral optical image	Support Vector	Urban area,	98.48%	-
			K-Nearest	Urban area,	98.23%	-
			Neural Network	Urban area,	97.72%	-
			Random Forest Classifier (RFC)	Urban area, Grasslands & Barren land	98.23%	-
Snehalata et al. [8] 2021	Nagjhari watershed, Bhatkuli block, Amravati, Maharashtra	Sentinel 2 and Landsat 5 and 8	Maximum likelihood Classification	Agriculture, Water Body	85.46%	0.85

Table 1: Methods and Techniques for LULC Classification

The Kappa coefficient, also known as Cohen's Kappa or simply Kappa, is a statistic that measures the inter-rater agreement between two annotators on a categorical classification problem. In the context of land use and land cover (LULC) and spatial analysis, the Kappa coefficient is often used as a measure of accuracy in the classification of remote sensing data. This measure is particularly useful in LULC and spatial analysis because it accounts for both the proportion of agreement between two annotators, as well as the proportion of agreement that would be expected by chance. A literature survey of studies [5] that have used the Kappa coefficient in LULC and spatial analysis highlights its usefulness in several ways. Firstly, the Kappa coefficient provides a more nuanced view of accuracy than simple overall accuracy, which is the proportion of samples that are classified correctly. The Kappa coefficient considers the possibility of matching by chance, which can lead to inflated overall accuracy scores in certain scenarios. Secondly, the Kappa coefficient can be used to compare different classification methods and algorithms, as it provides a consistent measure of accuracy that can be compared across studies.

In addition, the Kappa coefficient is usually utilized to assess the reliability of different annotators or classifiers. By computing the Kappa coefficient between two annotators, researchers can determine whether they agree with one another and to what extent. This information is found to be useful especially when working with bigger datasets. This is because, it allows researchers to identify areas of disagreement and to make adjustments to their classification algorithms or training data accordingly. In conclusion, the Kappa coefficient is an important tool in LULC and spatial analysis as it provides a robust and consistent measure of accuracy that accounts for the possibility of agreement by chance. By using the Kappa coefficient, researchers can compare different

ISSN 2394 - 7780

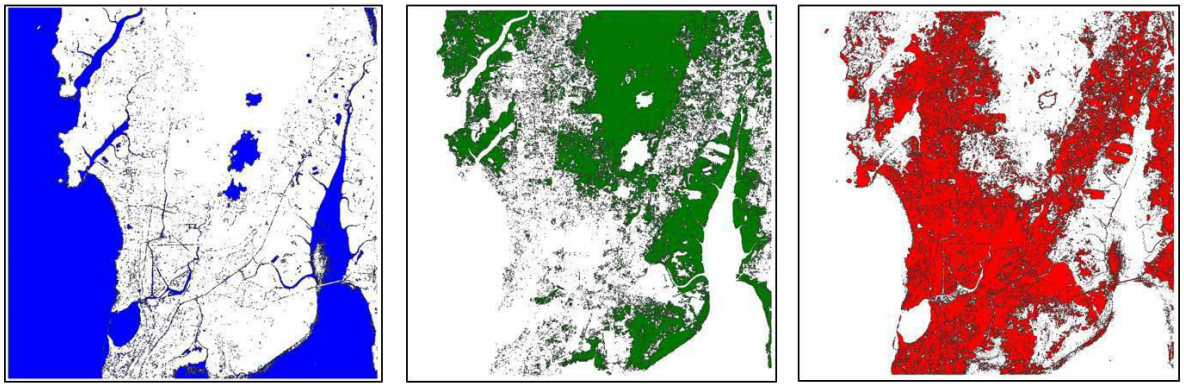
classification methods and assess the reliability of annotators, ultimately improving the quality of their results and contributing to a deeper understanding of the patterns and processes that drive LULC change.

The Kappa Coefficient Index is used to measure the accuracy of classified pixels in a LULC analysis. It is computed as the ratio of pixels which are correctly classified to the total number of pixels. A Kappa coefficient value of or above is considered to indicate excellent agreement, while a value between 0.4 and 0.6 is considered good agreement. If the Kappa coefficient is below 0.4, it is considered to indicate poor agreement. [5].

The data from the literature survey in Table 1 presents the results of different studies on land use and land cover (LULC) classification and change detection [5-8]. The studies were conducted in different parts of the world, including Rani Khola watershed of Sikkim Himalaya, Tanguar Haor in Bangladesh, Mumbai and its surrounding region in India, Islamabad in Pakistan, Haridwar district in Uttarakhand, and Nagjhari watershed in Maharashtra. The data shows that the studies used various datasets, including Landsat and Sentinel images, and different classification methods, such as maximum likelihood classification and MLPNN-based MCM.

The results [9] indicate that the highest accuracy was achieved in the Haridwar district study, with an overall accuracy of 93% and a Kappa coefficient of 0.91. The studies in Rani Khola watershed [5] and Tanguar Haor [6] also achieved high accuracy, with an overall accuracy of 85.67% and 89.13% respectively and Kappa coefficients of 0.8117 and 0.8543 respectively. The studies in Mumbai and its surrounding region [4], Islamabad [3], and Nagjhari watershed [8] also showed good accuracy with overall accuracy ranging from 85% to 89% and Kappa coefficients ranging from 0.85 to 0.89.

The results suggest that the use of remote sensing and GIS techniques can be useful in LULC classification and change detection. The Kappa coefficient was used as a measure of accuracy and agreement, and the results showed that most of the studies achieved excellent or good agreement. These findings highlight the importance of using



a) Water Region b) Forest and Trees c) Built-up Area

Figure 4: Land use Land Classfication using Minimum Distance-based Method

appropriate datasets and classification methods to achieve accurate results in LULC studies. Based on the results, we can observe that accuracy of land cover classification varies for different land cover types. In general, dense forest and open forests have the highest accuracy, with producer accuracy rates of 99.15% and 85.96% respectively. Additionally, land covers such as deep water, shallow water, and vegetation also have high accuracy rates. These results suggest that these land covers are well-distinguished by the classification method used. On the other hand, some land covers such as barren land, settlement and urban land have relatively low accuracy rates. This may indicate that these land covers are more difficult to distinguish or may be more prone to misclassification.

Some other sets of results show that land cover such as water bodies, orchards, and watershed have the highest accuracy rate. This can be concluded that these land covers are well distinguished by the classification method used. In contrast, other sets of results show that land cover types such as urban area, grasslands and barren land have high accuracy rates. This may indicate that these land covers are well- distinguished by the classification method used. Overall, it can be asserted that the accuracy of land cover classification varies for different land cover types, and it also depends on the classification method used. Further research and improvements in classification methods may help to improve the accuracy of land cover classification for certain land cover types.

ISSN 2394 - 7780

IV.EXPERIMENTS AND OBSERVATIONS

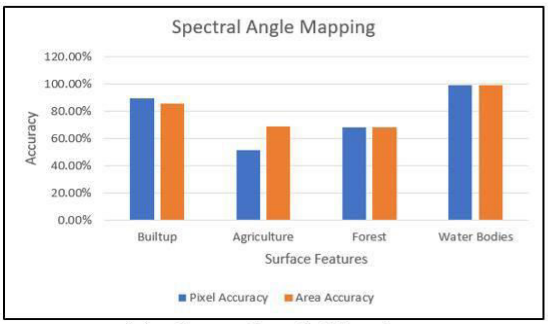
For experimatal verification and to explore the usefulness of the study, the analysis of LULC for parts of the Mumbai region was performed using experimental data obtained from the Bhuvan Portal. The spatial extent of the region of the study for experimental purpose was 73.0E19.0N- 73.25E19.25N - For Band 1 (Mumbai and Navi-Mumbai region). The goal of the study was to explore the possibility and utility of applying well-known, simple methods commonly used in Machine Learning to this problem. The methods used in the analysis included Spectral Angle Mapping (SAM), Minimum Distance (MD) and Maximum Likelihood (ML). The analysis of LULC from Satellite Images was conducted using three methods: Spectral Angle

Mapping (SAM), Minimum Distance (MD) and Maximum Likelihood (ML). Each of these methods has its own strengths and weaknesses and the selection of method depends on the specific requirements of the analysis.

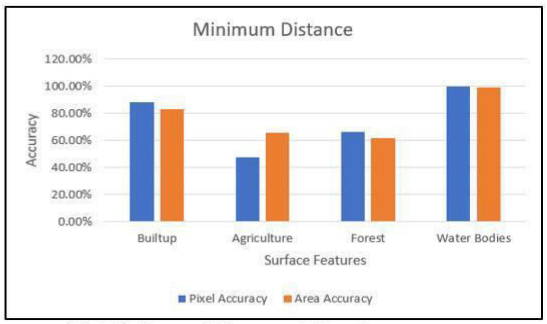
The results of the analysis of LULC from Satellite Images showed that the three methods used (Spectral Angle Mapping (SAM), Minimum Distance (MD) and Maximum Likelihood (ML)) had different levels of accuracy for different land cover classes. The Observations for LULC based on Minimum Distance based method is shown in Figure 4.

The Spectral Angle Mapping method is based on the idea that the spectral information contained in an image can be used to identify land cover classes. This method is relatively simple to implement and is effective in classifying land cover classes that have unique spectral signatures. The results of the Spectral Angle Mapping method showed a relatively high accuracy for the Water Bodies class, which is likely due to the unique spectral signature of water. However, the method performed less well in classifying the Agriculture class, which is often characterized by a more complex and hetrogeneous spectral signatures. The Spectral Angle Mapping method performed best for the Water Bodies class with a Pixel Accuracy of 99.30% and an Area Accuracy of 99.30%. This suggests that this method is effective in classifying pixels with unique spectral signatures, such as those of water bodies.

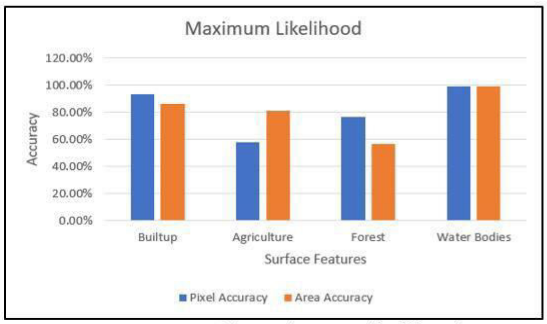
The Minimum Distance method is based on the idea that each land cover class can be represented by a mean signature and the classification of a pixel is determined by the minimum distance between the pixel signature and the mean signature of each class. This method is effective in classifying pixels with similar spectral signatures to the mean signature of a class. The results of the Minimum Distance method showed a relatively high accuracy for the Water Bodies class, which may be due to the homogeneous spectral signature of water. However, the method performed less well in classifying the Agriculture class, which often has a more complex and heterogeneous spectral signature. The Minimum Distance method showed a relatively high accuracy for the class representing Water Bodies with a Pixel Accuracy of 99.67% and an Area Accuracy of 99.56%. However, it performed less



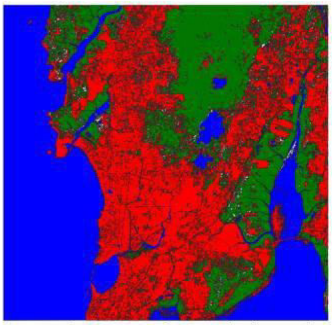
(a) Spectral angle Mapping



(b) Minimum Distance Mapping



(c)Maximum Likelihood



(d) LULC for Part of Mumbai Region

Figure 5: Performance of LULC using Different Method

ISSN 2394 - 7780

well in classifying the Agriculture class with a Pixel Accuracy of 47.41% and an Area Accuracy of 65.96%.

The Maximum Likelihood method is based on the idea that the probability of a pixel belonging to each class can be estimated based on the statistical distribution of the spectral information in each class. This method is effective in classifying pixels with a high degree of spectral variability, such as the Builtup area and Agriculture classes. The results of the Maximum Likelihood method showed a relatively high accuracy for the Builtup area and add justification for the observations citing the results. The Maximum Likelihood method performed best for the Builtup area and Agriculture classes, with a Pixel Accuracy of 93.64% and 58.06% respectively, and an Area Accuracy of 86.28% and 81.01% respectively. This suggests that this method is effective in classifying pixels with a high degree of spectral variability, such as those of builtup areas and agriculture lands. LULC based on the three methods is shown in Figure 5.

These results suggest that the choice of method will depend on the specific requirements of the analysis and the characteristics of the data being analyzed. For example, the Spectral Angle Mapping method may be more suitable for classifying pixels with unique spectral signatures, such as those of water bodies. Meanwhile, the Maximum Likelihood method may be more suitable for classifying pixels with a high degree of spectral variability, such as those of builtup areas and agriculture lands.

It is important to note that while this study did not generate any new theories, it provides valuable insights into the possibility and utility of applying well-known, simple methods to this problem. This study highlights the importance of choosing the right method for the specific requirements of the analysis and the characteristics of the data being analyzed. In conclusion, this study provides a proof-of- concept for the use of simple methods in Machine Learning for analyzing LULC using satellite imagery. Further research is needed to explore the use of other methods and to compare their results with those obtained in this study. Additionally, the use of multispectral images, as opposed to single-spectral images, may lead to improved results in land cover classification.

For further research, it is recommended to explore the use of other classification methods such as decision tree algorithms, support vector machines, and deep learning algorithms. It would also be beneficial to analyze the results using different accuracy metrics, such as the F1 score, which takes into account both precision and recall. Additionally, the use of multispectral images, as opposed to single-spectral images, may lead to improved results in land cover classification.

V. FUTURE RESEARCH POINTERS

The field of LULC research has vast potential for further exploration. Future research could focus on several key areas to improve our understanding of the impacts of human activities on the natural environment and inform effective land management efforts. Improving classification methods, such as through the integration of machine learning and deep learning techniques, as well as other data sources like social media and citizen science data, could increase the accuracy of land cover classification for certain types. Additionally, further research is needed to assess the impacts of land use change on biodiversity, ecosystem services, and climate change, including the development of new indicators that better reflect the complexity and dynamics of land use patterns.

Integrating remote sensing and GIS data with other data sources, such as sensor data and citizen science, through data mining and machine learning, is another area of future research. Climate change and its impact on land use and land cover also require further investigation. LULC change scenarios can be developed through agent-based models and other modeling techniques to understand their impact on biodiversity, ecosystem services, and climate change. Improving monitoring systems, such as through the use of drones, UAVs, and satellite imagery, is also important to track LULC patterns over time. Finally, integrated LULC management strategies, informed by multi-criteria decision analysis, are needed to balance competing LULC needs.

In addition to the existing research pointers, future research in the field of land use and land cover (LULC) could also focus on the following areas:

(a) Improving the accuracy of LULC mapping: Further research could be aimed at improving the accuracy of LULC mapping, particularly for land cover types that are challenging to distinguish.

(b) Evaluating the effectiveness of LULC management strategies: There is a need for further research to evaluate the effectiveness of existing LULC management strategies and identify ways to improve them.

(c) Studying the role of LULC in carbon sequestration: LULC can play a critical role in carbon

ISSN 2394 - 7780

sequestration, and future research could focus on understanding this relationship and developing strategies to enhance it.

(d) Understanding the relationship between LULC and water resources: LULC can also have significant impacts on water resources, and future research could focus on understanding these impacts and developing strategies to mitigate them.

(e) Exploring the use of LULC data for decision- making: LULC data can be used to inform decisions in a variety of fields, and future research could focus on exploring these applications and identifying new ones.

Our future research would focus on these key aspects in the field of LULC, with the goal of advancing our understanding of the relationship between land use, land cover, and the environment, and identifying new and effective strategies for managing and conserving these critical resources.

VI. CONCLUSION

The analysis of LULC from satellite images is a critical aspect of understanding the impact of human activities on the natural environment and identifying areas for conservation and land management efforts. The results of this study show that different methods of LULC analysis, including Spectral Angle Mapping (SAM), Minimum Distance (MD), and Maximum Likelihood (ML), have varying degrees of accuracy in classifying different land cover types such as Builtup, Agriculture, Forest, and Waterbodies. The findings suggest that ML had the highest accuracy rate for classifying water, forest, and agriculture. On the other hand, Builtup and Agriculture showed lower accuracy rates. The study also showed that Object-based Image Analysis (OBIA) performed better than pixel-based classification approaches due to its use of spectral, spatial, and contextual data.

It is recommended to use Multiple Perceptron Neural Network (MLPNN) based Markov Chain Model (MCM) for future LULC predictions as it can process numerous transitions at one time, reducing the calibration time of the model by producing several parameter values that require less data for training. While the results of this study provide a valuable insight into the accuracy of LULC classification, further research is necessary to improve the accuracy of classification methods for certain land cover types that may be more prone to misclassification. In general, LULC research continues to play a critical role in understanding the impact of human activities on the natural environment and guiding conservation and land management efforts.

REFERENCES

[1] TechTaget. (2023, January 01). Technical support: geospatial analysis. Retrieved January 05, 2023, https://tinyurl.com/2p8w2ytu,

[2] Rawal D., Gupta V., "Land Use Land Cover Change Modeling using Multi-Layer Perceptron-Markov Chain; A case Study of Ahmedabad City", International Journal of Science and Research (IJSR), Volume 10 Issue 3, March 2021, pp. 641-652, https://www.ijsr.net/get_abstract.php?paper_id=SR21309194442

[3] Hassan, Z., Shabbir, R., Ahmad, S. S., Malik, A. H., Aziz, N., Butt, A., & Erum, S. (2016). Dynamics of land use and land cover change (LULCC) using geospatial techniques: a case study of Islamabad Pakistan. SpringerPlus, Vol 5, Issue 1.

[4] Vinayak, B.; Lee, H.S.; Gedem, S. Prediction of Land Use and Land Cover Changes in Mumbai City, India, Using Remote Sensing Data and a Multilayer Perceptron Neural Network-Based Markov Chain Model. Sustainability 2021, 13, 471.

[5] Mishra, P. K., Rai, A., & Rai, S. C. (2019). Land use and land cover change detection using geospatial techniques in the Sikkim Himalaya, India. The Egyptian Journal of Remote Sensing and Space Science.

[6] Haque, M. I., & Basak, R. (2017). Land cover change detection using GIS and remote sensing techniques: A spatio-temporal study on Tanguar Haor, Sunamganj, Bangladesh. The Egyptian Journal of Remote Sensing and Space Science, 20(2), 251–263.

[7] Chowdhury, Arnab & Dwarakish, Prof. (2022). Selection of Algorithm for Land Use Land Cover Classification and Change Detection. 2. 15- 24. 10.48175/IJARSCT-2610.

[8] Ashwini K, Sil BS. Impacts of Land Use and Land Cover Changes on Land Surface Temperature over Cachar Region, Northeast India—A Case Study. Sustainability. 2022; 14(21):14087. https://doi.org/10.3390/su142114087

ISSN 2394 - 7780

[9] Kar, Rajmita & Reddy, G.P. Obi & Kumar, Nirmal & Singh, Surendra. (2018). Monitoring spatio-temporal dynamics of urban and peri-urban landscape using remote sensing and GIS – A case study from Central India. The Egyptian Journal of Remote Sensing and Space Science. 21. 10.1016/j.ejrs.2017.12.006.

[10] Tiwari, P.C. (2000). Land-use changes in Himalaya and their impact on the plains ecosystem: Need for sustainable land use. Land Use Policy. 17. 101-111. 10.1016/S0264-8377(00)00002-8.

[11] K. Patel, M. Jain, M. I. Patel and R. Gajjar, "A Novel Approach for Change Detection Analysis of Land Cover from Multispectral FCC Optical Image using Machine Learning," 2021 2nd International Conference on Range Technology (ICORT), Chandipur, Balasore, India, 2021, pp. 1-6, doi: 10.1109/ICORT52730.2021.9582057.

[12] "Monitoring Land Use/Cover Change in Navi Mumbai, Maharashtra, India Using Remote Sensing and Satellite Data.", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.6, Issue 1, page no.726-737, January-2019, Available:http://www.jetir.org/papers/JETIR1901594.pdf

[13] TechTaget. (2023, January 01). Technical support: geospatial analysis. Retrieved January 05, 2023, https://tinyurl.com/2p8w2ytu

[14] Alshari, E. A., & Gawali, B. W. (2021). Development of classification system for LULC using remote sensing and GIS. Global Transitions Proceedings, 2(1), 8–17.

[15] Shetty, S. (2019). Analysis of machine learning classifiers for LULC classification on Google earth engine.

[16] M.B. Patel, C.G. Desai, P.N. Omericar, Image classification tool for land use / land cover analysis: a comparative study of the maximum likelihood and minimum distance method, Int. J. Geol. Earth Environ. Sci. 2 (2012) 189–196

[17] M. Salwa Thasveen and S. Suresh, "Land - Use and Land - Cover Classification Methods: A Review," 2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS), Kollam, India, 2021, pp. 1-6, doi: 10.1109/ICMSS53060.2021.9673623.

[18] Shen, L. & Li, J. & Wheate, Roger & Yin, J. & Paul, Siddhartho. (2020). Multi-Layer Perceptron Neural Network and Markov Chain Based Geospatial Analysis of Land Use and Land Cover Change. Journal of Environmental Informatics Letters. 10.3808/jeil.202000023.

[19] Gupta, Rupesh. (2014). The Pattern of Urban Land-use Changes: A Case Study of the Indian Cities. Environment and Urbanization Asia. 5. 83-104. 10.1177/0975425314521539.

[20] K. Patel, M. Jain, M. I. Patel and R. Gajjar, "A Novel Approach for Change Detection Analysis of Land Cover from Multispectral FCC Optical Image using Machine Learning," 2021 2nd International Conference on Range Technology (ICORT), Chandipur, Balasore, India, 2021, pp. 1-6, doi: 10.1109/ICORT52730.2021.9582057.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

A REVIEW ON HAND GESTURE RECOGNITION FOR SPEECH IMPAIRED PATIENTS

Sunita Gupta[1], Swapnali Lotlikar[2], Mokshi Jain[3], Heenal Patel[4]
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, Maharashtra

ABSTRACT
In the past few years, the hand gesture recognition system has received much attention due to its diverse applications and ability to effectively communicate with the machine through human-computer interaction. This article provides an overview of current hand gesture recognition systems. The key issues of the hand gesture recognition system are presented with the challenges of the gesture system. Overview methods of the current attitude and gesture recognition system are also presented. All over the world, deaf and mute people face problems in expressing their feelings to other people. Sign language is the mother tongue of deaf people, which they use in their daily life, and it facilitates the process of communication between deaf people. The problem faced by deaf people is addressed using sign language technique. Hand gestures are one of the typical methods used in sign language. It is very difficult for hearing impaired people to communicate with the world. This project presents a solution that not only automatically recognizes hand gestures, but also converts them into speech and text output so that the disabled can easily communicate with normal people. Next, a summary of the results of the review of hand gesture methods, database and comparison of the main stages of gesture recognition is given. Finally, the advantages and disadvantages of the discussed systems are explained.

Keywords: Hand Posture, Hand Gesture, Human Computer Interaction (HCI), Segmentation, Feature Extraction, Classification Tools, Neural Network, Sensor, Communication.

INTRODUCTION
The most basic and most important type of interaction with someone is communication. Gestures or sign language are used to communicate with mute and deaf patients. The typical sign language used by stupid people is difficult for most people to understand. There is also no standardized sign language in the world. QPeople with speech and hearing impairments are unable to communicate with others in a normal way. Vocalizers translate sign language into a voice that can be understood by blind and sighted people.

Gesture Vocalizer is a tool developed to help mute, deaf and blind communities communicate with each other and with the general public.[1]The system is capable of dynamic reconfiguration, so it can act as a "smart gadget" that supports many sign languages. A data glove and a microcontroller that can detect virtually all hand movements and translate some specific gestures into audible human speech help compensate for the gesture vocalizer.

According to the World Health Organization, there are about 1 million mute people and 300 million deaf people in the world. The power of communication can be either a blessing or curse. It helps express thoughts and feelings[2][3].The power of communication can be both a blessing and a curse. Expressing thoughts and emotions is beneficial. It can be quite difficult for quiet people to communicate with mute people. Communication becomes extremely difficult as most individuals are not trained in hand sign language. To bridge the communication gap between stupid and normal people, this system is introduced. This system uses a method to design an electronic glove that would aid communication.

