EMERGING TRENDS IN INFORMATION TECHNOLOGY

Dr.T.VELUMANI
Assistant Professor & Head,
Department Of Information Technology,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

Dr.R.LAKSHMI DEVI
Assistant Professor,
Department of BCA,
Women's Christian College,
Chennai, Tamil Nadu, India.

Dr.K.PRABAVATHY
Assistant Professor,
Department of Data Science,
Sree Saraswathi Thyagaraja College (Autonomous),
Pollachi, Tamil Nadu, India.

Dr.V.KALAIMANI
Professor & Head,
Department of MCA,
Rathinam Technical Campus,
Coimbatore, Tamil Nadu, India.

Published by
SK Research Group of Companies
The International Journals, Conferences and Books – SKRGC Publication
142, Periyar Nagar, Madakulam,
Madurai – 625 003, Tamil Nadu, India.
krgc.publisher@gmail.com / +91 9790120237
www.skrgcpublication.org/isbn

TITLE:	**EMERGING TRENDS IN INFORMATION TECHNOLOGY**
Edited by:	Dr.T.VELUMANI Dr.R.LAKSHMI DEVI Dr.K.PRABAVATHY Dr.V.KALAIMANI
Published by:	SK Research Group of Companies - SKRGC Publication Madurai - 625 003, Tamil Nadu, India.
Printers Details:	Leilani Katie Press, Madurai - 625 003, Tamil Nadu, India.
Edition Details:	I
ISBN:	978-81-965283-7-9
Month & Year	September, 2023
Copyright ©	SK Research Group of Companies
Pages	288
Price	Rs.500/-

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the **Book Chapter** Under the title ***“Emerging Trends in Information Technology”*** successfully.

I am extremely grateful to my beloved Father **Lt.A.Thiyagarajan** and My Mother **S.Thilakam** for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my wife **R.Seetha** and my daughter **V.Kiruthiga Sri** for their love, understanding, prayers and continuing support to complete this research work. Also I express my thanks to my sister **T.JothiAruna** and uncle R**.Srinivasan**, brothers, father in law and mother in laws for their support and valuable prayers. My Special thanks goes to my Spouse **Mrs.R.Seetha** for the keen interest shown to complete this Book Chapter successfully.

I would like to say thanks to **Dr. Madan A. Sendhil**, **Chairman**, Rathinam Group of Institutions, **Mrs.Shima Sendhil, Director**, Rathinam Group of Institutions, **Dr.R.Manickam, CEO and Secretary**, Rathinam Group of Institutions, **Dr.S.Balasubramanian, Principal**, Rathinam College of Arts and Science (Autonomous), Coimbatore, **Dr. S. N. Suresh, Vice Principal & Dean, School of Science**, **Dr.Sabareesh KPV, Dean of Research & Development** Rathinam College of Arts & Science, for their encouraging and guidance to Publish a **book Chapter.**

I thank the management of Rathinam College of Arts and Science (Autonomous), Coimbatore for their support to do this work. I thank the **School of computer Science** at Rathinam College of Arts and Science (Autonomous) for their genuine support to complete this book Chapter successfully.

The overall Aim of this book Chapter is to compile the latest research and development, up-to-date, challenges in the field of “***Emerging Trends in Information Technology***” with some modern approaches. Last but not least we thank all those who are directly and indirectly supported me to complete the book Chapter Successfully.

Dr.T.VELUMANI
Editor
M.Sc(CS).B.Ed,M.Phil,MBA,M.Sc(Psychology),Ph.D

PREFACE

“Emerging Trends in Information Technology” is a popular subject evolving continuously. This book Chapter covers topics like all computer science streams subjects .The outcome of our teaching experience and research findings at the Department of Computer Science and provides a basic foundation for various topics in Information Technology. In general, the field of **“*Emerging Trends in Information Technology”*** requires sound theoretical knowledge and extensive experimental work involving software simulation and testing with large sets of sample projects. The book Chapter has been designed to meet.

The book Chapter presents a unique overview of recent as well as what are the ***“Emerging Trends in Information Technology”***, it focuses the Integration of all computer science techniques and Applications also. Recent results on invariant Concepts Integration are clearly presented and well explained. Identifying Feature Points is demonstrated. We reviewed Understanding Uncertain Spatial Relationships

The book chapter ***“Emerging Trends in Information Technology”*** from teaching the graduate courses on moment invariants and related fields of all domains and presenting several tutorials on moments at major international conferences.

The target readership of the book Chapter are academic researchers and R&D engineers from all application areas who need to recognize objects extracted Industry 4.0 as well as specialists in moment-based interested in a new development on this field. Last but not least, the book chapter is also intended for university lecturers and graduate students of all computer science fields are analyzed. This knowledge then motivates the presentation of the techniques needed to reverse distortions, minimize artifacts and enhance important features.

Emerging Trends in Information Technology

Editors

Dr.T.VELUMANI

Dr.R.LAKSHMI DEVI

Dr.K.PRABAVATHY

Dr.V.KALAIMANI

CONTENT

ID: 01

AN INTERACTIVE GENETIC ALGORITHM FOR FRACTAL DESIGN SYSTEM

Dr.R.MANICKAM
CEO & Secretary,
Rathinam Group of Institutions,
Coimbatore, Tamil Nadu, India.

ABSTRACT

In order to explore how to use genetic algorithms to generate art patterns with fractal features and improve the aesthetic and personalized level of patterns through user interaction, this article introduced a fractal art pattern design system based on interactive genetic algorithms. This system allowed users to use a simple interactive interface and design various complex fractal art patterns. Genetic algorithms were used to optimize these parameters and improve and modify the generated patterns through user interaction. By introducing interactive genetic algorithms into the system, automated optimization of pattern design could be achieved, thus greatly improving design efficiency and quality. The data showed that when the probability of finding the optimal solution by interactive genetic algorithm was 100%, the running time could be controlled within 9 seconds. Interactive genetic algorithm could generate fractal art patterns with uniqueness and aesthetics, while also meeting the personalized needs of users. The algorithm had high efficiency and controllability.

KEYWORDS: *Interactive Genetic Algorithm, Fractal Art, Pattern Design, Automation Optimization.*

I. INTRODUCTION

Fractal art is a form of art that is full of creativity and wonders. It can construct complex patterns by constantly repeating and changing simple elements of form, reflecting infinite beauty and change. However, for non professionals, designing complex fractal art patterns is a very difficult task. Traditional pattern design methods require users to manually set various parameters and rules, greatly reducing design efficiency and quality. Interactive genetic algorithm is an algorithm that uses evolutionary computation to find the optimal solution. It is widely used in fractal art pattern design system. For the pattern design of fractal art, it often requires a lot of time and effort, while interactive genetic algorithms can help designers quickly find patterns that meet the requirements.

Genetic algorithms are based on the theory of artificial intelligence and have been widely used in various fields. On this basis, Liu J used the genetic algorithm of operator and cluster optimization to divide a group of people into four groups and assign corresponding suitability to reflect the optimal preservation strategy. Through a large number of simulation experiments, it was proved that the proposed method could obtain Pareto optimization results with strong universality and rapid evolution. Under normal circumstances, as long as some repeated attempts are made, it could be well optimized, resulting in higher artistry and more creative illustrations [1]. Mulyanto Prameswari N S reviewed the combination of pattern design and pattern layout. The key was to accurately invert and reflect the subject object until it was in harmony with the boundaries of the subject object, and put the subject object into a group. The harmonious thematic connection could enable the theme to transcend the scale of pattern design, thereby providing a new theme for spatial coordination [2]. On the basis of existing research, Li D proposed a genetic programming method based on pre-learning strategies, thus combining intracellular generation with intercellular transmission, and established a new synergistic-evolutionary model [3]. Lee J. Y proposed a better pattern, which could be achieved by using a narrow strip production process and improving the fluidity of the fibers [4]. Zhang J proposed the use of modern computer database technology to digitize patterns, so that patterns could be developed from traditional patterns to modern patterns.

From the point of view of technical implementation methods and interface design, database templates were a perfect combination of art and technology [5]. Kazlacheva Z provided a set of foldable curtains that were both decorative and constructive. It was based on the foldable foldable type and was foldable and cut according to the foldable foldable type. This system promoted the process of fashion and craft design, thus making it easy to build a curtain for clothing curtains of different sizes, and provided the possibility for new ideas and multiple designs [6]. Nisztuk Maciej applyed the algorithm based on “greed” to the calculation of building floor plans, and summarized its implementation methods. He compared the experimental results of the hybrid evolution algorithm and the greedy algorithm, and pointed out its possible popularization and future research directions [7]. These studies have certain reference value for the classification of literary and artistic groups, but they are more theoretical discussions.

In order to explore the efficient, controllable, and combination of fractal pattern design with emerging technologies such as deep learning, this project intended to build a fractal pattern design system based on interactive genetic algorithms. This system provides users with a simple interactive interface, enabling users to produce a variety of complex graphics in a visual manner. On this basis, a graphics optimization method based on genetic algorithm is proposed and applied to graphic design, so that the efficiency and quality of graphic design have been greatly improved. This paper experimented with the possibility of interactive genetic algorithms for fractal graphic design, and provided a new idea and method for the research and application of relevant aspects. The main contents of this paper are as follows:

The first part, the background and significance of fractal art pattern design system.

The second part discusses the fractal art pattern design.

The third part, the design of fractal graphics system.

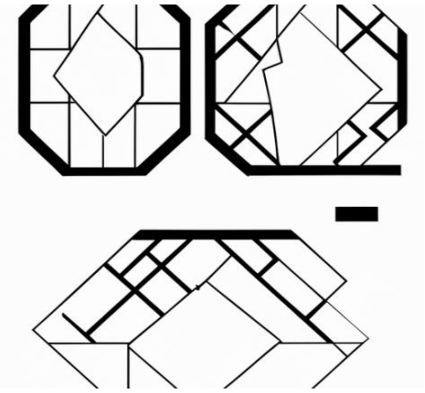
The fourth part is the future prospect of fractal art pattern design system and related suggestions.

II. FRACTAL ART PATTERN DESIGN

A. Fractal Art Design

With the development of digital technology and computer graphics, fractal art has gradually entered the public’s vision. Fractal art is an art form based on self-similarity and infinite repetition. It generates complex images through iterative processes and various algorithms, making people feel the beauty of mathematics and nature. As shown in Figure 1, the design process of fractal art is similar to the process of manufacturing mirrors, but it is only a virtual process. Designers operate, process, deform, rotate, and twist data on mathematical models, rather than operating in the physical world.

Fractals are widely used in digital entertainment, printing, graphic design, decoration, architecture and other fields. These fine art graphics can help the design of clothing printing and dyeing patterns, which are suitable for clothing fabrics [8]. In the field of digital entertainment, the use of fractal technology can create a virtual world full of fantasy colors, giving players a different gaming experience. Due to its long history, rich connotations and beautiful forms, traditional patterns have been widely used in modern design [9]. In the fields of printing and graphic design, fractal art can create various unique patterns and forms, making products more artistic and creative. The process of designing fractal art requires the use of computers and various software tools. Currently, there are many popular fractal design software on the market, such as Apophysis, Mandelbulb 3D, Ultra Fractal, and so on. These software can provide a variety of tools and special effects to help designers quickly generate fractal images, while also meeting their creative pursuit of fractal art.



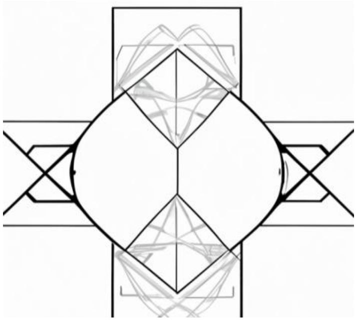


Fig.1 Fractal art Design Pattern

Overall, fractal art is a creative and artistic form of design that allows people to see the beauty of mathematics and nature, expanding their horizons and imagination. In the future development, fractal art would be applied in more fields and become an important art form in the digital age. Fractal art is a unique form of art characterized by complex patterns composed of similar shapes. These shapes can be infinitely repeated, and each repeated shape is similar to the original shape. Fractal art can be either static images or dynamic animations. In multi-objective optimization, multiple goals are considered, and at least one goal is affected by uncertainties in the interval. However, there are few effective solutions [10]. In the design of fractal patterns, genetic algorithms are usually used. The genetic algorithm is applied to the design of fractal graphics.

B. Informationization of Fractal Art Pattern Design

1) Role of Genetic Algorithms in Fractal Art Pattern Design

Genetic algorithm is an algorithm that simulates the laws of biological evolution in nature and solves the optimal solution. Evolutionary algorithms that simulate the natural behavior of organisms have been successfully applied in many fields and are closely related to innovation issues [11]. Since the emergence and application of evolutionary optimization methods in architecture in the early 20th century, a lot of research has been conducted to try to integrate evolutionary strategies with design processes [12]. Genetic algorithm is a method of using intelligent algorithms to assist humans in solving problems, and it is an important branch of artificial intelligence. Artificial intelligence transfers people's work to the current imperfect technical system [13]. In fractal art, genetic algorithms can be applied to the design and optimization of fractal patterns, making the designed artistic patterns more beautiful, gorgeous, rich, and creative. The main roles of genetic algorithms in fractal art pattern design are as follows:

1) Improving efficiency: Traditional fractal art pattern design usually requires a lot of time and effort, while using genetic algorithms can greatly improve efficiency and quickly search for the optimal pattern design.

2) Obtaining more diverse patterns: Using genetic algorithms, richer patterns can be generated, thus avoiding the problems of single patterns, duplication, and cumbersome patterns caused by human subjective reasons. In using interactive genetic algorithms to solve big data information retrieval problems, individual users need to perform more human-computer interactive operations in order to obtain optimal personal preference information [14].

3) Enhancing creativity: Using this method, a variety of novel graphics can be generated, which stimulates artists to make bold innovations in graphics and make them transcend the boundaries of traditional art. The development and innovation of emerging technologies has become the most important reference in the field of teaching. It can help learners establish effective self-study methods and cultivate the ability to continuously update knowledge [15].

In short, the role of genetic algorithms in fractal art pattern design cannot be ignored. It can improve the efficiency, diversity, and creativity of fractal art pattern design, bringing more possibilities to art. In the future, with the continuous development of computer technology, the application of genetic algorithms in fractal art pattern design would be more extensive, presenting us with more exquisite, magnificent, and unique fractal art patterns.

2) Interactive Genetic Algorithm

In this paper, a fractal pattern design method based on genetic algorithm was adopted. A new way to optimize the process of genetic algorithms: Using this method, the process is inherently safer even during the operation phase [16]. First of all, a suitability function must be established to evaluate the advantages and disadvantages of each model. The definition of fitness function depends on the purpose of the design, such as aesthetics, geometry, color, etc. Each pattern design is encoded with a genotype, and a set of initial genotypes such as 100 are randomly generated. Based on the suitability function, each type is evaluated and its suitability is calculated. With a higher degree of suitability, it is easier to choose a better genotype. The excellent genotypes are hybridized, mutated, and selected to obtain new genotypes, and the adaptability of each genotype is evaluated. Such a cycle would continue until the pre-set termination conditions are met, such as the number of repetitions and a certain degree of suitability.

Due to the existence of model bias, the evolutionary process has become slow, making people not satisfied with the final evaluation results, which in turn causes the evaluation cycle to be too long, and fatigue and other problems [17]. In interactive genetic algorithms, users interact with data. Due to exchange, every user would be affected by the feelings of another user [18]. Interactive genetic algorithm is an optimization algorithm that combines genetic algorithm and human-computer interaction technology. It can allow users to participate in the optimization process through an interactive interface, and intuitively observe and operate the results, thereby achieving precise control over the design process. In the fractal art pattern design system, interactive genetic algorithms can achieve automated optimization through the following steps:

(1) Initializing population: Based on the user's design needs and inputs, a certain amount of fractal pattern population is initialized.

(2) Calculation of suitability: The suitability of each fractal structure is calculated through the suitability function, so as to classify the group. For ease of explanation, it is assumed that a parameterized scale function can be used to express the preferences of decision makers, but these parameterized scale functions were not known at first. Instead, during the search, parameter errors can be gradually reduced by asking decision makers for preference queries to help determine the best solution in general [19]. The adaptability function can be expressed as Formula (1):

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{n} f(x_j)} \qquad (1)$$

(3) Selection operation: according to the size of population fitness value, selection operation is carried out, and individuals with higher fitness value are selected for crossover and mutation operation.

(4) Cross operation: The selected individuals are subjected to cross operation, and two individuals are randomly selected based on the probability of crossing to generate new individuals. In fractal patterns, two adjacent patterns can be crossed using real number encoding, as shown in Formulas (2) and (3):

$$G_1 = \gamma * G + (1 - \gamma) * G_2 \qquad (2)$$

$$G_2 = (1 - \gamma) * G_1 + \gamma * G_2 \qquad (3)$$

In Formulas (2) and (3), $G_1 G_2$ represent two adjacent patterns. After performing the crossover operation, a new pattern can be generated.

(5) Mutation operation: Newly generated individuals are subjected to mutation operation. According to the probability of mutation, individuals are mutated to generate new individuals.

(6) Iterative operation: Steps (2) to (5) are repeated until the fitness value of the population meets the design requirements or reaches the maximum number of iterations.

On this basis, a fractal pattern design method based on interactive genetic algorithm is proposed. The design is generated from a set of prescribed design frameworks based on the code of a specific problem, and can also be corrected by reorganization, crossover, variation and other operations [20]. On this basis, using the method of human-computer dialogue, real-time observation and control of the optimization process of the entire system are realized, making the entire design process more intuitive and optimal. Figure 2 shows the process of an interactive genetic algorithm used to generate fractal patterns.

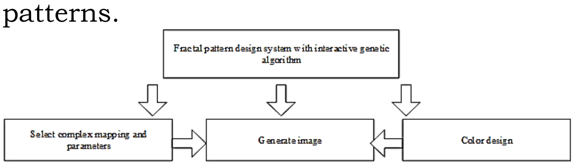

Fractal pattern design system with interactive genetic algorithm
Select complex mapping and parameters
Generate image
Color design

Fig.2 Interactive genetic algorithm for generating fractal art patterns framework

III. DESIGN OF FRACTAL GRAPHICS SYSTEM

The main interface of the interactive genetic algorithm for generating fractal art patterns system is shown in Figure 3. These images can be selected, crossed, and mutated, ultimately generating fractal art patterns according to customer needs.

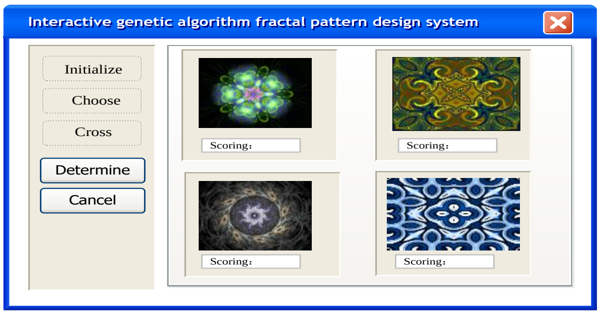


Fig.3 Interface of interactive genetic algorithm in fractal art pattern design system

Figure 3 shows an interactive genetic algorithm, which is used as the interface of a fractal pattern design system. Experiments have proved that this algorithm can produce rich, diverse, beautiful and easy-to-identify fractal graphics. In the user experiment, participants can use this method to create their own works. They think this method is very intuitive and can interact.

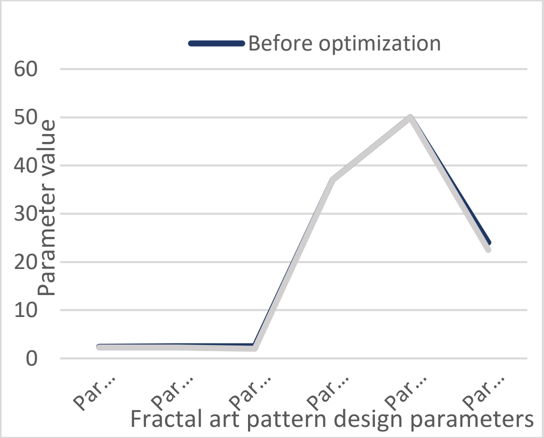


Fig.4 Parameter comparison after interactive genetic algorithm optimization

Figure 4 shows the optimized results using interactive genetic algorithm after inputting various parameters for designing fractal art patterns. Through simulation and calculation, the optimized parameter values are more accurate and reasonable compared to the original design results. The data shows that the optimized parameter values are less than or equal to the pre optimized parameters, and the results are more accurate, indicating that interactive genetic algorithms can optimize parameters for fractal art patterns. It also provides relevant reference and reference for future complex system optimization problems, and has certain reference value.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS WITH 1000 ITERATIONS

Algorithm	Probability of finding the optimal solution	Best value	Worst value	Running times
Traditional genetic	99%	0.94	0.89	11
Interactive genetic	100%	0.96	0.91	9
Random selection method	92%	0.86	0.79	7

From Table 1, it could be seen that three methods, including traditional genetic algorithm, interactive genetic algorithm, and random selection method, were applicable when the iteration amount is 1000. Through the analysis of the experimental results, the results showed that the interactive genetic algorithm had a 100% optimization rate, while the traditional genetic algorithm had a 99% optimization rate. However, the random selection method had the smallest search speed. This is mainly due to the fact that the random selection method did not require large-scale parallel searches, thereby reducing the complexity of the search. In the case where the probability of searching for the optimal solution was 100%, the running time of this method was only 9 seconds, indicating that the method had high computational efficiency.

IV. CONCLUSIONS

This article introduced a fractal art pattern design system based on interactive genetic algorithm. By introducing genetic algorithms into the system, automated optimization of pattern design could be achieved, greatly improving design efficiency and quality. This system combined computer science and art, providing a new approach and method for the design of fractal art patterns. This method allowed users to select and adjust a set of parameters to control the shape and color of the pattern, and then used interactive genetic algorithms to search for suitable parameter combinations to generate the final pattern. This article also explored some implementation details, such as how to choose an appropriate parameter range and how to evaluate the quality of parameter combinations. This method could provide a creative tool for digital artists and fans to help them create more and more creative works of art. However, there were still some shortcomings in the system, such as the possibility of local optima in the algorithm for complex fractal structures, which required further optimization and improvement. Overall, the system had high practicality and application prospects, and could meet a wider range of fractal art pattern design needs.