Although much research in gesture detection and sign language has progressed, none has been structured to a significant extent. Despite the positive results of many studies, actual success is still far from definitive due to a number of assumptions shared by virtually all researchers.[4] All previous attempts are limited to sterile laboratory environments and are unsuitable for real and encompassing environments[1][3]. The most common assumption of virtually all researchers is that there should be an appropriate contrast for images and stills with bright environments.

Requiring the participation of the speech, hearing and speech impaired population in various stages of designing real-time experimental systems, such as generating a huge data set, is always a challenge. The results are limited to fewer gestures, that is, datasets associated with basic needs and emergencies, although many academics contribute datasets in the form of videos and photographs to similar works on sign language recognition.[1] There are also IoT-based alternatives that use various pressure sensors attached to gloves, but these solutions are inconvenient for those who have trouble speaking clearly while wearing gloves and leave questions about the sensitivity and accuracy of the sensor output unanswered[2].The formatter will need to create these components, incorporating the applicable criteria that follow.

ISSN 2394 - 7780

LITERATURE REVIEW

In general, communication between people with disabilities and normal people takes place through synthesized speech, which is known as sign language. Sign language is a way for deaf and mute people to communicate with each other. It has been observed that it is very difficult for the disabled to communicate with the society. Normal individuals are unable to understand their sign language. To bridge this gap, the proposed system acts as an intermediary between disabled and normal people.[4] Various researches are conducted to analyze and evaluate how the device can reduce the difficulty of communication between people with hearing and speech impairments and to find out the limitations of the device compared to other technologies and devices working towards a similar goal. Their communication with others involves only using their hand movements and expressions, and they have created artificial talking mouths for mute people. This will also help other people to understand disabled people.[2]

In the past, many techniques have been used to translate hand gestures into text. However, they were limited in terms of their functionality. Many techniques required gloves with sensors, which not only complicated the application, but also made it more expensive. In another version, the system was limited to a certain background without any noise or interference.

The system is designed for infant patients who use gloves that convert finger flexion into voice using flex sensors. Hasan [8] applied multivariate Gaussian distribution to hand gesture recognition using non-geometric features. The input hand image is segmented using two different methods [9]; segmentation based on skin color using the HSV color model and thresholding techniques based on clustering [9]. Some operations are performed to capture the shape of the hand to extract the hand element; a modified directional analysis algorithm is adopted to find the relationship between statistical parameters (variance and covariance) [8] from the data and is used to calculate the object (hand) slope and trend [8] by finding the direction of the hand gesture [8] From the resulting Gaussian function, the image was divided into circular regions, in other words, the regions are made into a terraced shape to eliminate the effect of rotation [8][9]. The shape is divided into 11 terraces with a width of 0.1 for each terrace [8][9]. 9 terraces are the result of dividing 0.1 width which are; (1-0.9, 0.9-0.8, 0.8-0.7, 0.7-0.6, 0.6, 0.5, 0.5-0.4, 0.4 -0.3, 0.3-0.2, 0.2-0.1) and one terrace for a terrace that has a value less than 0.1 and the last for the outer area that jutted out from the outer terrace [8][9].

Kulkarni [10] recognizes the static posture of American Sign Language using neural networks algorithm. The input image is converted to HSV color model, resized to 80x64 and so on image preprocessing operations are applied to hand segmentation [10]from uniform background [10], features are extracted using histogram technique and Hough algorithm. Feed forward Neural networks with three layers are used for gesture classification. 8 samples are used for every 26 characters in sign language, 5 samples are used for training and 3 samples are used for testing for each gesture, the system achieved 92.78% recognition rate using MATLAB language.[10].Wysoski et al. [8] presented rotationally invariant positions using a boundary histogram. The camera used to acquire the input image, a skin color detection filter was used, and then a clustering process was used to find the boundary for each group in the clustered image using a common contour tracking algorithm. The image was gridded and the boundaries were normalized. The boundary was represented as a string of chord sizes, which were used as histograms, by dividing the image into N number of regions in a radial shape according to a specific angle. MLP neural networks and DP matching dynamic programming were used for the classification process. Many experiments implemented on different feature formats in addition to using different chord size histogram and FFT chord size. 26 static positions from American Sign Language used in the experiments. A homogeneous background was applied in the work. Stertorous [6] proposed a novel self-growing and self-organized neural gas network (SGONG) for hand gesture recognition. A color segmentation technique based on a skin color filter in the YCbCr color space was used to detect the hand region, an approximation of the shape of the hand using a network (SGONG) was detected; The three features were extracted using a finger identification process that determines the number of raised fingers and hand shape characteristics and a Gaussian distribution model used for recognition.

METHODOLOGY

Most researchers have classified the gesture recognition system mainly into three steps to get the input image from a camera, video, or even a device equipped with a data glove.

Method of Image Extraction and Processing

The segmentation process is the first process for hand gesture recognition. It is the process of dividing the input image into regions separated by boundaries. The segmentation process depends on the type of gesture, if it is a dynamic gesture, it is necessary to locate and track the hand gesture, if it is a static gesture (posture), only segment the input image. The hand should be located first, a bounding box is generally used to determine the

ISSN 2394 - 7780

dependence on skin color, and secondly, the hand must be tracked, there are two main approaches for tracking the hand; either the video is divided into frames and each frame must be processed separately, in this case the hand image is considered as a position and segmented, or using some tracking information such as shape, skin color using some tools such as Kalman filter, the color space uses in a particular application, plays a vital role in the success of the segmentation process, however color spaces are sensitive to changes in illumination, for this reason researchers tend to use only chrominance components and neglect luminance components such as r-g and HS color space[1][2].

Classification of Gestures

After modeling and analyzing the input image of the hand, a gesture classification method is used to recognize the gestures. The recognition process is influenced by the correct selection of feature parameters and a suitable classification algorithm. For example, edge or contour detection operators cannot be used for gesture recognition because many hand positions are generated and may cause mis-classification. Euclidean distance metric used to classify gestures. Statistical tools used for gesture classification, the HMM tool has demonstrated its ability to recognize dynamic gestures in addition to finite state machine (FSM), learning vector quantization, and principal component analysis (PCA). Neural network has been widely used in hand shape extraction and hand gesture recognition. Other soft computing tools such as Fuzzy C-Mean Clustering (FCM) and Genetic Algorithms GA.[4][5]

Extraction Function

A good segmentation process leads to a perfect feature extraction process and plays an important role in a successful recognition process. It contains a segmented image vector can be extracted in different ways according to the specific application. Different methods have been used to represent the features that can be extracted. Some methods used the shape of the hand as the outline and silhouette of the hand, while others used the position of the fingertips, the center of the palm, etc.created 13 parameters as an element vector, the first parameters represent the aspect ratio of the bounding box of the hand, and the remaining 12 parameters are the mean values of the brightness pixels in the image. A self-growing and self-organizing neural gas algorithm (SGONG) was used to capture the hand shape, then three features were obtained; Palm area, downtown Palm and Hand incline. calculated the center of gravity (COG) of the segmented hand and the distance from the COG to the farthest point in the fingers and extracted one binary signal (1D) to estimate the number of fingers in the hand region. Divide the segmented image into different size blocks and each block represents a brightness measurement in the image. Many experiments have been used to decide the correct block size that can achieve a good recognition rate.[2][3].

Recognition of Sign Language

Since sign language is used to interpret and explain a certain topic during a conversation, it is given special attention. Many gesture recognition systems have been proposed using different types of sign languages. For example, recognized American Sign Language ASL using boundary histogram, MLP neural network and dynamic programming. Recognized Japanese Sign Language JSL using Recurrent Neural Network, 42 alphabets and 10 words. Recognized Arabic Sign Language ArSL using two different types of neural network, partially and fully recurrent neural network. [4,6,7]

Robot Control

Controlling a robot using gestures is considered one of the interesting applications in this area. He designed a system that uses numbering to count five fingers to control a robot using hand position characters. The robot is given commands to perform a certain task where each character has a specific meaning and represents a different function, for example "one" means "move forward", "five" means "stop" etc[8][9]

Virtual Environment (VE)

One of the popular applications in gesture recognition system are virtual VE environments, especially for communication media systems [9]. Provided real-time 3D pointing gesture recognition for natural human computer Interaction HCI from binocular views. The proposed system is accurate and independent of user characteristics and environmental changes [2][9][10].

Controlling the TV

Hand positions and gestures are used to control the television set. In a set of hand gestures are used to control TV activities, such as turning the TV on and off, volume up and down, muting the sound, and changing the channel using an open and closed hand [16][17].

ISSN 2394 - 7780

3D Modeling
To build 3D modeling, the determination of hand shapes is needed to create, build and display the 3D hand shape [9]. Some systems built 2D and 3D objects using the silhouette of a hand. 3D manual modeling can also be used for this purpose, which is still a promising area of research.

ADVANTAGES
1) Useful for handling emergency conditioning.

2) Easy to implement.

3) It provides communication between dumb and blind.

4) Easy to operate: Anyone can operate it easily.

5) Easy to define gestures: we can add or define our own gestures.

DISADVANTAGES
In this section, the disadvantages of some discussed methods are explained: Orientation histogram the method used in has some problems which are; similar gestures may differ in addition, orientation histograms and different gestures can have similar orientation histograms that the proposed method was well achieved for all the objects that dominate the image, even though they are not hand gesture A neural network classifier was used for gesture classification but it is time-consuming and when the number of training data increases, so does the time required for classification are also increased[2]. V NN required several hours of learning 42 signs and four days to learn ten words. Fuzzy c-means clustering algorithm used in has some disadvantages; The problem of extracting the wrong object occurred if the objects are larger than hand. The performance of the recognition algorithm decreases when the distance is greater than 1.5 meters between the user and the camera. In addition, its variations change to light conditions and unwanted objects can overlap with the hand gesture. There is a variation on the system the ambient lighting changes, causing mis segmentation of the hand region [12]. HMM tools are ideal for dynamic gesture recognition but are computationally intensive.

Other limitations of the system, as noted in where gestures are only performed with the right hand, the arm must be vertical, the palm facing the camera, and the background flat and uniform [15]. In System limitations limit the application such as; gestures are made with the right hand only arm must be vertical, palm facing the camera, background uniform. In a system could only recognize numbers from 0 to 9. While a system designed to control a robot, can only count the number of active fingers regardless of which specific fingers they are active with a fixed set of praises [18].

Comparison
In case of vision- based gesture recognition systems, a lot of the digital signal processing has to be done. Also there require large programming. Because of this the response of the system is quite slow [13]. Also, the electric, magnetic fields, and any other disturbances may affect the performance of the system. In case of glove-based sign language recognition system that is available in the market lot of the hardware is required. Large numbers of cables have to be connected to the computers for monitoring the data. Hence the systems require lot of space. Also, the system is not handy. In case of our project flex sensors that we are using are of low cost. Also, the processor that we use in our process is very compact. Hence the space required for our system is very less compared to the other projects that are available in the market. Thus, the system is portable [15]. Performance of the system is not affected by the disturbances. Here we are converting hand signs in to the corresponding speech signal; hence the system is the proper means of effective communication [17]. As the hardware required in designing are low cost, the overall cost of the system is less compared to the other systems available in the market, and it is the system is flexible enough for a user to add, modify or delete a hand sign [19].

Comparison Chart

Methodology	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Image Extraction	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Extraction function	X	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Classification of gestures	X	X	X	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Recognition of sign language	X	X	X	✓	X	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X
Robot control	X	X	X	X	X	X	X	✓	✓	X	X	X	X	X	X	X	X	X	X
Virtual Environment	X	✓	X	X	X	X	X	X	✓	✓	X	X	X	X	X	X	X	X	X
Controlling the T.V	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	✓	✓	X	X
3D Modelling	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	✓	X

ISSN 2394 - 7780

RESULT AND DISCUSSION

Through sign languages we can communicate without the help of acoustic sounds. Sign language uses sign patterns such as body language, movement of hands to develop the understanding between the people etc[5] .Learning of sign languages requires the special training, so this research paper is useful to bridge the gap between the deaf and dumb people and the people who can understand their language. The main result of the project is to correctly recognize and respond to the gesture accordingly[7].

CONCLUSION

Sign language may be a helpful to ease the communication between the deaf or mute community and additionally the standard people. This project aims to overcome the communication gap between the dumb community and the standard world.[14] Hand is the most richest source for communication between the people. speech impaired people uses sign languages to interact with people A smart, low cost, portable, cost-effective, lightweight, easy to use system is designed to help a person who cannot speak as compared to the other proposed system. [17]

Hence this research provides an elucidation for all the obstacles faced by all speech impaired people, as from this they will be satisfied, motivated and gain self confidence that their feelings will also be understood by other people.

FUTURE SCOPE

The application can be integrated with other mobile and IoT devices to improve user interaction and make the system more robust. The accuracy of the program can be further improvised by using neural networks.

An alternate stress could be put on the use of the application in the fields of medicines, military, governance etc. A genuine blend of various technologies in mentioned fields could make way for power tools and applications which will serve the community around the world. Finally, the use can be further designed to make more accessible to the consumers. The whole point of making the solution as a commercially viable product for the users is to help the impaired community around the world.

REFERENCES

[1] G. R. S. Murthy, R. S. Jadon. (2009). “A Review of Vision Based Hand Gestures Recognition,”International Journal of Information Technology and Knowledge Management, vol. 2(2), pp. 405- 410.

[2] P. Garg, N. Aggarwal and S. Sofat. (2009). “Vision Based Hand Gesture Recognition,” World Academy of Science, Engineering and Technology, Vol. 49, pp. 972-977.

[3] Mokhtar M. Hasan, Pramoud K. Misra, (2011). “Brightness Factor Matching For Gesture Recognition System Using Scaled Normalization”, International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3(2).

[4] Joseph J. LaViola Jr., (1999). “A Survey of Hand Posture and Gesture Recognition Techniques and Technology”, Master Thesis, Science and Technology Center for Computer Graphics and Scientific Visualization, USA.

[5] Malima, A., Özgür, E., Çetin, M. (2006). “A Fast Algorithm for Vision-Based Hand Gesture Recognition For Robot Control”, IEEE 14th conference on Signal Processing and Communications Applications, pp. 1-4. doi: 10.1109/SIU.2006.1659822 ,

[6] Verma, R., Dev A. (2009).”Vision based hand gesture recognition using finite state machines and fuzzy logic”. IEEE International Conference on Ultra-Modern Telecommunications & Workshops (ICUMT '09), pp. 1-6. doi: 10.1109/ICUMT.2009.5345425

[7] V. S. Kulkarni, S.D.Lokhande, (2010) “Appearance Based Recognition of American Sign Language Using Gesture Segmentation”, International Journal on Computer Science and Engineering (IJCSE), Vol. 2(3), pp. 560-565.

[8] E. Stergiopoulou, N. Papamarkos. (2009). “Hand gesture recognition using a neural network shape fitting technique,” Elsevier Engineering Applications of Artificial Intelligence, vol. 22(8), pp. 1141– 1158, doi: 10.1016/j.engappai.2009.03.008

[9] W. T. Freeman and Michal R., (1995) “Orientation Histograms for Hand Gesture Recognition”, IEEE International Workshop on Automatic Face and Gesture Recognition.

ISSN 2394 - 7780

[10] https://www.researchgate.net/publication/354725141_Implementation_of_Hand_Gesture_System_for_Speech_Impaired_People

[11] https://www.researchgate.net/publication/329305443_Hand_Gesture_Detection_and_Conversion_to_Speech_and_Text

[12] https://www.researchgate.net/publication/3339054_Hand-Gesture_Computing_for_the_Hearing_and_Speech_Impaired

[13] https://www.irjet.net/archives/V3/i7/IRJET-V3I7214.pdf

[14] https://www.irjet.net/archives/V6/i9/IRJET-V6I9108.pdf

[15] https://www.ijraset.com/research-paper/hand-gesture-based-speech-recognition-system

[16] http://www.sooxma.com/abstracts/Design%20and%20development%20of%20hand%20gesture%20recognition%20system%20for%20speech%20impaired%20people.pdf

[17] https://www.ripublication.com/irph/ijisaspl2019/ijisav11n1spl_31.pdf

[18] https://www.hindawi.com/journals/cin/2022/8777355/

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

COMPARISON OF BERT, XLNET, ELECTRA AND DEBERTA LANGUAGE MODELS FOR NLP

Prashant Chaudhary and Nandan Chitaliya
Assistant Professor and Student, MSc-IT, Part I, SVKM'S Usha Pravin Gandhi College of Arts, Science & Commerce

ABSTRACT
Machines can now read, grasp, decode, and comprehend human languages thanks to Natural Language Processing (NLP). Language models are essential for developing NLP applications. It takes time to construct intricate NLP language models from scratch. Several pre-trained NLP models out there are divided into groups according on the function they perform. Speech-to-text is one of the process of NLP which accurately transforming audio input to text. This paper discusses and compares NLP models that are essential for speech recognition. Speech recognition is required by any programme that follows voice commands or responds to spoken inquiries. Machine understand the data it is absorbing, NLP tasks dissect human text and speech data including identifying the part of speech and grammar of a certain sentence or passage based on its use and context.

Keywords: NLP models, BERT, XLNeT, ELECTRA, DeBRETa, Language models.

I. INTRODUCTION

With the emergence of transfer learning and pretrained language models in natural language processing (NLP), the limits of language understanding and output have been expanded. The recent breakthroughs in research have mostly followed the trend of transfer learning and the use of transformers to various downstream NLP tasks. There is disagreement in the NLP community on the utility of the massive pretrained language models that dominate the leaderboards for research However, other NLP opinion leaders point out some advantages of the current trend, such as the potential for identifying the fundamental flaws in the current paradigm. Many AI experts agree with Anna Rogers that obtaining cutting-edge results simply by utilizing more data and computing power is not research news.

Nonetheless, the most recent developments in NLP language models seem to be driven not just by the tremendous advances in computing power but also by the discovery of creative methods for lowering model weight while maintaining outstanding performance. We have compiled research papers highlighting the major language models published over the past several years to make it easier for you to remain abreast of the most recent developments in language modelling.

II. LITERATURE REVIEW

Deep learning models that have been specifically trained to carry out NLP tasks using a sizable dataset are known as pre-trained models (PTMs) for NLP.PTMs may learn universal language representations when they are trained on a big corpus. This can aid with downstream NLP tasks and save the need to train new models from start. Thus, pre-trained models may be thought of as reusable NLP models that NLP developers can use to create an NLP application rapidly. For a range of NLP applications, including text categorization, question answering, machine translation, and more, Transformers provides a library of pre-trained deep learning NLP models.

These pre-trained NLP tasks are free to use and don't require any prior knowledge of NLP. Suitable word embeddings were taught to the first generation of pre-trained models.

III. METHODOLOGY

A. BERT

BERT or Bidirectional Encoder Representations from Transformers, is a state-of-the-art paradigm for Natural Language Processing (NLP) presented by the Google AI team. According to its architecture, the model can take into consideration the information from the left as well as from right sides of each word. Though theoretically straightforward, BERT achieves cutting-edge scores on eleven NLP tasks, including named entity identification, question answering, and other tasks involving general language comprehension. By randomly masking a portion of the input tokens, a deep bidirectional model may be trained without cycles where words can indirectly "see themselves." Additionally, by creating a straightforward binary classification assignment to determine if sentence B follows sentence A quickly, pre-training a sentence connection model will help BERT comprehend the links between phrases. Developing an extremely large model.

ISSN 2394 - 7780

B. XLNet

In 20 language tasks, the huge bidirectional transformer XLNet outperforms BERT in terms of prediction metrics thanks to superior training techniques, larger data sets, and more processing capacity. Similar to BERT, XLNet analyses the words before and after a given token to produce predictions about it. This technique is known as bidirectional context. To do this, XLNet optimizes the anticipated log-likelihood of a sequence in light of all potential factorization order permutations. Since XLNet is an autoregressive language model and does not rely on data corruption, it avoids the problems with BERT brought on by masking, including the pretrain-finetune gap and the assumption that unmask tokens are independent of one another.

To further improve architectural designs for pretraining, XLNet combines the relative encoding technique and recurrence mechanism from Transformer-segment XL. Carnegie Mellon University and Google have developed a new model, XLNet, for natural language processing (NLP) applications such as reading comprehension, text categorization, sentiment analysis, and others.

The generalized autoregressive pretraining technique known as XLNet takes advantage of the best aspects of both autoencoding and autoregressive language modelling while avoiding their drawbacks. The trials show that the new model performs at a state-of-the-art level on 18 NLP tasks, outperforming both BERT and Transformer-XL.

C. ELECTRA:

The computational efficiency of the masked language modeling-based pre-training algorithms is low since they only use a small subset of the tokens for learning. The researchers recommend an unique pre-training task called replacement token identification, where a model is pre-trained as a discriminator to distinguish between genuine and replaced tokens. Some tokens are replaced by samples from a tiny generator network. The proposed method, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), is not adversarial despite resembling GAN because the token replacement generator is trained with maximum likelihood. It allows the model to learn from all input tokens rather than just a small masked-out subset.

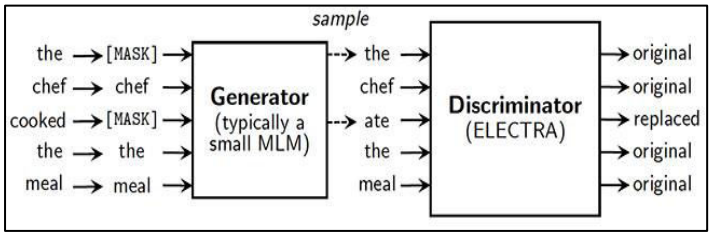


Fig 1 The fake tokens are sampled from a small masked language model that is trained jointly with ELECTRA.

D. DeBERTa:

(Decoding-enhanced BERT with disentangled attention) DeBERTa, suggested by Microsoft Research authors, has two key advantages over BERT: disentangled attention, and an improved mask decoder. DeBERTa has two vectors that each encode the content and relative position of a token or word. Whereas the self-attention in BERT is equal to merely having the first two components, the self-attention mechanism in DeBERTa processes self-attention of content-to-content, content-to-position, as well as position-to-content. To fully model relative positions in a series of tokens, the authors contend that position-to-content self-attention is also necessary. Moreover, DeBERTa has an improved mask decoder, which receives both the absolute position of the token or word and the relative information. For the first time, a single scaled-up DeBERTa variant outperforms the human baseline on the SuperGLUE benchmark. At the time of this publishing, the ensemble DeBERTa method on SuperGLUE is the best.

IV. RESULTS & DISCUSSION

In this study, we employed four models that make use of language creation and comprehension machine learning approaches. This aids in determining which characteristics have the most predictive potential. Models like BERT, XLNeT, ELECTRA, and DeBERTa are frequently utilised. The accuracy and glue score results of these classifiers are as follows:

1)**BERT:** This programme improves the state-of-the-art for 11 NLP tasks, including attaining 93.2% accuracy on SQuAD 1.1 and beating human performance by 2%. obtaining a GLUE score of 80.4%, which is a 7.6% absolute improvement over the prior best result.

ISSN 2394 - 7780

2) **XLNeT:** On 17 of the 24 tasks taken into consideration, the T5 model with 11 billion parameters produced state-of-the-art results, including: a GLUE score of 89.7 with significantly better performance on the CoLA, RTE, and WNLI tasks.A SQuAD dataset Exact Match score of 90.06.

A SuperGLUE score of 88.9, which is very close to human performance and a substantial improvement over the previous state-of-the-art result (84.6). (89.8)

3) **ELECTRA:** With a GLUE score of 79.9, tiny surpasses the similarly small BERT model (75.1) and the significantly bigger GPT model (78.8).

Only 25% of their pre-training computation is used by an ELECTRA model that outperforms XLNet and RoBERTa.

4) **DeBERTa:** DeBERTA model trained on half the training data achieves: an improvement of +0.9% in accuracy on MNLI (91.1% vs. 90.2%), an improvement of +2.3% in accuracy on SQuAD v2.0 (90.7% vs. 88.4%), and an improvement of +3.6% in accuracy on RACE (86.8% vs. 83.2%) compared to the current state-of-the-art method RoBERTa-Large.