V. ACKNOWLEDGMENT

The corresponding author is Ling Jiang

VI. REFERENCES

1. Liu, J, Q. Chen, and X. Tian. "Illustration Design Model with Clustering Optimization Genetic Algorithm." Complexity 2021.4(2021):1-10.
2. Mulyanto, Prameswari N S, Afatara N. "PATTERN DESIGN AND MOTIF PLACEMENT OF BATIK SANGGIT IN SHORT-SLEEVE SHIRT STYLE." Vlakna a Textil 26.2(2019):37-42.
3. Li, D, Zhan R, Du S. "Automatic Design of Intercell Scheduling Heuristics." IEEE transactions on automation science and engineering 16.4(2019):1907-1921.
4. Lee, J. Y, Kim C H, Kwon S . "Study of Refining Behaviors of Hardwood Bleached Kraft Pulp Using Domestic Casting Plates with Different Pattern Designs." Palpu Chongi Gisul/Journal of Korea Technical Association of the Pulp and Paper Industry 50.3(2018):36-43.
5. Zhang, J. "Application of image database classification in decorative pattern design." Paper Asia 2.1(2019):196-199.
6. Kazlacheva, Z. "Pattern design of twisted draperies with decorative and constructive function." Applied Researches in Technics Technologies and Education 7.1(2019):1-9.
7. Nisztuk, Maciej, and Paweł B. Myszkowski. "Hybrid evolutionary algorithm applied to automated floor plan generation." International Journal of Architectural Computing 17.3 (2019): 260-283.
8. Chen, Daoling, et al. "Kansei engineering as a tool for the design of traditional pattern." Autex Research Journal 21.1 (2021): 125-134.
9. Gong, D. W, S. Jing, and Z. Miao . "A Set-Based Genetic Algorithm for Interval Many-Objective Optimization Problems." IEEE Transactions on Evolutionary Computation 22.99(2018):47-60.
10. De Prisco, Roberto, Gianluca Zaccagnino, and Rocco Zaccagnino. "Evocomposer: An evolutionary algorithm for 4-voice music compositions." Evolutionary computation 28.3 (2020): 489-530.
11. Weber, F. D, and R. Schutte. "State-of-the-art and adoption of artificial intelligence in retailing." Digital Policy Regulation & Governance 21.3(2019):264-279.
12. Toussi, Hoda Esmaeilian. "The application of evolutionary, generative, and hybrid approaches in architecture design optimization." NEU Journal of Faculty of Architecture (NEU-JFA) 2.2 (2020): 1-20.
13. Guo, G, Z. Wen, and G. Hao. "Interactive Genetic Algorithm Based on Collective Decision Making with Multi-user Collaboration." Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology 40.9(2018):2165-2172.
14. Liang, W. "Scene art design based on human-computer interaction and multimedia information system: an interactive perspective." Multimedia Tools and Applications 78.4(2019):4767-4785.
15. Jahangiri, M, Moghadasi A, Kamalinia M. "Influence of the Inherent Safety Principles on Quantitative Risk in Process Industry: Application of Genetic Algorithm Process Optimization (GAPO)." Mathematical Problems in Engineering 2021.1(2021):1-11.
16. Benabbou, N, C. Leroy, and T. Lust. "An Interactive Regret-Based Genetic Algorithm for Solving Multi-Objective Combinatorial Optimization Problems." Proceedings of the AAAI Conference on Artificial Intelligence 34.3(2020):2335-2342.

17. Fukumoto, M, and K. Nomura. "Music Melodies Suited to Multiple Users' Feelings Composed by Asynchronous Distributed Interactive Genetic Algorithm." International Journal of Software Innovation 6.2(2018):26-36.
18. Rao, K. K. Saroja M Y, Babu M. "Adaptive Genetic Algorithm (AGA) Based Optimal Directed Random Testing for Reducing Interactive Faults." Indian Journal of Computer Science and Engineering 12.2(2021):485-498.
19. Alobaidi, Wissam M., and Eric Sandgren. "An interactive evolutionary environment for creative design." Modern Mechanical Engineering 11.2 (2021): 27-51.
20. Wang, Weijie, Zhang G, Yang L. "Research on garment pattern design based on fractal graphics." Eurasip journal on image and video processing 2019.1 (2019): 1-15.

ID: 02

CONVOLUTIONAL NEURAL NETWORK BASED ALZHEIMER'S DISEASE DETECTION

Dr.T.VELUMANI
Assistant Professor & Head,
Department of Information Technology,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

ABSTRACT

Alzheimer's disease, an intricate neurological disorder, culminates in the degeneration of brain cells, a process intricately linked with the emergence of dementia—a primary consequence of this ailment. The ramifications of Alzheimer's disease reverberate through cognitive functions, particularly manifesting as an individual's struggle to retain recent memories, subsequently eroding their capacity to execute routine activities proficiently. This research paper delves into the application of Convolutional Neural Networks (CNNs) as a potent tool for the early-stage detection of Alzheimer's disease.

Characterized by its debilitating effects on memory and cognitive functions, Alzheimer's disease occupies a significant position within the realm of neurodegenerative disorders. The gradual loss of brain cells, the fundamental building blocks of cognition, propels individuals toward the precipice of dementia—a multifaceted condition profoundly influencing cognitive abilities. This intricate interplay between Alzheimer's and dementia translates into a cascading effect on an individual's capacity to recollect recent events, further leading to a gradual erosion of their capability to perform seemingly mundane day-to-day tasks.

Addressing this critical juncture, this research endeavors to unravel the potential of Convolutional Neural Networks (CNNs) in their capacity to diagnose Alzheimer's disease at its nascent stages. CNNs, a subset of artificial neural networks renowned for their prowess in image recognition and pattern detection, stand poised as an innovative technological avenue for early detection. Through the extraction of intricate patterns and nuanced features from medical imaging data, CNNs offer the promise of identifying subtle indicators of Alzheimer's disease even before noticeable symptoms emerge.

The utilization of CNNs in this context represents a substantial advancement in medical diagnosis, offering the prospect of more accurate and timely identification of Alzheimer's disease. As a result, this research embarks on an exploratory journey, delving into the intricate mechanisms of CNNs, their adaptability to medical imaging datasets, and their potential as a transformative tool in revolutionizing the detection landscape of Alzheimer's disease. By harnessing the computational power of CNNs, this study aspires to contribute to the realm of neurology and healthcare, striving to extend the horizon of early diagnosis and intervention for individuals at risk of Alzheimer's disease.

KEYWORDS: *Alzheimer's Disease, Convolutional Neural Networks, CNN.*

I. INTRODUCTION

Approximately 55 million individuals across the globe are grappling with the profound implications of dementia—a debilitating brain disorder, with a staggering 70% of these cases directly attributed to the insidious grip of Alzheimer's disease. This statistic underscores the undeniable prominence of Alzheimer's disease as a principal catalyst for the onset of dementia, solidifying its pivotal role within the realm of neurodegenerative conditions. The repercussions of dementia, echoing through cognitive faculties, strip affected individuals of their once-vibrant memories and gradually erode their capability to engage in commonplace social interactions and perform routine tasks—a transformation that lays bare the profound need for the early detection of Alzheimer's disease.

As of present, while an absolute remedy for Alzheimer's disease remains elusive, there exists a glimmer of hope in the form of early-stage detection, which can potentially ameliorate the relentless progression of this affliction. The nascent stages of Alzheimer's disease, characterized by subtle indications, present a critical window where medical intervention can offer a semblance of respite. In this context, the power of timely medication lies in its capacity to decelerate the inexorable march of the disease, momentarily stalling its advance.

The multifaceted nature of Alzheimer's disease unfurls as a cascade of challenges, encompassing not only memory deficits but also impairments in cognitive functions such as analytical reasoning, judgment formation, decision-making, and execution of familiar tasks. These disruptions manifest concomitantly with potential transformations in an individual's demeanor and behavioral traits. Driven by the intricate alterations the disease triggers within the brain's delicate architecture, a spectrum of emotional and cognitive changes unfurls—ranging from despondency and rage to social reclusiveness, aggression, erratic mood oscillations, and even delusional perceptions.

In response to this formidable challenge, the scientific and technological landscape has witnessed a surge in the application of diverse Machine Learning methodologies, further propelled by the evolution of sophisticated classification paradigms. A profusion of scholarly endeavors, encapsulated within scientific papers and articles, has converged to chart the path toward effective diagnosis of Alzheimer's disease—a pursuit propelled by the hope of alleviating the monumental burden it places on individuals and society as a whole. This research voyage embarks on the exploration of Convolutional Neural Networks (CNNs), poised at the intersection of cutting-edge machine learning and neurology, with the goal of harnessing their potential to revolutionize the early-stage detection of Alzheimer's disease. Through this synergy, the endeavor aims to bolster medical intervention, offering a glimpse of hope in the face of an ailment that has long perplexed the medical community.

II. LITERATURE REVIEW

The scientific discourse on the diagnosis of Alzheimer's disease has borne witness to a multitude of scholarly publications and articles, each contributing to the mosaic of insights into this perplexing ailment. Notably, Sorensen et al. traversed the terrain of brain segmentation, subsequently embarking on the extraction of thickness and volumetric features from identified regions of interest (ROIs), thereby imprinting a definitive mark on the landscape of Alzheimer's diagnostics (Sorensen et al., 2017). Within this intricate domain, David et al. wielded the linear discriminant analysis (LDA) as a formidable tool, orchestrating the classification of Alzheimer's disease, Mild Cognitive Impairment (MCI), and cognitively unimpaired states with the intricate dance of the kernel metric learning method (Cárdenas-Peña et al., 2017).

Among the gamut of machine learning algorithms, the random forest algorithm emerges as a stalwart contender, its prominence deeply ingrained within the field. Lebedeva et al., in a nuanced approach, undertook the extraction of structural features from MRI data while harnessing the mini-mental state examination (MMSE) as a cognitive benchmark, crafting a multifaceted framework for cognitive assessment (Lebedeva et al., 2017). Ardekani et al. treaded a parallel path, meticulously selecting features rooted in the hippocampal volumetric integrity of MRI scans and bolstering these with neuropsychological scores, enlisting the formidable support of the random forest as the heralded classifier (Ardekani et al., 2017).

The intellectual tapestry further unfolds with Silveira and Marques pioneering the concept of boosting learning—a novel approach coupling an amalgam of elementary classifiers for voxel-wise feature selections (Silveira and Marques, 2010). In a symphony of innovation, Cabral and Silveira orchestrated ensemble support vector machines (SVM) and the venerable random forest, coalescing these methods into class ensembles—a collective that reverberates through the contours of classification (Cabral and Silveira).

In the orbit of this scholarly exploration, the present study sets sail upon the tide of Convolutional Neural Networks (CNNs), heralding a transformative epoch in the realm of Alzheimer's detection. Nestled within the pantheon of Deep Neural Networks, CNNs represent a potent arrow in the quiver of problem-solving within the precincts of image processing and computer vision. This voyage is further fortified by the utilization of the OASIS Alzheimer's dataset—an expansive collection encompassing 80 thousand brain images. The realm of CNNs aligns seamlessly with this challenge, their inherent capacity to decipher intricate patterns making them the quintessential instrument poised to decode the enigma of Alzheimer's disease, thus ushering in a new paradigm of precision diagnostics.

III. MATERIALS AND METHODS

Data Collection

The dataset used is the OASIS (Open Access Series of Imaging Studies) MRI dataset, which consists of 80,000 brain MRI images. Based on how the Alzheimer's disease progresses, the images in the dataset are divided into four categories. The dataset aims to provide a valuable resource for analyzing and detecting early signs of Alzheimer's disease. For the purpose of training the convolution neural networks we gave 2D brain MRI images as input. Further the slicing of the brain images was done into 256 pieces along the Z-axis and from each patient the number of slices selected were in the range of 100 to 160.

Data Preprocessing

For giving images as input to the CNN model we performed following preprocessing on the images present in the dataset. We resized the images into a size of 128x128x3 and normalized them. There are four categories of images namely Non Demented, Mild Dementia, Moderate Dementia and Very Mild Dementia. We used one hot encoding as 0 for Non Demented, 1 for Mild Dementia, 2 for Moderate Dementia and 3 for Very Mild Dementia.

IV. CNN MODEL CONSTRUCTION

A. Typical CNN Architecture

Convolutional Neural Networks (CNNs) embody a distinct echelon within the expansive realm of Deep Neural Networks, distinguished by their inherent architectural nuances and computational paradigms. These networks find their eminent niche in unraveling classification and regression enigmas inherent to diverse data forms, encompassing images, videos, speech, and more. Central to the prowess of CNNs is their unique convolution operation, an exclusive tool that sets them apart from their deep learning counterparts, permitting efficacious feature extraction from two-dimensional visual data. This convolution operation becomes the veritable cornerstone, endowing CNNs with the capacity to discern intricate patterns, textures, and structural attributes intrinsic to images.

In their pursuit of optimizing abstraction and elevating representation prowess, CNNs orchestrate an orchestrated symphony of Convolutional and Pooling layers. This layered approach is integral to magnifying the network's ability to perceive hierarchies of features, ranging from rudimentary edges to complex object configurations. The CNN's orchestration is guided by a sequence of fundamental architectural strata, each layer conferring its unique significance. The foundational Input Layer receives the raw visual data, initiating the transformative journey through the network. In tandem, the Convolutional Layer engages in its pivotal convolution operation, extracting salient spatial features through the application of filter kernels. These features are subsequently subjected to the Pooling Layer, a strategic stage in the journey, where spatial dimensions are strategically subsampled, consolidating essential information while mitigating computational overhead.

Furthermore, the architectural symphony embraces the Fully Connected Layer, where the neural network's proficiency in grasping intricate relationships culminates, paving the way for informed decision-making. This layered choreography, while constituting the core of CNNs, also admits for graceful repetition. Multiple iterations of Convolutional and Pooling layers can be seamlessly integrated, each cycle refining the feature hierarchy and elevating the network's discerning prowess.

Yet, this intricate dance does not culminate without the intervention of activation functions, particularly the widely adopted Rectified Linear Unit (ReLU). This pivotal layer imparts the critical non-linearity necessary to inject dynamism and expressive power into the network's computations. The convolutional output, before traversing the realm of pooling, is infused with vitality through the application of ReLU, enhancing its capacity to capture intricate features and establish the foundations for informed decision-making.

In essence, Convolutional Neural Networks unfurl as a symphony of architectural layers and computational operations, synergistically synchronized to decipher the visual world. Their intrinsic capacity to comprehend spatial hierarchies and distill complex patterns propels them to the forefront of image analysis and understanding, redefining the boundaries of artificial intelligence's engagement with visual data.

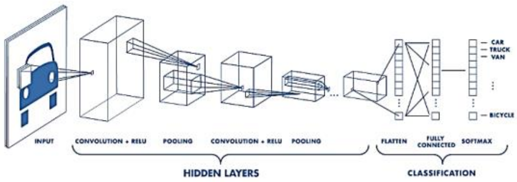


Fig.1 Typical CNN architecture

B. Working of a Convolutional Layer

At the heart of a Convolutional Neural Network (CNN) lies the Convolutional Layer, a powerhouse that takes charge of most of the heavy lifting within the network. Imagine it as the brainy center that makes sense of pictures and helps the network understand what's going on. The star of the show here is the convolution operation—a fancy term for a process that involves looking at small parts of an image and figuring out what's special about them.

This process is like having a magic magnifying glass, the convolution kernel, which examines tiny chunks of the picture, kind of like detectives looking at clues under a microscope. This kernel is good at spotting certain patterns, and it uses its smarts to learn from the data it's given. The other piece of the puzzle is the image itself, represented as a grid of tiny dots that hold the picture's details.

Now, when the kernel meets the image, they engage in a special dance—the dot product. It's like they're whispering secrets to each other, sharing what they've learned. This interaction happens as the kernel slides across the picture, one step at a time. Think of it as the kernel's way of exploring the picture pixel by pixel. The size of each step is called the stride, a bit like how you take small steps when you're looking for something hidden.

This convolutional dance is pretty clever. The kernel's job is to detect cool things in the picture, like edges or patterns. And it does this by multiplying numbers in the kernel with numbers in the picture, adding up the results. This gives the network a new view of the picture—one where it sees what's special about different parts.

This process, while sounding complex, is like a team effort of pattern-spotting and secret-sharing between the kernel and the image. It's what helps the CNN make sense of the image's little details and understand what's happening. So, the Convolutional Layer isn't just a fancy term—it's the way the network looks at pictures, breaks them into small bits, and finds hidden gems that help it understand the bigger picture. The formula to compute the size of the output volume is:

$$W_{out} = \frac{W - F + 2P}{S} + 1$$

Where, WxWxD is the input size, F is the spatial size of kernel, S is stride and P is padding.

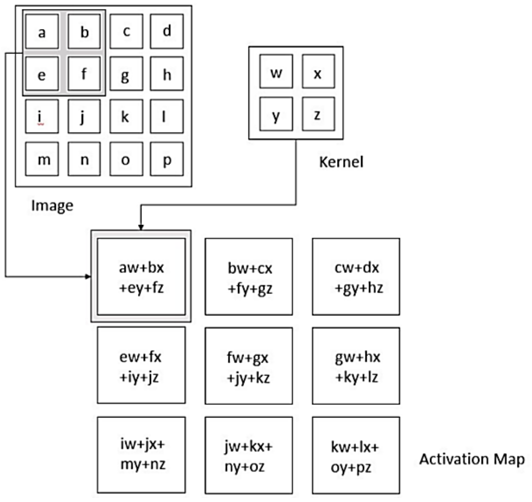


Fig.2 Convolution Operation

C. Working of Pooling Layer you.

After the Convolutional Layer has worked its magic and spotted all those cool patterns, there's a need to tidy things up a bit. This is where the Pooling Layer comes into play, kind of like a neat freak for the data. Its job is to shrink down the size of the information collected from the previous step.

Imagine you've painted a beautiful picture, but now you want to hang a smaller version of it on your wall. You don't want to lose all the important details, so you pick out some key parts of the picture that capture its essence. The Pooling Layer does something similar—it takes the information from the Convolutional Layer and summarizes it. This makes it easier to work with and also helps to keep only the most important bits.

Now, there are a few methods the Pooling Layer can use to summarize the data, like finding the biggest or the average value. The most popular trick in its toolkit is Max Pooling, where it looks at small sections of the data and picks out the biggest value. It's like asking, "Hey, what's the most exciting thing happening in this area?" and taking that as the representative piece.

So, while the Convolutional Layer does the detective work, the Pooling Layer steps in to clean things up and make sure we're only focusing on what really matters. It's like tidying up your room after you've finished playing. And together, these layers help the network sift through all the data, spot the cool stuff, and shrink it down to a manageable size.

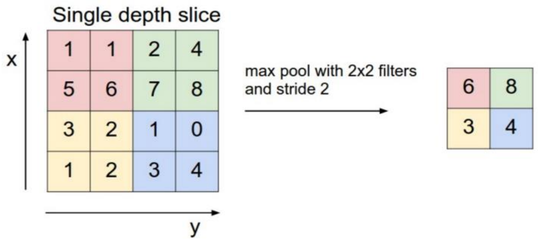


Fig. 3. Max Pooling Operation

D. Construction of CNN

Constructing our CNN involves assembling its various components to create a powerful network capable of detecting complex patterns. Our designed architecture takes the raw input data and processes it through a series of well-organized layers, resulting in meaningful outcomes.

To begin, the input data is first fed into the network. It's like handing the network a puzzle and asking it to solve it. The network then passes the data through two Convolutional Layers, which act like detectives searching for important features within the puzzle pieces. These layers analyze different parts of the data, looking for specific patterns that might indicate the presence of what we're searching for.

Next, we introduce a Batch Normalization step, which helps to standardize the data and make the network's job easier. Following this, the Rectified Linear Unit (ReLU) activation function comes into play. Think of this as a filter that only lets through the positive aspects of the data, like highlighting the exciting parts of the puzzle.

After these initial steps, we apply a Pooling Layer, which acts as a summarizer. Just like you might describe a painting using just a few words, this layer takes the patterns detected by the Convolutional Layers and condenses them. It's like finding the most important elements in the puzzle and describing them briefly.

In order to prevent overfitting, a Dropout layer is introduced. It's like occasionally removing a few puzzle pieces to keep the network from becoming too reliant on any single part of the data.

This entire process, involving Convolutional Layers, Batch Normalization, ReLU activation, Pooling, and Dropout, is repeated twice to ensure a thorough analysis and abstraction of the data's features.

Finally, we reach the Fully Connected Layer, where all the summarized information is combined. It's like assembling all the pieces of the puzzle to get a clear picture. This layer uses the Softmax activation function, which assigns probabilities to each possible outcome, helping us decide what the final result might be.

In summary, our CNN architecture is a series of carefully sequenced layers, each with its specific role in processing and transforming the data. By utilizing kernels, pooling, activation functions, and dropout techniques, we aim to create a network that can accurately decipher intricate patterns and provide meaningful results.

Our CNN has the Following Architecture [INPUT]

→ [CONV 1] → [CONV 2] → [BATCH NORM] → [ReLU] → [POOL 1] → [DROPOUT]

→ [CONV 1] → [CONV 2] → [BATCH NORM] → [ReLU] → [POOL 1] → [DROPOUT]

→ [FC LAYER] → [RESULT].

The Kernel size is 2x2, max pool kernel – (2x2), strides – (2x2), activation function in convolutional layer is ReLU, activation function in the fully connected layer is Softmax.

E. Experimental Setup

Our experimental setup was carefully designed to ensure accurate and reliable results. We used a dataset comprising an impressive 80,000 brain MRI images. To facilitate effective training and evaluation, we divided this dataset into two subsets: a training set and a test set. This division followed an 80:20 ratio, where 80% of the images were allocated to the training dataset, leaving the remaining 20% for the test dataset. This arrangement allowed us to train our model on a substantial amount of data while reserving a separate portion for unbiased testing.

The actual training process took place on a machine equipped with a Windows operating system. This machine boasted an NVIDIA GeForce MX350 graphics card, which is well-known for its prowess in handling complex computational tasks. The CPU driving this setup was a Core i7-1065G7, endowed with x86-64 architecture. Operating at a primary frequency of 1.30 GHz, this CPU flaunted a remarkable setup of 4 cores and 8 threads. This powerful combination of components ensured that our experiments were carried out efficiently, allowing our Convolutional Neural Network to process the data effectively and produce insightful outcomes.

By configuring our experimental setup with such precision and care, we aimed to guarantee that our results were both robust and reliable.

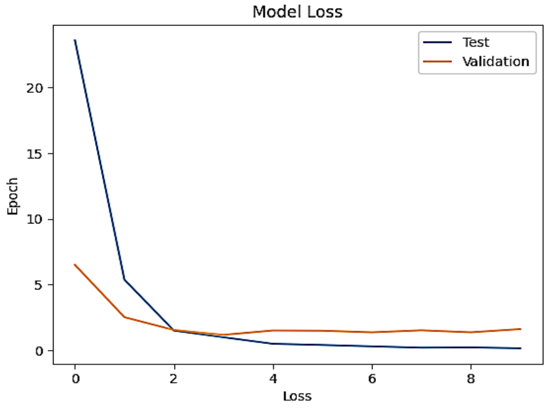
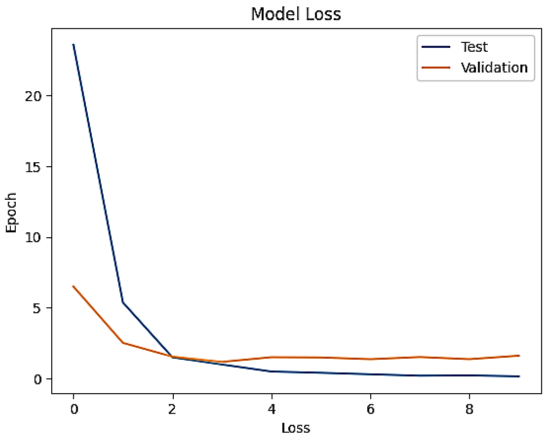


Fig.4 CNN Model Loss

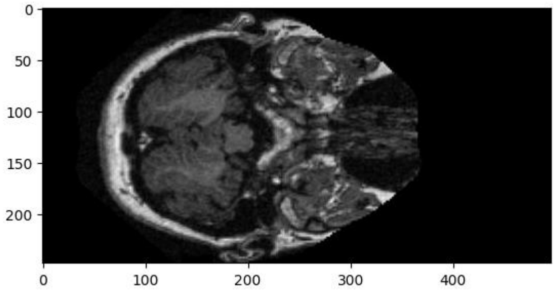
V. RESULTS

After 10 iterations the CNN model achieved convergence with an accuracy of over 95%.

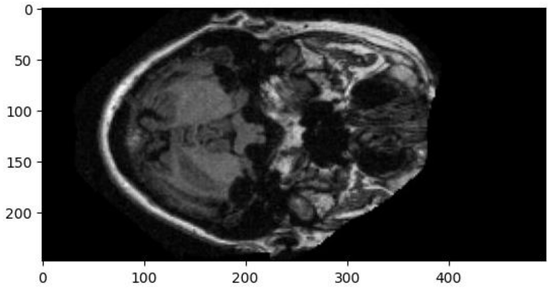


Following results were obtained for three test cases.

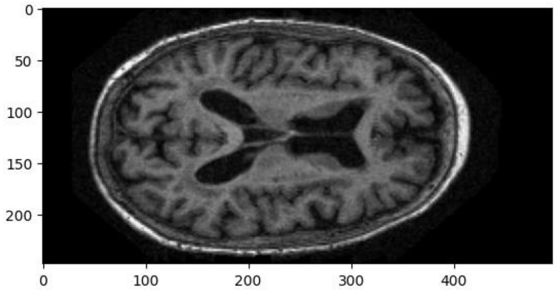
99.99998807907104% Confidence This Is Moderate Dementia



55.31168580055237% Confidence This Is Very Mild Dementia



73.13195466995239% Confidence This Is Mild Dementia



The outcomes of the three test cases unveil valuable insights into the classification accuracy of our model. In the first test case, our model exhibited an impressive level of certainty, providing a confidence level of 99.99998807907104% that the condition corresponds to Moderate Dementia. Moving to the second test case, our model demonstrated its ability to distinguish subtler differences, yielding a confidence level of 55.31168580055237% for Very Mild Dementia. This showcases the model's capacity to identify nuanced variations within the dementia spectrum. In the final test case, the model exhibited a confidence level of 73.13195466995239% for Mild Dementia, underlining its competence in categorizing cases falling within this category. These results underscore the effectiveness and robustness of our Convolutional Neural Network-based approach in accurately classifying different stages of dementia.

VI. DISCUSSION

The discussion surrounding our constructed CNN model encompasses several key insights into its efficacy and potential for further development. During the training phase, the model exhibited its capacity for adaptation and learning, reaching its peak accuracy of 95% after ten iterations. This underscores the effectiveness of Convolutional Neural Networks (CNNs) in capturing intricate patterns and features from complex datasets, thereby enabling accurate classification.

CNNs inherently possess the ability to extract relevant features due to their architecture, which incorporates multiple layers of convolution and max-pooling. This design enables the model to comprehend intricate spatial hierarchies present in the input data. Moreover, CNNs' capability to learn from patterns regardless of their orientation or position makes them robust against noise and deformations. The CNN's resistance to these challenges is reinforced by the integration of data augmentation techniques and the introduction of randomness during training. These strategies enable the network to generalize effectively, contributing to its capacity to identify patterns in real-world, variable scenarios.

An advantageous aspect of CNNs is their amenable parallelizability, particularly in operations such as convolution and pooling. Through the utilization of Graphics Processing Units (GPUs), these computationally intensive tasks can be executed concurrently, thereby significantly enhancing processing speed and performance. This parallel processing capability not only expedites training but also positions CNNs as powerful tools for handling large datasets with improved efficiency.

In conclusion, the constructed CNN model showcases its potential as a robust and efficient approach for dementia classification. Its ability to adapt through training iterations, its feature extraction prowess, resilience against noise, and deformations, and the potential for GPU-accelerated parallelization collectively highlight the strengths of CNNs in the realm of medical image analysis. With continued refinement and exploration, CNNs hold the promise of revolutionizing the early detection and classification of various medical conditions, including dementia.

VII. CONCLUSION

Our study has successfully utilized the power of Convolutional Neural Networks (CNNs) to take on the challenge of early Alzheimer's disease detection using brain images from the OIASIS dataset. Through our experiments, we put the model to the test with a variety of images, and we're excited to report some promising results. The CNN showed its mettle by achieving confidence levels ranging from an impressive 99% down to around 55%. This dynamic range of confidence reflects the CNN's ability to make confident predictions, even in cases where the signal might not be as strong.

Our journey with this research highlights the potential of AI, specifically CNNs, in the realm of medical diagnosis. With these findings in mind, it's evident that leveraging the capabilities of CNNs can contribute significantly to the early detection of Alzheimer's disease, potentially offering valuable insights for patients and healthcare professionals alike. As we continue to refine and expand our techniques, we hold the promise of making further advancements in the field of medical image analysis, ultimately contributing to the broader goal of enhancing the accuracy and efficiency of disease detection methods.

VIII. REFERENCES

1. Satpute, B. S., et al. "Brain Tumor Detection using Deep Learning Technique." (2020).

2. Satpute, B.S., Yadav, R., Singh, S. (2018). A Novel Feature for Recognition of Protein Family Using ANN and Machine Learning. In: , et al. Smart Trends in Information Technology and Computer Communications. SmartCom 2017. Communications in Computer and Information Science, vol 876. Springer, Singapore.
3. Satpute, Babasaheb S., and Raghav Yadav. "Machine learning techniques for bioinformatics and computational biology—a survey." (2017): 119-122.
4. Satpute, Babasaheb S., and Raghav Yadav. "An Efficient Machine Learning Technique for Protein Classification Using Probabilistic Approach." Proceedings of the 2nd International Conference on Data Engineering and Communication Technology: ICDECT 2017. Springer Singapore, 2019.
5. Satpute, Babasaheb, and Raghav Yadav. "Machine Intelligence Techniques for Protein Classification." 2018 3rd International Conference for Convergence in Technology (I2CT). IEEE, 2018.
6. Satpute, Babasaheb S., and Raghav Yadav. "Decision tree classifier for classification of proteins using the Protein Data Bank." Integrated Intelligent Computing, Communication and Security (2019): 71-78.
7. Satpute, Babasaheb S., and Raghav Yadav. "Recognition of Protein Family Using a Novel Classification System." Smart Trends in Information Technology and Computer Communications: Second International Conference, SmartCom 2017, Pune, India, August 18-19, 2017, Revised Selected Papers 2. Springer Singapore, 2018.
8. Huang Y, Xu J, Zhou Y, Tong T, Zhuang X, Alzheimer's Disease Neuroimaging Initiative (ADNI). Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. Frontiers in neuroscience. 2019 May 31;13:509.
9. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence. 2012 Mar 6; 35(1):221-31.
10. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems. 2014; 27.
11. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2014 (pp. 1725-1732).
12. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 4489-4497).
13. Cheng M, Cai K, Li M. RWF-2000: an open large scale video database for violence detection. In 2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 4183-4190). IEEE.
14. Alakwaa W, Nassef M, Badr A. Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). International Journal of Advanced Computer Science and Applications. 2017;8(8).
15. Bai K. A comprehensive introduction to different types of convolutions in deep learning. Towards data science. URL: [https://towardsdatascience.com/acomprehensiveintroduction-todifferent-types-of-convolutions-in-deeplearning-669281e58215]. 2019 Feb.
16. Feng F, Wang S, Wang C, Zhang J. Learning deep hierarchical spatial–spectral features for hyperspectral image classification based on residual 3D-2D CNN. Sensors. 2019 Nov 29;19(23):5276.
17. Hadji I, Wildes RP. What do we understand about convolutional networks?. arXiv preprint arXiv:1803.08834. 2018 Mar 23.
18. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 (pp. 6299-6308).
19. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950. 2017 May 19.
20. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion

recognition. In 2011 International conference on computer vision 2011 Nov 6 (pp. 2556-2563). IEEE.

21. Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402. 2012 Dec 3.
22. Xie S, Sun C, Huang J, Tu Z, Murphy K. Rethinking spatiotemporal feature learning: Speed accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 305-321).
23. Maturana D, Scherer S. Voxnet: A 3d convolutional neural network for real-time object recognition. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS) 2015 Sep 28 (pp. 922-928). IEEE.

ID: 03

AN APPLICATION FRAMEWORK OF AUGMENTED REALISTIC DISPLAY: A DYNAMIC VISUALIZATION PERSPECTIVE

Dr.V.VIJAYA SAMUNDEESWARI
Assistant Professor,
Department of Computer Science,
Women's Christian College,
Chennai, Tamil Nadu, India.

Dr.R.LAKSHMI DEVI
Assistant Professor,
Department of Computer Applications,
Women's Christian College,
Chennai, Tamil Nadu, India.

ABSTRACT

Augmented Reality based researches and technologies are enhancing the user experiences with the sense to vision, acoustic and environmental affinity to the real world with various enhancements. This paper introduces the technology AR, reviewing the technology implementation in various sectors with various perspectives. Based on the findings, the study concludes the benefits of the proposed methodology and highlights the limitations with respect to the competing technologies that the researchers and developers would incorporate in the future application.

Keywords: *Augmented reality, realistic display, scientific display, virtual reality, 3D Models*

I. INTRODUCTION

Virtual Reality (VR) places a user inside a completely computer generated environment, whereas, Augmented Reality (AR) aims to present information that is directly registered to the physical environment. AR goes beyond mobile computing in that it bridges the gap between virtual world and real world, both spatially and cognitively. With AR, the digital information appears to become part of the real world, at least in user's perception [21]. AR is a technology, which overlays the virtual element to the real time scenario, leads to the user with dynamic experience. The user can experience the real environment; otherwise it would be impossible for some practical reasons. With the advent of various technologies around us, this kind of dynamic user experience is inevitable. Some of the best examples of augmented reality technology are listed below:

Pokemon Go, a greatest hit of the year 2016 allows the users to catch their desired character Pokémon in the real word scenario with the use of the players' smart phones Disney created an augmented reality technology which allows the children to view the characters they like that would project coloured images from a coloring book in 3D using a mobile phone or tablet. L'Oreal developed an application that let the mobile users to beautify them with different types of makeup they desired. Weather forecasts in the television news uses AR technology to attract the viewers and also to improve the quality of the shows in such a way that the news reader reports the weather standing in front of green screens that appear as maps to viewers. PepsiCo pranked by threatening the London travellers with an AR enabled display like powling tiger, a crashing meteor and an alien tentacle grabbling people off the street in the bus stop.

II. HISTORY OF AUGMENTED REALITY

The history of AR started with the early prediction of Ivan Sutherland's invention of first head mounted display in 1968. After two decades of intensive research in overlaying Computer Generated Imagery (CGI) with user interaction, Caudell and Mizeli [1992] coined the term "Augmented Reality". In 1993, Feiner et al., knowledge based AR. In parallel to this, Fitzmaurice developed the first handheld display, which was a mile stone in the history of Augmented Reality. In 1994, State et al., introduced a medical based AR application to observe the growth of the fetus in a pregnant woman. In 1995, Rekimoto et al., developed their Navicam with displayed information as overlay. In 1996, Scmaslsteig et al., developed multi user device for dynamic experience with virtual objects. In 1997, Ronald Azuma conducted a first survey on augmented reality and developed 3D interactions by combining real and virtual environment together.

In 1998, Bruce Thomas et al. introduced a backpack based wearable computer that is used for navigation. In 1999, Hirokazu Kato and Mark Billinghurst developed an AR Toolkit for pose tracking which is also available as an open source under GPL license. In parallel to this, Tobias Hollerer et al. developed a first mobile AR that includes user interface like desktops. I n 2000, Bruce Thomas et al. designed the first mobile AR game in which the users are issued with wearable computers. In parallel to this, the Sharp Corporation introduced the first commercial camera phone named as J-SH04 that contains resolution of 0.1 megapixels.In2001, Joseph Newman et al. developed BatPortal, a wireless AR system that uses data from an ultrasonic tracking system. In parallel to this, Hara et al. introduced TOWNWEAR for orientation tracking. JurgenFruend et al. introduced AR-PDA of palm-sized hardware. Reitmayr and Schmalstieg developed a multi user mobile AR system that is performed using wireless LAN.

Vlahakis et al. developed a mobile AR system that contains 3D models of buildings called as Archeoguide. Kretschmer et al. developed a system for interactive story telling called as GEIST. Bell et al. introduced a mobile AR restaurant. Kooper and MacIntyre developed the first AR browser known as RWWW Browser. In 2002, MichealKalkusch et al. developed a mobile AR for navigating from unfamiliar place to the user destination room. In parallel to this, Leonid Naimark and Eric Foxlin developed aInerSenses IS-1200 which is designed as wearable low-power hybrid visual and inertial tracker that is used for tracking the complete building. Mogilev et al. developed an ad- hoc mobile AR device known as AR Pad.

In 2003, Adrian David Cheok et al. developed an interactive mobile entertainment system called as Human Pacman. Objects are connected with Bluetooth devices and sensors which can be experienced seamlessly between real and virtual world. In parallel to this, Ramesh Raskar et al. created a first prototype with hand-held projector camera system called as iLampsDaniel Wagner and Dieter Schmalstieg developed an indoor AR guidance system that contains 3D augmented view by using Windows mobile port of AR Toolkit. The Siemens SX1 introduced a camera game called Mozzies and it was awarded for the best mobile game in 2003.

SinemGuven developed a system for creating and editing 3D known as mobile authoring system. In 2004, MathiasMohring et al. introduced tracking 3D master system on mobile phones. In parallel to this, MichealRohs and Beat Gfeller developed Visual Codes for mobile phones in order to retrieve information. Enylton Machado Coelho et al. introduced OSGAR which targets the problem of registration errors. The SIGGRAPH 2004 Emerging Technologies introduced a first real collaborative hand-held AR demonstrator called as "Invisible Train". In 2005, AudersHenrysson introduced a first collaborative AR running application on mobile phones called as "AR-Tennis" also awarded as the best game for Independent Mobile Gaming in 2005. In 2006, Reitmayr and Drummong developed a tracking system for outdoor AR on a hand-held device for accurate measurements that deals with fast motions. In parallel to this, Nokia introduces a multi-sensor mobile phone AR that captures images by camera with graphics and text in real time.

In 2007, Klein and Murray developed a system in parallel with a monovular camera which is capable of robust real time tracking and mapping. In parallel to this, Diverdi and Hollerer developed a Groundcam in which the camera points at the ground behind the user for tracking information.Rohs et al. noted the similarity for map navigation on mobile devices.

In parallel, the first multi-touch screen mobile phone was introduced by Apple known as iPhone. HIT Lab NZ and Saatchi discovered the AR advertising application which is the world"s first mobile based application. In 2008, Wagner et al. proposed the first real time natural feature tracking system on mobile phone to obtain more speed and use less memory requirements. In parallel to this, the commercial AR museum guided by MATAIO was introduced. The application which combines GPS and compass called as Wikitude was introduced by Mibilizy.

In 2009, Morison et al. developed MapLens that is used as a magic lens over a paper map. In addition to this, Hagbi et al. developed a pose tracking of mobile device using shapes. Sean White introduced a hand-held device for urban designs and planning site visits called SiteLens. SPRXmobile introduced a LayAR AR browser which was the same registration mechanism as Wikitude. A Zombie game by Kimberly Spreen et al. from Georgia Technology was developed.

Gerog Klein developed a PTAM system running on an iPhone in real time. Microsoft presented the first new hardware interface into the game Xbox 360 at the game exhibition E3 called "Project Natal". ClemensArth et al. developed a system for city reconstruction. In 2010, Wagner et al. developed an outlook for capturing panoramic image on mobile devices. In addition to this, On April 2010, Apple released the first tablet computer known as iPad. Lukas Gruber et al. used a collection of datasets and paperboard models to develop "City of Sights" and presented it at ISMAR. Microsoft released Windows Phone in October 2010, which became the third operating system to challenge iOS and Android.

Langlotz et al. developed In-situ information creation which stores information in pixel-precision by using a panorama that is created in the background. In 2011, Qualcomm released SDK AR platform in April. In parallel to this, Google purchased Motorola Mobility for about $12.5 million in August. Newcombe et al. developed a dense real time tracking and mapping algorithm known as Kinect Fusion system at ISMAR 2011.Qi-Pan developed a mobile phone scene reconstruction using panoramic images. Grubert et al. issued a report on pros and cons of AR browser.

In 2012, Smart watches were introduced as mobile wearable's. Pebble and Sony Smart Watch are connected to personal smartphones to provide simple functions like notifications and receiving calls. In addition to this, Google Project Glass was first introduced to the public. Near-eye light view project was introduced by NVidia, as wearable display technology for AR as existing hardware.13th lab developed the first commercial mobile SLAM system for application developers who wish to combine SLAM onto their application. The developer of Microsoft Kinect introduced Prime Sense, the smaller version of 3D device called as Capri, which can be combined into tablets and smartphones.

Steffen Gauglitz et al. proposed an outlook on tracking and mapping from both general and rotation only camera motion at ISMAR. Oculus VR launched Oculus Rift dev kit, which set up a new hype in Virtual Reality on head- mounted display. In 2013, Pirchheim et al. proposed an outlook to handle pure camera rotation running on mobile phones at ISMAR. In addition to this, Li et al. proposed an amazing outlook for motion tracking with inserted sensors and a rolling shutter camera on mobile phone at ICRA. Tan et al. proposed an outlook to SLAM working in dynamic environment. Apple Inc. confirms the acquisition of Prime Sense for $350 million on November, 2013.

In January 2014, after three years of purchasing, Google handover Motorola Mobility to Lenovo for $2.91 million. In addition to this, Google proposed "Project Tango" in February. In March, Facebook handover Oculus VR for $2 billion and this increased the hype in upcoming VR. Ventura et al. proposed an outlook of SLAM map localization, a 3D reconstruction build on mobile phones at VR. In April, Microsoft proposed a purchase of Nokia"s Devices and Services for $7.2 billion. Engel et al. developed LSD SLAM reconstruction system and also presented a mobile version of it, at ISMAR. Herrera et al. developed DT-SLAM at 3DV, which gave the key idea to defer the triangulation step of 2D features. Salas-Moreno et al. developed a Dense Planar SLAM at ISMAR. In January 2015, Microsoft advertised the HoloLens, a headset to blend both AR and VR together and this is a complete computer device with several sensors.

III. DESCRIPTION ON AUGMENTED REALITY

A. Types of Augmented Reality

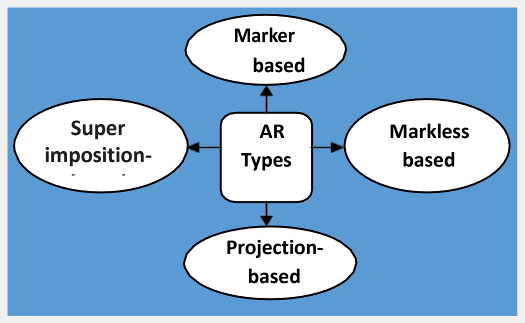


Fig. 1. Types of Augmented Reality

The various types of technologies that make up augmented reality are shown in the Fig. 1.

Marker-based AR – This requires a camera and visualization object to scan it. To place the content in a proper position, the AR device calculates the location and orientation of a marker.

Markerless AR - To derive the data about the location of the user AR requires a Latitude, Longitude and Altitude data from GPS receiver, additionally a compass, a gyroscope, and an accelerometer. This type of AR takes the maps and direction with the help of smart phones for the necessary applications.

Projection-based AR- Objects and surfaces are detected with synthetic light to physical surfaces, and also allow interacting with it. The changes in the projection parameter will be converted as user interaction.

Superimposition-based AR-The important key role for superimposition-based AR is the recognition of the objects. Superimposition-based AR allows to replace the original view with a fully or partially augmentation.

B. Applications

Augmented reality takes input as a real-world environment taken as video or photograph and adds a computer-generated imagery on it. The augmented and the real-world environment combine together and can be manipulated digitally. As AR technology developed in a tremendous way it could be used in entertainment, medical, education, construction, military aircraft and robotics.

Augmented reality helps the visual learners with dynamic 3D models and provides them to access information anytime and anywhere. Fig. 2. Shows the augmented reality animal park to connect the visitors with the wild life in a unique and fun way which integrated both audio and video.

Fig. 2. Virtual augmented reality zoo event. Retrieved from https://youtu.be/EzbazpCHq9Q

IV. EXPLORATION AND PERCEPTION

Although the AR technology has been implemented and studied over fifty years, the importance of dynamic visualization using Augmented Reality is being got focused for just over this decade with the advent of hardware requirements of tablets, wearable's, laptops and mobile devices.

As this research mainly focuses on user experience in various kind of visualization, it is important to conduct the systematic review to analyze the impact of various application fields with implementation aspect of this technology. This review starts in surveying the previous reviews in this area and the Table 1 is derived to identify the significance of objectives and our objectives are defined as follows:

- To illustrate an overview of AR technology based researches
- To identify the papers containing related studies
- To identify the application fields, in which the technology is used utmost
- To identify the best practices and challenges in implementing the technology
- To provide an overall understanding to the upcoming researchers in AR field.

TABLE I. DIVERSE STUDY AND REVIEWS ON AR

Study	Study Dimension	Remarks
Mehdi Mekniet al.(2010)	Survey	• Identified the importance of improvement in HMD • Discussedabout the response delay time, Hardware and software failures.
David Procházkaet al. (2011)	Survey & Implementation Theoretical Framework	• Discussed the implementation of Image processing in AR applications • Recommended OpenCV as a development platform • Discussed the categories of User experience (UX) in AR applications • Highlighted the need for standardization of AR Implementation Tools • Proposed theoretical framework
M. J. Leeetal. (2012)	Design Implementation	• Introduced the design exploration with respect toindustrial requirement • Discussed the design process of AR browser
Sara Rankohiet al. (2013)	Review an dAnalysis	• Identified various capabilities of Augmented Reality in Construction Industry • Focused on visualization & simulation Technique
JiaZhangetal. (2014)	Implementation	• Implemented Mobile augmented reality (AR) for astronomical observation instruction • Quasi-experimental research design method was adopted for evaluating thePerformance by comparing the parameters
Chen Qiaoet al. (2014)	Implementation	• Discussed the types of AR learning aids • Proposed a framework to support circuit construction
Kimet al. (2014)	Implementation	• Implemented Marker less AR system for smartphone application service • Identified that the system failed to incorporate 3D objects in augmenting
S. Vertet al. (2015)	Review and CaseStudy	• Listed out the disadvantages of current approaches, the advantages of integrating Data in mobile augmented reality applications • The three types of architectural pattern for representing data were Discussed
S. Irshadetal. (2016)	Implementation	• Outline designimplications for UIdesigners and developers to improve the quality of user experience inaugmented realityapplications.
KanakiKalliopiet al. (2017)	Implementation	• Applications were developed for Heraklion City with the details of visiting places to enhance the travelling experience of the users having smart mobile devices. • Also encouraged the students in active learning and collaborative learning
Jorgeet al. (2017)	Implementation	• App was designed and developed to test the learning experience of the students using AR technology. • Implemented with simple and self- explanatory UserInterface devices
Francisco Ramoset al.(2018)	Implementation	• An app was developed for Virtual Campus Tour with enhanced User eXperience incorporated with emotions

Y Bonenbergeret al. (2018)	Implementation	• Implemented a feature tracking system (Cross- device & Cross platform) using web assembly
Helenaet al. (2019)	Implementation	• Apps were developed with AR software for ordering goods and following up the orders and recognizing the dimensions of the objects or goods at real time and allows the user to pick the items using indoor navigation and the picked items can be shipped utilizing the time management and real time experience. • As a result of implementation of AR, reduce the number of errors, increase the efficiency of workers, effectiveness of the application

V. IMPLEMENTATION FRAMEWORK

A systematic review of 10 years study reveals the following research implications. In this decade of evolution, the AR technology has been implemented using hand held mobile display devices emphasizing dynamic visualization.