V. CONCLUSION

Four models have been compared as a result, and each one has its own distinctive qualities. Each of the models performed well in the necessary NLP application, but the XLNet and ELECTRA mode was best for real-time deployment because to its combination of speed and accuracy. The other two models are also quite good and have various use case scenarios. It may be inferred that all of these models are suitable to be turned into marketable products and that it is quite likely that computer applications will be used in real-time.

REFERENCES

[1] Zekeriya Anil Guven, Murat Osman Unalir, Natural language based analysis of SQuAD: An analytical approach for BERT,Expert Systems with Applications, Volume 195,2022,116592,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.116592.

[2] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, Yu Sun, Pre-Trained Language Models and Their Applications,Engineering,2022,,ISSN 2095-8099, https://doi.org/10.1016/j.eng.2022.04.024.

[3] Mercier, F. (n.d.). Efficient transfer learning for NLP with electra - researchgate.net. https://www.researchgate.net. Retrieved April 18, 2021.

[4] Yao, Mariya. "10 Leading Language Models for NLP in 2022." TOPBOTS,17June2022,www.topbots.com/leading-nlp-language-models-2020.

[5] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." https://arxiv.org/, 11 Oct. 2018, arxiv.org/abs/1810.04805.

[6] Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arxiv.org, 19 June 2019, arxiv.org/abs/1906.08237.

[7] Clark, Kevin, et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." arxiv.org, 23 Mar. 2020, arxiv.org/abs/2003.10555.

[8] He, Pengcheng, et al. "DeBERTa: Decoding-enhanced BERT With Disentangled Attention." arvix.org, 2020, arxiv.org/abs/2006.03654.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

BIG DATA ANALYSIS IN E-COMMERCE: A SURVEY PAPER

Smruti Nanavaty[1] and Suneeti Kargutkar[2]
[1]Assistant Professor, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce
[2]Student-Msc IT Part 1, SVKM's Usha Pravin Gandhi of Arts, Science and Commerce

ABSTRACT
In recent times we have seen ever increasing demand of online markets, this paper helps us study in depth the role of big data analysis (bda) in online markets. Big data analysis is a vast topic which helps applications like e-commerce etc to run smoothly without worrying about data corruption, data decay etc. While data analytics provides a lot of positive outcomes it also contains few setbacks, this paper tries to deal with the setbacks. In all 5 papers are chosen to analysis big data analytics application on e-commerce platforms.

Keywords: Big data analysis, online-platform, e-commerce, business insights.

INTRODUCTION
Abundant data is generated on a regular basis while using online platforms like shopping sites, social media portals etc, this data can be classified based on the 5 majors' V' of data can be called as big data. This data can be integrated, altered according to the needs of an organisation to produce optimised business results.

The convenience provided by internet to use any online platform at any time and any place has resulted in growth of e-commerce. Moreover, a lot of youth prefers e-markets over local markets for various results. In laymen terms e-commerce can be defined as a online platform where goods/services can be sold, exchanged and purchased. In past few years the use of ecommerce websites is growing for many reasons like easy gui,time saving etc. Moreover it peaked when the pandemic hit the world and everyone was in quarantine , this time resulted as a huge profit period for ecommerce platforms who used bda to provide easy resource to costumers based on data such as sites visited by the costumers, products recently viewed etc.

Big data analysis provides a lot of positive business insights to ecommerce organisations. This review paper will helps us understand various methods used in big data analysis in e-commerce by various researchers. This paper will help the bda to improve in e-commerce platform, it will also help to compare various researchers ideology based on various bda methods usage in ecommerce platform. This paper is structured as follows: section 2 literature review, section3 summary table, section 4 conclusion.

LITERATURE REVIEW
S.S Alrumiah, Mohd. Hadwan et al[1]. This research paper gives us idea about vendor and customer based on big data analysis. Types of data in online retail is defined and how factors such as product quality, number of reviews etc influence costumers buying habits. The methodology used in this paper is based on extraction and screening. Fig 1 gives example of extraction and screening method of how papers are selected for a given purpose on basis of extraction and screening mythology. The authors draw the conclusion that integrating BDA skills into internet-retail initiative improves online sales efficiency and sellers' profits. Furthermore, by providing specialised services and products, businesses can fascinate customers by being aware of their desires and practices. BDA enhanced the online shopping experience for both customers and sellers.

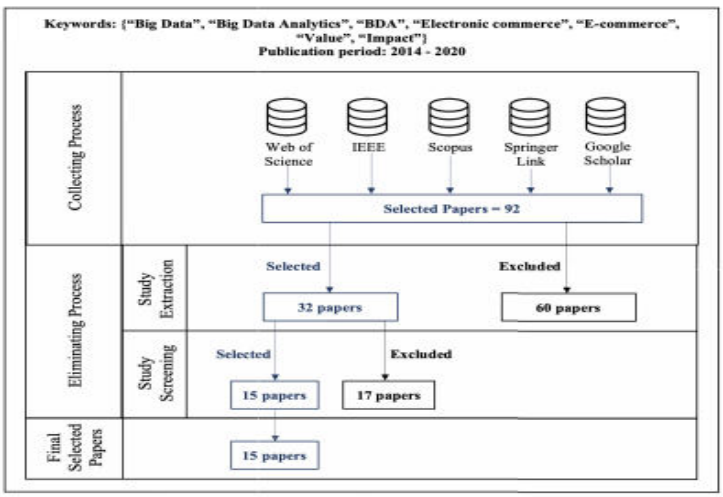


Fig1: Paper selection process [6]

ISSN 2394 - 7780

Rongrui Yu et al [2]. This research paper proposes the importance of cloud in big data analysis. The cloud environment provides the ecommerce platform which basic architecture, data storage sub-design. This article shows the use of data mining algorithm, map reduce for large data. We use JavaEE technology to design a set of full-featured web mall systems built on top of Hadoop and HBase. This document enables communication between the web system, the HBase image storage system, and the product recommendation system through web service communication technology and organizes them into an organic whole. Finally, we analyse mobile e-commerce based on impact of cloud computing. This includes relevant theory, service modes, architecture, core technologies, and e-commerce applications and other content. Gain a comprehensive understanding of the performance benefits of cloud computing.

Yuhao Gao et al [3]. This research paper suggests various types analysis like descriptive, prescriptive, business predictive analysis. This paper also discusses big data issues and how it can be managed. The paper also cover the wide range of data analysis and its future , how there is still lot to explore in big data analysis while mentioning data techniques and how data processing is an important data analysing technique. At the end of this article, we will explain how an online store can benefit from its store in the future if big data is properly analysed and manipulated. In addition, possible Expert Systems use data from big data processing to make marketing decisions that can be much better than those of an individual decision maker.

Venky Shankar et al [4]. This research paper explains how vendors can profit with the helps of big data analysis. How for data analysis descriptive and predictive models are used which provide business insights and help in increasing revenue. This research paper focuses on how big data is going to take over markets in near future and how it is already dominating with its analysis to provide personalised recommendation and offers which prove to be beneficial for both seller and buyer.

Kumud Arya,Tapan Kumar,Manoj Kumar Jain et al [5]. This research paper provides a survey-based view of various e-commerce organization of how to they try to increase their overall revenue with data analysis. This paper discusses how big data analysis is growing worldwide and how it is benefiting e-commerce platforms Hadoop and Rapidminer's Customer Churn analytics are used for valuable business insights.It also explains how, by utililizing the dynamics of people, processes and technologies to transform data into insights for sound decision-making and solutions to business problems, big data analytics (BDA) gradually adds value of e-commerce firms.

Fig 2: BDA benefits [7]

ISSN 2394 - 7780

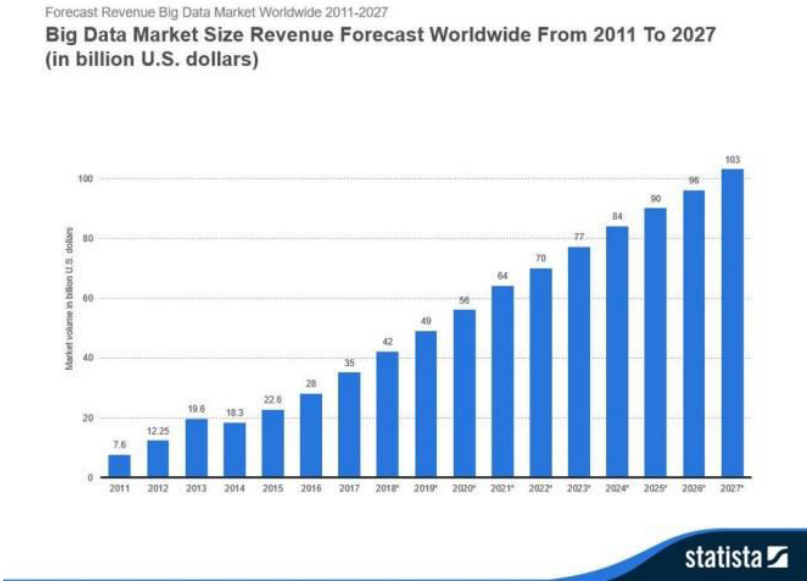


Fig3: Revenue graph [8]

Summary Table

Year	Author	Paper	Methodology/Tools	Conclusion
2021	S.S Alrumiah,Mohd.Hadwan	Implementing big data analytics in e-commerce: vendor and customer view	Screening study and extraction study.	Behaviour of vendor and customer is analysed using bda which helps in overall profit.
2021	Rongrui Yu,Chunqiong Wu,Bingwen Yan,Baoqin Yu,Xiuko Zhou,Yanliang Yu,and Na Chen	Analysis of impact of big data on e-commerce in cloud computing environment.	MapReduce	Analysis of existing electronic commerce system and its summarization is done which helps solve current problems faced by the e-commerce system.
2021	Yuhao Gao	Big data analysis on e-commerce platform	Automatic/semi-automatic data processing.	How analysis and manipulating data can lead to more superior business decision then regular methods.
2019	Venky Shankar	Big data and analytics in retailing.	Predictive model, descriptive data model.	Basic questions regarding bda and how e-vendors can use bda to increase profit are covered.
2016	Kumud Arya,Tapan Kumar,Manoj Kumar Jain	Big data analytics of global e-commerce organisations: a study survey analysis.	Hadoop, RapidMiner's customer churn analysis.	Explains how global organisations are increasing revenue with bda application

ISSN 2394 - 7780

CONCLUSION

In this review paper, we have surveyed various recent big data analyses in e-commerce research papers. We have come across various big data tools used by an e-commerce organization for top-quality business insight which will benefit the organization in increasing revenue. This research paper helps us understand how big data is used for every tiny to large business decision-making process.

We have seen the growth of prominent e-commerce websites like NetFlix, google, and Nykaa, etc and how they benefit themselves with the use of data analysis to increase their profit. The study reflects that once Big Data Analytics and its scope are well defined; distinctive characteristics and types of big data are well understood; and challenges are properly addressed, the Big Data Analytics application will maximize business value through facilitating the pervasive usage and speedy delivery of insights across organizations.

REFERENCE

[1] S.S Alrumiah,Mohd.Hadwan Implementing big data analytics in e-commerce: vendor and customer view. 2021.

[2] Rongrui Yu,Chunqiong Wu,Bingwen Yan,Baoqin Yu,Xiuko Zhou,Yanliang Yu,and Na Chen Analysis of impact of big data on e- commerce in cloud computing environment. 2021

[3] Yuhao GAO Big data analysis on e-commerce platform.2021.

[4] Venky Shankar Big data and analytics in retailing.2019.

[5] Kumud Arya,Tapan Kumar,Manoj Kumar Jain Big data analytics of global e-commerce organisations: a study survey analysis. 2016.

[6] S.S Alrumiah,Mohd.Hadwan Implementing big data analytics in e-commerce: vendor and customer view.2021 fig 3 pg3

[7] Whizlabs blog

[8] www.forbes.com

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

COMPARISON OF YOLOV3, YOLOV5S AND MOBILENET-SSD V2 FOR CURRENCY DETECTION FOR VIRTUAL IMPARED PERSON

Sunita Gupta, Darshan Dhuri and Dheeraj Mistry
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce Maharashtra, Mumbai, 400056

ABSTRACT
Deep learning is now demonstrating its full potential with a wide range of applications and is fundamental to numerous technical fields. The tracking and recognition of objects is one of the more popular applications of deep learning. Recent developments in this regard have yielded encouraging results. This study examines and contrasts various systematic. The main feature of this app will be that it can identify the money using a camera and read out the value of the note. Three popular machine learning algorithms—YOLOv3, YOLOv5, and MobileNet-SSD V2— are used to recognize currency in app. Each model examines a currency, and they are evaluated according to how accurately they do so and how quickly the output can be processed.

Keywords: YOLOv3; YOLOv5; MobileNet-SSD.

1. INTRODUCTION
A technology known as currency detection is used to recognise and locate banknotes or coins based on their pictures. Since it is used to automate the process of handling and counting currency, it is a crucial task in the financial sector. Applications for currency detection systems include automated teller machines (ATM), vending machines, and point-of-sale systems. The issue of currency recognition can be approached in various ways, including by using both machine learning and image processing techniques. One of the most common methods is to identify and categorise coins or bills in an image using object identification algorithms like YOLOv3, YOLOv5s, and MobileNet-SSD V2.These algorithms may learn the characteristics specific to each denomination by training on a dataset of photographs of coinsor banknotes. After being taught, the algorithm can be used to find coins or bills in fresh photos. The effectiveness of YOLOv3, YOLOv5s, and MobileNet-SSD V2 for the task of cash identification will be compared in this comparison. To identify which method is most appropriate for this task, we will compare their accuracy, speed, and resource needs.

2. LITERATURE REVIEW
The two components of object detection are localization andcategorization. The selectable features (Haar, HOG, and convolutional layer) are first extracted as part of the detection pipeline before the item is classified using a localizer or classifier. These localizers and classifiers often operate on an image using a region proposal approach or an in-sliding window technique. Methods like Deformable Parts Models (DPM) are examples of the sliding window approach, and methods like R-CNN employ the region proposal approach to generate bounding boxes before applying a classifier to the defined bounding boxes. Then duplicate bounding boxes are filtered out during post-processing. Because each component used in such systems is taught independently, the pipeline methods is difficult to optimize and exceedingly complicated. However, by combining into a single network, methods like YOLO have turned object detection as another, single error. As a result, in systems such as YOLO, the algorithm performs equations on to an image to recognize items and classify them. Furthermore, MobileNetV2's low latency and low power models have shown good accuracy. In order to determine the most effective algorithm for currency detection system, this paper analyses the performance of the YOLOv3, YOLOv5s, and MobileNet-SSD V2 systems.

3. EXISTING ALGORITHMS

3.1 Mobilenet SSD
Cnn models are used to construct a design with several layers capable of classifying input objects into any of the known structures. These items are now detectable. Because of recent advances in deep learning for image processing, higher resolution feature maps are being used. An object detection model known as Mobile net SSD uses an input picture to calculate the output bounding box and class of an item. The Single Shot Detector (SSD) object recognition model can swiftly identify things that are designed for mobile devices thanks to Mobile net as its foundation.

3.1.2 One stage Detector
When learning the bounding box coordinates and class probabilities from an input, object detection is a straightforward regression problem. The one stage detector employs YOLO, YOLO v2, SSD, and Retina Net, among other technologies. In object detection, a complicated form of image classification, a neural network

ISSN 2394 - 7780

anticipates items in a picture and detects those using coordinates. The major purpose of our study is to evaluate the deployable efficacy and precision of the object identification systems YOLO and MobileNet SSD in a range of settings, as well as to highlight some of the key features that distinguish our study from others. In figure 1.1 one stage detector diagram

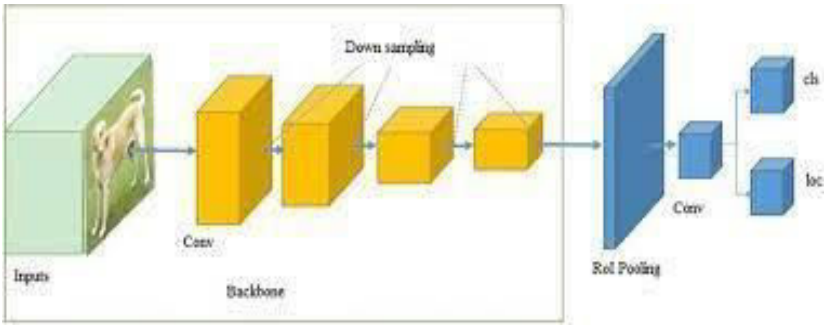


Figure (1): one stage detector diagram

3.1.3 Two Stage Detector

This requires two steps to finish the detection. In the first stage, regions of interests with a high likelihood of becoming an object are generated using a region proposal network. The second phase is object detection, during which objects are classified definitively and subjected to bounding box regression. Some examples of two stage detectors include RCNN, Fast RCNN, SPPNET, Faster RCNN, etc.

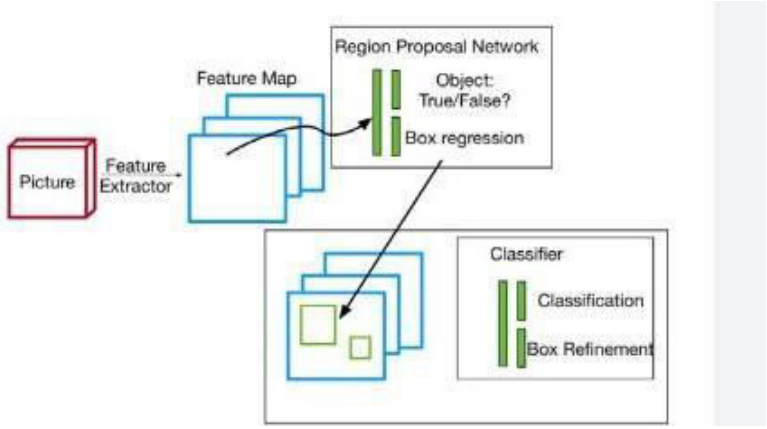


Figure 2: Two stage detector diagram

3.1.4 SSD

The word SSD refers to a single shot detector. To get the final detections, the SSD technique employs a non-maximum suppression phase following the generation of a fixed-size collection of bounding boxes and scores for the existence of object class instances in those boxes by a feed-forward convolutional network[9].Scores include confidence values for each category of object, with the number 0 denoting the backdrop. This presents detecting techniques with many citations and resolutions. Multi-reference approaches generate a number of frames at varying points in a image, each with its own size and aspect ratio, and then predict the detection box using these references. The capacity to recognize objects at different scales and on different network layers enables multi resolution techniques. An SSD network implements a method for recognizing multiple item classes in pictures by generating confidence scores for the presence of each object category in each default box. Furthermore, it modifies boxes to better fit the shape of the objects. The CNN-based SSD Architecture detects entity goal classes in two steps: (1) con vnet feature extraction and (2) sensor application. SSD extracts feature maps using VGG16. The Conv4 3 layer of the VGG16 is then used to detect objects. The class with the highest score is chosen as the one for the bounded item in each prediction, which consists of a bounding box and 21 scores for each class (plus one additional class for no object) [3].During training, the main aim is to get a high class confidence score by matching the default boxes with the ground truth boxes.

3.2 Machine Learning Algorithm and ImageProcessing- based system

The dataset has been standardised in this system due to the wide range of feature values. The K-fold cross-validation method was then used to partition the data. Utilizing K- Nearest Neighbors, Support Vector Classifier, and Gradient, the prediction model was trained.

ISSN 2394 - 7780

Amplifying Classifier By evaluating a data point's closest neighbours and assigning each one a score based on the distance between them, KNN categorises the provided data point. Based on the distance, the nearest data points are given more weight. SVM classification is carried out by selecting the best-fitting hyper-plane for splitting the category. Gradient boosting classifiers are created using a paradigm akin to a decision tree, in which layers of yes-or-no questions are put out to create a prediction model. All three classifiers in this system offer more than 97% accuracy in the task of classifying currencies.

3.3 Edge Detector-Based System

Here, a camera or other device has been used to take a pictureof the currency [8]. Then the image is resized and converted to the grayscale format. Then the edges of the image have been detected using an edge detector. Then the image is segmented using various machine learning and clustering algorithms. Some dimensionality reduction approaches are then utilized to highlight the key elements ofthe image. The generated picture is then compared to the data set already in place as the last step to determine if thenote is authentic or fraudulent (counterfeit).

3.4 Canny Edge Detection Algorithm-basedsystem

The reference image's security properties are saved in the firststage. After the user presses the "Compare" button, the GUI software will compare the note under test to the reference note [9]. The GUI makes it evident how these two notes differ in the comparison. Here, the Canny edge detection technique is used to improve and sense images. The parameters of this multiphase edge detection technique are an upper and lower threshold. Some Indian banknotes have Optically Variable Device (OVD) patches, a special security feature that detects counterfeit notes better than the traditional three-wayrecognition of bills.

4. METHODOLOGY

4.1 Dataset

The Data was imported from kaggle to make sure the modelsare fed with a wide variety of variations for smooth operationafter deployment, a repertory of photos featuring different views of the cash was used. All appropriate photos were gathered, and then they were cropped to a 1:1 aspect ratio to prevent lossy training owing to dimensional variances. For arelatively short training procedure in this instance, the photoshave been cropped to a resolution of 416x416. The length of time needed to train the models also increases as this resolution is increased.

4.2 Training

The models were trained using tensor flow. By carrying out this on the platform, the training period was significantly reduced. The data collection contains about 1000 training photos for each testing data point. There are approximately 50,000 training photos for 50 testing data with varied visual scenarios. This is done to guarantee that findings from the picture categorization are highly accurate. In the second stage of building the CNN model approach using Tensor Flow, the parameter with 10,000 iterations was handled with multiple epochs. the training precision for epochs of 10 and 100. In the last step of image classification performance, numerous photographs were tested, and an accuracy % based on CNN image classification was calculated.

5. RESULTS & DISCUSSION

In this study, we employed three machine learning-based algorithms for classification and regression. This aids in determining which characteristics have the greatest predictive potential. The most popular models are the YOLOv3, YOLOv5s, and MobileNet-SSD-V2. We assessed these classifiers' performance using measures for accuracy and f-beta score. To study how the size of the training data impacts the prediction scores, we trained our classifier modelusing a range of training data sizes. For the test data as well as a portion of the training data, prediction scores were produced. We also determined how long training and prediction would take. Finally, we used grid search to furthernarrow the model that was chosen.

ISSN 2394 - 7780

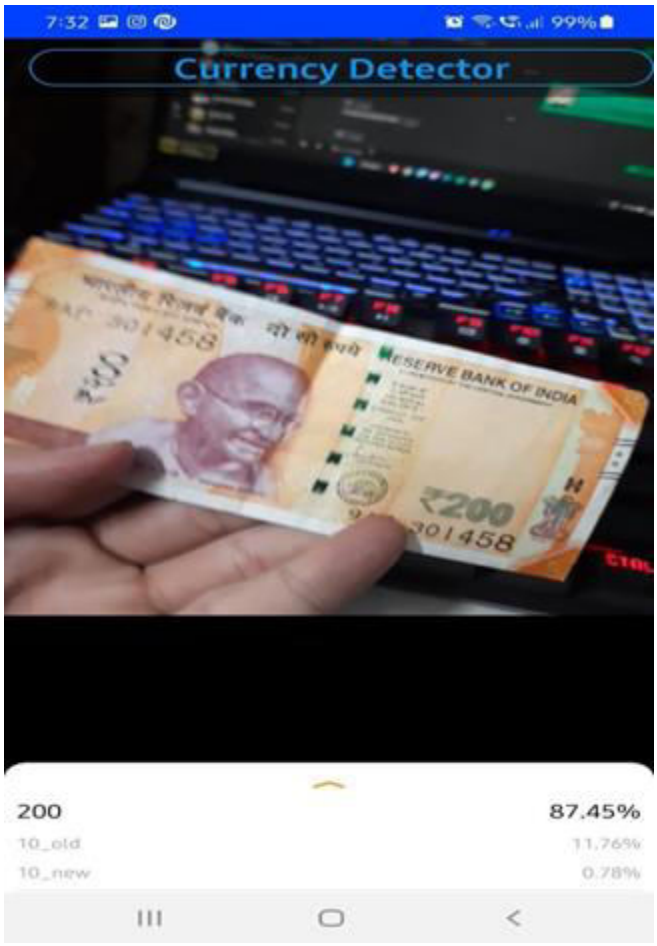


Figure -3: Outputs of the MobileNet-SSD V2 Architecture. 200 Indian Currency is detected.