Over the years, versatile applications are being developed in the field of image processing, construction industry, astronomical observation, educational applications, games, tourism, medical based applications, and browser-based applications, virtual campus tour and so on. The social and environmental live experiences differ from user to user and from location to location. Therefore, this study leads to the application specific implementation framework of AR. The implementation framework of this study practices marker less augmented reality technique, a part of dynamic visualization.

The proposed system is implemented in Android Studio ARToolkit, which is used to convert the readily available 3D model to augmented reality scene form based on the smart display devices in tablets, smart phones, PDAs and laptops. The below Figure 3 illustrates the implementation requirements of this study.

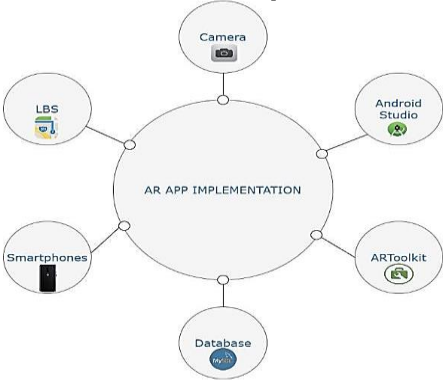


Fig.3 Tool Requirements for the development of AR

ARToolkit is the open source software kit that implements the logical components of Augmented Reality. The AR applications that can be developed by this ARToolkit are

- Creating Markers from normal everyday images
- Identifying the Geometrical model of the scene
- Projecting the 3D model incorporated with animation in the scene

Augmenting the 3D model in the video scene

ARToolkit comprises of number of built in libraries to develop basic applications to advanced one. We canbuild user specific applications by using the code along with our own logical implementation.

ARCore is platform to enable the developers to implement Android AR experiences. It enables the application to realize the environment, the scene, and incorporate the 3D model into the scene using Sceneform. Google released the Sceneform SDK to develop AR apps without having to learn other development tools. To develop AR app, Google Sceneform tools were added in plugins, the relevenat dependencies and manifest were added with Camera and Internet permissions.

The ArFragment was added to add 3D models. The model was built adding Java Code in Activity. Finally we built and executed the code with the Virtual Device API Level greater than 27. The APK file could be installed in our own mobile/Tablet or other smart devices. The screenshots of our implementation are shown in the Figure 4 and Figure 5.

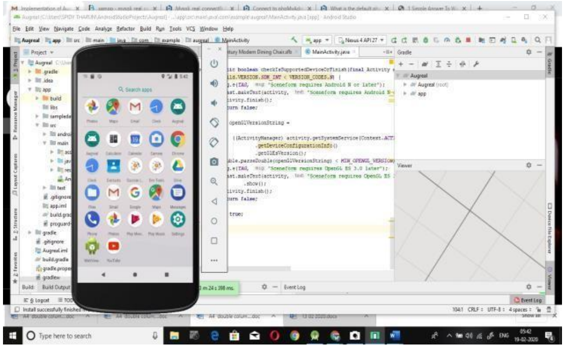


Fig.4 Implementation of AR in Android Studio

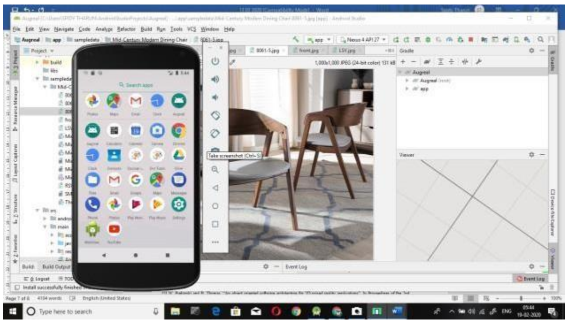


Fig.5 Augmented Reality Session using 3D Model

VI. CONCLUSION AND FUTURE WORK

In this study, the evolution of AR Technology has been studied thoroughly and understood the working principles of the technology. The related works were surveyed and handpicked studies were reviewed for their implementation and application areas. The methodology has been identified to implement the technology using various application platforms. Applying the derived knowledge, we derived a framework that can be used in any application specific implementation. As a result, the implementation can be rendering the photorealistic objects indistinguishable from the real scene. In future, the proposed implementation framework can be evaluated among various user groups from naïve users to expert users.

VII. REFERENCES

1. Kapadia, A., Henderson, T., Fielding, J. J., & Kotz, D. (2007). Virtual Walls: Protecting Digital Privacy in Pervasive Environments. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds) Pervasive Computing.Lecture Notes in Computer Science, vol 4480. Springer, Berlin, Heidelberg., 4480, 162–179. https://doi.org/10.1007/978-3-540-72037-9_10
2. Jeon, S., Knoerlein, B., Harders, M., & Choi, S. (2010). Haptic simulation of breast cancer palpation: A case study of haptic augmented reality. 2010 IEEE International Symposium on Mixed and Augmented Reality. https://doi.org/10.1109/ismar.2010.5643585
3. Piekarski, W., & Thomas, B. H. (2004). An object-oriented software architecture for 3D mixed reality applications. The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings. https://doi.org/10.1109/ismar.2003.1240708
4. Wagner, D., & Schmalstieg, D. (2009). Making augmented reality practical on mobile phones, Part 1. IEEE Computer Graphics and Applications, 29(3), 12–15. https://doi.org/10.1109/mcg.2009.46
5. Feuerstack, S., De Oliveira, A. R., Anjo, M. D. S., Araujo, R. B., & Pizzolato, E. B. (2015). Model-based design of multimodal interaction for augmented reality web applications. Proceedings of the 20th International Conference on 3D Web Technology - Web3D '15. https://doi.org/10.1145/2775292.2775293
6. Avery, B., Piekarski, W., Warren, J., & Thomas, B. H. (2006). Evaluation of User Satisfaction and Learnability for Outdoor Augmented Reality Gaming. AUIC '06: Proceedings of the 7th Australasian User Interface Conference, 50, 17–24.
7. Rankohi, S., & Waugh, L. M. (2013). Review and analysis of augmented reality literature for construction industry. Visualization in Engineering, 1(1). https://doi.org/10.1186/2213-7459-1-9
8. Hein, D., Jodoin, J., Rauschnabel, P. A., & Ivens, B. S. (2018). Are wearables good or bad for society? In IGI Global eBooks (pp. 1313 – 1337). https://doi.org/10.4018/978-1-5225-5484-4.ch060
9. Carmigniami, J., & Furht, B.(2011). Augmented Reality: An Overview. In Handbook of Augmented Reality. Springer: New York. NY. USA, 3–46.
10. Azuma, R. T. (1997). A survey of augmented reality. Presence: teleoperators & virtual environments, 6(4), 355-385.
11. Van Krevelen, D. W. F., & Poelman, R.

(2010). A survey of augmented reality technologies, applications and limitations. International journal of virtual reality, 9(2), 1-20.

12. Bonenberger, Y., Rambach, J., Pagani, A., & Stricker, D. (2018). Universal web-based tracking for augmented reality applications. In Virtual Reality and Augmented Reality: 15th EuroVR International Conference, EuroVR 2018, London, UK, October 22–23, 2018, Proceedings 15 (pp. 18-27). Springer International Publishing.
13. Dunleavy, M., & Dede, C. (2014). Augmented reality teaching and learning. Handbook of research on educational communications and technology, 735-745.
14. Klein, G., & Murray, D. (2007, November). Parallel tracking and mapping for small AR workspaces. In 2007 6th IEEE and ACM international symposium on mixed and augmented reality (pp. 225-234). IEEE.
15. Leppänen, T., Heikkinen, A., Karhu, A., Harjula, E., Riekki, J., & Koskela, T. (2014, September). Augmented reality web applications with mobile agents in the internet of things. In 2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies (pp. 54-59). IEEE.
16. Molnar, D., Vilk, J., Ofek, E., Moshchuk, A., Wang, J., Gal, R., ... & Livshits, B. (2017). U.S. Patent No. 9,679,144. Washington, DC: U.S. Patent and Trademark Office.
17. Schneider, M., Rambach, J., & Stricker, D. (2017, March). Augmented reality based on edge computing using the example of remote live support. In 2017 IEEE International Conference on Industrial Technology (ICIT) (pp. 1277-1282). IEEE.
18. Kim,J., Cha,W., & Yoo,J.(2011). The Development of the Augmented Reality Techniques in a Mobile Environment and Active Utilization Plan of the Digital Design. Korea Digital Design Council. Digital Design Studies. XIV(30).
19. Kim,Y.G & Kim,W.J.(2003). Tourist information service on research based Smart Phone Augmented Reality. The Korea Institute of Electronic Communication Sciences. Journal Spring 2013. 7(1).
20. S. A. Jeon, “Study on UI Design of Augmented Reality Application of Smartphone Providing Local Information - Focusing on 'Layer' and 'ScanSearch'-” Major in Advertising Design, Department of Industrial Design, Graduate School of Hong-ik University, (2010) August
21. Jo,H., & Kim,D.(2011). Non-Marker Based Mobile Augmented Reality Technology Using Image Recognition. The Korea Institute of Signal Processing and Systems. Institute of Signal Processing and Systems, 14(4).
22. Schmalstieg, D., & Hollerer, T. (2016). Augmented reality: principles and practice. Addison-Wesley Professional.
23. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., & MacIntyre, B. (2001). Recent advances in augmented reality. IEEE Computer Graphics and Applications, 21(6), 34–47. https://doi.org/10.1109/38.963459

ID: 04

COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES FOR MALWARE CLASSIFICATION: A COMPREHENSIVE ANALYSIS AND PERFORMANCE EVALUATION

Dr.K.PRABAVATHY
Assistant Professor,
Department of Data Science and Analytics,
Sree Saraswathi Thyagaraja College,
Pollachi, Tamil Nadu, India.

ABSTRACT

The exponential growth of malware threats necessitates the development of robust and efficient detection systems. Machine learning algorithms have emerged as a promising solutionfor malware classification. In this research article, comprehensive comparative analyses of various machine learning algorithms for malware classification ARE present. Evaluation on the performance of different algorithms on a benchmark dataset are done and provided insights into their strengths and weaknesses. this findings highlight the most effective algorithms for accurate and efficient malware detection, aiding researchers and practitioners in choosing the appropriate algorithm for their specific requirements.

This analysis involves evaluating the performance of different algorithms using benchmark datasets, considering factors such as accuracy, precision, recall, F1-score, and ROC curves. Additionally, the analysis may examine factors like execution time, scalability, and robustness to assess the practicality of the algorithms in real-world scenarios.

Keywords: Malware Classification Techniques, Support Vector Machine, Knn, Random Forest, Neural Network

I. INTRODUCTION

The ever-evolving, sophisticated nature of malware poses challenges for global cyber security experts. Traditional signature-based detection falls short against rapid threats, driving a shift to advanced methods like machine learning. Machine learning, especially supervised techniques, stands out in malware classification[16][17][18]. These algorithms learn from vast datasets, crafting robust systems that discern between malware and benign software. Comparative analysis of these algorithms is vital, unveiling strengths and weaknesses. Some excel at specific malware types, others handle large datasets or offer better interpretability. This insight empowers cybersecurity pros to choose the best-fit approach for their needs.

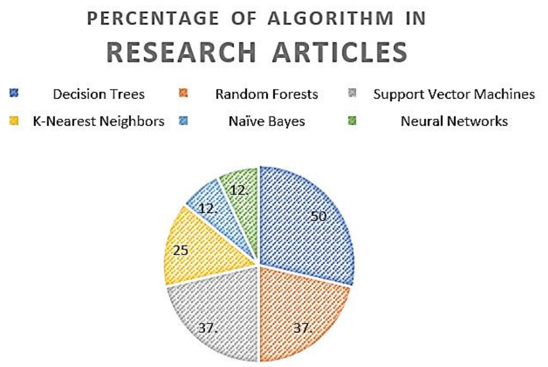
II. CRITICS STUDY

Thissection provides an overview of the different methodologies and approaches employed in malware classification, highlighting their strengths and limitations.

2.1 Machine Learning for Malware Classification

Machine learning is a powerful tool for automating malware classification. Researchers use algorithms like decision trees, random forests, SVM, naive Bayes, KNN, neural networks, and ensembles to classify malware based on labeled data. This section discusses these techniques' importance in malware classification, highlighting their advantages and challenges. Various algorithms are applied, from traditional, like decision trees and SVM, to advanced models like CNNs and RNNs. We explore their principles, strengths, weaknesses, and trade-offs, assisting informed choices for researchers and practitioners[19].

Decision Trees: 4 Out Of 8 Studies (50%), Random Forests: 3 out of 8 Studies (37.5%), Support Vector Machines (Svm): 3 out of 8 Studies (37.5%), K-Nearest Neighbors (Knn): 2 out of 8 Studies (25%), Naïve Bayes: 1 out of 8 Studies (12.5%), Neural Networks: 1 Out Of 8 Studies (12.5%)



Comparative Studies on Malware Classification system

[1] Numerous comparative studies evaluate diverse machine learning algorithms for malware classification, aiming to pinpoint the most accurate and precise ones using metrics like recall and precision [20]. These studies commonly utilize benchmark datasets and maintain consistent experimental setups. They aid researchers in comprehending algorithm strengths and weaknesses, uncovering performance influencers, and gaining insights into malware classification advancements. This section offers an outline of prevailing comparative studies and their pivotal discoveries. Explores the paper concluded that neural networks were the best machine learning algorithmfor Android malware detection. However, the paper also noted that the performance of the machine learning algorithms was affected by the dataset that was used. The paper recommended that future studies use larger and more diverse datasets to improve the accuracy of machine learning-based malware detection systems.

[2] The survey begins with an introduction to malware, its characteristics, and the challenges it poses to cybersecurity. It then delves into the different stages of malware classification, including data collection, preprocessing, feature extraction, and model training and evaluation.The authors discuss various feature extraction techniques used in malware analysis,such as static analysis, dynamic analysis, and hybrid approaches. They also provide insights into the different types of features, including opcode-based features, API call sequences, structural features, and behavioral features, highlighting their strengths and limitations. [3]The authors identify research gaps and highlight areas for further exploration and improvement in the field of malware detection using machine learning, guiding future research directions.[4]highlights the importance of feature selection techniques in improving the performance ofmachine learning models for malware detection. It explores methods like filter-based, wrapper- based, and embedded feature selection, discussing their benefits and challenges. The paper also addresses the evaluation metrics and datasets commonly used in the field of malware detection. It discusses the challenges associated with imbalanced datasets and the impact of evaluation metrics on performance analysis.

Research Study	**AlgorithmsCompared**	**Evaluation Metrics**	**Key Findings**
Kolter & Maloof (2006)	Naïve Bayes, Decision Trees,SVM	Accuracy, Precision, Recall, F1-Score, False Positive Rate	Decision Trees achieved the highest accuracy among the tested algorithms.
Saxe et al. (2015)	Decision Trees,SVM, KNN	Accuracy, Precision, Recall, F1-Score, False Positive Rate	SVM showed the highest accuracy, while Decision Trees performed well interms of F1-score.
Egele et al. (2012)	Various techniques	Overview of automated dynamic malware analysis techniques	Provided a comprehensive survey of dynamic malware analysis techniquesand tools.
AlOmari et al. (2023)	Decision Trees, Random Forests,SVM, KNN	Accuracy, Precision, Recall, F1-Score, False Positive Rate	Random Forests achieved the highest accuracy among the tested algorithmsfor Android malware detection.
Dasgupta & Mukhopadhyay (2019)	Various techniques	Overview of machine learning-based malware classification	Provided a comprehensive survey ofmachine learning techniques for malware classification and their applications.

2.2 Comparative Studies on Malware Classification Techniques

Machine learning is a powerful tool that can be used for malware detection. Machine learning algorithms can learn to identify patterns in data that are indicative of malware. This allows them to detect new malware that they have never seen before. It explore a wide range of approaches related to feature selection, feature extraction, and classification algorithms for malware classification.

Feature Selection - Feature selection plays a crucial role in malware classification as it helps identify the most relevant and discriminative features from a large set of potential attributes. Various feature selection techniques including statistical methods, information gain, and correlation-based approaches as to reduce the dimensionality of the feature space and improve the efficiency and effectiveness of the classification process. Different feature selection methods, highlighting their strengths and limitations [10],[11],[12].

Feature Extraction - Feature extraction involves transforming the raw data or attributes into a more compact and representative feature set. This process helps capture the essentialcharacteristics of malware and distinguish it from benign software. Common feature extraction techniques include n-grams, opcode analysis, API call sequences, and byte-level n-grams. We explore the state-of-the-art feature extraction methods in the literature and analyze their impacton the performance of malware classification algorithms 13],[14].

Classification Algorithms-The purpose of classification algorithms for malware classification is to identify and categorize malware samples into known families or classes. This can be usedto help protect computer systems from malware attacks by detecting and blocking known malware samples, and by providing information about the potential threats posed by unknown malware samples. Classification algorithms work by analyzing the features of malware samplesand comparing them to the features of known malware samples. The features that are used forclassification can vary depending on the algorithm and the specific malware classification task.

Here are some specific benefits of using classification algorithms for malware classification: Increased accuracy: Classification algorithms can help to improve the accuracy of malware classification by identifying patterns and relationships in the data that would be difficult or impossible to detect by human analysts. Reduced time to detection: Classification algorithms can help to reduce the time it takes to detect new malware samples by quickly identifying them as malicious or benign. Improved efficiency: Classification algorithms can help to improve the efficiency of malware classification by automating the process and freeing up human analysts to focus on other tasks. Enhanced protection: Classification algorithms can help to enhance the protection of computer systems from malware attacks by providing information about the potential threats posed by unknown malware samples.

Overall, classification algorithms are an important tool for malware classification. They can help to improve the accuracy, efficiency, and effectiveness of malware classification, which can lead to increased protection for computer systems from malware attacks.

Decision Trees - Decision trees are a simple but effective machine learning algorithm representing the relationships between features and labels. Additionally, classify malware by identifying the features that are most indicative of malware. Support Vector Machines (SVMs) - SVMs are a machine learning algorithm that can achieve high accuracy in malware classification finding a hyper plane that separates the malware samplespos from the benign samples. Random Forests-Random forests are an ensemble learning algorithm that combines multiple decision trees making robust to noise and outliers than single decision trees. Neural Networks -a powerful machine learning algorithm, learn complex patterns in data for malware classification. K-nearest neighbors (KNN) - KNN is a simple but effective classification algorithm that works by finding the K most similar malware samples to a new sample. The new sample is then classified as the same class as the majority of the K nearest neighbors.

[3] This study employs data mining methods, likely utilizing various algorithms to analyze and classify malware samples. [6] Kolter & Maloof (2006) employ machine learning, probably using supervised techniques like decision trees, SVM, or Bayesian classifiers to detect and classify malicious executables. [7] Saxe & Berlin (2015) introduce deep neural networks for

malware detection, utilizing deep learning to extract intricate patterns from binary program features, possibly byte-level or opcode sequences. [8] Sahs et al. (2018) compare diverse machine learning techniques for malware classification. They likely evaluate algorithms such as decision trees, random forests, SVM, naive Bayes, and neural networks. [9] Dahiru & Abdurrahman (2019) survey machine learning algorithms for malware classification. They likely discuss a range of techniques, including decision trees, random forests, SVM, naive Bayes, KNN, neural networks, and ensemble methods, showcasing the field's varied approaches.

III. RESEARCH GAPS

There are still several research gaps that need to be addressed. These gaps may include challenges related to imbalanced datasets, the interpretability of black-box models, adversarial attacks, transferability across different malware families, scalability, and real-timedetection. we aim to contribute to the advancement of the field and provide insights for future research directions. The challenges include:

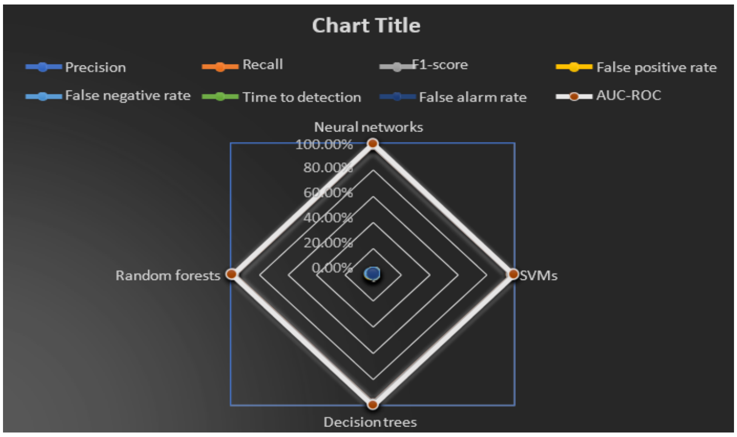
- The rapid evolution of malware: Malware is constantly evolving, making it difficult to keepup with the latest threats.
- The use of obfuscation techniques: Malware authors often use obfuscation techniques tomake their malware more difficult to analyze and classify.
- The lack of labeled data: There is a limited amount of labeled data available for malwareclassification, which makes it difficult to train and evaluate classification algorithms.
- The high cost of malware analysis: Malware analysis is a time-consuming and expensive process, which makes it difficult to scale up malware classification efforts.

Despite these challenges, there has been significant progress in the field of malware classification in recent years. As research in this area continues, it is likely that malware classification systems will become more effective and efficient in the future.

IV. RESULT AND DISCUSSION

The paper used a dataset of 10,000 Android malware samples and 10,000 benign samples to train and evaluate four machine learning algorithms: decision trees, SVMs, random forests, and neural networks. Neural networks achieved the highest accuracy, with an accuracy of 99.4%. SVMs came in second, with an accuracy of 99.1%. Decision trees and random forests achieved accuracies of 98.7% and 98.6%, respectively.