Currencydetection	YOLOv3	YOLOv5S	MobileNET-SSD V2
Train setaccuracy	0.9867	1.0	. 1.0
Test set accuracy	0.98	0.9945	0.9982
mAP (%)	. 33.7	10.7	54.4

6. CONCLUSION

In this project, we have proposed a mobile application for currency recognition that identifies Indian currency to assist blind people in their daily lives and address the common aiming problem for blind users. Regional audio is the outputformat for this project. Better performance and recall values can be found in the Mobile NET-SSD V2 algorithm. This work will be expanded to apply the categorization to distinguish between genuine currency and counterfeit money. Foreign languages that are widely spoken can be included. To create a voice note in the user's native language that can recognize cash notes on a low-end mobile phone for visuallyimpaired people.

7. REFERENCES

[1] Ashwani Kumar , Sonam Srivastava ,Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector,Third International Conference on Computing and Network Communications (CoCoNet19.)

[2] Alexey Bochkovskiy , Chien-Yao Wang,,Hong-YuanMark Liao,YOLOv4: Optimal Speed and Accuracy of ObjectDetection,arXiv:2004.10934v1,23 April 2020.

[3] Ã• ngel Morera , Ã• ngel SÃ¡nchez , A. BelÃ©n Moreno , Ã• ngel D. Sappa and JosÃ© F. VÃ©lez SSD vs. YOLO for Detection of Outdoor Urban Advertising Panels under Multiple Variabilities, Sensors 2020, 20, 4587; doi:10.3390/s20164587.

[4] Mark Sandler, Andrew Howard , Menglong Zhu , Andrey Zhmoginov and Liang-Chieh Chen ,MobileNetV2: Inverted Residuals and Linear Bottlenecks,arXiv1801.04381v4.

[5] Mohit Phadtare , Varad Choudhari , Rushal Pedram and Sohan Vartak, Comparison between YOLO and SSD MobileNet for Object Detection in a Surveillance Drone, IJSREM, 2021.

ISSN 2394 - 7780

[6] Harshal Honmote , Pranav Katta , Shreyas Gadekar and Prof. Madhavi Kulkarni,Real Time Object Detection and Recognition using MobileNet-SSD with OpenCV, International Journal of Engineering Research & Technology(IJERT), ISSN: 2278-0181, Vol. 11 Issue 01, January-2022.

[7] Lu Tan, Tianran Huangfu, Liyao Wu and Wenying Chen,Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identifcation, Tan et al. BMC Medical Informatics and Decision Making (2021) 21:324 https://doi.org/10.1186/s12911-021-01691.

[8] Andrew G. Howard, Menglong Zhu, Bo Chen,Dmitry Kalenichenko Weijun Wang,Tobias Weyand, Marco Andreetto and Hartwig Adam, MobileNets:EfficientConvolutionalNeuralNetworksforMobileVisionApplications,arXiv:1704.04861v.

[9] Wei Liu , Dragomir Anguelov , Dumitru Erhan, Christian SzegedyScottReed,ChengYangFu,andAlexanderC.Berg,SS D:SingleShotMultiBoxDetector,arXiv:1512. 02325v5.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

OPTIMIZING THE NETFLIX STREAMING EXPERIENCE WITH DATA SCIENCE

Palak Katrodia
Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Maharashtra, Mumbai-400056

ABSTRACT
The streaming industry has seen a significant growth in recent years, with Netflix being one of the leading players in the market. The company's success can be attributed to its vast library of content and personalized recommendations. However, as the number of subscribers continues to grow, it becomes increasingly important for Netflix to optimize the streaming experience for its users. In this paper, we will explore how data science can be used to optimize the Netflix streaming experience by analysing user behaviour and preferences, and by extracting information from the content itself. The goal is to make recommendations that are tailored to each individual user, leading to a more enjoyable streaming experience and an overall improvement in the performance of the platform.

Keywords: Netflix, Streaming experience, User behaviour, Data Science

I. INTRODUCTION
Netflix was founded in 1997 as a service that shipped DVDs to customers by mail. In 2007, Netflix launched the product we're most familiar with today: streaming movies and TV over the Internet. However, as the number of subscribers continues to grow, it becomes increasingly important for Netflix to optimize the streaming experience for its users. This can be achieved by using data science to analyse user behaviour and preferences, and to make recommendations that are tailored to each individual user.

The use of data science in the streaming industry is not a new concept, but as the industry continues to grow and evolve, it becomes more important to use data science to stay competitive. With the vast amount of data available, data science can be used to gain insights into user behaviour and preferences, to improve the overall performance of the platform and to make more accurate recommendations.

In this paper, we will explore how data science can be used to optimize the Netflix streaming experience. We will examine the different ways data science can be used to analyse user behaviour and preferences, as well as the content itself. We will also discuss the potential benefits of using data science in this way, including a more enjoyable streaming experience for users and an overall improvement in the performance of the platform.

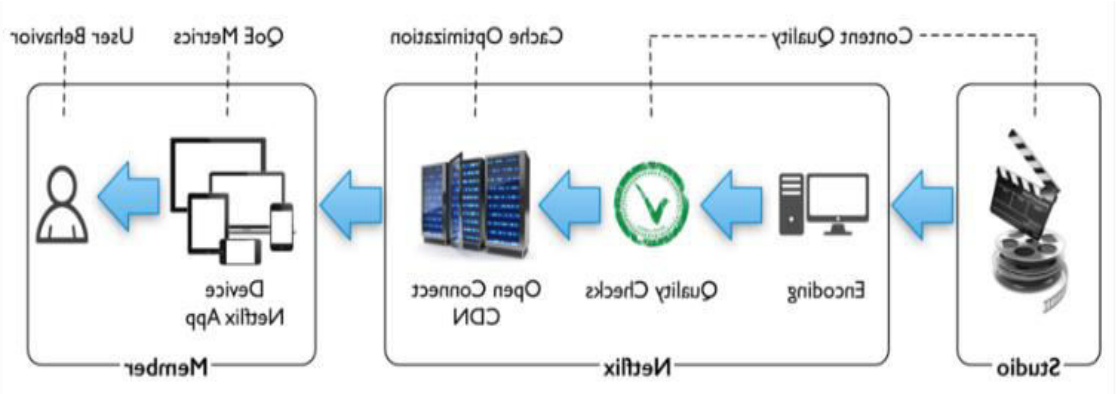


Figure 1. The Netflix Streaming Supply Chain: opportunities to optimize the streaming experience exist at multiple points

By using data science to analyse user behaviour and preferences, as well as the content itself, Netflix can make more accurate recommendations that are tailored to each individual user. This can lead to a more enjoyable streaming experience, as users are more likely to find content that they are interested in. Additionally, by understanding user behaviour, Netflix can also improve the overall performance of the platform, such as by reducing buffering times and improving video quality.

ISSN 2394 - 7780

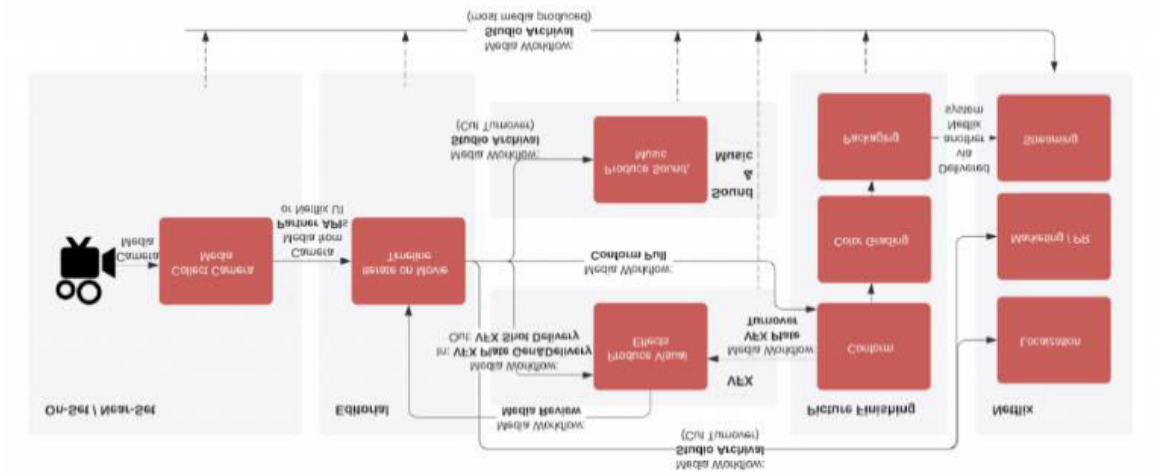


Figure 2. Data Science at Netflix: How Advanced Data & Analytics helps Netflix Generate Billions

II. METHODOLOGIES & ALGORITHMS

Netflix used data and algorithms to personalize movie recommendations for its users. By 2007, the company had developed an algorithm called Cinematch that was able to predict which movies a user would enjoy based on their viewing history and ratings. This algorithm was a key factor in the success of Netflix's movie rental business.

A. Cinematch Algorithm

Cinematch is a recommendation algorithm developed by Netflix that is based on collaborative filtering. The algorithm uses the preferences and behaviour of similar users to make recommendations to a target user. Cinematch takes into account a user's past ratings of movies and TV shows, as well as their viewing history, to recommend new content that they are likely to enjoy.

The algorithm uses a combination of user-user and item-item collaborative filtering to make recommendations. User-user collaborative filtering compares a target user's preferences with those of similar users, while item-item collaborative filtering compares a target user's preferences with those of users who have rated similar movies or shows. By combining these two methods, the Cinematch algorithm can make more accurate recommendations to users.

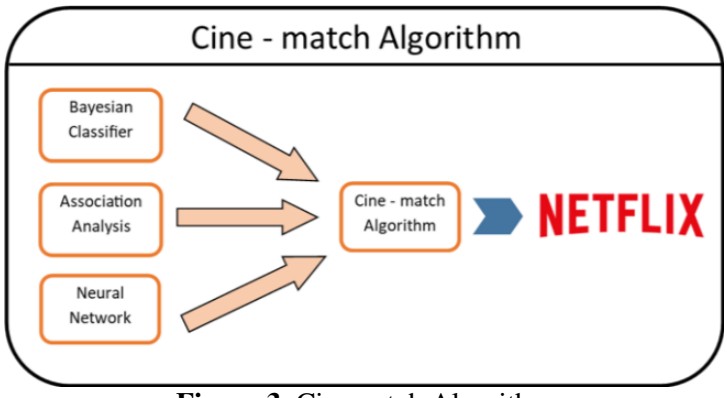


Figure 3. Cinematch Algorithm

One of the key features of the Cinematch algorithm is its ability to learn over time. As users rate more movies and TV shows, the algorithm updates its recommendations based on the new information. This allows the algorithm to continuously improve its recommendations over time, resulting in a more enjoyable streaming experience for users.

There are several methodologies that can be used in optimizing the Netflix streaming experience with data science:

a) Collaborative Filtering

Collaborative filtering is a technique that can be used in optimizing the Netflix streaming experience with data science. It is based on the idea that users who have similar preferences and behavior will also have similar interests in movies and shows. Collaborative filtering can be used in two ways: user-user collaborative filtering and item-item collaborative filtering.

ISSN 2394 - 7780

1. **User-User Collaborative Filtering:** This method looks at the preferences of similar users to make recommendations to a target user. Netflix can use user-user collaborative filtering to recommend shows and movies to a user based on what similar users have watched in the past. For example, if two users have similar viewing habits and both have watched a particular show, Netflix can recommend that show to the target user.
2. **Item-Item Collaborative Filtering:** This method looks at the preferences of users for similar items to make recommendations to a target user. Netflix can use item-item collaborative filtering to recommend shows and movies to a user based on what similar shows and movies they have watched in the past. For example, if a user has watched a lot of romantic comedies, Netflix can recommend other romantic comedies to the user.

Collaborative filtering is a powerful technique because it can make recommendations based on the preferences and behavior of similar users. This can lead to more accurate recommendations for the target user, resulting in a more enjoyable streaming experience. Netflix can use this method by collecting the data of the user's behavior and preferences and then apply machine learning algorithms to identify the patterns and make recommendations based on that. Additionally, Netflix can also use item-item collaborative filtering by analyzing the metadata of the movies and shows and then recommend the similar items.

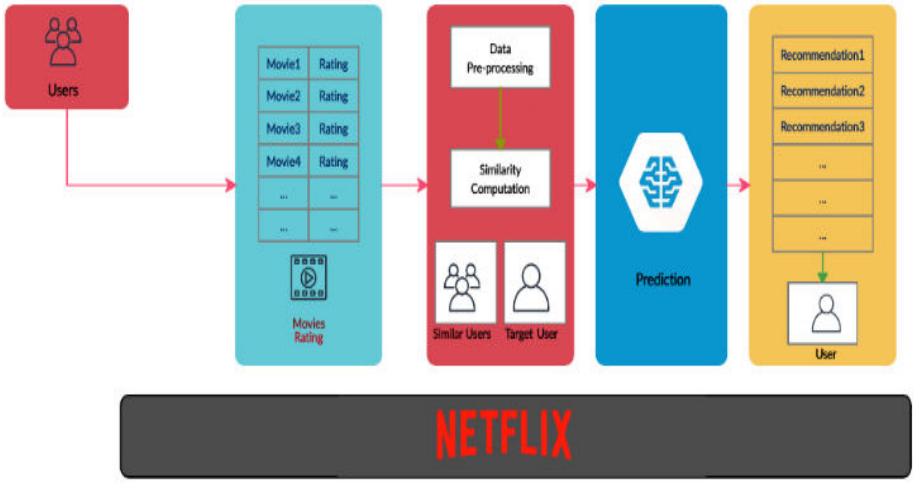


Figure 4. Collaborative Filtering used by Netflix

b) Content-Based Filtering

Content-based filtering is another technique that can be used in optimizing the Netflix streaming experience with data science. It is based on the idea that users will be interested in content that is similar to what they have already watched. This method uses the characteristics of the content, such as genre, plot, and main characters, to make recommendations to users.

1. **Analyzing Movie and Show Metadata:** Netflix can use data science to extract information such as genre, plot, and main characters from each movie and show. By analyzing this metadata, Netflix can make recommendations to users based on their interests and preferences. For example, if a user has watched a lot of action movies, Netflix can recommend other action movies to the user.
2. **Natural Language Processing (NLP):** NLP is a technique that can be used to extract information from the description, reviews, and other written content of a movie or show. By using NLP, Netflix can analyze the text data and make recommendations based on the user's interests and preferences. For example, if a user has watched a lot of movies with a certain actor, Netflix can recommend other movies with that actor to the user.
3. **Analyzing Audio and Visual Features:** Netflix can also use data science to analyze the audio and visual features of a movie or show. This can include analyzing the color, lighting, and camera angles used in a movie, as well as the tempo and rhythm of the soundtrack. By analyzing these features, Netflix can make recommendations based on the user's interests and preferences.

ISSN 2394 - 7780

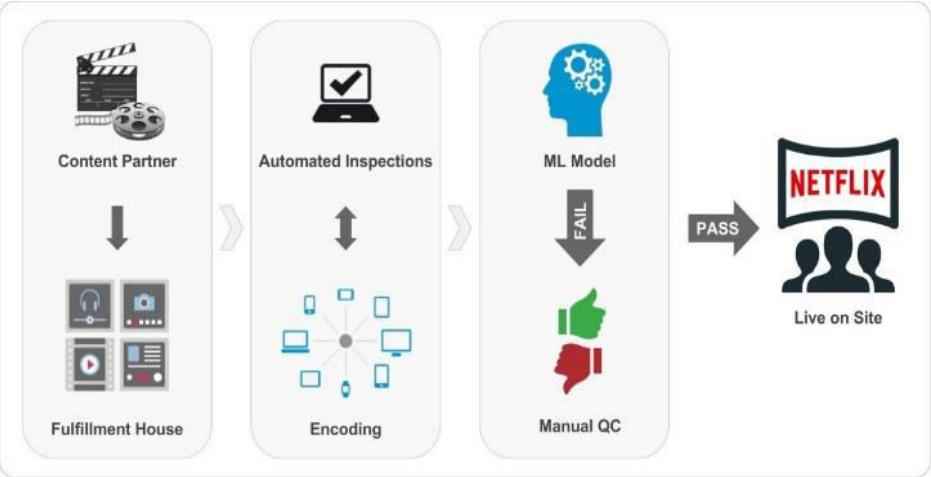


Figure 5. Optimizing Content Quality Control at Netflix

C) A/B Testing

A/B testing is a methodology that can be used to optimize the Netflix streaming experience with data science. It involves experimenting with different versions of an algorithm or recommendation system to determine which one performs the best. This can help Netflix to continuously improve its recommendations over time.

1. **Experimenting with Different Algorithms:** Netflix can use A/B testing to experiment with different recommendation algorithms. For example, the company can test a collaborative filtering algorithm against a content-based filtering algorithm to see which one performs better. By comparing the performance of different algorithms, Netflix can determine which one is the most effective at making accurate recommendations to users.
2. **Experimenting with Different User Interfaces:** Netflix can also use A/B testing to experiment with different user interfaces. For example, the company can test a new layout for its recommendation page to see if it leads to more users clicking on the recommended shows and movies. By experimenting with different user interfaces, Netflix can determine which one is the most effective at getting users to engage with the content.
3. **Experimenting with Different Variables:** Netflix can also use A/B testing to experiment with different variables. For example, the company can test the effect of different recommendation algorithms on different segments of users (e.g. users from different regions, ages, etc.). By experimenting with different variables, Netflix can determine which one is the most effective at making accurate recommendations to different segments of users.

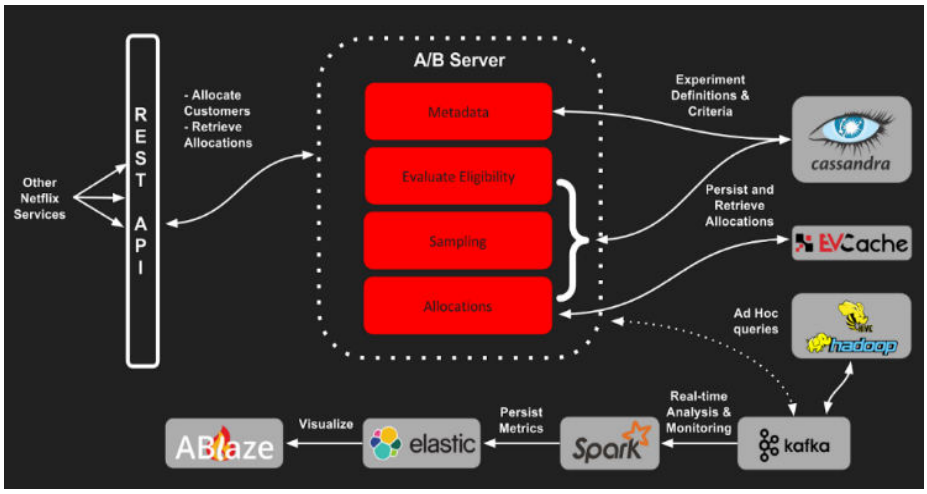


Figure 6: A/B Testing at Netflix

III. NETFLIX REVENUE AND SUBSCRIBER GROWTH

As of 2021, Netflix had over 208 million subscribers worldwide, with the majority of subscribers located in the United States. In 2020, the company reported a revenue of $25.9 billion, an increase of 21% from the previous year. This growth can be attributed to an increase in the number of subscribers, as well as an increase in the average revenue per subscriber.

ISSN 2394 - 7780

In the first quarter of 2021, Netflix added 4 million new subscribers, bringing the total number of subscribers to 209.18 million, despite the pandemic.

Since June 2020, the number of Netflix paying members has grown by 8.4%.

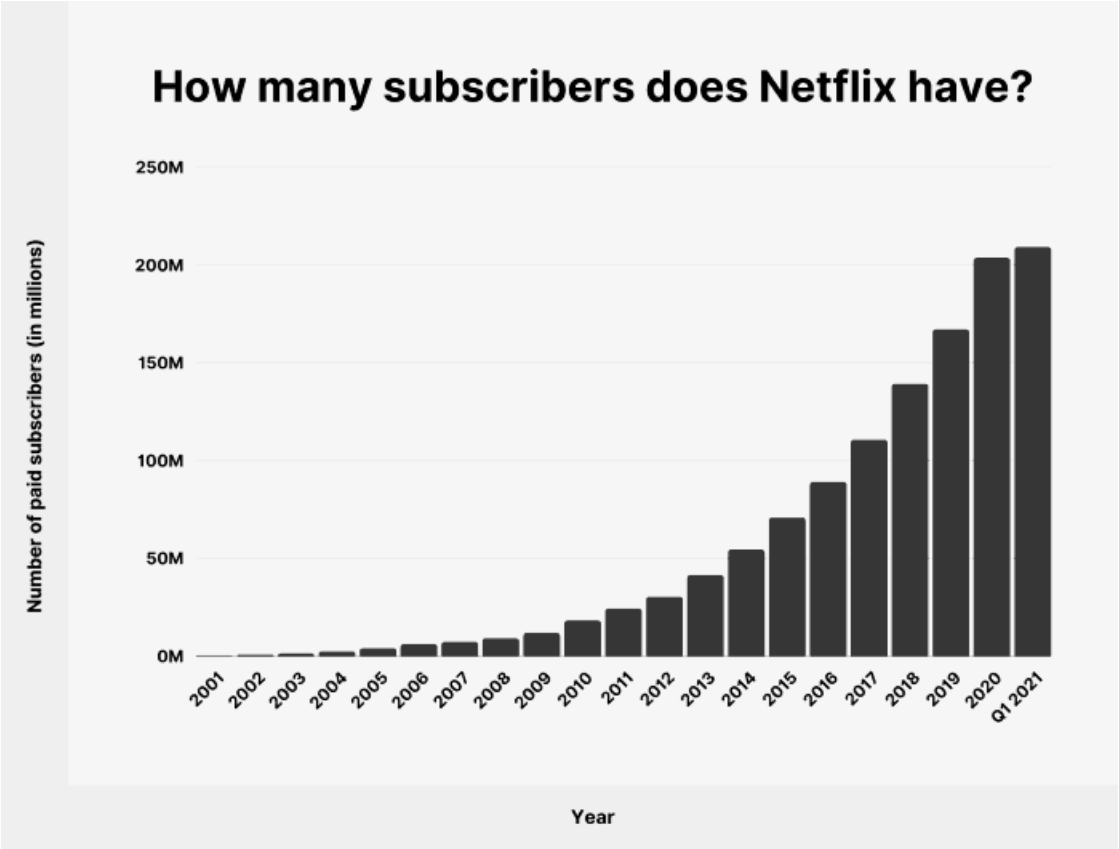


Figure 7. Subscriber Growth

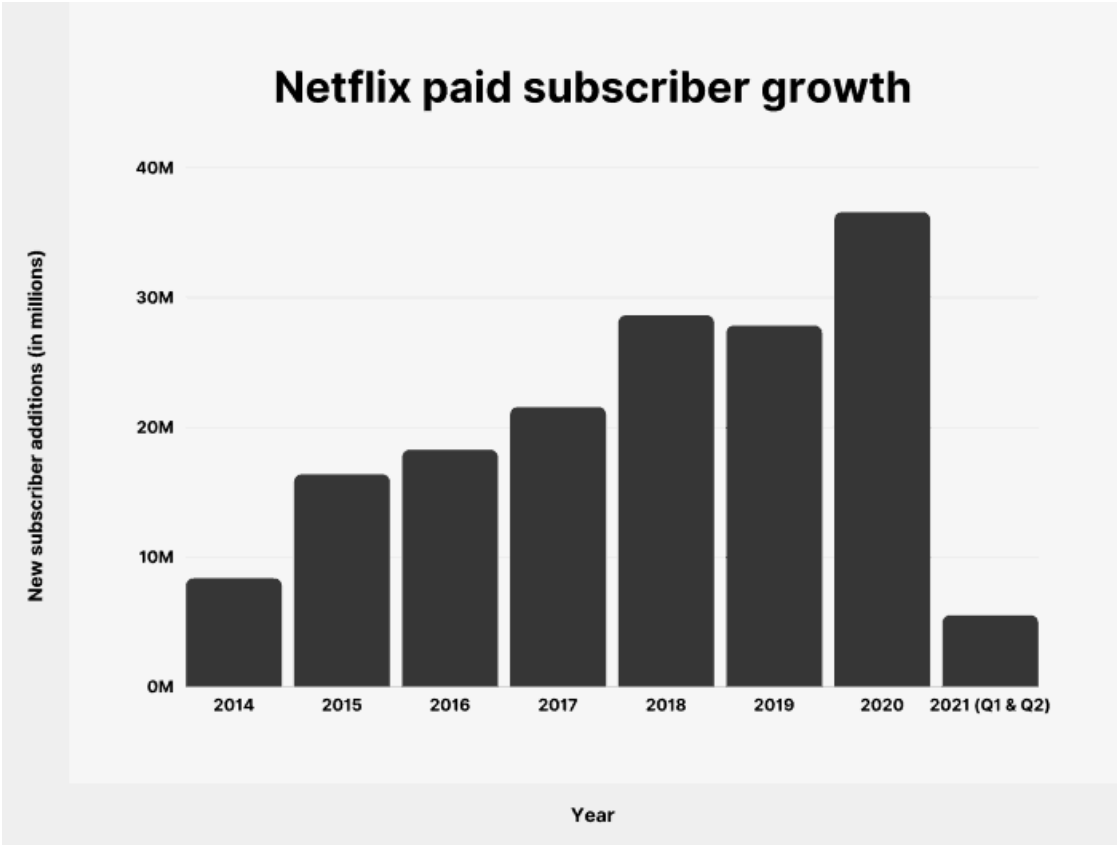


Figure 8. Paid Subscriber Growth

ISSN 2394 - 7780

Below is charted the growth of Netflix's subscriber base since 2001:

Year	Number of paid subscribers
2001	400 thousand
2002	796 thousand
2003	1.41 million
2004	2.48 million
2005	4.02 million
2006	6.15 million
2007	7.32 million
2008	9.16 million
2009	11.89 million
2010	18.26 million
2011	24.30 million
2012	30.36 million
2013	41.43 million
2014	54.47 million
2015	70.83 million
2016	89.09 million
2017	110.64 million
2018	139.25 million
2019	167.09 million
2020	203.66 million
Q1+Q2 2021	209.18 million

Table 1. Netflix's Subscriber Growth Chart

In terms of revenue growth, Netflix has consistently seen year-over-year growth in the last 5 years. In 2020, the company reported revenue of $25.9 billion, an increase of 21% from the previous year. In 2021, the company's revenue reached $7.16 billion in Q1, up 24% YoY, and $7.2 billion in Q2, up 22% YoY.