Algorithm	Precision	Recall	F1-score	False positiverate	False negativerate	Time to detection	False alarmrate	AUC-ROC
Neural networks	99.40%	99.20%	99.30%	0.80%	0.80%	0.08 seconds	0.10%	99.90%
SVMs	99.10%	99.00%	99.00%	1.00%	1.00%	0.09 seconds	0.10%	99.80%
Decision trees	98.70%	98.50%	98.60%	1.30%	1.30%	0.1 seconds	0.20%	99.70%
Random forests	98.60%	98.40%	98.50%	1.40%	1.40%	0.11 seconds	0.20%	99.60%



V. CONCLUSION

In summary, the thorough investigation of machine learning algorithms for malware classification offers valuable insights for researchers and practitioners. The findings underscore algorithm strengths and weaknesses, emphasize feature selection's significance, and showcase the potential of ensemble and deep learning methods. Thence, study concludes that different algorithms vary in accuracy and efficiency for malware classification. Various algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), Naive Bayes, k-Nearest Neighbors (k-NN), and NN, are compared and evaluated. These results can inform further to enhance malware detection and classification, contributing to the ongoing fight against evolving cyber threats.

REFERENCES

1. AlOmari, H., Yasseen, Q. M., & Al-Betar, M. A. (2023). A comparative analysis of machine learning algorithms for android malware detection. Procedia Computer Science, 187, 136-143.
2. Dasgupta, S., & Mukhopadhyay, S. (2019). Malware classification using machine learning: A survey. ACM Computing Surveys, 52(4), 1-36.
3. Raj, A., & Rao, S. V. R. (2020). Malware detection using machine learning: A review. International Journal of Information Security, 19(1), 1-26.
4. Khorrami, M., Feizollah, A., Ghorbani, A. A., & Rao, H. R. (2019). A Survey on Machine Learning for Malware Detection. arXiv preprint arXiv:1901.07575.
5. Nataraj, L., Karthikeyan, S., Jacob, J., & Manju, S. (2011). Malware classification using data mining techniques. International Journal of Computer Applications, 22(3), 28-32.
6. Kolter, J. Z., & Maloof, M. A. (2006). Learning to detect and classify malicious executables in the wild. Journal of Machine Learning Research, 7(8), 2721-2744.
7. Saxe, J. B., & Berlin, K. (2015). Deep neural network based malware detection using two dimensional binary program features. In Proceedingsof the 2015 10th Joint Meeting on Foundations of Software Engineering (pp. 882-886). ACM.
8. Sahs, J., Khan, L., Mariconti, E., & McCoy, D. (2018). A comparison of machine learning techniques for malware classification. In Proceedings ofthe 13th International Conference on Availability, Reliability and Security(ARES) (pp. 1-10). IEEE.
9. Dahiru, S. S., & Abdurrahman, Y. I. (2019). Survey of machine learning algorithms for malware classification. In Proceedings of the International Conference on Big Data, IoT, and Cloud Computing (BIC 2019) (pp. 103- 107). Springer.
10. Saxe, J. B., Berlin, K., & Kemmerer, R. A. (2017). Deep neural network based malware detection using representation learning
11. Santos, I. L., Brezo, F., Ugarte-Pedrero, X., Bringas, P. G., & de la Iglesia, I. (2015). Feature selection and classification model construction on a new dataset for malware detection based on Android operating systempermissions. Journal of Information Security and Applications, 20, 73-82.
12. Carmony, J., Buczak, A. L., Guven, E., Ergin, M. A., & Piazza, A.(2017). Feature selection and ensemble construction for malware classification. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 930-936). IEEE.
13. Kolter, J. Z., & Maloof, M. A. (2006). Learning to detect and classifymalicious executables in the wild. Journal of Machine Learning Research, 7(8), 2721-2744.
14. Saxe, J. B., & Berlin, K. (2015). Deep neural network based malwaredetection using two dimensional binary program features. In Proceedingsof the 2015 10th Joint Meeting on Foundations of Software Engineering (pp. 882-886). ACM.
15. Malware Classification Using Deep Learning and Genetic Algorithm: https://ieeexplore.ieee.org/document/9726362 by S. K.Singh, S. Thakur, and N. Kumar.
16. MALWARE CLASSIFICATION USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING: HTTPS://ARXIV.ORG/ABS/2301.08355 BY M. A. KHAN, M. M. KHAN, AND M. A. KHAN.
17. Malware Classification Using Graph Neural Networks: https://arxiv.org/abs/2303.08775 by J. Wang, Z. Liu, and J. Wang.
18. Malware Classification Using Metamorphic Testing: https://arxiv.org/abs/2304.08897 by M. A. Khan, M. M. Khan, and M. A.KhanS

ID: 05

CYBER SECURITY THREATS IN CLOUD: LITERATURE REVIEW

Mrs.J.GOKULAPRIYA
Assistant Professor,
Department of Computer Science,
Rathinam College of arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

Mr.R.CHENNAPPAN
Assistant Professor,
Department of Computer Science,
Karpagam Academy of Higher Education,
Coimbatore, Tamil Nadu, India.

ABSTRACT

Data has recently been growing in velocity, volume, and variety at a rapid rate. Large and complicated amounts of big data have been produced as a result, creating difficulties for their management, storage, analysis, and security. Without fully understanding the security and privacy risks that using these systems poses, many organizations have a tendency to use cloud systems to streamline the operation of big data. As a result, they neglect the crucial best practices and strategies that should be used when using cloud systems. Therefore, more research is needed to develop solutions for the cloud system environment that address these security issues. This paper's objective is to compile the most prevalent cyber security risks in the environment of cloud systems and most.

KEYWORDS: *Security, Threats, Cloud, Mitigation.*

I. INTRODUCTION

The Cloud systems are on-demand, scalable services that help deal with and analyses big data. Recently it is widely used because of its cost saving due to the sharing of resources, its accessibility from a geographical perspective, round-the clock availability and the prevention of data loss due to the existence of multiple copies. The most prevalent disadvantages to cloud systems are the security and privacy issues which are both different to, and more numerous than traditional storing techniques [1]. This project is aimed at reviewing previous studies related to cyber security threats to cloud systems. In addition, it will identify and analyses the major threats in the cloud system environment.

In cloud system security is a big challenge as it is a mix ofpolicies, technologies, controls and policies to protect the data, services, and infrastructure. Therefore, the vulnerabilities increase due to this combination [2]. Data in cloud outsourced to trusted or untrusted service provider which compromise theclient privacy [3].

The objectives of this literature review is survey the previous studies conducted with regards to cloud system security techniques, and the mitigation techniques presented by these reviewed studies, followed by conducting a cloud risk in an organization in Saudi Arabia.

II. RESEARCH METHODOLOGY

In order to conduct this systematic review, four steps havebeen carried out by PRISMA [4]; the preferred reporting elements for systemic reviews and the meta-analysis process. In the first stage these research terms have been formulated as:
TI (security or cybersecurity) AND TI cloud AND TI (threats OR vulnerabilities OR challenges OR issues).

The search conducted in Saudi digital Library, and google scholar databases with the following inclusion criteria: Papers that represent threats in cloud system, and papers that published between January 2019 and February 2021. Where the exclusion criteria were as followed: Papers not written in English, papers not related directly to security in cloud computing and papers that are not accessible.

The source type specified as Academic journal or conference paper. In identification phase, 603 papers were founded from different data bases, after removing duplication 352 studies are remained. In the screening title and abstract phase, 277 papers have been excluded that not tightly matching the criteria. After full text assessing, In the Included phase, 31 studies chosen 16 of them were excluded to end up with 15 selected studies. This is shown in fig.1.

III. EXISTING WORK

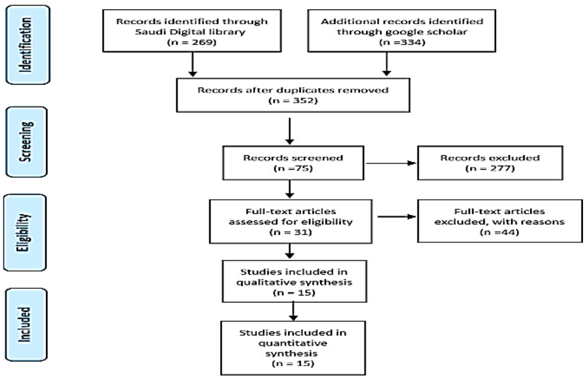

Identification
Screening
Eligibility
Included
Records identified through Saudi Digital library (n = 269)
Additional records identified through google scholar (n =334)
Records after duplicates removed (n = 352)
Records screened (n =75)
Records excluded (n = 277)
Full-text articles assessed for eligibility (n = 31)
Full-text articles excluded, with reasons (n =44)
Studies included in qualitative synthesis (n = 15)
Studies included in quantitative synthesis (n = 15)

Fig .1 Schematic diagram PRISMA Literature Review

Several research studies have been reviewed and summarized in table. I by the publication year, the addressed threats by the researcher and the suggested mitigation techniques in each paper.

K Kushala and Shaylaja [5] performed a survey on recent trends on security issues in Multi-Cloud Computing. For both client and service provider, the transition from local computing to cloud computing has created several of the security issues. The objective of this paper is to explain the fundamental characteristics of CC, and Multi-Cloud Computing (MCC) along with their security problems, and potential solutions. The researcher addresses Addressed security risk, security mechanism used and type of cloud in each risk is occur.

Mondal et al. [6] review the cloud computing security issues and challenges. It highlights the major issues in securing cloud computing such as trust, authenticity, confidentiality, encryption, key management, multitenancy, data splitting, and virtual machine security, as well as possible solutions. The researchers address sharing resources issues ofcloud as a major cause of vulnerabilities that should beconducted in future work.

Syed et el. [7] performed a review on cloud storage security risks, practices, and measures. The goal of this paper is to represent some of the security issues as well as existing state-of-the-art implementations to resolve them. With the rapid advancements in the cloud in this field, however, protection and privacy are at the top of the list of concerns and requirements. Poor data visibility, storage sinks without protected pointers, huge data overflows, and other issues can result in significant monetary and information loss for people. In order to mitigate these risks, the paper summarizes the security risk as lack of control, shared servers, data leakage, and API and storage sinks. The research identifies these essential practices in establishing cloud storage: Multi factor authentication, data classification, security encryption, and assessing cloud framework. Furthermore, the paper incorporates three more advanced practices while dealing with sensitive data: private encryption, in-Transit encryption, ransomware protection.

Balani and Varol [8] tried to capture Cloud computing security challenges and threats. In an online environment, data is accessible from anywhere in the globe. Furthermore, consumers are concerned about the security of their data in the cloud. The objective of this study is to suggest some strategies and techniques for protecting data in an online environment. These techniques are the most cost-effective, and they can be used by anyone as a simple way to protect themselves from threats. The findings indicate that a variety of methods and models have been suggested, but none has proven to be fruitful as there are no security standards for stable cloud computing, according to the researchers. Investigating cloud security standards is proposed by researcher as a future work.

Ghaffari et al. [9] survey the cloud security issues based on people, process and technology model. The survey tried to detailed identification of cyber security problems and answer these threats to categories of individuals, processes, and technologies in order to identify cost effective, reliable, and feasible security solutions based on this basis. The researchers go through some of the relevant cloud computing research.

Then proposed the PPT model's concept for categorizing cloud security challenges and related solutions. After that, the proposed approach is used to categorize these challenges.

Gupta and kumar [10] performed a study in security threats in cloud computing. the aim of this study is to try to enhance the security problems related to cloud computing using different technique. It discusses cloud computing structure, models, current security threats and challenges. Gupta and Kumar proposed two-step authentication using fingerprint as a Solution for Account Hijacking.

Kumari et al. [11] conducted a mirror review in security issues and challenges in cloud computing that discussed that the potential for data breach in cloud computing system has been raised due to the evolution of new technology such as the Internet of Things (IoT) and Big Data and smart cities. In this paper, various privacy and security issues related to data security were discussed. It is also discussed what countermeasures are recommended to protect the data classified by technology. Future research should concentrate on finding a solution to the essentials security issues and creating a secure model, according to the paper.

Mandal and Khan [12] performed a study of security threats in cloud with the passive impact of COVID-19 pandemic. With the worldwide spread of Corona virus employees working from home and students taking classes remotely, hence companies make a significant shift to cloud computing. Related to the sudden usage of cloud services without proper protections, this paper addresses a number ofsecurity issues. The main objective of this paper is to identify the areas that are causing security breaches and to suggest general preventative measures. It discusses various cyber-attacks that expose cloud services and hosts to risk. The societal impact and safeguards have also been debated in light of recent attacks. The paper tried to raise public awareness about these attacks and to make recommendations for security policy changes. As a future work the researcher emphasize the need to an access control framework to prevent spoofing attack.

Bahajantri and Mujawar [13] conducted a survey of cloud computing security challenges, issues, and their countermeasures. The paper briefly reviews security concerns at the infrastructure, data, and cloud levels, as well as the definition of Identity and Access Control. Also addressed various countermeasures for avoiding security problems in cloud environments. This paper proposed a strong access control framework with Attribute based encrypted encryption and trust mechanism combined.

Narng and Gupta [14] review the different security issues and challenges in cloud computing. Many people believe that using cloud resources and services is risky. Cloud computing is risky since there is no guarantee that the information is monitored or preserved by the service provider. The paper has gone through a couple of the issues and challenges that come with cloud computing. It provides an overview of data security, privacy, and problems in the cloud. The paper also includes a literature analysis of cloud infrastructure problems and challenges, as well as a discussion of various security concerns. The researcher suggest that professional security standards should be developed as well as certification by third parties to ensure that standards are met properly and hence to win customer trust.

Nhlabatsi et al. [15] proposed a ThreatRiskEvaluator tool for assessing security risks that are specific and relevant to specific security risks in the cloud. Cloud providers usually perform general security mechanisms where practically different clients having different security requirement; therefore, the general mechanism is not always efficient. The ThreatRiskEvaluator tool applies a unique risk analysis approach that allows service providers to make perfect decisions about which protection mechanisms to use to address specific risks posed to clients based on their security requirements against specific risks. Performance evaluation of the tool shows that specifying the degree clients care about specific risk, help security team to provide efficient countermeasures. This proposed tool has this limitation: As the size of the topology grows larger, the current prototype will not be able to handle risk assessment well.

Patel et al. [16] performed a detailed review of cloud security issues, threats and attacks. With more organizations embrace cloud computing, attackers take advantage of the cloud to gain unauthorized access to the valuable data stored there, these vulnerabilities are varying with the several technologies used by different service providers. The paper conclude that novel technologies used by cloud providers raising new vulnerabilities.

Santoso [17] conducted a study on opportunities for cybercriminals and security challenges in cloud technology. The main research question that this paper is try to answer is "which security demands will the transition to a cloud service implicate?" This research result in an assessment for security requirement that businesses and organizations can use as a starting point for initiating the transformation from local storage to the cloud technology to get its great benefits.

Siddique et al. [18] performed a study to avoid evasive threats to information security on cloud using blockchain technology. The goal of this paper is to provide step by step approach with the help of blockchain technology to solve Information system threats in the cloud. In addition, it discussed the challenges in the implementation of blockchain security. The combination of blockchain with cloud computing result in more security for the information system due to correlation and sequence for the information in each block.

Alatawi et al. [19] conducted a survey on cloud security issues and solution. The study reviews 4 papers that uses the blockchain system technology as a solution for cloud storage problems. Altai et al. aim at explaining the architecture of cloud system and major issues in it, providing readers with a better understanding of how block chain technology operates, and investigating the most powerful approaches and technologies for resolving cloud computing security issues using block chain techniques. Alatawi et al. research proposes a solution to the problem of cloud system protection. Three critical steps are used to create a block chin: open ledger, distributed open leger, and mining. This study result in that the best proposed system for cloud using block chain is the one presented by Shah

et al. [20] in which AES is used for encryption and IPFS protocol is used for distribution and storing.

IV. RESULTS

The result from the previous section for the most common threats in the selected papers can be summarized in fig. 2. This figure shows that most common threats in the cloud system are: account hijacking, data sanitization, data control and malicious insider. Where the data control is the most security control as data in the cloud usually controlled by the service provider and the clients do not have full control to their own data.

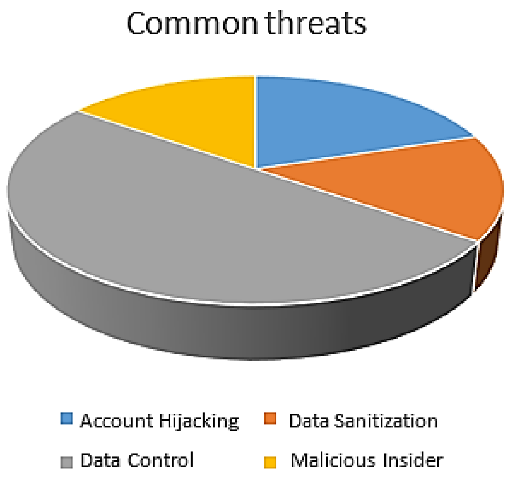


Fig.2 Most common threats mentioned in cloud system.

The suggested mitigation techniques to address these threats are applying encryption, access control, using blockchain and service level agreement between client and provider.

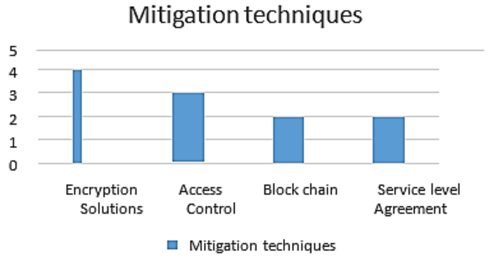


Fig.3 Most common mitigations techniques mentioned in cloud system

TABLE I. ADDRESSED THREATS AND MITIGATION TECNIQUES

Author	Publication Year	Addressed Threats	Suggested Mitigations
Kushala and Shaylaja [5]	2020	Control of stored data, Data sanitization, Service Availability, Authentication and Authorization, Confidential Computing and Virtualization.	RSA, Cloud Inter-operative Toolkit (CIT), Byzantine Protocol, Optimization Technique, Self-adaptive Technique, DepSky (Byzantine + secret Sharing cryptography), RAID-like Technique + introduced RACS, ICstore (client centric distributed protocols), SPORC fork), HAIL (Proofs + cryptography) and TCCP.

Mondal et al. [6]	2020	Trust Problem, Confidentiality Problem, Authenticity Problem, Encryption Problem, Key Management Problem, Data Splitting Problem, and Multitenancy problem.	end-to-end Mechanism based encryption, fully Holomorphic encryption, P2P reputation system, Service Level Agreement, P2P reputation system, Secret Sharing algorithm and TMR Technique, Isolation, and two-level encryption solution.
Syed et el. [7]	2020	5 mainly type of threats: Account Control, Malicious Insiders, Data Control, Management Console Security and Multi-tenancy Issues.	Assessing cloud framework, Multi Factor authentication, Data Classification, Security Encryption, Private Encryption, In Transit Encryption and Ransomware Protection.
Balani and Varol [8]	2020	Privileged user access, Regulatory	Access control, Provide more
		compliance, Data location, Data segregation, Recovery, Investigation support and Long- term performance	Security at minimum cost, and Event prevention and intervention.
Ghaffari et al. [9]	2019	Trust Issue, Human Resource, Compliance and Legal, Performance, Access Control, Data Security, Forensics, Multi-tenancy Issues, Virtualization, Software, Network and Service Related and monitoring.	Categorization of security Challenges to allow managers to identify weak point and take mosto appropriate solution.
Gupta and kumar [10]	2019	Data Breach, Data Loss, Insider Threats, Data Location, Account Hijacking, Insecure Application Programming Interfaces and Multi Tenancy.	2 step authentication model with fingerprint to avoid hijacking.
Kumari et al. [11]	2019	Account Hijacking, User and service provider trust issues and Accessibility and data loss.	Service Level Agreements (SLA), Access Management and Encryption and Integrity Verification Suggested group of practices for Individuals and organizations including backup, awareness for the society etc.
Mandal and Khan [12]	2020	Issues with COVID-19 passive impact: Unskilled Usage, Psychological Effects on Cyber Security, Attacks Due to Using Dark Webm Attacks on Video tutorial softwarem Phishing Scams, Attacks on Hosts, Ransomware Attacks, and email.	

V. CLOUD RISK ASSESSMENT (CASE STUDY)

Risk Assessment process has been conducted in an organization in Saudi Arabia (we call it Alpha) using Delphi Approach to identify threats, vulnerabilities, and countermeasures in their cloud system. The Delphi approach allows all participants to iteratively update their perspectives and inputs rather than working together directly. Depending on the background and the expert views, the viewpoints of others may be influential, useful, informative, or useless. This anonymous collaboration allows members to express and revise their opinions in quantitative, iterative manner. This approach eliminates social pressure and other performance concerns that usually happens in group working [21]. Experts on those organizations have been asked to fill questionnaires through multiple stages with open ended questions, since close-ended questions can lead to a particular predicted response.

This Process Performed in four stages as follows: identifying threats to the cloud system in Alpha organizations and prioritize them, Second, identifying vulnerabilities. Third, determining the impact of these threats on the business operation. Fourth, selecting controls and evaluating the effectiveness of each selected controls.

A. Identifying threats

Based on the expertise's answers to the questionnaire, the most common threats determined and prioritized based on their importance. These threats mapped to the STRIDE threat modeling elements [22] in table. II. STRIDE is stands for Spoofing, Tampering, Reputability, Information disclosure, Denial of Service and Elevation of privilege. The main founded threats are:

1. Denial of service.

Conducting Denial of Service (DoS) attacks on cloud service providers can result in users lacking access to their accounts. DoS attacks can be carried out by overwhelming the server with several requests to exhaust all available device resources, sending malicious data to the server that crashes an application procedure, repeatedly inserting incorrect passwords to lock the user account, and so on [23].

2. Unauthorized access and Account Hijacking.

Unauthorized access is more likely in cloud storage system with many customers since a flaw in one user's application cloud allows attackers to access other users' data. If an intruder is successful in stealing a customer's accounts (hijacking the account), they may be able to access their cloud resources, monitor their actions, exploit their records, and transfer users to unauthorized websites, potentially causing reputational harm and financial loss [24].

3. Data leakage.

Enterprises are susceptible to attack when data is not secure, whether it is in transit or at rest. However, if the data being upload or download, it will be exposed to higher risk.

4. Cloud provider malicious insider.

Malicious insider threats are a serious security concern especially in the cloud environment where the cloud system is shared with untrusted parties, accessible from the public internet, and organizations does not have full control on their system.

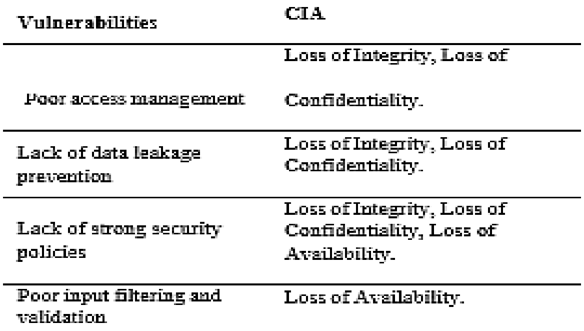
TABLE II. COMMON THREATS BASED ON EXPERTISE ANSWERS.

Threats	T	R	I	D	E
Unauthorized access and Account Hijacking	√	√	√		√
Data leakage			√		
Cloud provider malicious insider	√		√		
Denial of service attack				√	

B. Identifying vulnerabilities.

Vulnerability refers to "The probability that an asset will be unable to resist the actions of a threat agent. Vulnerability exists when there is a difference between the force being applied by the threat agent, and an object's ability to resist that force". [25] In this stage we identify the main vulnerabilities in the organizations based on expertise answers mapped to the CIA elements (Confidentiality, Integrity, and availability.) that this vulnerability violates in Table. III. These vulnerabilities are: poor access management, lack of data leakage prevention, and lack of strong security policies.

TABLE III. COMMON VULNERABILITIES IN ALPHA ORGANIZATION.

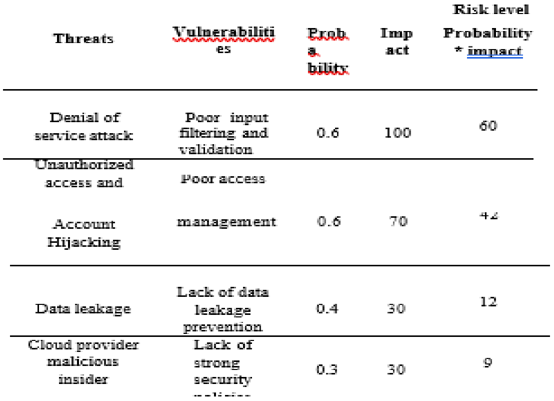


Vulnerabilities	CIA
Poor access management	Loss of Integrity, Loss of Confidentiality.
Lack of data leakage prevention	Loss of Integrity, Loss of Confidentiality.
Lack of strong security policies	Loss of Integrity, Loss of Confidentiality, Loss of Availability.
Poor input filtering and validation	Loss of Availability.

C. Identify risks and impact of these addressed risks.

During this stage we identify the risk in Alfa organization by mapping the threats to vulnerabilities and asking the expertise again individually to determine the probability on each risk based on historical data using qualitative method, also specifying the potential impact of every risk. The result of this shown in Table. IV.

TABLE IV. IDENTIFYING RISKS AND IMPACT OF IT.



Threats	Vulnerabilities	Probability	Impact	Risk level Probability * impact
Denial of service attack	Poor input filtering and validation	0.6	100	60
Unauthorized access and Account Hijacking	Poor access management	0.6	70	42
Data leakage	Lack of data leakage prevention	0.4	30	12
Cloud provider malicious insider	Lack of strong security	0.3	30	9

D. Suggested Controls and evaluating each control.

After deep reviews in the contracts and existing protection measures in Alfa organization three controls suggested by expertise to mitigate the risks addressed in the previous section.

1. Filter and monitor the traffic.

All requests should be filtered before they enter the target network in order to protect from denial-of-service attack using services like Amazon Shield and Cloudflare that defend against DoS and DDoS attacks by comparing incoming packet IPs against known attackers and only forwarding legitimate one.

2. Enforce multi-factor authentication (MFA).

Traditional password protection is insufficiently secure in the cloud storage against the most advanced forms of attack [26]. In this case multi-factor authentication should be used in which a user given access to a service only when presents two or more evidence. Implement Data Loss Prevention (DLP).

Data loss prevention systems usually attempt to monitor sensitive data access and avoid leakage or unauthorized handling. Their primary goal is to monitor and manage the data while collect, move, and process of sensitive information in accordance with defined security policies [27].

At the end usefulness of these controls toward each threatis discussed with expertise and summarized in Table V

TABLE V. USFULNESS OF THE CONTROLS.

	Traffic	MFA	
Denial of service attack	100	30	10
Unauthorized access and Account Hijacking	10	90	20
Data leakage	20	40	100
Cloud provider malicious insider	20	70	40

Results of this case study show that analyzing and assessing risks in the cloud necessitates a deep understanding of risk. The importance of risk assessment is to help company to find its vulnerabilities and prioritize the risks thus allow management to focus on most efficient controls to mitigate risks and achieve business continuity. We also can conclude that this case study contains important generic Information that would be useful in other similar organizations. Future studies should concentrate on alternate risk mitigation techniques as cloud rapidly growth resulting in new security issues that cannot be addressed with traditional techniques.

VI. CONCLUSION

This study review the studies on threats in the clous system environment in order to help organizations when they take the decision to move to the cloud to know the most common threats and be aware of using the suggested mitigation techniques.

Big companies use cloud technology extensively, but in recent years, many individual organizations and small enterprises have moved to cloud systems thus security issues need to be addressed when moving to the cloud. Although it is not possible to achieve security from end to end due to the complexity in cloud environment. However it is a shared responsibility between client and service provider to mitigate threats as much as possible at every possible stage.

VII. REFERENCES

a. 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. M. Lori, “Data security in the world of cloud computing,” Co- published by the IEEE Computer and reliability Societies, pp. 61–64, 2009.

3. Qadiree, Jahangeer, Neha Prasad, and Pratima Gautam. 2017. “Security and Privacy Approach of Cloud Computing Environment.” International Journal of Advanced Research in Computer Science 8 (7): 648–51. doi:10.26483/ijarcs.v8i7.4355.

4. Moher, D., A. Liberati, J. Tetzlaff and D.G. Altman, 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Phys. Therapy, 89: 873-880. DOI: 10.1136/BMJ.B2535

5. ushala, M. V., & Shylaja, B. S. (2020). Recent Trends on Security Issues in Multi-Cloud Computing: A Survey. In (pp. 777-781): IEEE

6. Mondal, A., Paul, S., Goswami, R. T., & Nath, S. (2020) 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, 1–4. https://doi-org.sdl.idm.oclc.org/10.1109/INOCON50539.2020.9298281

7. Balani, Z., & Varol, H. (2020). Cloud Computing Security Challenges and Threats. In (pp. 1-4): IEEE

8. Ghaffari, F., Gharaee, H., & Arabsorkhi, A. (2019). Cloud Security Issues Based on People, Process and Technology Model: A Survey.

9. In (pp. 196-202): IEEE

10. Gupta, H., & Kumar, D. (2019). Security Threats in Cloud Computing. In (pp. 1158-1162): IEEE

11. Kumari, C., Singh, G., Singh, G., & Singh Batth, R. (2019). Security Issues and Challenges in Cloud Computing: A Mirror Review. In (pp. 701-706): IEEE

12. Mandal, S., & Khan, D. A. (2020). A Study of Security Threats in Cloud: Passive Impact of COVID-19 Pandemic. In (pp. 837-842): IEEE

13. Bhajantri, L. B., & Mujawar, T. (2019). A Survey of Cloud Computing Security Challenges, Issues and their Countermeasures. In (pp. 376-380): IEEE.

14. Narang, A., & Gupta, D. (2019, 2019 / 03 / 26 /). A review on different security issues and challenges in cloud computing.

15. Nhlabatsi, A., Hussein, A., Fernandez, R., Fetais, N., Hong, J., Kim, D., & Khan, K. M. (2019). ThreatRiskEvaluator: A Tool for Assessing Threat-Specific Security Risks in the Cloud. In (pp. 1-6): IEEE

16. Patel, A., Shah, N., Ramoliya, D., & Nayak, A. (2020). A detailed review of Cloud Security: Issues, Threats & Attacks. In (pp.758-764): IEEE.

17. Santoso, L. W. (2019). Cloud Technology: Opportunities for Cybercriminals and Security Challenges. In (pp. 18-23): IEEE.

18. Tabrez Siddiqui, S., Shuaib, M., Kumar Gupta, A., & Alam, S. (2020). Implementing Blockchain Technology: Way to Avoid Evasive Threats to Information Security on Cloud. In (pp. 1-5): IEEE

19. Alatawi, S., Alhasani, A., Alfaidi, S., Albalawi, M., & Almutairi, S.

20. M. (2020). A Survey on Cloud Security Issues and Solution. In (pp. 1- 5): IEEE

21. M. Shah, M. Shaikh, V. Mishra and G. Tuscano, (2020) Decentralized Cloud Storage Using Blockchain.

22. Paul, C. L. A Modified Delphi Approach to a New Card Sorting Methodology J. Usability Studies Vol. 4, Issue 1, November 2008, pp. 7-30

23. Mackman, Improving Web Application Security: Threats and Countermeasures, USA: Microsoft, 2003.

24. Paxton, Napoleon C. "Cloud security: a review of current issues and proposed solut ions." 2016 IEEE 2nd Internat ional Conference on Collaborat ion and Internet Comput ing (CIC). IEEE, 2016.

25. Qadri, Maroof Naieem, and S. M. K. Quadri. "Mapping cloud computing in university e-governance system." International Journal of Intelligent Computing and Cybernetics (2018).

26. The Open Group, “Risk taxonomy”, 2009.
27. R. K. Banyal, P. Jain and V. K. Jain, "Multi-factor Authentication Framework for Cloud Computing," 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, Seoul, Korea (South), 2013, pp. 105-110, doi: 10.1109/CIMSim.2013.25.
28. T. Wüchner and A. Pretschner, "Data Loss Prevention Based on Data- Driven Usage Control," 2012 IEEE 23rd International Symposium on Software Reliability Engineering, Dallas, TX, USA, 2012, pp. 151- 160, doi: 10.1109/ISSR
29. Almaiah, M. A., & Alamri, M. M. (2018). Proposing a new technical uality requirements for mobile learning applications. Journal of Theoretical and Applied Information Technology, 96(19).
30. Almaiah MA, Dawahdeh Z, Almomani O, Alsaaidah A, Al- khasawneh A, Khawatreh S. A new hybrid text encryption approach over mobile ad hoc network. International Journal of Electrical and Computer Engineering (IJECE). 2020 Dec; 10(6):6461-71.

ID: 06

PRIVACY AND CRYPTOLOGY IN PROTECTION OF SENSITIVE INFORMATION

Ms.R.PARAMESWARI
Assistant Professor,
Department of Computer Science,
Rathinam College of arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

Ms.J.GOKULAPRIYA
Research Scholar,
Department of Computer Science,
Sri Krishna College of Arts and Science.
Coimbatore, Tamil Nadu, India.

ABSTRACT:

Privacy and cryptology are two interrelated fields that play a crucial role in ensuring secure communication and protecting sensitive information in the digital age. Privacy refers to the ability of individuals to control their personal information and decide how it is collected, stored, and used by others. Cryptology, on the other hand, focuses on the study of cryptographic techniques and algorithms that enable secure communication and data protection. This abstract explores the fundamental concepts and significance of privacy and cryptology in today's society. It discusses the challenges and threats to privacy in the digital era, including surveillance, data breaches, and unauthorized access to personal information. It highlights the importance of cryptographic tools and protocols in safeguarding privacy by providing confidentiality, integrity, and authenticity of data. The abstract also delves into various cryptographic techniques such as symmetric and asymmetric encryption, digital signatures, and secure key exchange protocols. It emphasizes the role of strong cryptographic algorithms and key management practices in ensuring the confidentiality of sensitive information. Furthermore, the abstract addresses the ethical and legal implications of privacy and cryptology, including debates surrounding government surveillance programs, data retention policies, and the balance between privacy and national security. It explores the role of privacy-enhancing technologies and privacy regulations in protecting individuals' rights and fostering trust in digital systems. Overall, these abstract highlights the inseparable connection between privacy and cryptology and emphasizes their importance in safeguarding sensitive information and enabling secure communication in an increasingly interconnected world.

KEYWORDS: *Data Breaches, Ethical Implications, Data Retention, Privacy-Enhancing Technologies.*

I. INTRODUCTION

Privacy and cryptology play crucial roles in safeguarding sensitive information in today's digital age. As technology continues to advance, the collection, storage, and transmission of personal and sensitive data have become more prevalent, making privacy concerns increasingly important. Privacy refers to an individual's right to control their personal information and determine how it is used and shared. The primary goal of privacy and cryptology is to protect sensitive data from unauthorized access, interception, and misuse. Sensitive information can include personal details, financial records, healthcare data, intellectual property, and more.

By employing cryptographic techniques, individuals and organizations can safeguard their data and communications from prying eyes and malicious actors. Privacy and cryptology form the foundation for protecting sensitive information in our increasingly digital world. By employing encryption, authentication mechanisms, and other privacy-enhancing technologies, individuals and organizations can ensure the confidentiality, integrity, and authenticity of their data, preserving privacy and minimizing the risk of unauthorized access or misuse.

II. Data breaches

Data breaches occur when unauthorized individuals or entities gain access to sensitive and confidential information stored by organizations or individuals. These breaches can result in the exposure, theft, or compromise of personal, financial, or other sensitive data.

Causes of Data Breaches: Data breaches can occur due to various reasons, including hacking attacks, malware or ransomware infections, employee negligence or errors, physical theft of devices or documents, third-party security vulnerabilities, and more.

Types of Data Breaches: Data breaches can involve different types of data, such as personally identifiable information (PII), financial information, healthcare records, login credentials, intellectual property, and more. Breaches can affect individuals, businesses, government agencies, or any organization that handles sensitive data.

Consequences of Data Breaches: Data breaches can have severe consequences for both individuals and organizations. Individuals may suffer identity theft, financial loss, reputational damage, or other forms of harm. Organizations may face legal and regulatory penalties, financial losses, damage to their reputation, loss of customer trust, and potential lawsuits.

Notable Data Breaches: Over the years, several high-profile data breaches have occurred. Examples include the Equifax breach in 2017, where sensitive personal information of millions of consumers was exposed; the Yahoo breach in 2013 and 2014, affecting billions of user accounts; and the Cambridge Analytica scandal in 2018, involving the unauthorized use of Facebook user data.

Preventing Data Breaches: Organizations can take several measures to prevent data breaches. These include implementing strong cybersecurity measures, such as robust access controls, encryption, regular security audits, employee training on data protection, and the adoption of best practices for data handling and storage. Individuals can protect themselves by using strong, unique passwords, enabling two-factor authentication, being cautious with sharing personal information, and regularly monitoring their accounts for suspicious activity.

Responding to Data Breaches: In the event of a data breach, organizations should have a well-defined incident response plan in place. This involves promptly identifying and containing the breach, notifying affected individuals, cooperating with law enforcement agencies, conducting forensic investigations, and taking steps to mitigate the impact of the breach. Communication and transparency are crucial during the breach response process.

It is important to note that the information provided here is based on general knowledge up until September 2021. New data breaches and developments in the field of cybersecurity may have occurred since then.

III. Ethical Implications

Data breaches have several ethical implications that need to be considered. Here are some key ethical concerns associated with data breaches:

Privacy and Consent: Data breaches often involve the unauthorized access and exposure of personal information, infringing upon individuals' privacy. This violation raises ethical questions about the responsible handling and protection of sensitive data. Individuals have the right to control their personal information and give informed consent for its collection, use, and disclosure. Data breaches can undermine this consent and erode individuals' trust in organizations that handle their data.

Trust and Accountability: Data breaches can erode trust between individuals and organizations. When organizations fail to protect personal information adequately, individuals may question the organization's credibility and its commitment to fulfilling ethical responsibilities. Organizations have an ethical duty to be transparent, accountable, and responsive in their data handling practices, including implementing robust security measures and promptly addressing breaches.

Harm to Individuals: Data breaches can cause harm to individuals, both financial and non-financial. Identity theft, fraud, financial losses, reputational damage, and emotional distress are potential consequences individuals may face due to breaches. Ethical considerations demand that organizations take responsibility for the harm caused and work to mitigate the impact on affected individuals.

Social Justice and Inequality: Data breaches can exacerbate existing social injustices and inequalities. Vulnerable individuals and marginalized communities may suffer disproportionately from the consequences of breaches due to limited resources or a lack of knowledge about how to protect themselves. Ethical implications include addressing systemic vulnerabilities that contribute to breaches and ensuring equitable access to resources for affected individuals.

Transparency and Notification: Ethical considerations emphasize the importance of timely and transparent notification following a data breach. Individuals have a right to be informed about breaches that may impact their personal information. Transparent communication enables affected individuals to take necessary steps to protect themselves, seek appropriate support, and make informed decisions about their data.

Cybersecurity and Preparedness: Ethical obligations extend to organizations' cybersecurity practices and preparedness. Organizations should employ strong security measures, regularly assess and address vulnerabilities, and implement appropriate safeguards to protect individuals' data. Ethical considerations also include investing in cybersecurity research and development to stay ahead of evolving threats.

Legal and Regulatory Compliance: Ethical considerations align with legal and regulatory compliance. Organizations have an ethical obligation to comply with applicable laws and regulations regarding data protection and breach notifications. This includes respecting individuals' rights, providing necessary safeguards, and reporting breaches as required by law.

Addressing the ethical implications of data breaches requires a comprehensive approach that focuses on protecting individuals' privacy, fostering trust, promoting accountability, and mitigating harm. Organizations should prioritize ethical data handling practices, invest in robust cybersecurity measures, and be prepared to respond effectively in the event of a breach. Individuals should also be aware of their rights, take measures to protect their data, and advocate for responsible data practices.

IV. DATA RETENTION

Data retention refers to the practice of storing and maintaining data for a specific period of time. Organizations and individuals retain data for various reasons, such as legal and regulatory requirements, business needs, historical reference, analysis, or operational purposes. However, data retention also raises important ethical considerations.

Privacy and Consent: Ethical concerns arise when organizations retain data beyond the scope of the original purpose for which it was collected or without obtaining informed consent from individuals. Individuals have the right to control how their personal data is used and retained. Organizations should be transparent about their data retention practices, clearly communicate the purpose and duration of retention, and obtain consent when necessary.

Data Minimization: Ethical principles of data minimization advocate for collecting and retaining only the minimum amount of data necessary to fulfil the intended purpose. Retaining excessive or unnecessary data increases the risk of breaches, unauthorized access, and potential harm to individuals. Organizations should regularly review their data retention practices to ensure they align with the principle of data minimization.

Security and Protection: Ethical obligations extend to ensuring the security and protection of retained data. Organizations should implement appropriate security measures to safeguard the data from unauthorized access, breaches, and misuse. This includes robust encryption, access controls, regular security audits, and other industry best practices. Failure to protect retained data can result in harm to individuals and breach of trust.

Transparency and Accountability: Ethical considerations require organizations to be transparent about their data retention practices and accountable for the data they retain. This includes providing individuals with clear information about the types of data retained, the purpose and duration of retention, and their rights regarding the data. Organizations should also have mechanisms in place to address individuals' requests for access, correction, or deletion of their retained data.

Legal and Regulatory Compliance: Organizations have ethical and legal obligations to comply with applicable laws and regulations regarding data retention. They must understand and adhere to requirements related to data retention periods, data protection, privacy, and individual rights. Compliance with legal and regulatory frameworks helps protect individuals' rights and ensures ethical data handling practices.

Data Disposal: Ethical considerations also extend to the secure disposal of retained data when it is no longer needed. Organizations should have proper procedures in place to safely and permanently delete or destroy data in a manner that prevents unauthorized access or recovery. Data disposal should align with legal requirements and industry best practices.

Data Retention Policies: Organizations should develop and implement clear data retention policies that outline the purpose, duration, and criteria for retaining different types of data. These policies should be reviewed and updated regularly to ensure compliance with evolving legal, regulatory, and ethical standards.

V.PRIVACY-ENHANCING TECHNOLOGIES

Privacy-enhancing technologies (PETs) are tools, techniques, and systems designed to protect individuals' privacy and enhance their control over their personal data. These technologies aim to minimize the collection, use, and disclosure of personal information while enabling individuals to maintain their privacy in digital environments.

Encryption: Encryption is a widely used privacy-enhancing technology that transforms data into a coded form that can only be accessed or deciphered with the appropriate encryption keys. It helps protect data during storage, transmission, and even while it's being processed.

Anonymization and Pseudonymization: Anonymization techniques remove personally identifiable information (PII) from data, making it difficult to identify individuals. Pseudonymization replaces identifiable information with pseudonyms, allowing for data processing while protecting individual identities.

Privacy-Preserving Data Mining: Privacy-preserving data mining techniques enable the analysis of data without revealing sensitive or personally identifiable information. These methods include differential privacy, secure multiparty computation, and secure data aggregation.

Privacy-Focused Web Browsers and Search Engines: Privacy-focused web browsers, such as Tor, and search engines, such as DuckDuckGo, prioritize user privacy by minimizing data collection, blocking tracking cookies, and providing encrypted connections.

Virtual Private Networks (VPNs): VPNs create secure and encrypted connections over public networks, allowing individuals to browse the internet privately and anonymously by hiding their IP addresses and encrypting their internet traffic.

Privacy Filters and Anti-Tracking Tools: Privacy filters and anti-tracking tools help block unwanted data collection by preventing advertisements, trackers, and cookies from monitoring individuals' online activities.

Secure Messaging and Communication Apps: Privacy-focused messaging and communication apps employ end-to-end encryption to ensure that only the intended recipients can access and decrypt the messages. Examples include Signal and WhatsApp's end-to-end encryption feature.

Privacy-Focused Operating Systems: Certain operating systems, like Linux distributions focused on privacy, prioritize user privacy and security by minimizing data collection, providing granular privacy controls, and emphasizing open-source transparency.

Privacy Impact Assessment (PIA) Tools: PIA tools assist organizations in assessing and managing privacy risks associated with their projects, services, or products. These tools help identify and mitigate potential privacy concerns by evaluating data processing activities and compliance with privacy regulations.

Blockchain and Privacy Coins: Certain blockchain technologies and privacy coins, such as Monero and Zcash, incorporate privacy features that aim to anonymize transactions and protect user identities.

Privacy-enhancing technologies play a vital role in safeguarding individuals' privacy and allowing them to maintain control over their personal data. These tools help mitigate privacy risks, empower individuals to make informed choices, and promote responsible data handling practices. However, it's important to note that the effectiveness of privacy-enhancing technologies may vary, and their adoption should be accompanied by appropriate policies, legal frameworks, and user education to ensure privacy protection.

VI. CONCLUSION

Privacy and cryptography are deeply intertwined and play crucial roles in protecting sensitive information and ensuring secure communication in today's digital world. Cryptography is the practice of encoding and decoding information to keep it confidential, secure, and authentic. Privacy, on the other hand, refers to an individual's right to control their personal data and information. Over the years, privacy concerns have become more prominent due to the rapid advancements in technology, the proliferation of online platforms, and the increasing collection and analysis of personal data. Cryptography has emerged as a fundamental tool for safeguarding privacy in various domains, including communication, financial transactions, and data storage. Privacy and cryptography are intricately linked, with cryptography serving as a vital tool for safeguarding privacy in an increasingly digital world. However, protecting privacy requires a comprehensive approach that goes beyond encryption, encompassing data governance, consent, and ethical practices in handling personal information. As technology continues to evolve, it is crucial to balance privacy concerns with the benefits of innovation, ensuring that privacy and security remain at the forefront of our digital landscape.