Netflix has recorded a compound annual revenue growth rate of 35.68% since 2001.

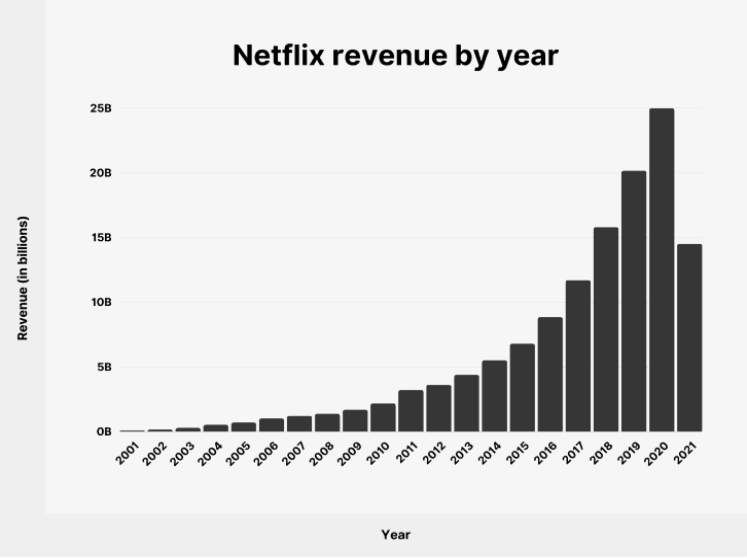


Figure 9. Revenue Growth

In terms of profit, Netflix has also seen significant growth in recent years. In 2020, the company reported a net income of $1.7 billion, an increase of 72% from the previous year.

ISSN 2394 - 7780

Here's a full breakdown of Netflix's revenue by year:

Year	Revenue
2001	$75.91 million
2002	$152.80 million
2003	$272.24 million
2004	$506.22 million
2005	$682 million
2006	$997 million
2007	$1.2 billion
2008	$1.36 billion
2009	$1.67 billion
2010	$2.16 billion
2011	$3.2 billion
2012	$3.6 billion
2013	$4.37 billion
2014	$5.5 billion
2015	$6.78 billion
2016	$8.83 billion
2017	$11.69 billion
2018	$15.79 billion
2019	$20.15 billion
2020	$24.99 billion
2021 (Q1 & Q2)	$14.5 billion

Table 2. Netflix's Revenue Growth chart

IV. CONCLUSION

In conclusion, optimizing the Netflix streaming experience with data science is a critical aspect of the company's success. By analysing user behaviour and preferences, as well as the content itself, Netflix can make more accurate recommendations that are tailored to each individual user. This can lead to a more enjoyable streaming experience for users and an overall improvement in the performance of the platform.

The use of machine learning algorithms, natural language processing, user segmentation, A/B testing, and predictive modelling are some of the ways data science can be used to optimize the streaming experience. Collaborative filtering, content-based filtering, hybrid filtering, matrix factorization, deep learning, and clustering algorithms are some of the algorithms that can be used to make the recommendations.

Netflix's Cinematch algorithm, initially developed in-house, is an example of how data science and collaborative filtering are used to make recommendations to users. Additionally, this company's revenue and subscriber growth statistics, which are consistently growing, showing the importance of the optimization and personalization of the streaming experience.

Overall, the use of data science in the streaming industry is critical to staying competitive. With the vast amount of data available, data science can be used to gain insights into user behaviour and preferences, to improve the overall performance of the platform and to make more accurate recommendations. This can lead to a more enjoyable streaming experience for users, and ultimately drive the success of the platform.

ISSN 2394 - 7780

V. REFERENCES

[1] Ko, H.; Lee, S.; Park, Y.; Choi, A. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. Electronics 2022, 11, 141. https://doi.org/10.3390/electronics11010141

[2] Ko, Hyeyoung, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields" Electronics 11, no. 1: 141. https://doi.org/10.3390/electronics11010141

[3] Beheshti, A.; Yakhchi, S.; Mousaeirad, S.; Ghafari, S.M.; Goluguri, S.R.; Edrisi, M.A. Towards Cognitive Recommender Systems. Algorithms 2020, 13, 176. [Google Scholar] [CrossRef]

[4] Walek, B.; Fojtik, V. A Hybrid Recommender System for Recommending Relevant Movies Using an Expert System. Expert Syst. Appl. 2020, 158, 113452. [Google Scholar] [CrossRef]

[5] Wu, M.-L.; Chang, C.-H.; Liu, R.-Z. Integrating Content-Based Filtering with Collaborative Filtering Using Co-Clustering with Augmented Matrices. Expert Syst. Appl. 2014, 41, 2754–2761. [Google Scholar] [CrossRef]

[6] Goldberg, D.; Nichols, D.; Oki, B.M.; Terry, D. Using Collaborative Filtering to Weave an Information Tapestry. Commun. ACM 1992, 35, 61–70. [Google Scholar] [CrossRef]

[7] Fayyaz, Z.; Ebrahimian, M.; Nawara, D.; Ibrahim, A.; Kashef, R. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Appl. Sci. 2020, 10, 7748. [Google Scholar] [CrossRef]

[8] Kawasaki, M.; Hasuike, T. A Recommendation System by Collaborative Filtering Including Information and Characteristics on Users and Items. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence, Honolulu, HI, USA, 27 November–1 December 2017; pp. 1–8. [Google Scholar]

[9] Gomez-Uribe, C.A.; Hunt, N. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 2016, 6, 1–19. [Google Scholar] [CrossRef][Green Version]

[10] Bell, R.M.; Koren, Y. Lessons from the Netflix Prize Challenge. SIGKDD Explor. Newsl. 2007, 9, 75–79. [Google Scholar] [CrossRef]

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

REVIEW OF COMPARISON OF IMAGE CLASSIFICATION TECHNIQUES

Harsh Dhiraj Jethwa
Student, MScIT, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
Images can be recognized by machines in the same way as we humans do. Image classification is used to narrow the gap between the computer vision and the human vision. In general, so many recent advances in computer vision and machine learning have been achieved through extensive pre-training on large datasets. For example, the most popular image classification dataset, ImageNet-1k, contains thousands of classes, each consisting of hundreds or thousands of training examples. Key Words: Image classification, CNN, KNN, ResNet, SVM

INTRODUCTION
Images, including photos and videos, represent a large part according to a single criterion. On the other hand, in many training data mainly labelled images of cats and dogs. In simple terms, most networks breakdown the images into abstract shapes and colours which are used to form a hypothesis regarding the image content.

An image classifier takes the numerical pixel values of an image, passes it through a CNN and get a final output. This result can be a single class or the probability of a class that best describes the image. The magic happens in the hidden layers of CNNs. Image classifications can be single-character or multi-character. With single-label classification, each image has only one label or annotation. As a result, for each image, the model visualizes, analyzes and classifies the image of global data generation. It uses artificial intelligence-based image classification to interpret and organize this data. Image classification is one of the buzzwords in automation, the task of classifying and labeling images or vectors within images by grouping them according to one or more criteria, image classification analyzes photos with deep learning models based on artificial intelligence, which can identify and recognize a variety of criteria, from the content of the image to the time of day. Imagine the classical example, you are given a set of images each of which depicts a cat or a dog. Instead of labelling the pictures all on your own, you want to use an algorithm to do the work for you. It looks at the whole picture and outputs possibilities for each of the classes it was trained on. This is usually made possible through training neural networks. As in other applications of supervised learning the network is fed with the sufficient amount of indications. Depending on the classification, some images may have more than one symbol. An image with all symbols used simultaneously. Many aspects affect success, efficiency and impact. However, choosing the right tool is one of the most important steps. Properly classifying your images can save you time and money, while achieving the best results. Whether you make stationary toys, sell vintage clothing, or sell image classification software, we can help you improve the accuracy and efficiency of your process.

Table 1. Summary of the related works of classification systems

Research No	Name/Year	Title of project	Purpose	Method Used	Result
Research 1	Gregor, Danihelka, Graves, Rezende, & Wierstra (2015)	DRAW: A Recurrent Neural Network for Image Generation	➢ Train neural network for image classification ➢ Trained complex images with MNIST models	Artificial Neural Network (ANN)	Classification improved even naked eye cannot distinguish it with main data
Research 2	Rastegari, Ordonez, Redmon, & Farhadi (2016)	XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks	➢ Balanced number of face images and non-face images are used for training ➢ Employing the bi-scale CNN 120 trained with the auto-stage training	Convolutional Neural Network	Current state of the art achieves about 80% detection rate with 50 false positives.
Research 3	Kamavisdar, Saluja, & Agrawal (2013)	A Survey On Image Classification Application Techniques	➢ Multiple dataset that being located under each of Hierarchical classifier ➢ Rejection of the class on the intermediary stage	Decision Tree	Considered very simple and high rate of efficiency
Research 4	Pasolli, Melgani, Tuia, Pacifici, & Emery (2014)	SVM Active Learning Approach for Image Classification Using Spatial Information	➢ Combining spatial information from sequential process of trial process with spectral	Support Vector Machine (SVM)	Two (2) of the images have high resolution in terms of effectiveness of regularly
Research 5	Korytkowski, Rutkowski, & Scherer (2016)	Fast Image Classication by Boosting Fuzzy Classifiers	➢ Simply boosting Meta knowledge where local characteristic can be mostly found	Fuzzy Classifiers	Testing process give short period of time where it produce 30% shorter compared to the previous one.

ISSN 2394 - 7780

RELATED WORKS

The neural network architecture (NNA) framework is a combination of facial expressions from the two pairs of human eyes and automatic encoding of transform sequences. Many complex images are generated, but the system is slowly refining the MNIST model during this study. We're also testing with the Street View house number dataset, which improves results because it's indistinguishable with the naked eye. By using a two-layer CNN with 120 trained data points, and stepwise machine learning, image classification can achieve a

detection rate of 81.6% with 6 false positive triggers. The Face Detection Dataset and Benchmark (FDDB) is currently stateof-the-art and achieves a detection rate of approximately 80% with 50 false positives.

EXISTING TECHNIQUES

3.1 Deep Convolutional Neural Network Architecture: -

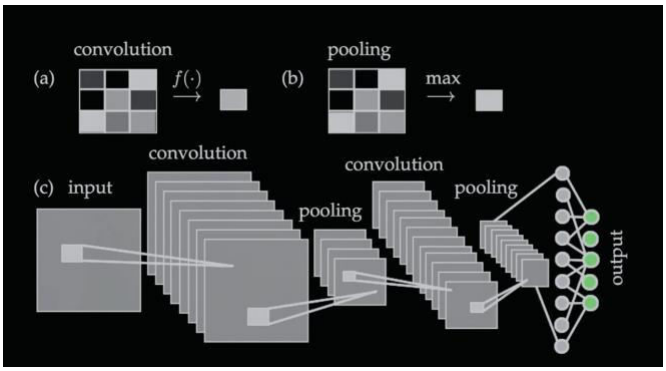


Fig 1.1 A typical deep CNN architecture

Training Neural Networks is one of the most common, or what people find successful with neural networks, is image classification or image classification. [1]In 2012, the ability to train neural networks at scale on deep neural architectures with image networks became possible thanks to advanced datasets and computing power that allowed us to actually see and begin to exploit the structure of deep neural networks. Thus, image networks represented a truly significant transition that revolutionized modern deep learning architectures. The above architecture is a standard type of architecture that humans have come to develop for image tagging and classification. It usually has many layers. You can see many different architectures, structures and styles on the internet, but most of them have this basic scheme. So you usually have an input, and the input is an image. The goal of most image processing is to take an input and label it. So what is common in an image grid is a series of images with a rating on the back. From this input we go to the output signal, to the output. This is also very important in the development of cameras for self-driving cars. A self-driving car's camera must map the entire scene, whether it's pedestrians, cyclists, other vehicles or the road. All this is achieved in this type of architecture. The two types of visual functions are the idea of convolution and convergence. Building computer-based algorithms often has two main components. You usually start with an image, which is a large image that you want to map to a tag space, and the tag space can be anything. You can have thousands of tags, and your goal is to take a photo, run it through this network, and create an exact tag for that image.

One of the interesting things about what convolution does is look at parts of an image. That means a convolution window that can slide over the entire image. The idea is, for example, in a window that looks for features. If it's a picture and there's a puppy in the right corner, it might not be the most helpful to see the whole picture. However, I have a scrolling window that goes through the entire image and checks the label of each small subblock and what properties the associated subblock has. You can build this curved window in several layers. Randomly running this network can create a filter from the input space to the convolutional space. Each individual pixel in the convolution window is produced by a non-linear output function (which can be a RELU), and the sigmoid will produce an output that gives an estimate of what's going on in that convolution window and will train the weights. give yourself a little representative types of images from input to output. You can do this multiple times to create different filters for your angular layer. By convolution with different activation functions, we can generate a set of functions representing the input image in this convolutional space. Typically in image processing you control a polar window. The sum is to take the square and take the average. This is called the maximum amount. This returns only the maximum value. And notice what happens along the way. It's compression. So you take a block, transform it into the average or maximum of the block, compress the image into a convolution layer, then do another convolution layer on top of another convolution layer, and so on. Therefore, most of these network architectures start with a large convolutional window and slowly

ISSN 2394 - 7780

subdivide it through several layers of this network down to a very small group of servers, eventually expanding to fully externally connected networks. These structures can have billions of network weights, so the only way to really use the full power of a deep CNN is having an equally large training dataset to pour data from. This is why imaging is so important.

Deep Residual Learning for Image Recognition : -

[2]ResNet is a powerful CNN that won the image net challenge in 2015. [3]ResNet was actually able to achieve pretty powerful performance with around 3.57% error which is pretty incredible. The main key point behind ResNet is right now they are essentially able to stack layers maybe around 152 layers approximately within the ResNet and overcome a problem known as the vanishing gradient problem. Basically we use the gradient to try to go back to the network and update it's weights and one of the challenges when you stack all of this different layers on top of each other the gradient becomes very small and essentially the network performance becomes extremely poor and you are not able to train the network anymore if you have a very deep CNN. Deep neural networks are difficult to train because of the problem of vanishing gradients when updating the weights. We need to use back propagation. ResNets include 'skip connections' feature which enables training of multiple deep layers (152 layers) without vanishing gradient issue.

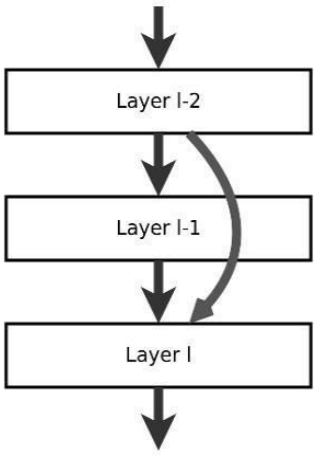


Fig 1.2 Skip connections

ResNet use the concept of skip connections, which can solve the vanishing gradient problem. Figure 1-2 shows three convolutional layers stacked one after the other.These are traditional networks. This is called skip connection and using this ResNet solves the gradient loss problem. Some layers are skipped so that the weights due to missing links are not too small.

K -Nearest Neighbor Based Image Classification : -

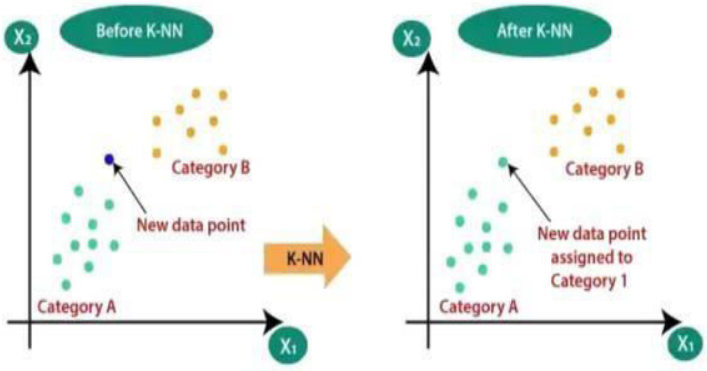


Fig 1.3 a) Before applying b) after applying

One of the simplest machine learning algorithms based on supervised learning methods is K closest neighbour. In order to classify a new data point based on similarity, it stores all of the existing data. With the KNN algorithm, fresh data can be quickly and accurately categorised into a suitable category. Although the KNN technique can be used for both classification and regression, it is often employed for classification problems. KNN is a non-parametric method since it does not make any assumptions about the underlying data. It stores

ISSN 2394 - 7780

the dataset instead of instantly learning from the training set, and when it comes time to classify, it makes a change to the dataset. When there is minimal to no information available regarding the distribution of data, this algorithm is typically utilised. It may simply be thought of as a prediction-making algorithm. The parameter K in KNN denotes the number of closest neighbours to a specific data point that will be taken into consideration while reaching a decision. The result relies on the class to which the majority of these nearby points belongs, hence this is the primary determining factor. KNN uses a variety of distance metrics, including the Hamming, Manhatten, and Euclidean distances, to calculate distances. Comparatively speaking, it is more accurate than most categorization methods. **3.4 Support Vector Machine : -**

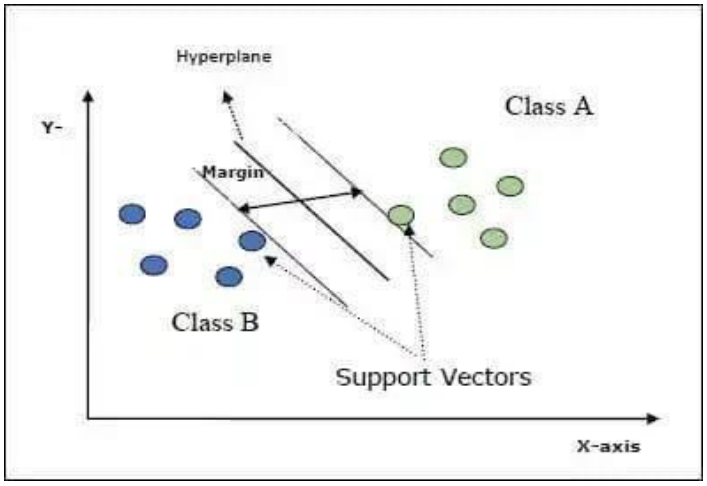


Fig 1.4 Support Vector Machine

One of the most important aspects in machine learning is the classification of a huge collection of objects into two or more categories. Is that a dog or a cat in the picture? The stock is up or down. Support vector machines, sometimes known as SAMs, are among the most straightforward and, perhaps, elegant systems for categorization. Any object you want to categorize has coordinates that are presented as a point in an n-dimensional plane; these coordinates are often referred to as features. With all the points belonging to one category being on one side and all the points belonging to the other side being on the other, SVMs generate a hyperplasia, which can be either a line in 2D or a plane in 3D, to do the classification test. SVM looks for the feature that maximizes the margin to points in the other category and so best separates the two categories, even if there may be more than one of these features. The support vectors are the points that exactly fall on the margin, and the margin is the space in between these points. SVMs create a hyperplasia—either a line in 2D or a plane in 3D—with all the points belonging to one category falling on one side and all the points belonging to the other category falling on the other in order to perform the classification test. Despite the possibility of more than one of these qualities, SVM aims to find the one by expanding the distance to points within the other category that divides the two groups the most efficiently.To use SVM, you only need to load a Python library, prepare your training data, and feed it to the fit function before calling predict to give a new object the appropriate category. The major benefit of SVMs is their ease of understanding, application, and interpretation. Also, they perform well even with very little training data. Another issue with SVMs is their ease of use. The points cannot be separated by a hyperplane in many situations. In this situation,

a) Adding some nonlinear features to the data that are already there is a frequent workaround.

b) In this higher dimensional space, locate the separating hyperplane; and

c) Project back to the initial space

SVM can be used for text recognition, face recognition, and spam filtering.

RESULTS & DISCUSSION

[7] When compared to other approaches being employed in this application, the suggested fish species picture classification method's accuracy is exceptionally high at 96.29%. [6] The authors suggested a brand-new method for differentiating among normal and abnormal CT scans. Pre - processing stage, extraction of features, feature reduction, and classification are the four steps of the suggested technique. They applied support vector machines, k-nearest neighbours, and artificial neural networks to achieve classification likelihood of success of 92%, 97%, and 98%, respectively. [5] 26 positive cases have been employed to test the suggested system. CNN is one of the best deep learning techniques. The author collected 26 cases of brain tumors. KNN, ANN, SVM, and CNN were ranked with an accuracy of 89%, 90%, 91%, and 95%, respectively. The proposed method was

ISSN 2394 - 7780

evaluated against a dataset of 26 patents. With small modifications, the system could also be developed to classify other types of cancer.

Table 2. Classification Accuracy for the used classifiers

Classifier	KNN	ANN	SVM	CNN
Accuracy	89%	90.11%	91.11%	95%
Error	11%	7.89%	8.89%	5%

In this paper, we used four algorithms which uses techniques for classification. The techniques mainly used are Deep CNN, ResNet, KNN and SVM respectively

CONCLUSION

In this, we have provided brief summary of three architectures and each of them demonstrates their own unique characteristics. Each architecture was able to apply the desired image classification. We can conclude that it is possible to use real-time computer vision applications for image classification.