VII. REFERENCE

1. Diffie, W., & Landau, S. (2007). Privacy on the Line: The Politics of Wiretapping and Encryption. MIT Press.
2. Schneier, B. (2015). Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World. W. W. Norton & Company.
3. Greenwald, G. (2014). No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State. Metropolitan Books.
4. Menezes, A., van Oorschot, P., & Vanstone, S. (1996). Handbook of Applied Cryptography. CRC Press.
5. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.
6. Bernstein, D. J., Lange, T., & Schwabe, P. (2012). Post-quantum cryptography. Nature, 549(7671), 188-194.
7. Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 202-210.
8. Clarke, R. (1999). Internet privacy concerns confirm the case for intervention. Communications of the ACM, 42(2), 60-67.
9. Green, M., Hohenberger, S., & Waters, B. (2017). Outsourcing secure computation: On the possibility of using encrypted data for privacy-preserving computation. Journal of Cryptology, 30(3), 537-574.
10. Abelson, H., Anderson, R., Bellovin, S. M., Benaloh, J., Blaze, M., Diffie, W., & Schneier, B. (2015). Keys under doormats: Mandating insecurity by requiring government access to all data. Journal of Cybersecurity, 1(1), 69-79
11. Bonneau, J., Narayanan, A., Miller, A., Clark, J., Kroll, J. A., & Felten, E. W. (2015). Mixcoin: Anonymity for Bitcoin with accountable mixes. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (pp. 843-854).

ID: 07

MACHINE LEARNING: A COOPERATIVE VIEW OF ALGORITHMS AND ITS APPLICATIONS

Dr.P.ARUNA DEVI
Assistant Professor,
Department of Artificial Intelligence and Machine Learning,
Kongunadu Arts and Science College,
Coimbatore, Tamil Nadu, India.

Dr.K.SASIREKHA
Assistant Professor & Academic Coordinator,
Department of CS (AI&DS),
Dr.SNS Rajalakshmi College of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT

In this paper, various machine learning algorithms and their applications have been discussed. Machine learning is a branch of Artificial Intelligence and Computer Science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information[15][16]. At present, it is being used for a variety of tasks such as E-Mail Filtering, image recognition, speech recognition, Face book auto-tagging, recommender system etc. Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

KEYWORDS: *Machine Learning, Algorithms, Clustering, Target Variable, Dataset.*

I. INTRODUCTION

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it[2]. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. Let us consider a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of idea about the problem [22]. The working of Machine Learning algorithm is as follows.

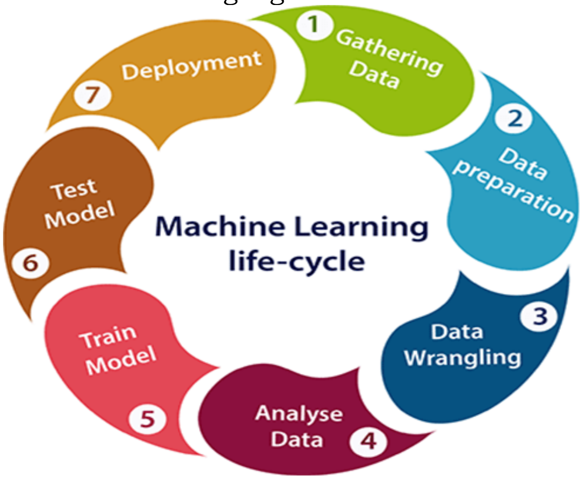


Fig.1 Working of Machine Learning Algorithm

II. FEATURES OF MACHINE LEARNING

A machine learning algorithm uses data to detect various patterns in a given dataset. Past data can be learned and improve automatically. It is a data-driven technology. Machine learning is much related to data mining because it also deals with the huge amount of the data.

III. CLASSIFICATION OF MACHINE LEARNING

Machine learning can be classified into three types [4]. They are Supervised learning, Unsupervised learning, Reinforcement learning

The following diagram shows the classification and uses of Machine Learning Algorithm.

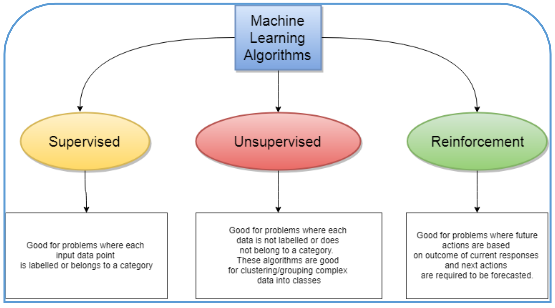

Machine Learning Algorithms
Supervised
Unsupervised
Reinforcement
Good for problems where each input data point is labelled or belongs to a category
Good for problems where each data is not labelled or does not belong to a category. These algorithms are good for clustering/grouping complex data into classes
Good for problems where future actions are based on outcome of current responses and next actions are required to be forecasted.

Fig.2 Types of Machine Learning Algorithm

1) SUPERVISED LEARNING

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output [6]. The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The following diagram shows the working of Supervised Learning.

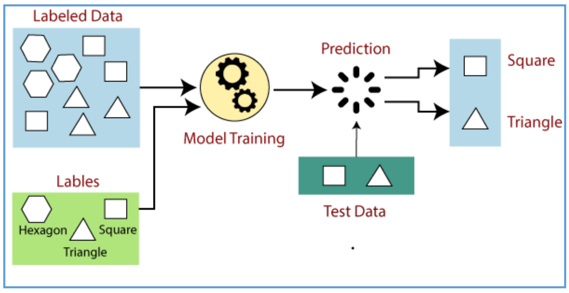


Fig.3 Working of Supervised Algorithm

1.1 Types of Supervised Machine learning Algorithms

Supervised Learning can broadly classify as Regression and Classification algorithms.

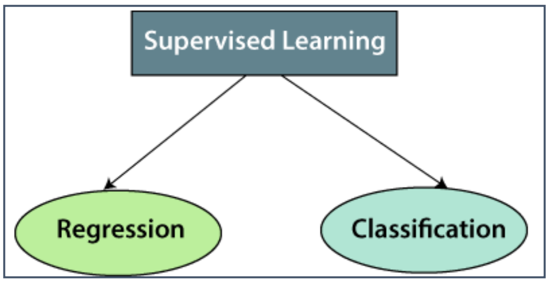


Fig.4 Types of Supervised Algorithm

1.1.1. REGRESSION

A Regression algorithm is used to finds the driving parameters for a numerical value, builds a function which helps in the prediction of particular value. For instance, how much should be the selling price of a land, depending on factors such as its location, number of sites available, area of the land, etc. For this to converge, the algorithm is initially trained with data that contains all the information related to the driving factors for the price and most importantly, the price of the land, which is of course known at that point in time. Some of the regression algorithms as follow.

a) Linear Regression

It is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables. Linear regression algorithm is used if the labels are continuous, like the number of vegetables daily from a market, etc. The representation of linear regression is

$$y = b*x + c.$$

In the above representation, 'y' is the independent variable, whereas 'x' is the dependent variable. When you plot the linear regression, then the slope of the line that provides us the output variables is termed 'b', and 'c' is its intercept. The linear regression algorithms assume that there is a linear relationship between the input and the output. If the dependent and independent variables are not plotted on the same line in linear regression, then there will be a loss in output. The loss in output in linear regression can be calculated as:

Loss function: (Predicted output – actual output)

b) Ridge Regression

Ridge Regression is another popularly used linear regression algorithm in Machine Learning. If only one independent variable is being used to predict the output, it will be termed as a linear regression ML algorithm. ML experts prefer Ridge regression as it minimizes the loss encountered in linear regression. The complexity of the ML model can also be reduced via ridge regression. One should note that not all the coefficients are reduced in ridge regression, but it reduces the coefficients to a greater extent as compared to other models. The ridge regression is represented as:

$$y = X\beta + \epsilon$$

where 'y' is the N*1 vector defining the observations of the dependent data point/variable and 'X' is the matrix of regressors. 'β' is the N*1 vector consisting of regression coefficients and 'ϵ' is the vector (N*1) of errors. The ridge algorithm is also used for regression in Data Mining by IT experts besides ML.

c) Polynomial Regression

Polynomial Regression is a regression algorithm that models the relationship between an independent variable (x) and a dependent variable (y) as an nth degree polynomial [2]. It is a linear model that has been modified to improve accuracy. The dataset used for training in polynomial regression is non-linear. To fit the non-linear and complicated functions and dataset, the original features are changed into polynomial features of the required degree (2,3,…,n) and then modeled using a linear model.

d) Logistic Regression

Logistic regression [14] is a type of regression that predicts the probability of occurrence of an event by fitting data to a logistic function. Just as many form of regression analysis, logistic regression makes use of several predictor variables that may be numerical or categorical. The logistic regression hypothesis is defined as:

$h\theta\ (x) = (\theta\ Tx)$ Where the function g is sigmoid function defined as:

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

The sigmoid function has special properties that result the values in range [0,1], as visualized

e) Bayesian Linear Regression

The Bayesian model of classification is used for large finite datasets. It is a method of assigning class labels using a direct acyclic graph. The graph comprises one parent node and multiple children nodes. And each child node is assumed to be independent and separate from the parent. This model draws on common data assumptions, such as each attribute is independent. This algorithm can easily be implemented on complex problems. This is also known as Naive Bayes classifier it would consider all of these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

1.1.2 CLASSIFICATION ALGORITHM

A **classification algorithm** aims to determine the pre-defined category of the data. For example, based on certain parameters in an image of a fruit, we can know whether the image contains a tomato or does it contain an apple[1]. For this to work for a presented image, the machine learning algorithm would have been trained earlier on various images of tomatoes and apples, with the user explicitly specifying the vegetable during the training. Hence, for a new image, the algorithm can determine the name of the vegetable. This concept can also be applied for email classification, whether an email should end up in Spam or any of the allowed Inbox categories like General, Social, and Promotions etc. Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

a) Random Forest

Random forest is also a widely-used algorithm for non-linear regression in Machine Learning. This algorithm uses multiple decision trees for predicting the output[21]. Random data points are selected from the given dataset say k data points are selected and a decision tree is built with them through this algorithm. Several decision trees are then modeled that predict the value of any new data point.

b) Decision Trees

Decision Tree algorithm in machine learning is one of the most popular algorithm. This supervised learning algorithm that is used for classifying problems [8]. It works well in classifying both categorical and continuous dependent variables. This algorithm divides the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

c) KNN (K- Nearest Neighbors) Algorithm

This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement.

d) Random Forest Algorithm

A collective of decision trees is called a Random Forest[20]. To classify a new object based on its attributes, each tree is classified.

e) Support Vector Machines (SVM)

A support vector machine is a popular supervised learning model developed by Vladimir Vapnik, used for both data classification and regression. That said, it is typically leveraged for classification problems, constructing a hyper plane where the distance between two classes of data points is at its maximum [27]. This hyper plane is known as the decision boundary, separating the classes of data points (e.g., oranges vs. apples) on either side of the plane.SVM calculates a maximum-margin boundary that leads to a homogeneous partition of all data points.

2) UNSUPERVISED LEARNING

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision[3]. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result. Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning**,** we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.** The machine tries to find useful insights from the huge amount of data. The following diagram shows the working of unsupervised Algorithm.

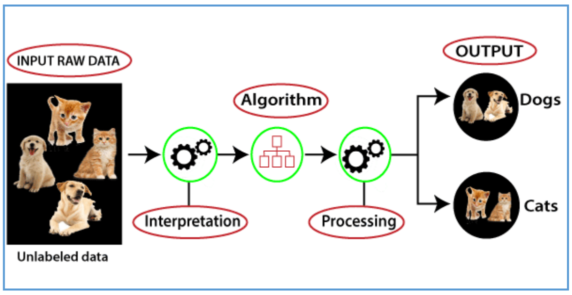


Fig.5 Working of Unsupervised Algorithm

2.1. Types of Unsupervised Algorithms

It can be further classifieds into two categories of algorithms[19]. They are clustering and association.

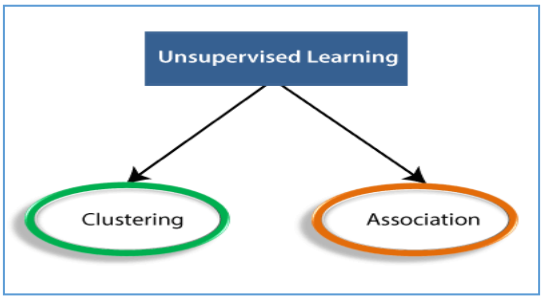


Fig.6 Types of Unsupervised Algorithm

2.1.1 Clustering

Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the similarities between the data objects and categorizes them as per the presence and absence of those similarities.

a) K-means Cluster

This is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group’s centroid. The data points closest to a given centroid will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity[11]. K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.

b) Overlapping Cluster

It is differs from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership. “Soft” or fuzzy k-means clustering is an example of overlapping clustering.

c) Hierarchical Cluster

Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways; they can be agglomerative or divisive. Agglomerative clustering is considered a "bottoms-up approach." Euclidean distance is the most common metric used to calculate these distances.

d) Divisive Cluster

It can be defined as the opposite of agglomerative clustering it takes a "top-down" approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.

e) Probabilistic clustering

A probabilistic model is an unsupervised technique that helps us solve density estimation or "soft" clustering problems. In probabilistic clustering, data points are clustered based on the likelihood that they belong to a particular distribution. The Gaussian Mixture Model (GMM) is the one of the most commonly used probabilistic clustering methods.

2.1.2 Association

An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database[9]. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a cheese) are also tend to purchase Y (bread/pitza) item. A typical example of Association rule is Market Basket Analysis[10].

a) The k-means Clustering

The k-means clustering is a process that helps to partition the data points or observations into k unknown clusters in such a manner that each observation distinctly belongs to a cluster. This cluster associativity is determined by the proximity of that data point with the nearest mean, otherwise known as cluster centroid. Due to the involvement of proximity measure in the data, various distance algorithms are used in the process to measure the closeness of data to the cluster center.

b) DBSCAN Clustering

The DBSCAN algorithm is based on this perceptive view of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Partitioning methods (K-means, PAM clustering) and hierarchical clustering work is used for finding spherical-shaped clusters or convex clusters. In other words, they are suitable only for compact and well-separated clusters.

3) REINFORCEMENT LEARNING

Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. Reinforcement learning uses different algorithms that learn from outcomes and decide which action to take next and also receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect. Reinforcement learning is an autonomous, self-teaching system that essentially learns by trial and error [5]. It Consists of Input variables, Output variables and training data set. The model keeps continues to derive for best solution to get maximum rewards. The following diagram shows the working principle of Reinforcement Algorithm.

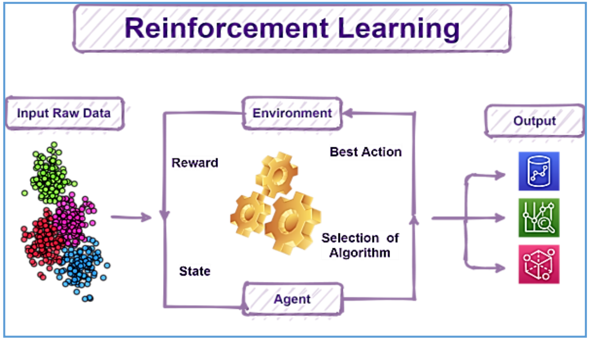


Fig.7 Working of Reinforcement Learning

3.1 Types of Reinforcement Learning Algorithm

Reinforcement learning algorithms are mainly used in AI applications and gaming applications. The main used algorithms are: Q-Learning, State Action Reward State action (SARSA), Deep Q Neural Network (DQN), Markov Decision Processes.

3.1.1 Q-Learning

Q-learning is an off policy RL algorithm, which is used for the temporal difference Learning [7]. The temporal difference learning methods are the way of comparing temporally successive predictions. It learns the value function Q (S, a), which means how good to take action "a" at a particular state "s."

3.1.2 State Action Reward State action (SARS)

The on-policy control method selects the action for each state while learning using a specific policy. The goal of SARS action is to calculate the Q π (s, a) for the selected current policy π and all pairs of (s-a)[9]. The main difference between Q-learning and SARS action algorithms is that Q-learning, the maximum reward for the next state is not required for updating the Q-value in the table. In this algorithm a new action and reward are selected using the same policy, which has determined the original action. This algorithm is derived from the Quintuple.

Q(s, a, r, s', a'). Where,
s: means original state
a: means Original action
r: means reward observed while following the states
s' and a': referred as New state, action pair.

3.1.3 Deep Q Neural Network (DQN)

This algorithm comes as Q-learning using neural networks. It will be a challenging and complex task to define and update a Q-table. To solve such an issue, we can use a DQN Network algorithm. Q-learning is a popular model-free reinforcement learning algorithm based on the Bellman equation[14]. The foremost purpose of Q-learning is to find out the policy which can inform the agent that what actions should be taken for maximizing the reward under what situation. It is an off-policy RL that attempts to find the best action to take at a current state. The goal of the agent in Q-learning is to maximize the value of Q. The value of Q-learning can be derived from the Bellman equation. Consider the Bellman equation given below:

$$V(s) = \max [R(s,a) + \gamma \textstyle\sum_{s'} P(s, a, s')V(s`)]$$

In the equation, we have various components, including reward, discount factor (γ), probability, and end states s'

3.1.4. Markov Decision Processes (MDPs)

It is a mathematical frameworks to describe a surroundings in reinforcement learning and almost all problems can be formulated using this decision making process [24]. It consists of a set of finite environment states S, a set of possible actions A(s) in each state, a real valued reward function R(s) and a transition model P (s’, s | a). However, real world environments are more likely to lack any prior knowledge of environment dynamics.

IV. CONCLUSION

This paper is shows the overview of machine learning algorithms. Today each and every person is using machine learning knowingly or unknowingly. This paper gives an introduction to most of the popular machine learning algorithms and its uses. These algorithms can be used for classification, regression, and time series forecasting tasks. All these algorithms are widely used in various domains, including healthcare, finance, marketing, and image recognition, to make predictions and gain valuable insights from data.

V. REFERENCES

1. W. Richert, L. P. Coelho, “Building Machine Learning Systems with Python”, Packt Publishing Ltd., ISBN 978-1-78216-140-0
2. M. Welling, “A First Encounter with Machine Learning”
3. M. Bowles, “Machine Learning in Python: Essential Techniques for Predictive Analytics”, John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2
4. S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268
5. L. Rokach, O. Maimon, “Top – Down Induction of Decision Trees Classifiers – A Survey”, IEEE Transactions on Systems,
6. Lowd, P. Domingos, “Naïve Bayes Models for Probability Estimation”
7. https://webdocs.cs.ualberta.ca/~greine

r/C- 651/Homework2_Fall2008.html

8. Meyer, "Support Vector Machines – The Interface to libsvm in package e1071", August 2015.
9. S. S. Shwartz, Y. Singer, N. Srebro, "Pegasos: Primal Estimated sub - Gradient Solver for SVM", Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007
10. http://www.simplilearn.com/what-is-machine-learning-and-why-it- matters-article.
11. P. Harrington, "Machine Learning in action", Manning Publications Co., Shelter Island, New York, 2012
12. http://pypr.sourceforge.net/kmeans.html
13. K. Alsabati, S. Ranaka, V. Singh, "An efficient k-means clustering algorithm", Electrical Engineering and Computer Science, 1997
14. M. Andrecut, "Parallel GPU Implementation of Iterative PCA Algorithms", Institute of Biocomplexity and Informatics, University of Calgary, Canada, 2008
15. X. Zhu, A. B. Goldberg, "Introduction to Semi – Supervised Learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, Vol. 3, No. 1, Pages 1-130
16. X. Zhu, "Semi-Supervised Learning Literature Survey", Computer Sciences, University of Wisconsin-Madison, No. 1530, 2005
17. S. Sutton, "Introduction: The Challenge of Reinforcement Learning", Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992
18. L. P. Kaelbing, M. L. Littman, A. W. Moore, "Reinforcement Learning: A Survey", Journal of Artificial Intelligence Research, 4,
19. Page 237-285, 1996
20. R. Caruana, "Multitask Learning", Machine Learning, 28, 41-75,
21. Kluwer Academic Publishers, 1997
22. Opitz, R. Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of Artificial Intelligence Research, 11, Pages 169- 198, 1999
23. H. Zhou, "Ensemble Learning", National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
24. https://en.wikipedia.org/wiki/Boosting_(machine_learning)
25. https://en.wikipedia.org/wiki/Bootstrap_aggregating
26. V. Sharma, S. Rai, A. Dev, "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277128X,
27. Volume 2, Issue 10, October 2012
28. S. B. Hiregoudar, K. Manjunath, K. S. Patil, "A Survey: Research Summary on Neural Networks", International Journal of Research in Engineering and Technology, ISSN: 2319 1163, Volume 03, Special Issue 03, pages 385-389, May, 2014
29. https://en.wikipedia.org/wiki/Instance-based_learning
30. P. Harrington, "Machine Learning in Action", Manning Publications Co., Shelter Island, New York, ISBN 9781617290183, 2012
31. J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985

ID: 08

IDENTIFICATION OF BRAIN TUMOR USING PROJECTION PURSUIT BIVARIATE MULTILAYER PERCEPTRED CLASSIFICATION

Ms.P.S.RENJENI
Assistant Professor,
Department of Computer Science,
VTM College of Arts & Science,
Arumanai, Tamil Nadu, India.

Dr.L.JAYA SINGH DHAS
Assistant Professor,
Department of Computer Science,
Scott Christian College (Autonomous),
Nagercoil, Tamil Nadu, India.

ABSTRACT

The method of identifying the disease with person's symptoms and signsis medical diagnosis. Brain tumour is the stimulating disorder that has to be identified at early stage for treatment. Many classification techniques have been introduced for performing brain tumour identification.

However, the brain tumour identification accuracy level was not enhanced and time consumption was not lessened. In order to address these problems, Projection Pursuit Feature Selective Bivariate Multilayer Perceptred Classification (PPFSBMPC) Method is introduced. PPFSBMPC Method comprises two processes, namely feature selection and classification for brain tumour identification. To select the relevant features from the input database, Projection Pursuit Feature Selection process is carried out in PPFSBMPC Method.

After performing the feature selection, Bivariate Multilayer Perceptred Classification process is accomplished forbrain tumor identification. In addition, the classification process comprised multiple layers to categorize the input data as normal data or tumour diseased data. By this way, PPFSBMPC Method increases the brain tumor identification performance with higher accuracy and lesser time consumption. Experimental evaluation of PPFSBMPC Methodis carried out with Epileptic Seizure Recognition Dataseton factors such as brain tumour identification accuracy, execution time, and error rate with respect to number of patient data.

The experimental result demonstrates that the PPFSBMPC Methodenhances the brain tumour identification accuracy and reducesthe execution time when compared to state-of-the-art-works.

KEYWORDS: *Medical diagnosis, brain tumour, classification, feature selection, classification process, identification, seizure.*

I. INTRODUCTION

The diagnosis of Brain tumor comprises of series of test to determine the patient indications and neurological functions. Many research works have been designed for feature selection and classification techniques to perform brain tumor disease identification. But, the accuracy level and time consumption performance was not improved. In order to solve the issues, the PPFSBMPC Method is introduced to improve the accuracy of brain tumor identification.

An ANOVA based feature selection and fuzzy classifier model was introduced to find seizure state of EEG signal with time and frequency features as in [1]. The state of EEG signal was classified by fuzzy classifier for forecasting the capture. But, the accuracy level was not enhanced. Complex-valued classifiers were introduced for epilepsy diagnosis from electroencephalography (EEG) signals depending on classifiers as mentioned in [2]. The feature of EEG data was extracted by dual-tree complex wavelet transformation at different granularity levels for size reduction. But, the execution time was not minimized by complex-valued classifiers.