REFERENCES

[1] Karol Gregor , Ivo Danihelka , Alex Graves , Danilo Rezende , Daan Wierstra , DRAW: A Recurrent Neural Network For Image Generation. arXiv:1504.04623 , 2015.

[2] Alex Krizhevsky , Ilya Sytskever , Geoffrey E. Hinton , ImageNet Classification with Deep Convolutional Neural Networks NIPS'12;Proceesings of the 25th International Conference on Neural Information Processing Systems, 2012.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren & Jian Sun , Deep Residual Learning for Image Recognition IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2016.

[4] Lorenzo Brigato , Bjorn Barz , Luca Iocchi , Joachim Denzier , Image Classification with Small Datasets: Overview and Benchmark , 2022.

[5] Hema Rajini N , Automatic Classification of MR Brain Tumor Images using KNN, ANN, SVM and CNN , 2017. https://doi.org/10.21744/irjeis.v3n1.895

[6] Zhang, W. L. & Wang, X. Z , Feature extraction and classification for human brain CT images , International Conference on Machine Learning and Cybernetics , 2007.

[8] Dhruv Rathi, Sushant Jain, Dr. S. Indu , Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning arXiv:1805.10106 , 2018.

[9] Mohd Azlan Abu , Nurul Hazirah Indra , Abdul Halim Abd Rahman , Nor Amalia Sapiee , Izanoordina Ahmad , A Study on Image Classification based on Deep Learning and Tensorflow ISSN 0974-3154, pp. 563-569 , 2019.

International Journal of Advance and Innovative Research
Volume 10, Issue 2 (XI): April - June 2023
ISSN 2394 - 7780

EARLY PREDICTION OF CANCER USING AI/ML TECHNIQUES - A CRITICAL STUDY

Mayisha Lubana Hussain, Pranjyoti Hazarika and Dr. Dibya Jyoti Bora
Department of IT, SCS, The Assam Kaziranga University, Jorhat, Assam, India

ABSTRACT
When a tumor initially develops, it is usually classified as benign or malignant. Commonly malignant tumors are referred to as cancer. Therefore, early detection is crucial in order to stop the tumor's unusual growth. In this study, we have examined previously published research papers that suggested a machine-learning techniques to detect cancer at an early stage. The papers have analyzed their findings on numerous malignancies, including skin, lung, breast, colorectal, etc and thereby presenting a precise review on the same.

Keywords: AI, ML, malignant, benign, tumor

I. INTRODUCTION
AI and ML have recently played a significant role in the development and upkeep of the healthcare system, greatly aiding it. Cancer as defined by The National Cancer Institute is a disorder in which some body cells develop out of control and spread to other body regions. Typically, a malignant tumor, as it is also known, causes cancer. The uncontrollable growth and spread of malignant tumors affect the area directly afflicted and, in a later stage, can affect other areas as well. Due to the disease's unnoticed aspect, which makes early diagnosis difficult to detect, the challenge with the treatment is that it begins in the late stages, making it tough to successfully treat cancer. Machine learning techniques can be very useful in detecting cancer at an early stage. Several machine learning techniques have been extensively used to develop cancer prediction models, resulting in effective and accurate decision making. Additionally, patients' harmful radiation exposure can be prevented. The tumor cells may be accurately identified and non-invasively diagnosed using machine learning.

In the papers we reviewed, it is found that generally, supervised learning algorithms are employed to predict cancerous cell growth at an early stage. They analyse the performance based on the sensitivity, specificity, and accuracy of the model. The predicted accuracy is determined by using the testing set, which provides an estimate of the conversion errors. The training and testing datasets used are sufficiently large and independent to provide accurate information on the model's predictive ability.

II. LITERATURE REVIEW
Maclin, et.al suggested using an artificial neural network [1] that was conditioned using numerical ultrasound data from 52 confirmed cases. On an 80386 microcomputer, using the conventional backpropagation methodology, the nonlinear artificial neural network was trained on software. In 52 patients (17 malignant, 30 cystic, and 5 other) at a hospital in Memphis, the ANN used ultrasound data. The trained prototype worked flawlessly on 47 cases that weren't included in the training set of data. This study has to be expanded to include more situations in order to validate this prototype.

Breast cancer diagnosis using an SVM-based technique and feature selection was proposed by Akay, et. al [2]. The Wisconsin breast cancer dataset has been split into various training-test divisions for experiments. The method's performance is measured using classification accuracy, sensitivity, specificity, positive and negative predictive values, receive operating characteristic (ROC) curves, and a confusion matrix. The findings demonstrate that the SVM model with five features achieves the maximum classification accuracy (99.51%)

Bottaci, et. al suggested using artificial neural networks to recognize patterns in intricate biological datasets which are impossible to find with traditional linear statistical analysis [3]. The researchers created six types of neural networks which can predict death of cancer within 9, 12, 15, 18 and 24 months. And the model was trained and validated using 5 years follow-up information of 334 patients having colorectal cancer. The 12 months trained neural network was then applied to 2 years follow-up data from another institution's patients. For the prediction of mortality for specific patients at an institution Iwithin9, 12, 15, 18, 21, and 24months, all six neural networks achieved an overall accuracy of more than 80%. When the neural network trained to predict death within 12 months was applied to data from the second institution. It achieved an overall accuracy of 90%.

Vaka, et al proposed a Deep Neural Network with Support Value (DNNS) mode and a new algorithm to detect breast cancer [4]. The proposed method improves the effectiveness and image quality for more accurate breast cancer diagnosis and prediction. The mathematical equations for calculating the histogram value, sigmoid function, and histo-sigmoid function in the proposed DNNS approach have been modified and rewritten and the existing equations are comparable to conventional breast cancer detection techniques [5].The dataset was

ISSN 2394 - 7780

obtained from the M.G Cancer Hospital & Research Institute in Visakhapatnam. To address the issue of restricted data size, data augmentation is used to expand the dataset. The DNNS approach is divided into three stages: Preprocessing, Feature extraction, and Histo-sigmoid fuzzy clustering. The author has found that the benefits of the suggested algorithm include improved performance, effectiveness, and image quality.

A method to identify breast cancer was proposed by Hadidi et al. [6], first part consists of image processing techniques, and the second part consists of two supervised learning models one is Back Propagation Neural Network (BPNN) and the other is Logistic Regression (LR). Image Processing techniques are used to convert breast cancer images into a different format and to extract features. Image cropping was used for extracting a specific part of the image, Weiner filter was used for noise reduction and grey scale format was kept for the image. The image was transformed from the time domain to the frequency domain using Discrete Wavelet Transformation (DWT) and the output matrices were fed into the algorithm. Logistic Regression (LR) and Backpropagation Neural Network (BPNN) are used for classification. 209 images from 50 patient's cases were taken in order to train, test, and validate methods each dataset contained 110 images with breast cancer and 96 regular images, which were organized as vector and were used in Logistic Regression (LR) and Backpropagation Neural Network (BPNN). The result of these models is Logistic Regression using a higher number of features and Backpropagation Neural Network (BPNN) crossed 93% regression with 240 features.

Using the concept of transfer learning, Khan, et al. proposed a new deep-learning framework for the identification and classification of breast cancer in breast cytology images [7].The proposed framework uses previously trained CNN architectures such as GoogLeNet, Visual Geometry Group Network (VGGNet) and Residual Networks (ResNet), to extract features from images. These architectures are then fed into a fully connected layer for average pooling classification of malignant and benign cells. The proposed framework's accuracy has been found to be superior to all previous deep learning architectures. Researchers used two datasets one dataset is a benchmark dataset and the other was created locally. In both the first augmentation techniques are used to produce 8000 images overall by scaling, rotating, translating, and modeling colors. 200 of these 8000 images are used for testing, while 6000 images are used to train the architecture.

Jeyaraj,et al designed and developed a partitioned Convolution Neural Network (CNN) to detect benign and cancerous images [9]. The developed method is compared with SVM and Deep Belief Network (DBN) to validate the effectiveness of classification. Bagging and boosting methods are used to improve performance accuracy. The suggested model is assessed using sensitivity, specificity, and classification accuracy. The proposed model had an accuracy of 91.4, specificity of 0.94, and sensitivity of 0.91which performed better than the compared method (SVM and DBN).

Lu, et al. proposed machine learning techniques that were successful in correctly classifying Benign Ovarian Tumors (BOT) and Ovarian Cancer (OC) [10]. The dataset includes of 49 different features, such as demographics, blood tests as part of standard care, general chemistry, and tumor markers, present in 349 Chinese patients. Data for the 235 patients (89 BOT and 146 OC) were subjected to the machine learning Minimum Redundancy - Maximum Relevance (MRMR) feature selection method, from which a straight forward decision tree model was built. The remaining 114 patients were used to test the model (89 BOT and 25OC).The outcomes were compared to the predictions generated by the logistic regression model and the risk of ovarian malignancy algorithm (ROMA). Furthermore, the model produces more accurate predictions than ROMA.

Alhazmi, et al proposed a model which employed artificial neural network (ANN) [11].Dataset is obtained between 2017 and 2020 from the oral pathology lab located at Prince Mohammed Bin Nasser Hospital. A total of 138 cases with pathologic reports were received. Out of which 65 cases were excluded due to missing values. 10000 interactions were trained for the model. Researcher's used10-foldcross-validationto evaluates the analysis's generalizability to a different dataset, and then several threshold values are evaluated in order to determine the sensitivity and specificity. The model's ability to accurately predict oral cancer was78.95%.

Xiao et al proposed a multi model deep learning-based ensemble method for cancer prediction [12]. The model has two stages, in the first stage, five classification methods such as: k-nearest neighbor (KNN), SVM, decision tree (DT), Random forest and Gradient-boosting decision tree. After applying the 5-fold cross-validation technique, the average prediction is derived from these classification methods. For the integration of all the first stage prediction a deep neural network is used and in the second stage multi-model ensemble method is employed. Lung Adenocarcinoma (LUAD), Stomach Adenocarcinoma (STAD), and Breast Invasive Carcinoma (BRCA) datasets are used, and gene expression data is obtained. While testing LUAD, STAD, and BRCA datasets, researchers obtained an accuracy of 99.20%, 98.78%, and 98.41%, respectively.

ISSN 2394 - 7780

Wan et al used Cross-validation (CV) techniques to train and test ML models to achieve more sensitivity and precision in a massive colorectal cancer dataset, with the majority of cases being early-stage [13]. To standardize each feature, large outliers were swapped out with the 99th percentile value. The standardized data were then optionally subjected to principal component analysis (PCA) and truncated singular-value decomposition (SVD). Two classification algorithms- Logistic regression and support vector machine (SVM) were taken into consideration for training. Five distinct CV approaches, including k-fold, binnedage, kbatch, balanced k-batch and ordered k-batch are used to get estimates of model performance. Except for the binnedage model, all models' CVs were computed using five folds (k=5). OncfDNA extracted from plasma samples (N = 546 colorectal cancer and 271 non-cancer controls), whole-genome sequencing was carried out. Reads that aligned to the bodies of protein-coding genes were removed and read counts were normalised. IchorCNA was used to determine the cfDNA tumor fraction. Models were trained using k-fold cross-validation and confounder-based cross-validation to assess generalization performance In a colorectal cancer sample significantly weighted towards early-stage cancer (80% stage I/II), the authors achieved a mean AUC of 0.93 at 85% specificity. Sensitivity increased with tumor stage and tumor proportion in most cases.

Table 1: Summary of the Reviewed Paper

AUTHOR	METHOD ANDMODELUSED	DATASET	RESULT
Maclin, et al [1]	The model is based on Artificial Neural Network and back propagation algorithm is used to train the model	52 actual cases with numeric ultrasound data are used outofwhich30 benign cases, 17 cases of renal cell carcinoma, and 5 cases where other conditions were found	The network's quickest training session, which consisted of 51 facts, produced 2373 passes and 51 facts with a tolerance of 0.005 and examined 123,455 total facts in 1 hour, 14 minutes, and 58 seconds. The model is said togiveanoverallaccuracyof99.5 to99.9%
Akey, et al[2]	Support Vector Machine in combination with Feature Selection to improve the model's accuracy	Wisconsin Breast Cancer Dataset (WBCD) from UCI machine learning repository	Nine models were created for nine features and out of these, the highest classification accuracy is 98.53%, 99.02%,99.51% for 50-50%,70- 30%, and 80-20% respectively.
Bottari, et al [3]	Fully connected multilayer feel forward network	5 years follow-up dataset of 334 patients is used	Giving an overall accuracy greater than 80%. The Probability that the model would correctly predict death within a 3 - month period varied from 61% to 71%.
Vaka, et al[4]	Deep Neural Network with Support Value for improving cancer image quality and optimizing other performance parameters.	M.G Cancer Hospital & Research Institute, Visakhapatnam, India	The author stated that the method they have proposed based on Deep Neural Network and Support Value give an accuracy of 97.21% which is more than previously used method.
Hadidi, et al [6]	Linear Regression (LR) is implemented for the classification of mammography images, Backpropagation Neural Network (BPNN)	209 images from which 50 patients have breast cancer	The model with linear regressionand backpropagation neural network exceeds 93%.
Khan, et al[7]	CNN architecture-based framework is proposed and to extract	Two datasets are used for breast cancer prediction and classification image dataset of	The model gives an accuracy of 97.52% and the author has found the model to give

ISSN 2394 - 7780

	features from images, GoogLeNet, GGNet, and ResNet are employed.	breast and the other is developed locally	an excellent result with regard to accuracy which improves classification accuracy.
Jeyaraj, et al [9]	Convolution Neural Network is used on pre-processed cancer images	BioGPS data portal, TCIA Archive, GDC dataset	The model gives an accuracy of 91.4 and is stated by the author that by employing a large number of cancer subject datasets for training, accuracy was increased by 4.5%.
Lu, et al.[10]	A simple decision tree model was built using the Minimum Redundancy-Maximum Relevance (MRMR) feature selection method	The dataset includes 49 different features, such as demographics, blood tests as part of standard care, general chemistry, and tumour markers, present in 349 Chinese patients.	The outcomes were compared to the predictions generated by the logistic regression model and the algorithm for calculating the risk of ovarian cancer (ROMA). Furthermore, the model produces more accurate predictions than ROMA.
Alhazmi, et al [11]	Artificial Neural Network was employed in the model	Dataset was acquired between 2017 and 2020 from the oral pathology lab located at Prince Mohammed Bin Nasser Hospital	The accuracy of the model in predicting oral cancer was 78.95%.
Xiao, et al[12]	During the first stage techniques such as K-Nearest Neighbour, Support Vector Machines, Decision Tree, Random Forest and Gradient Boosting are used in the first stage and the result of the first stage is used on the Multi-model ensemble method.	Datasets used are Lung Adenocarcinoma (LUAD), Stomach Adenocarcinoma (STAD) and Breast Invasive Carcinoma (BRCA) and gene expression data is gathered	The model is tested with LUAD, STAD and BRCA dataset and it gives an accuracy of 99.20%, 98.78% and 98.41% respectively.
Wan, et al[13]	Cross-validation techniques for training and testing the model is employed. K-fold, binnedage, kbatch, balancedk-batch and ordered k-batch to get estimates of model performance.	The blood samples of cell free DNA (ctDNA) is used for training and testing purposes. 546 colorectal cancer and 241 non-cancerous are presented on the dataset.	In a colorectal cancer sample, the author achieved a mean AUC of 0.92 (95% CI 0.91-0.93), significantly weighted towards early-stage cancer (80%stageI/II), with a mean sensitivity of 85% (95%CI83-86%) at 85% specificity. Sensitivity typically increased with tumor stage and tumor proportion.

Researchers have published their method to detect cancer using machine learning technologies which have used different supervised learning. They analyzed and predicted cancers such as breast cancer, colorectal cancer, oral cancer, ovarian cancer, lung cancer, stomach cancer, and others using images of malignant cells that had been pre-processed using a variety of pre-processing techniques. The training dataset for the model was very limited. Further exploration of the data could produce more diverse and better results. The model's ability to develop is constrained by the fixed number of sources that provide the dataset for testing and training. It is preferable to use multiple datasets from various sources.

CONCLUSION

In this review, we've discussed a variety of machine learning (ML) approaches and how they can be used to predict cancer at an early stage. We have observed the various machine learning techniques being applied, the datasets collected the various types of malignancies, and the general effectiveness of these techniques in

ISSN 2394 - 7780

predicting the prognosis of cancer. The majority of the studies that have been suggested concentrated on creating prediction models utilising supervised ML techniques in an effort to predict accurate disease outcomes. Most models currently lack enough data when it comes to predicting cancer outcomes. But developing ever-more-accurate models as datasets grow and improve in quality will likely increase the widespread usage of machine learning techniques in many clinical and medical settings in the future.

REFERENCES

1. Maclin PS, DempseyJ, BrooksJ, RandJ. Using neural networks to diagnose cancer. Journal of medical systems. 1991Feb;15(1):11-9
2. AkayMF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7
3. Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, LeePW, Macintyre IM, Duthie GS, Monson JR. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. The Lancet. 1997 Aug 16;350(9076):469-72
4. Vaka AR, Soni B, Reddy S. Breast cancer detection by leveraging Machine Learning. ICT Express. 2020 Dec 1;6(4):320-4
5. Abdhelhalim Lalami, Christophe Combastel, in Fault Detection, Supervision and Safety of Technical Processes 2006, 2007
6. Alarabeyyat A, Alhanahnah M. Breast cancer detection using k-nearest neighbor machine learning algorithm. In 2016 9th International Conference on Developments in eSystems Engineering (DeSE) 2016 Aug1 (pp.35-39). IEEE
7. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJ. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters. 2019 Jul 1;125:1-6
8. F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, IEEETrans. Biomed.Eng.63 (7) (2016) 1455–1462
9. Jeyaraj PR, Samuel Nadar ER. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. Journal of cancer research and clinical oncology. 2019 Apr;145(4):829-37
10. Lu M, Fan Z, Xu B, Chen L, Zheng X, Li J, Znati T, Mi Q, Jiang J. Using machine learning to predict ovarian cancer. International Journal of Medical Informatics. 2020 Sep 1;141:104195.
11. Alhazmi A, Alhazmi Y, Makrami A, Masmali A, Salawi N,Masmali K, Patil S. Application of artificial intelligence and machine learning for prediction of oral cancer risk. Journal of Oral Pathology & Medicine. 2021 May;50(5):444-50.
12. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer methods and programs in biomedicine. 2018 Jan 1;153:1-9.
13. Wan N, Weinberg D, Liu TY, Niehaus K, Ariazi EA, Delubac D, Kannan A, White B, Bailey M, Bertin M, Boley N. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC cancer. 2019 Dec;19(1):1-0.

MANUSCRIPT SUBMISSION

GUIDELINES FOR CONTRIBUTORS

1. Manuscripts should be submitted preferably through email and the research article / paper should preferably not exceed 8 – 10 pages in all.
2. Book review must contain the name of the author and the book reviewed, the place of publication and publisher, date of publication, number of pages and price.
3. Manuscripts should be typed in 12 font-size, Times New Roman, single spaced with 1" margin on a standard A4 size paper. Manuscripts should be organized in the following order: title, name(s) of author(s) and his/her (their) complete affiliation(s) including zip code(s), Abstract (not exceeding 350 words), Introduction, Main body of paper, Conclusion and References.
4. The title of the paper should be in capital letters, bold, size 16" and centered at the top of the first page. The author(s) and affiliations(s) should be centered, bold, size 14" and single-spaced, beginning from the second line below the title.

First Author Name1, Second Author Name2, Third Author Name3

1Author Designation, Department, Organization, City, email id

2Author Designation, Department, Organization, City, email id

3Author Designation, Department, Organization, City, email id

5. The abstract should summarize the context, content and conclusions of the paper in less than 350 words in 12 points italic Times New Roman. The abstract should have about five key words in alphabetical order separated by comma of 12 points italic Times New Roman.
6. Figures and tables should be centered, separately numbered, self explained. Please note that table titles must be above the table and sources of data should be mentioned below the table. The authors should ensure that tables and figures are referred to from the main text.

EXAMPLES OF REFERENCES

All references must be arranged first alphabetically and then it may be further sorted chronologically also.

- **Single author journal article:**

Fox, S. (1984). Empowerment as a catalyst for change: an example for the food industry. *Supply Chain Management*, *2*(3), 29–33.

Bateson, C. D.,(2006), 'Doing Business after the Fall: The Virtue of Moral Hypocrisy', Journal of Business Ethics, 66: 321 – 335

- **Multiple author journal article:**

Khan, M. R., Islam, A. F. M. M., & Das, D. (1886). A Factor Analytic Study on the Validity of a Union Commitment Scale. *Journal of Applied Psychology*, *12*(1), 129-136.

Liu, W.B, Wongcha A, & Peng, K.C. (2012), "Adopting Super-Efficiency And Tobit Model On Analyzing the Efficiency of Teacher's Colleges In Thailand", International Journal on New Trends In Education and Their Implications, Vol.3.3, 108 – 114.

- **Text Book:**

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (2007). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies* (3rd ed.). New York: McGraw-Hill.

S. Neelamegham," Marketing in India, Cases and Reading, Vikas Publishing House Pvt. Ltd, III Edition, 2000.

- **Edited book having one editor:**

Raine, A. (Ed.). (2006). *Crime and schizophrenia: Causes and cures.* New York: Nova Science.

- **Edited book having more than one editor:**

Greenspan, E. L., & Rosenberg, M. (Eds.). (2009). *Martin's annual criminal code:Student edition 2010.* Aurora, ON: Canada Law Book.

- **Chapter in edited book having one editor:**

Bessley, M., & Wilson, P. (1984). Public policy and small firms in Britain. In Levicki, C. (Ed.), *Small Business Theory and Policy* (pp. 111–126). London: Croom Helm.

- **Chapter in edited book having more than one editor:**

Young, M. E., & Wasserman, E. A. (2005). Theories of learning. In K. Lamberts, & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 161-182). Thousand Oaks, CA: Sage.

- **Electronic sources should include the URL of the website at which they may be found, as shown:**

Sillick, T. J., & Schutte, N. S. (2006). Emotional intelligence and self-esteem mediate between perceived early parental love and adult happiness. *E-Journal of Applied Psychology, 2*(2), 38-48. Retrieved from http://ojs.lib.swin.edu.au/index.php/ejap

- **Unpublished dissertation/ paper:**

Uddin, K. (2000). A Study of Corporate Governance in a Developing Country: A Case of Bangladesh (Unpublished Dissertation). Lingnan University, Hong Kong.

- **Article in newspaper:**

Yunus, M. (2005, March 23). Micro Credit and Poverty Alleviation in Bangladesh. *The Bangladesh Observer*, p. 9.

- **Article in magazine:**

Holloway, M. (2005, August 6). When extinct isn't. *Scientific American, 293*, 22-23.

- **Website of any institution:**

Central Bank of India (2005). *Income Recognition Norms Definition of NPA.* Retrieved August 10, 2005, from http://www.centralbankofindia.co.in/ home/index1.htm, viewed on

7. The submission implies that the work has not been published earlier elsewhere and is not under consideration to be published anywhere else if selected for publication in the journal of Indian Academicians and Researchers Association.

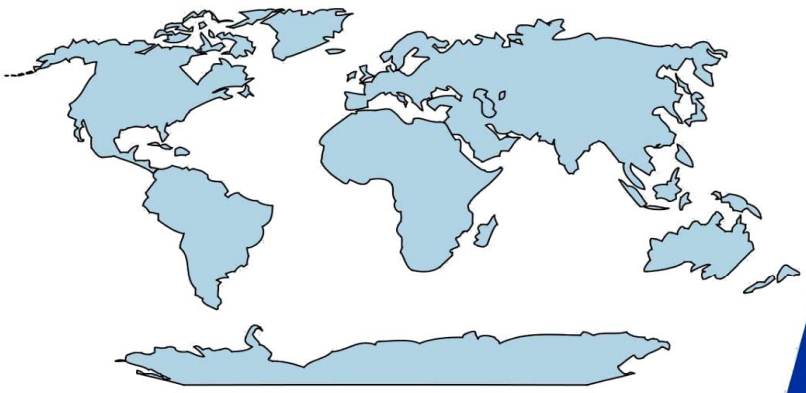
8. Decision of the Editorial Board regarding selection/rejection of the articles will be final.

www.iaraedu.com

Journal

ISSN 2322 - 0899

INTERNATIONAL JOURNAL OF RESEARCH IN MANAGEMENT & SOCIAL SCIENCE



Volume 8, Issue 2
April - June 2020

www.iaraedu.com

Journal

ISSN 2394 - 9554

International Journal of Research in Science and Technology

Volume 6, Issue 2: April - June 2019

Indian Academicians and Researchers Association
www.iaraedu.com

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXX

is admitted as a

Fellow Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

Date: 27.01.2020

Director

President

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXXXX

is admitted as a

Life Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

Date: 27.01.2020

Director

President

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXX

is admitted as a

Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Date: 27.01.2020

Director

President

IARA Organized its 1st International Dissertation & Doctoral Thesis Award in September'2019

1st International
Dissertation & Doctoral Thesis Award
(2019)

Organized By

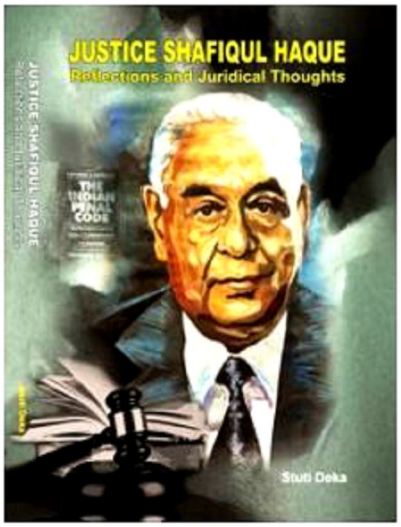
Indian Academicians and Researchers Association (IARA)

EMPYREAL PUBLISHING HOUSE

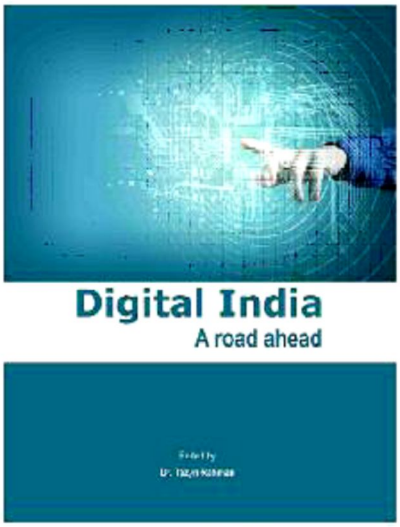
www.editedbook.in

Publish Your Book, Your Thesis into Book or Become an Editor of an Edited Book with ISBN

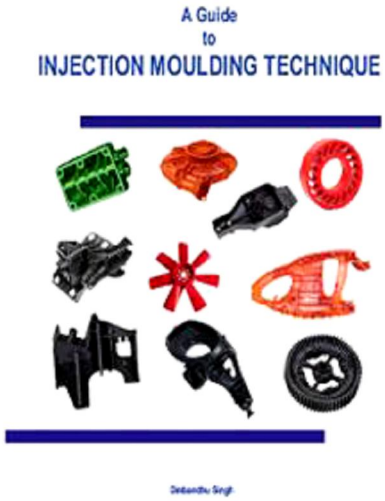
BOOKS PUBLISHED



Dr. Stuti Deka
ISBN : 978-81-930928-1-1

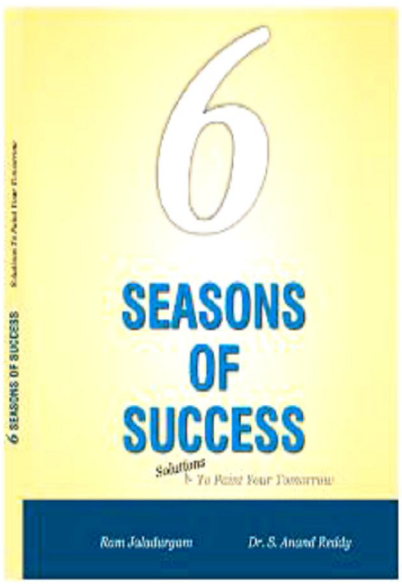


Dr. Tazyn Rahman
ISBN : 978-81-930928-0-4

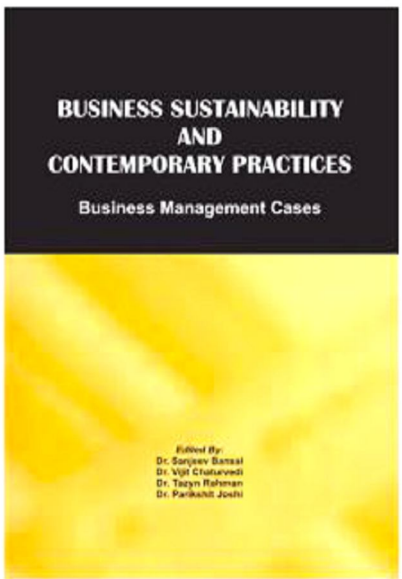


Mr. Dinbandhu Singh
ISBN : 978-81-930928-3-5

Dr. Ismail Thamarasseri
ISBN : 978-81-930928-2-8

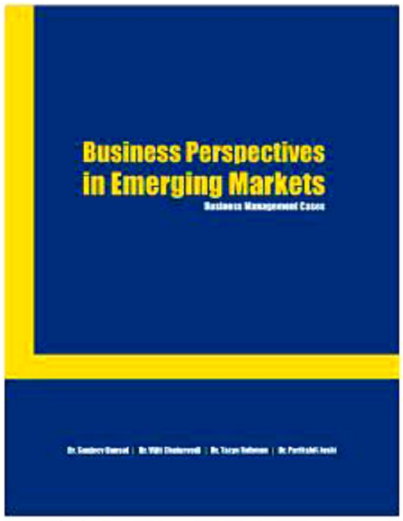


Ram Jaladurgam
Dr. S. Anand Reddy
ISBN : 978-81-930928-5-9



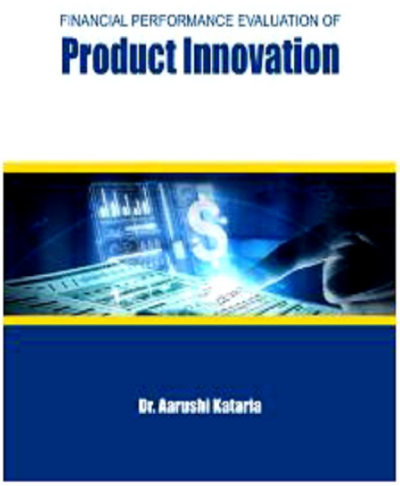
Dr. Sanjeev Bansal, Dr. Vijit Chaturvedi
Dr. Tazyn Rahman, Dr. Parikshit Joshi
ISBN : 978-81-930928-6-6

Ashish Kumar Sinha, Dr. Soubhik Chakraborty
Dr. Amritanjali
ISBN : 978-81-930928-8-0

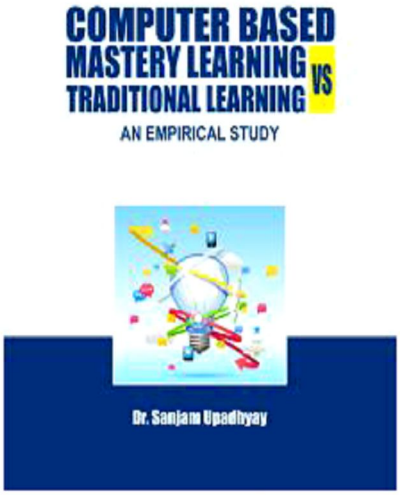


Dr Sanjeev Bansal, Dr. Vijit Chaturvedi
Dr. Tazyn Rahman, Dr. Parikshit Joshi
ISBN : 978-81-936264-0-5

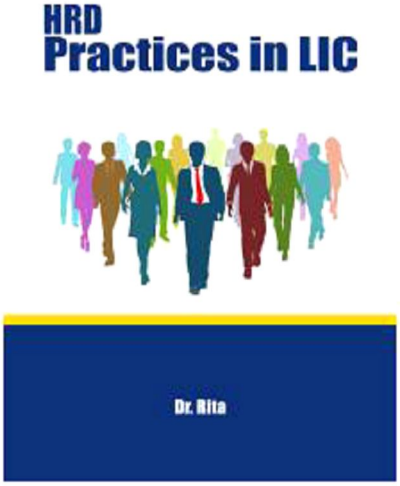
Dr. Jyotsna Golhar
Dr. Sujit Metre
ISBN : 978-81-936264-6-7



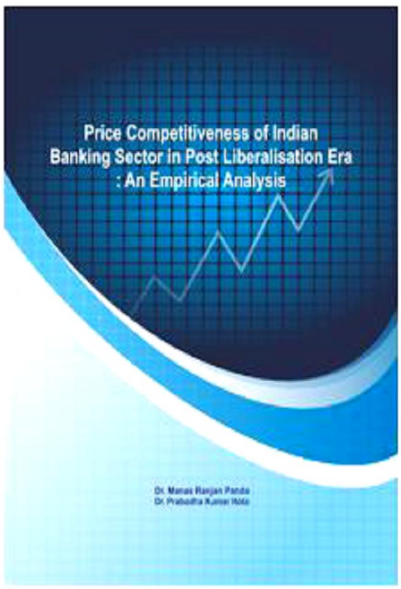
Dr. Aarushi Kataria
ISBN : 978-81-936264-3-6



Dr. Sanjam Upadhyay
ISBN : 978-81-936264-5-0



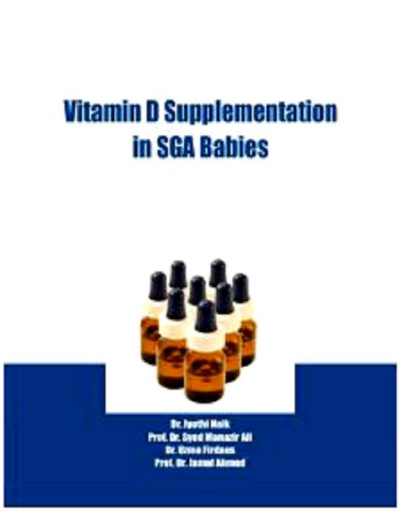
Dr. Rita
ISBN : 978-81-930928-7-3



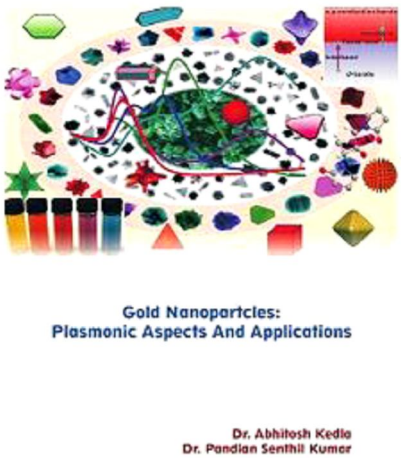
Dr. Manas Ranjan Panda, Dr. Prabodha Kr. Hota
ISBN : 978-81-930928-4-2

Poornima University
ISBN : 978-8193-6264-74

Institute of Public Enterprise
ISBN : 978-8193-6264-4-3



Dr. Jyothi Naik, Prof. Dr. Syed Manazir Ali
Dr. Uzma Firdaus, Prof. Dr. Jamal Ahmed
ISBN : 978-81-936264-9-8



Dr. Abhitosh Kedia
Dr. Pandian Senthil Kumar
ISBN : 978-81-939070-0-9

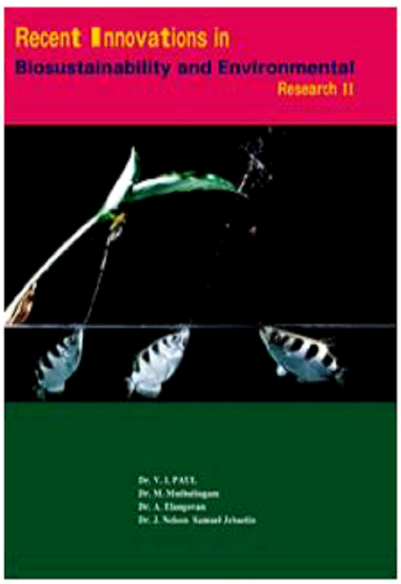
Dr. Vinod
S. Chandwani
ISBN : 978-81-939070-2-3

Prof. Dr. Dhananjay
Awasarikar
ISBN : 978-81-939070-1-6

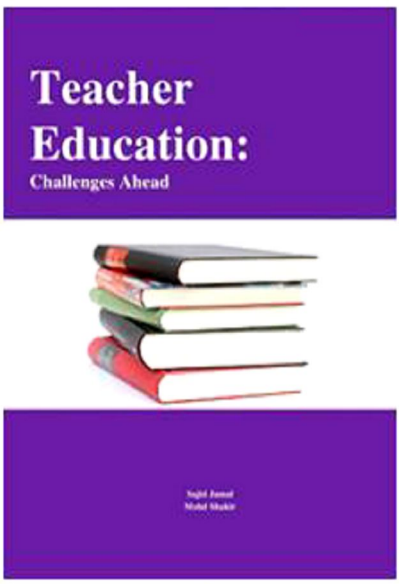
Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-4-7

Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-6-1

Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-7-8

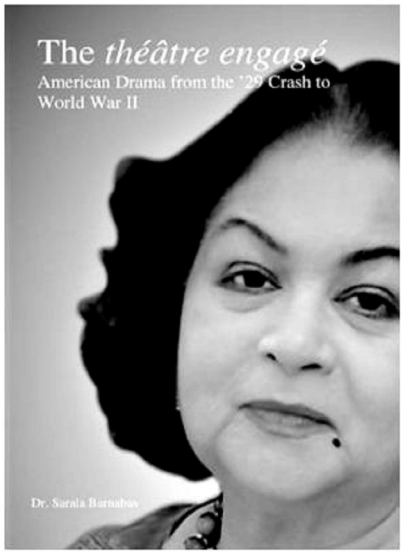


Dr. V. I. Paul, Dr. M. Muthulingam
Dr. A. Elangovan, Dr. J. Nelson Samuel Jebastin
ISBN : 978-81-939070-9-2



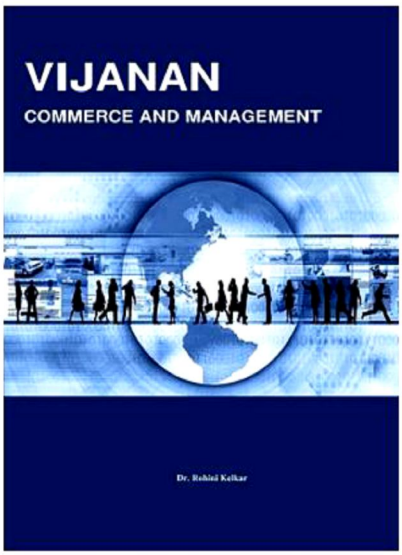
Sajid Jamal
Mohd Shakir
ISBN : 978-81-939070-8-5

Dr. R. Emmaniel
ISBN : 978-81-939070-3-0

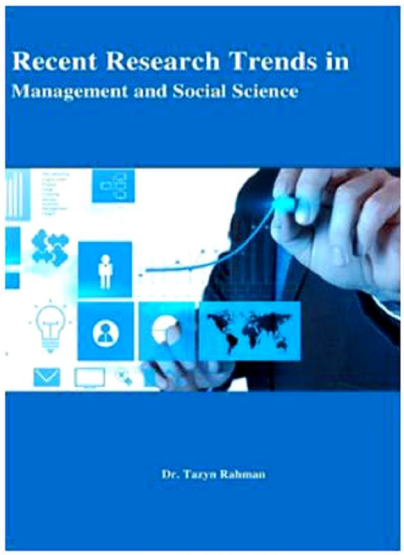


Dr. Sarala Barnabas
ISBN : 978-81-941253-3-4

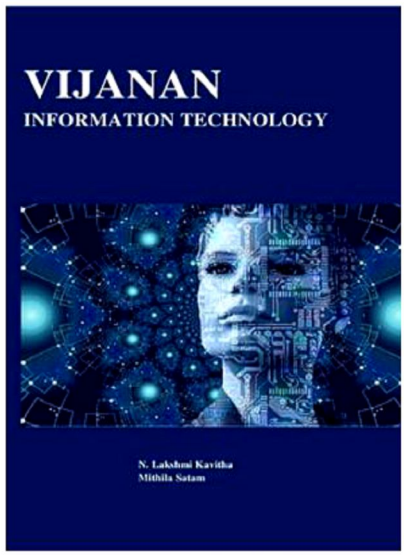
Dr. M. Banumathi
Dr. C. Samudhra Rajakumar
ISBN : 978-81-939070-5-4



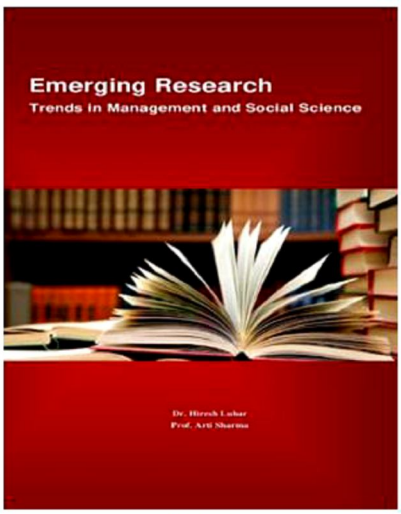
Dr. (Mrs.) Rohini Kelkar
ISBN : 978-81-941253-0-3



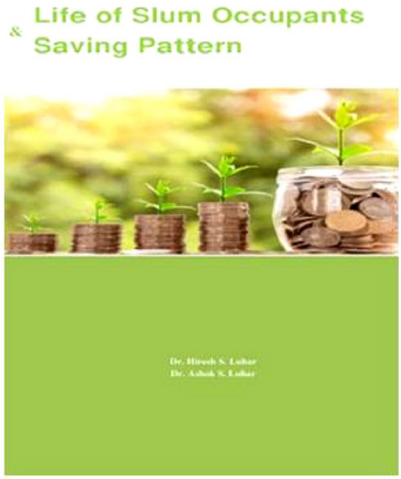
Dr. Tazyn Rahman
ISBN : 978-81-941253-2-7



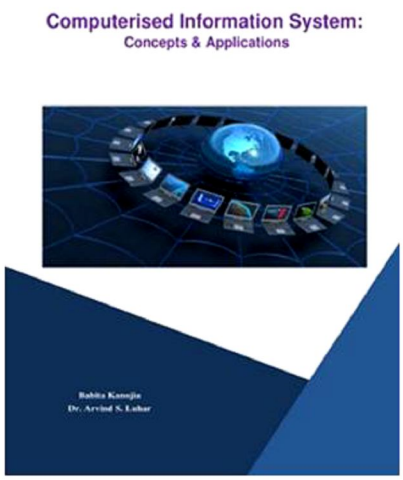
Dr. N. Lakshmi Kavitha
Mithila Satam
ISBN : 978-81-941253-1-0



Dr. Hiresh Luhar
Prof. Arti Sharma
ISBN : 978-81-941253-4-1



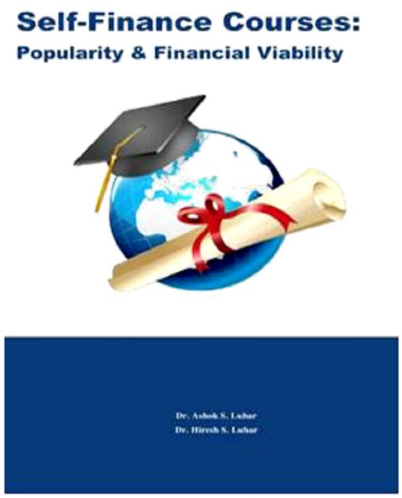
Dr. Hiresh S. Luhar
Dr. Ashok S. Luhar
ISBN : 978-81-941253-5-8



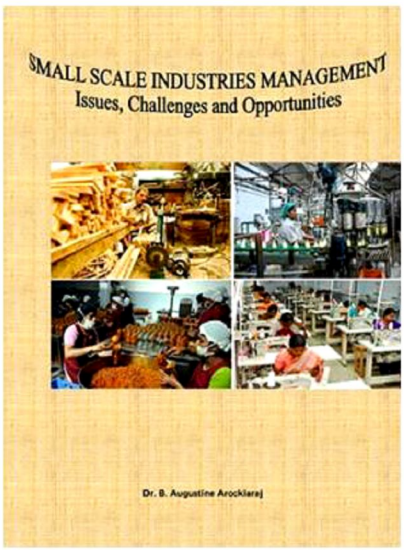
Dr. Babita Kanojia
Dr. Arvind S. Luhar
ISBN : 978-81-941253-7-2

Dr. Sw Rajamonaharane
SK Nathan
ISBN : 978-81-942475-0-0

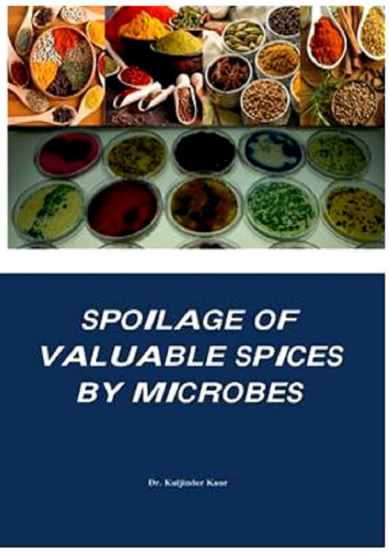
Aditi Sharma
ISBN : 978-81-941253-8-9



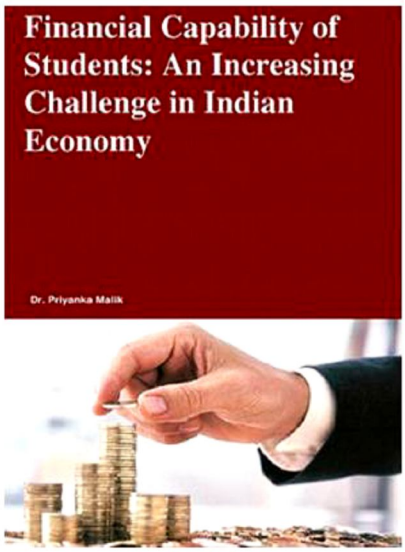
Dr. Ashok S. Luhar
Dr. Hiresh S. Luhar
ISBN : 978-81-941253-6-5



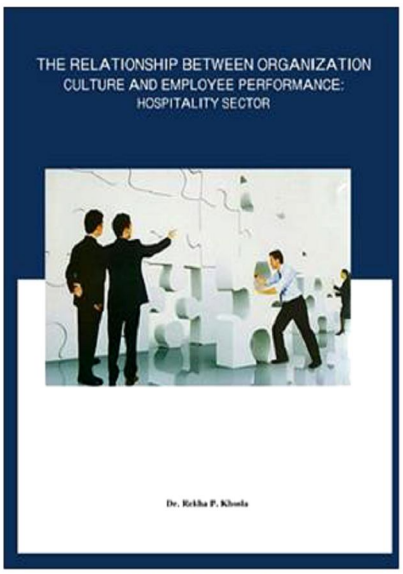
Dr. B. Augustine Arockiaraj
ISBN : 978-81-941253-9-6



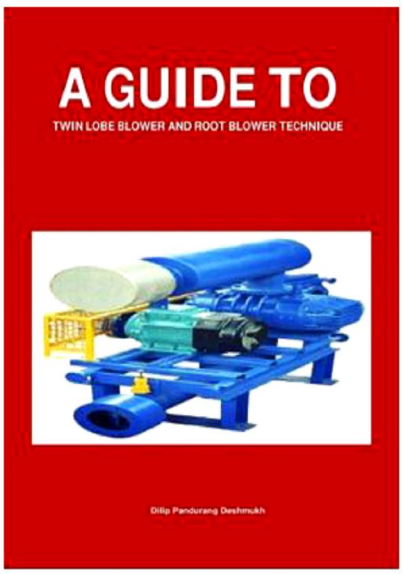
Dr. Kuljinder Kaur
ISBN : 978-81-942475-4-8



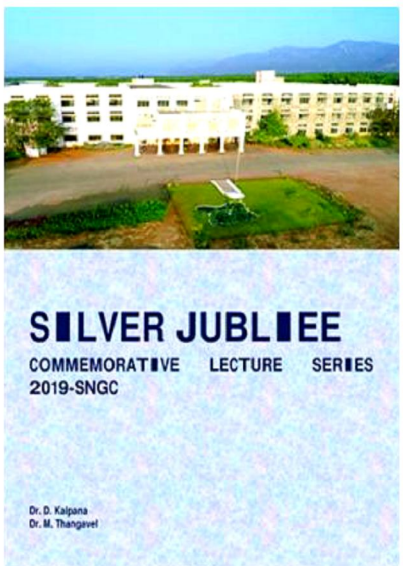
Dr. Priyanka Malik
ISBN : 978-81-942475-1-7



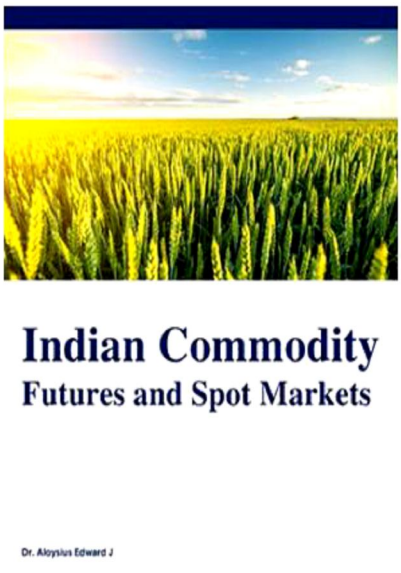
Dr. Rekha P. Khosla
ISBN : 978-81-942475-2-4



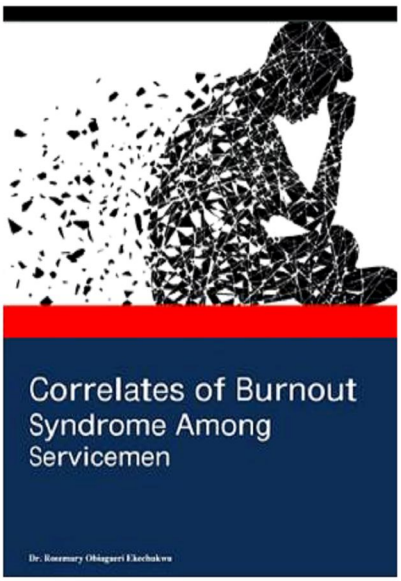
Dilip Pandurang Deshmukh
ISBN : 978-81-942475-3-1



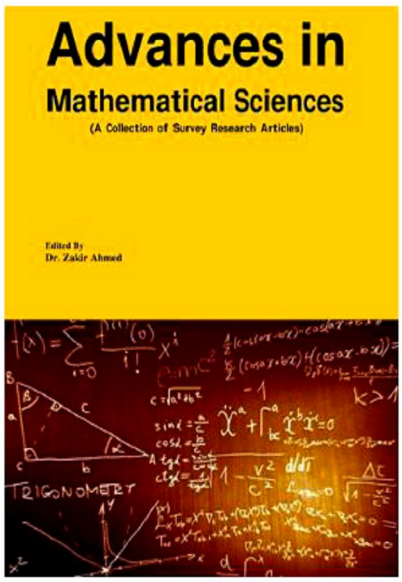
Dr. D. Kalpana, Dr. M. Thangavel
ISBN : 978-81-942475-5-5



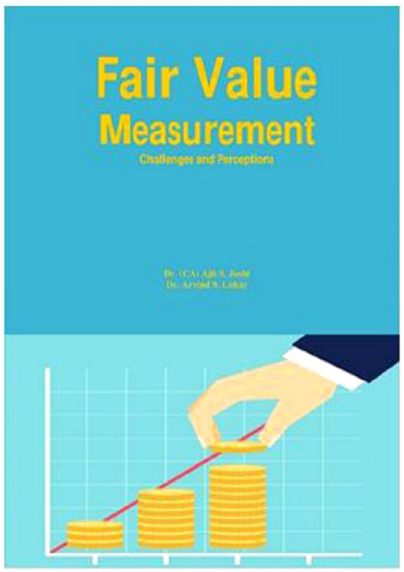
Dr. Aloysius Edward J.
ISBN : 978-81-942475-7-9



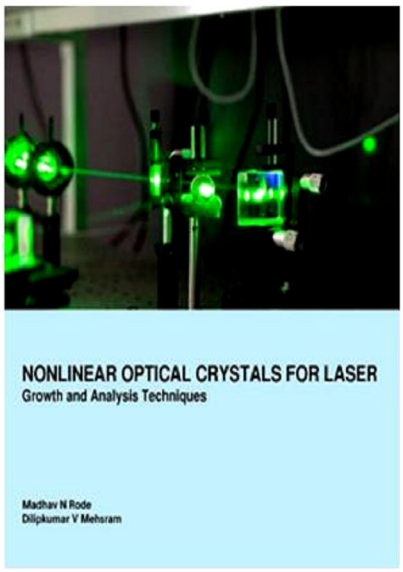
Dr. R. O. Ekechukwu
ISBN : 978-81-942475-8-6



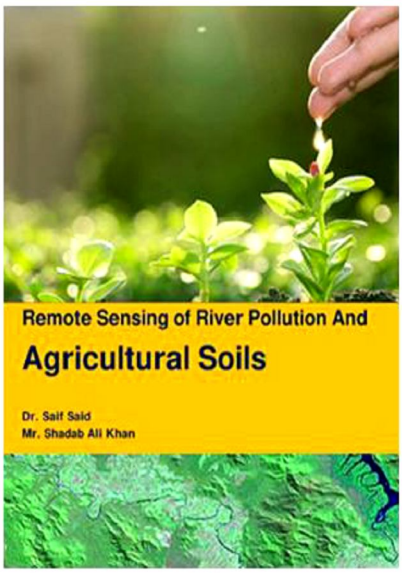
Dr. Zakir Ahmed
ISBN : 978-81-942475-9-3



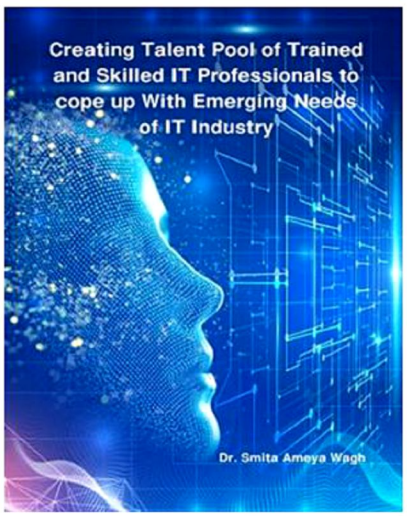
Dr. (CA) Ajit S. Joshi
Dr. Arvind S. Luhar
ISBN : 978-81-942475-6-2



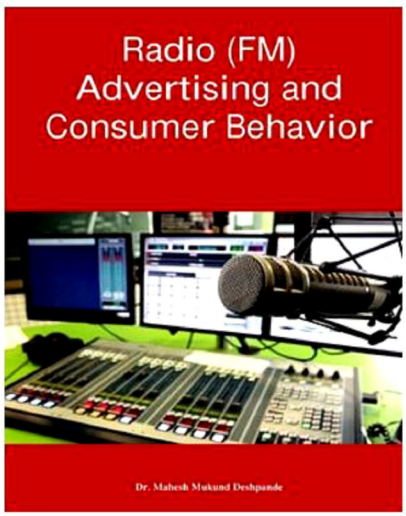
Madhav N Rode
Dilip Kumar V Mehsram
ISBN : 978-81-943209-6-8



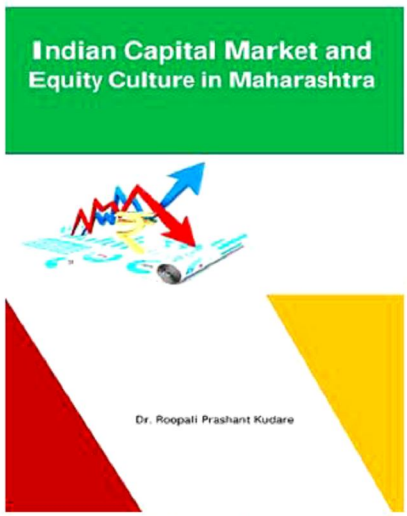
Dr. Saif Said
Shadab Ali Khan
ISBN : 978-81-943209-1-3



Dr. Smita Ameya Wagh
ISBN : 978-81-943209-9-9

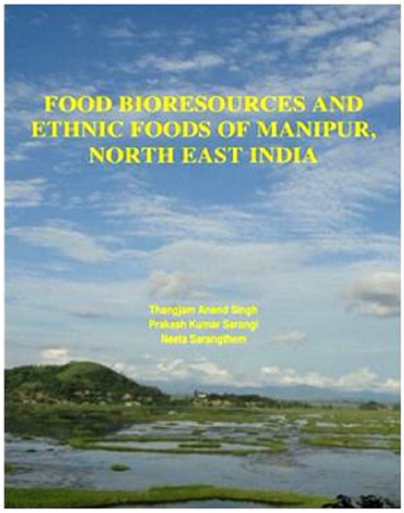


Dr. Mahesh Mukund Deshpande
ISBN : 978-81-943209-7-5

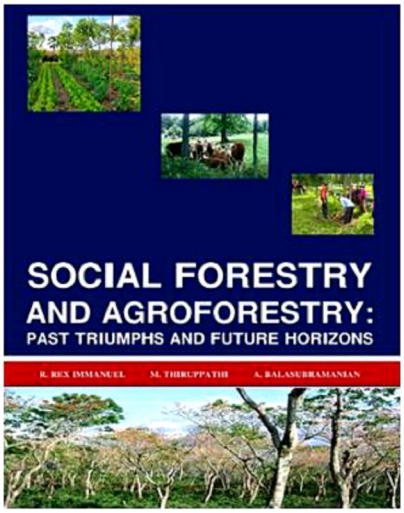


Dr. Roopali Prashant Kudare
ISBN : 978-81-943209-3-7

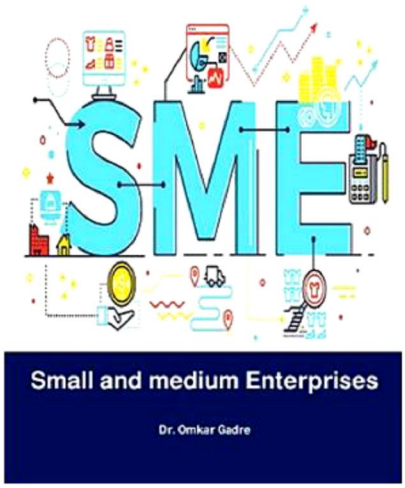
M. Thiruppathi
R. Rex Immanuel
K. Arivukkarasu
ISBN : 978-81-930928-9-7



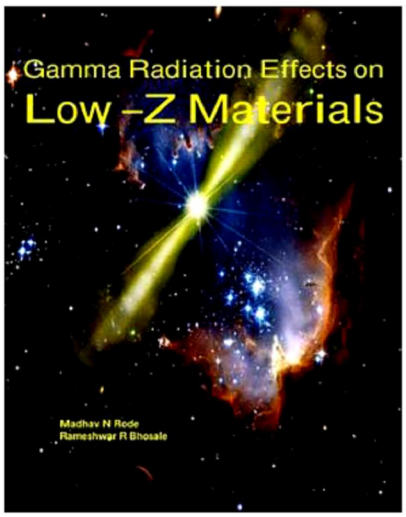
Dr. Th. Anand Singh
Dr. Prakash K. Sarangi
Dr. Neeta Sarangthem
ISBN : 978-81-944069-0-7



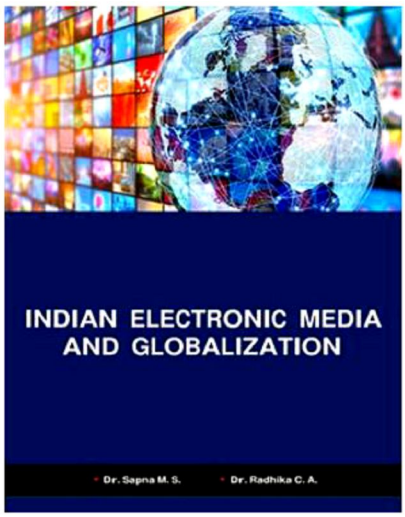
R. Rex Immanuel
M. Thiruppathi
A. Balasubramanian
ISBN : 978-81-943209-4-4



Dr. Omkar V. Gadre
ISBN : 978-81-943209-8-2

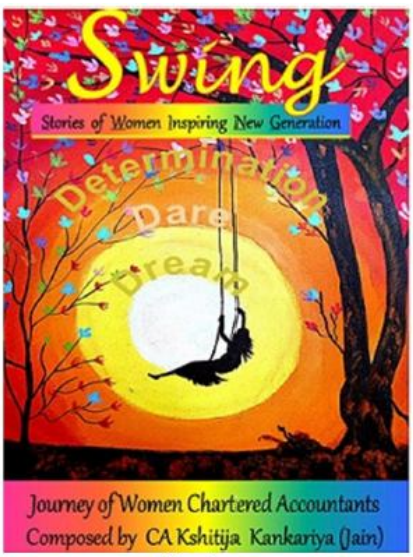


Madhav N Rode
Rameshwar R. Bhosale
ISBN : 978-81-943209-5-1

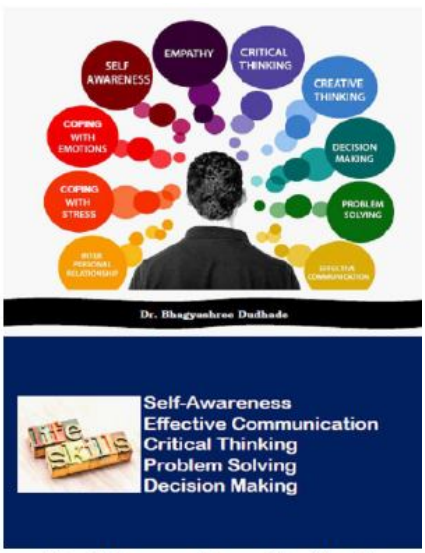


Dr. Sapna M S
Dr. Radhika C A
ISBN : 978-81-943209-0-6

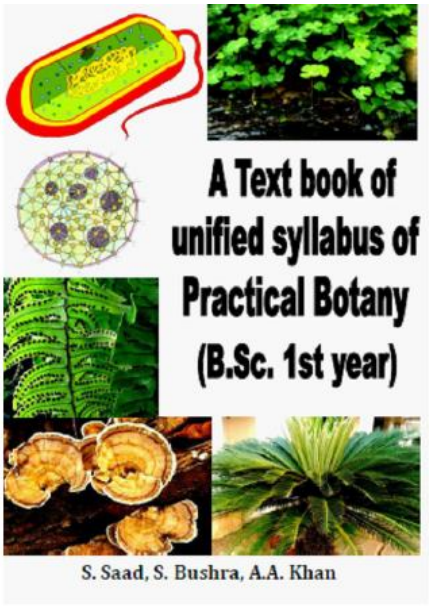
Hindusthan College
ISBN : 978-81-944813-8-6



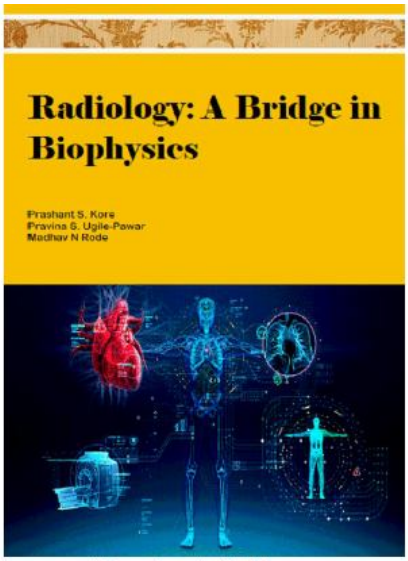
Swing
ISSN: 978-81-944813-9-3



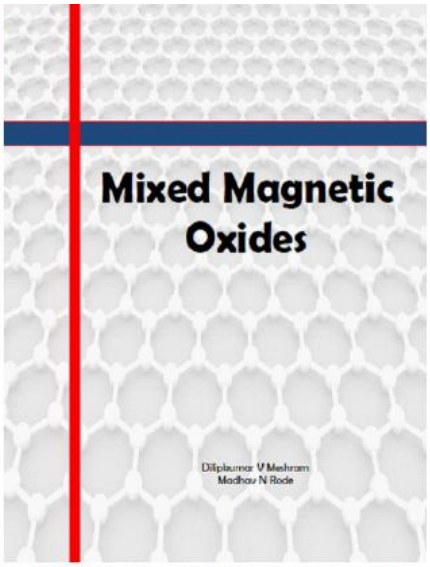
Dr. Bhagyashree Dudhade
ISBN : 978-81-944069-5-2



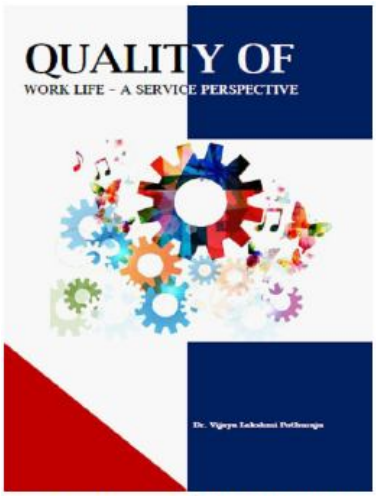
S. Saad, S. Bushra, A. A. Khan
ISBN: 978-81-944069-9-0



Prashant S. Kore
Pravina S. Ugile-Pawar
Madhav N Rode
ISSN: 978-81-944069-7-6



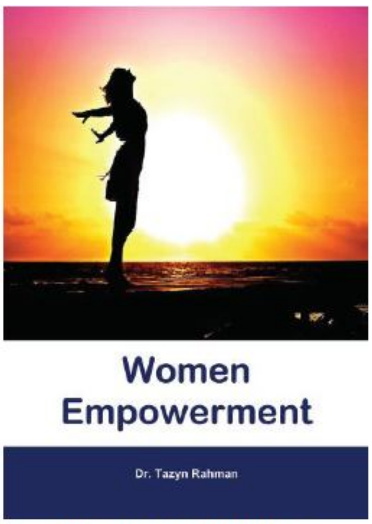
Dilipkumar V Meshram and
Madhav N Rode
ISSN: 978-81-944069-6-9



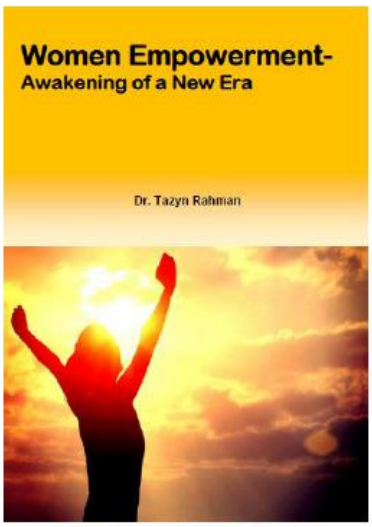
Dr. Vijaya Lakshmi Pothuraju
ISBN : 978-81-943209-2-0

Pratibha College
ISBN : 978-81-944813-2-4

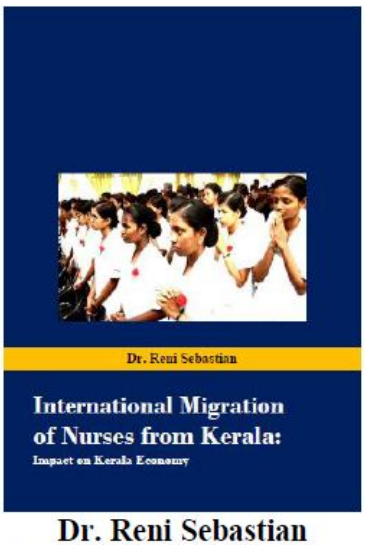
Pratibha College
ISBN : 978-81-944813-3-1



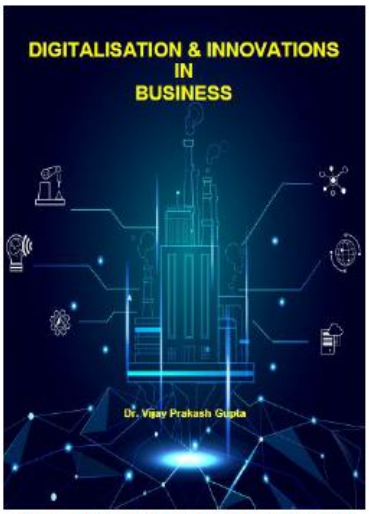
Dr. Tazyn Rahman
ISBN : 978-81-936264-1-2



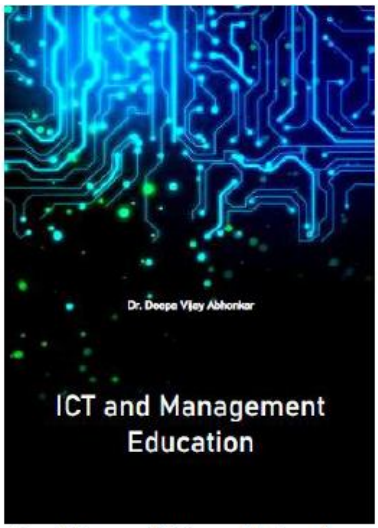
Dr. Tazyn Rahman
ISBN : 978-81-944813-5-5



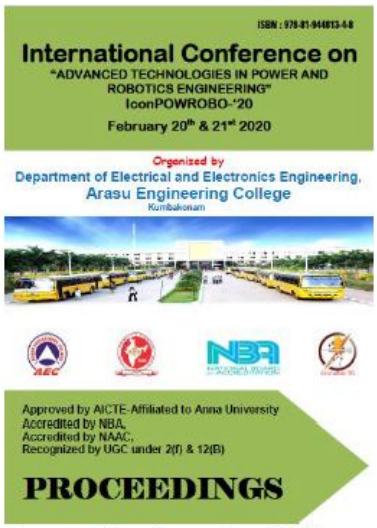
Dr. Reni Sebastian
ISBN : 978-81-944069-2-1



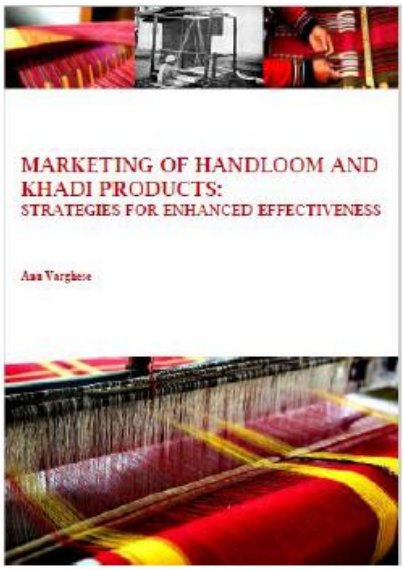
Dr. Vijay Prakash Gupta
ISBN : 978-81-944813-1-7



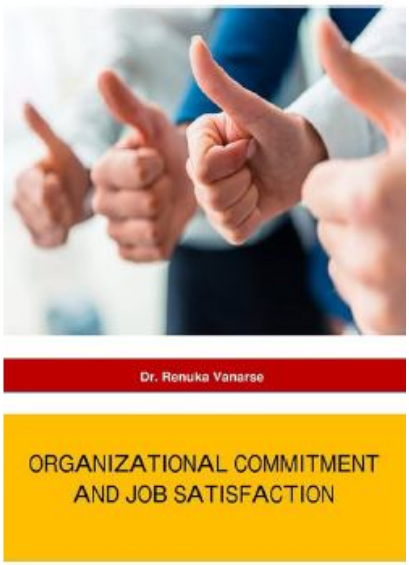
Dr. Deepa Vijay Abhonkar
ISBN : 978-81-944813-6-2



Arasu Engineering College
ISSN: 978-81-944813-4-8



Dr. Anu Varghese
ISBN : 978-81-944069-4-5



Dr. Renuka Vanarse
ISBN : 978-81-944069-1-4

Indian Academicians and Researchers Association
www.iaraedu.com