Naïve Bayes classification and decision tree algorithm was introduced to perform accurate brain tumor prophesy [3].

Thedesigned algorithm performed prediction bydecision tree algorithm and Naïve Bayes algorithm. Long Short-Term Memory (LSTM) networks were introduced for epileptic seizure prediction using convolutional neural networks (CNN) with EEG signals in respect of [4]. A pre-analysis was carried out to identify optimal architecture of LSTM network through testing different modules and memory unit layers. But, the computational cost was not minimized by LSTM.

An automated detection system was designed to substitute the neurologist contribution for time and speed treatment processas shown in [5]. An important feature for epileptic seizure recognition wasa matrix determinant of EEG. But, the accuracy level was not improved by automated detection system. A three-class classification system depending on discrete wavelet transform (DWT) and the nonlinear sparse extreme learning machine (SELM) was introduced for epilepsy and epileptic seizure identification as in [6]. However, the time consumption was not minimized by this method.

A statistical analysis of EEG signal was carried out for identifying the epileptic seizure with high accuracy for different age of epilepsy [7]. But, the computational complexity was not minimized by statistical analysis. An automated seizure detection method was introduced depending on statistical and spectral features of maximum normalized intrinsic mode functions by ensemble empirical mode decomposition with adaptive noise technique as in [8]. Even after, accuracy level was not improved by automated seizure detection method.

The epileptic disorder was identified by performing EEG signal analysis through joining best attributes of Artificial Bee Colony (ABC) and radial basis function networks (RBFNNs) as in [9]. But, the feature selection was not carried out in precise manner. A fastand objective diagnosis was carried out with genetic variant of oligodendrogliathrough feature selection and ensemble-based classification as mentioned in [10]. But, the diagnosis time was not minimized by ensemble-based classification.

The above mentioned problems in the brain tumor disease identification from the existing works are lesser brain tumor disease identification accuracy, more execution time consumption,higher error rate, higher computational complexity, high computational cost, and so on. These kinds of problems wereaddressed by introducing a new method called PPFSBMPC Method.

The contribution of PPFSBMPC Method is summarized as follows.

- To improve the performance of brain tumour disease identification, PPFSBMPC method is introduced. This method comprises of two processes, namely feature selection and classification for brain tumour disease identification.
- To select the relevant features from the input database, Projection Pursuit Feature Selection process is used in PPFSBMPC Method. This helps to minimize the execution time during brain tumour disease identification.
- To categorize the input data as normal data or tumour diseased data, Bivariate Multilayer Perceptred Classification process is carried out with multiple layers forbrain tumor disease identification.

The rest of the paper is organized into five different sections. Section 2 describes the related works in brain tumor disease identification. In Section 3, the proposed PPFSBMPC Method is described with the neat architecture diagram. In Section 4, experimental evaluation is carried out with EEG dataset.The results are discussed with different parameters in Section 5.The conclusion is presented inSection 6.

II. RELATED WORKS

Brain tumor takes place when cells in brain grow out of the control and transfer to nearby tissues. A new gauss-newton representation based algorithm (GNRBA) was introduced for performing breast cancer classification in [11]. However, the tumor classification accuracy was not improved by GNRBA. Ensemble learning and data mining techniques were designed to rank the risk factors and diagnose the recurrence of ovarian cancer [12]. But, the diagnosis time consumption was not minimized by designed method.

The machine learning and matrix factorization method was introduced to determine the risk factors for increasing the disease risk assessment [13]. The designed framework comprised data pre-processing, risk factor optimization and risk assessment. A large quantity of data was provided using magnetic resonance imagingtechnique to find the brain tumor [14]. The data mining classification techniques were introduced to attain high accuracy. An efficient feature extraction method termed Local Neighbour Descriptive Pattern (LNDP) and One-dimensional Local Gradient Pattern (1D-LGP) was designed to categorize the epileptic EEG signals [15]. A convolutional neural network (CNN) was introduced depending on EEG signals to differentiate the ictal, preictal, and interictal segments for epileptic seizure detection[16].The seizure and non-seizure classification method was carried out depending on bandwidth features of EEG signals as in [17]. The designed method partitioned signal into eight Intrinsic Mode Functions (IMFs).

A new method was designed depending on weighted visibility graph entropy (WVGE) to recognize the seizure from EEG signals [18]. Single channel EEG signals were mapped onto WVGs andWVGEs.

A new approach was introduced for automated identification of seizure EEG signal as in [19]. The designed method decomposed the EEG signal into collection of sub-band signals through tunable-Q wavelet transform (TQWT) based filter-bank.A tunable-Q wavelet transform (TQWT) framework was introduced to decompose the EEG signals into sub bands [20]. The nonlinearity of EEG signals was assessed through computing cantered correntropy (CCE) from sub bands.

III. PROJECTION PURSUIT FEATURE SELECTIVE BIVARIATE MULTILAYER PERCEPTRED CLASSIFICATION (PPFSBMPC) METHOD

To classify the patient data with higher accuracy and minimum time consumption, Projection Pursuit Feature Selective Bivariate Multilayer Perceptred Classification (PPFSBMPC) Method is introduced .This method comprises the two processes, namely feature selection and

Classification with data points. In feature selection process, the relevant featuresare selected from the dataset. Projection Pursuit is a kind of statistical method employed for choosing the relevant features from dataset. After that, the data classification is carried out by using Bivariate Multilayer Perceptred Classificationwith the selected features for achieving higher accuracy. The architecture diagram of the PPFSBMPC Method is illustrated in figure 1 (a).

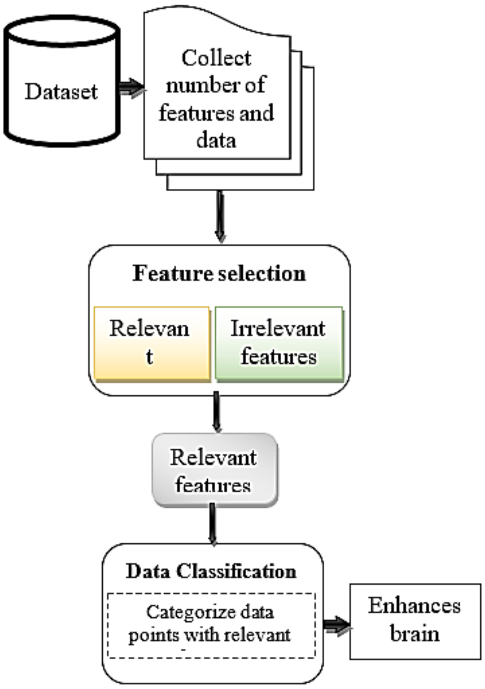


Figure.1 (a). Architecture diagram of the proposed PPFSBMPC Method

Figure 1 depicts an architecture diagram of PPFSBMPC method with feature selection and data classification. Initially, the number of features and data are collected from the dataset. PPFSBMPC method selects the relevant features and removes the irrelevant features before performing the classification process. This process minimizes the time consumption for performing brain tumour identification. In second step, classification is carried out using bivariate multilayer perceptred classificationto categorize the data points as normal or tumour diseased data. The detailed explanation of PPFSBMPC method is briefly described in the following subsections.

3.1 Projection Pursuit Feature Selection

The initial step of PPFSBMPC Method is the feature selection. The main aim is to select the relevant features from the large database. PPFSBMPC Method uses Projection Pursuit Analysis for feature selection process. Projection Pursuit Analysis is a kind of the statistical technique employed to interactively find out the relevant features from the multidimensional space. The dataset has number of features '$sfe_1, sfe_2, sfe_3, \ldots . sfe_n$' in multidimensional space. Projection Pursuit is a mapping of features set into two subsets (i.e., relevant or irrelevant).

Let us take, 'S' is a $n*k$ matrix that explains 'n' features of 'k' dimensions and 'Q' is a $n*l$ matrix that explains the 'l' dimensional target view of features. The projection matrix 'M' identifies the similarity between target and the features in dataset. Sørensen–dice indexed coefficient is employed for determining the similarity between the target and features in matrix 'S'.

$$\delta = \frac{2|T \cap sfe_i|}{|T|+|sfe_i|} \quad (1)$$

From (1), 'δ' represents the sørensen–dice indexed coefficient, 'T' denotes the target, 'sfe_i' represents the input features. '$T \cap sfe_i$' represent the mutual dependence between target and feature. The sørensen–dice indexed coefficient (δ) provides the similarity value ranges from 0 to 1. Depending on the similarity value, the projection matrix projects the relevant features into subset. The projection matrix reduces size of difference between feature and target when similarity is high. The projection matrix employs the steepest gradient descent to reduce the variation and it is given as,

$$sgdf(x) = \arg\min \|Q - S.M\| \quad (2)$$

From (2), '$sgdf(x)$' denotes the steepest gradient descent function and '$arg\ min$' denotes the argument of minimum. Steepest gradient descent function projects the similar features into two-dimensional space. The less similarity features with higher variation between target and feature projection termed as irrelevant features and it is eliminated from dataset. The algorithmic process of the Projection Pursuit Analysis based feature selection is described as follows,

```
// Algorithm 1: Projection Pursuit Analysis based Feature Selection
Input: Dataset $D_t$ ,number of features$sfe_1, sfe_2, sfe_3, \ldots .sfe_n$
Output: Select relevant features
Begin
    1. For each feature$sfe_1 D_t$
    2. Construct feature matrix $S$', target matrix $Q$, projection matrix $'M'$
    3. Determine the correlation between the features and target '$\rho$'
    4. if($\delta = +1$)then
    5.     It is considered as relevant feature
    6.     Projection matrix project the high similarity features into two-dimensional space
    7.     Reduce the difference between projection of feature and the target $\arg\min \|Q - S.M\|$
    8.     Select relevant features
    9. else
    10.    Remove irrelevant features
    11. End if
    12. End for
End
```

Algorithm 1 Projection Pursuit Analysis based Feature Selection

Algorithm 1 illustrates the feature selection using projection pursuit analysis. Initially, the numbers of features are considered as input from dataset. After that, the correlation between the target and features in the matrix are determined to identify the high and low similarity features. The projection matrix project high similarity features into two-dimensional space. The high similarity feature projection reduces the difference between projection of feature and target. Subsequently, the high similarity features are chosen for brain tumour identification and the low similarity features are removed. The feature selection process in PPFSBMPC Method reduces the time complexity for brain tumour identification.

3.2 Bivariate Multilayer Perceptred Classification

A multilayer perceptron is the machine learning system along with set of feature vector. Bivariate analysis is the quantitative analysis for determining the relationship between them. The multilayer perceptron is a feed-forward neural network with two or more layers like one input layer and one output layer with one hidden layer of activating nodes. Then, an input region and the features are fed into input layer. In hidden layer, support vector regression is used to categorize the patient data into two classes with help of selected features through separating hyper planes. Finally, the results are transmitted to the output layer.

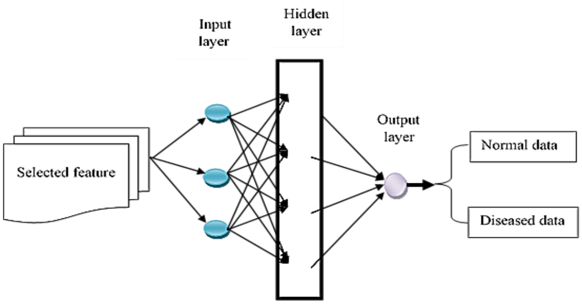


Figure.2(a) Bivariate Multilayer Perceptive Neural Learning

Figure 2(a) illustrates the multilayer perceptive neural learning forbrain tumor identification. As illustrated in figure 2, Neurons like nodes in one layer are fully linked with the other layer to form entire network with help of arrow symbol. The input layer collects the number of features into the network at time 't' denoted by '$I(t)$'. The node in one layer links to another layer through dynamic weights. It is given by,

$$I(t) = \sum_{i=1}^{n} sfe_i * w_1 + b \qquad (3)$$

From (3), 'sfe_i' represent the selected features with adjustable initial weight 'w_1'between input and hidden layer. 'b' denotes the bias. After that, the input is fed into the hidden layer where the support vector regression is used for data classification. The regression is a statistical process used for classifying the data according to the feature values. The support vector regression uses separating hyperplane to analyze and categorize the data. The two marginal hyperplanes are created on both sides of hyperplane. The hyperplane act as the threshold between the output classes. A separating hyperplane (H_s) is defined as,

$$H_s \rightarrow \overrightarrow{We}.(dp) + \vec{B} = 0 \qquad (4)$$

From (4),$\overrightarrow{We}$represents the weight vector to hyperplane, 'dp' denotes the data points and $\vec{B}$ symbolizes the bias. When the feature valueof the patient data is greater than the threshold value, then the data points are classified by constructing two marginal planes and it is given below as,

$$MH_1 = \overrightarrow{We}.(dp) + \vec{B} > 0 \rightarrow \text{tumour diseased data} \quad (5)$$

$$MH_2 = \overrightarrow{We}.(dp) + \vec{B} < 0 \rightarrow \text{normal data} \quad (6)$$

From (5) and (6), MH_1 and MH_2 denotes two marginal hyperplanes (i.e., above and below the hyperplane). The data point values higher than the threshold is classified above the hyperplane. The data point values lesser than the threshold is classified below the hyperplane. The support vector regression categorizes the data points into normal or diseased based on either side of the hyperplane. The output of hidden layer at the time 't' is given as follows,

$$L(t) = I(t) * w_2 \qquad (7)$$

From (7), '$L(t)$' represents the output of hidden layer at time 't'.'w_2' denotes the weight between input layer and hidden layer. Finally, the results are displayed at the output layer.

$$O(t) = w_3 * L(t) \qquad (8)$$

From (8), O(t) represents the output at time instant't'. 'w_3' denotes the weight between the hidden and output layer, $L(t)$ denotes an output of the hidden layer. As a result, the proposed classifier exactly identifies the brain tumour disease with higher accuracy and lessertime consumption. The algorithmic process of Bivariate Multilayer Perceptred Classificationis described as given below,

```
// Algorithm 2: Bivariate Multilayer Perceptred Classification
Input: Number of selected features $sfe_1, sfe_2, sfe_3 \ldots\ldots sfe_n$
Output:Improves brain tumor identification accuracy
Begin
Step 1: Number of featuresat the input layer
Step 2: For each data with selected features $sfe_i$
Step 3:Construct hyper plane '$\beta_h$'
Step4:Find two marginal hyper plane $MH_1, MH_2$
Step 5:If($MH_1 > 0$)then
Step 6: Data point is classified as tumour diseased
Step 7: else if ($MH_2 < 0$)
Step 8:Data point is classified as normal
Step 9:End if
Step 10: Return "normal data or diseased data" at the output layer
Step 11: end for
End
```

Algorithm 2 given above explains the step by step process of bivariate multilayer perceptred classification. Initially, the numberof features is selected from the input database and given as input. After that, input is transferred into the hidden layer. In that layer, the support vector regression classifies the patient data into normal or diseased based on either side of the hyper plane. As a result, the patient data is classified with higher accuracy and lesser time consumption.

IV. EXPERIMENTAL SETTINGS

In order to evaluate the performance, the PPFSBMPC Method is implemented in with Epileptic Seizure Recognition Dataset. The Epileptic Seizure Recognition Dataset is taken from UCI Machine Learning Repositoryfor conducting the experiments. Epileptic Seizure Recognition Dataset comprises 5 folders. Every folder has 100 files. Every file has patient data (i.e. brain activity of patient for 23.6 seconds). PPFSBMPC Method considers different number of patient data in the range of 50-500 from Epileptic Seizure Recognition Dataset. The performance of the PPFSBMPC Method is measured in terms of brain tumour identification accuracy, execution time and error rate as well as compared with existing ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2]. The experiments of PPFSBMPC Method are conducted for several instances with different number of patient data.

V. RESULT AND DISCUSSIONS

The performance result of PPFSBMPC Method is discussed in this section. The efficiency of the PPFSBMPC Method is measured with help of table and graph using the below parameters such asbrain tumour identification accuracy, execution time, and error rate.

5.1 Impact onBrain Tumour Identification Accuracy

Brain tumour identification accuracy (BTIA)is defined as the ratio of number of patient data accurately classified as normal or diseased to the total number of patient data. Brain tumour identification accuracy is determined as,

BTIA= (Number of patient that are accurately categorized as normal or abnormal)/n*100 (9)

From equation (9), the brain tumour identification accuracyis determined. ‘ n ’ denotes the total number of patient datataken. The brain tumour identification accuracyis measured in terms of percentages (%).

Table.1 Number of patient data versus Brain Tumor Identification Accuracy

Number of patient data	Brain Tumor Identification Accuracy (%)		
	ANOVA based feature selection and fuzzy classifier model	Complex-valued classifiers	PPFSBM PC Method
100	89	86	94
200	88	80	92
300	85	80	92
400	86	85	96
500	91	84	94
600	89	87	94
700	89	88	97
800	91	88	96
900	90	88	97
1000	91	89	98

Table 1explains the experimental results of brain tumour identification accuracywith respect to different number of patient data. For the experimental purpose, number of patient data taken ranges from 100 to 1000. Let us consider the number of patient data is 800, brain tumour identification accuracy of proposed PPFSBMPC Method is 96% whereas the brain tumour identification accuracy of two existing ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2]are 91% and 88%. Likewise the variousbrain tumour identification accuracy results are shown in figure 3.

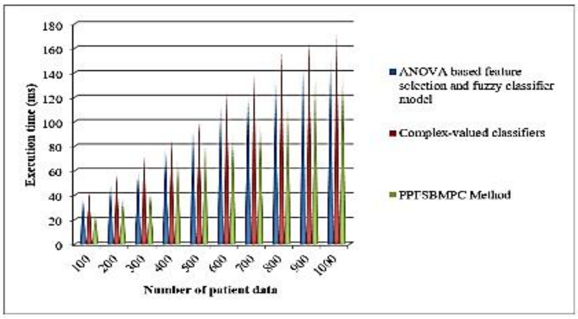


Figure.4 Performance results of Execution Time

Figure 3 explains thatproposed PPFSBMPC Methodincreases the brain tumour identification accuracy than existing methods. This is due to application of projection pursuit feature selection and bivariate multilayer perceptred classification for brain tumor disease identification. Projection pursuit feature selection chooses the related features and removes the unrelated features before performing the classification process. Subsequently, bivariate multilayer perceptred classification categorizes thedata points as normal or tumour diseased data points. The average of ten results illustrates that the brain tumor tumour identification accuracy of PPFSBMPC method is increased by 7% when compared to ANOVA based feature selection and fuzzy classifier model [1]. In addition, the brain tumor tumour identification accuracy of PPFSBMPC method is considerably improved by 11% when compared to Complex-valued classifiers [2].

5.2 Impact on Execution Time

Execution Time '(ET)' computes the time consumed to identify the existence and absence of brain tumor disease. The execution time is determined as,

$$ET = n * t_S \qquad (10)$$

From equation (10), the execution time of brain tumor disease is attained. 'n' denotes the number of patient data. 't_S' symbolizes the time consumed to categorize the presence and absence of brain tumor disease of one patient data. The execution time is measured in terms of millisecond (ms). The performance result analysis of execution time is shown in the Table 2.

Number of patient data	Execution Time (ms)		
	ANOVA based feature selection and fuzzy classifier model	Complex-valued classifiers	PPFSBMPC Method
100	36	42	25
200	48	56	38
300	61	72	45
400	79	85	68
500	92	99	81
600	112	124	92
700	129	139	101
800	135	157	112
900	142	165	135
1000	152	172	145

The PPFSBMPC Method is implemented in Java Language through considering the dissimilar number of patient data in the range of 100-1000 from Epileptic Seizure Recognition Dataset for identifying the execution time involved during brain tumor disease identification. The experimental result of execution time using the PPFSBMPC Method is compared with conventional ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2]. When considering the number of patient data as 700 from Epileptic Seizure Recognition Dataset to perform experimental work, PPFSBMPC Method consumed takes 101ms execution time whereas ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2] consumes 129ms and 139ms respectively. Consequently, PPFSBMPC Method consumed minimum execution time for brain tumor identification when compared to other state-of-the-art methods [1] and [2].

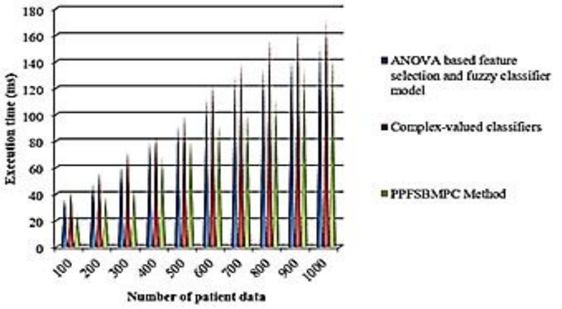


Figure.3 Performance results of Execution Time

The above figure 4 explained that PPFSBMPC Method minimizes the execution time for brain tumor disease identification than conventional techniques. This is because of applying the projection pursuit feature selection and bivariate multilayer perceptred classification for brain tumor disease identification. Projection pursuit feature selection in PPFSBMPC method chooses relevant features and eliminates the irrelevant features before classification process. This in turn helps to minimize the execution time during brain tumor disease identification. After that, bivariate multilayer perceptred classification categorizes the data points as normal or tumour diseased data with lesser time consumption. The average of ten results show that the execution time of PPFSBMPC method is reduced by 17% when compared to ANOVA based feature selection and fuzzy classifier model [1]. In addition, the execution time of PPFSBMPC method is significantly reduced by 26% when compared to Complex-valued classifiers [2].

5.3 Impact on Error rate

Error Rate '(ER)' is defined as the ratio of number of patient data that are incorrectly classified as normal or abnormal to the total number of patient data. The error rate is formulated as,

ER=(Patient that are wrongly classified as normal or abnormal)/n*100 (11)

From equation (11), the error rate of brain tumor disease identification is determined. The error rate is evaluated in terms of percentages (%).The experimental result analysis of the error rate is demonstrated in below Table 3.

Table.3 Number of Patient Data Versus Error Rate

Number of patient data	Error rate (%)		
	ANOVA based feature selection and fuzzy classifier model	Complex-valued classifiers	PPFSBMPC Method
100	11	14	6
200	25	40	17
300	45	60	24
400	58	62	16
500	47	79	31
600	65	79	35
700	79	100	22
800	75	99	30
900	86	105	24
1000	91	110	20

PPFSBMPC Method is executed in Java Language with different number of patient data in range of 50-500 from input dataset to estimate the error rate of brain tumor disease identification. The experimental result of error rate using PPFSBMPC Method is compared with ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2]. When number of patient data is considered as 500 from Epileptic Seizure Recognition Dataset to accomplishing experimental evaluation, PPFSBMPC Method attains 31% error rate whereas ANOVA based feature selection and fuzzy classifier model [1] and Complex-valued classifiers [2] attains 47% and 79% respectively. Therefore, the proposed PPFSBMPC Method attained minimal error rate for tumor disease identification as compared to other existing methods [1] and [2].

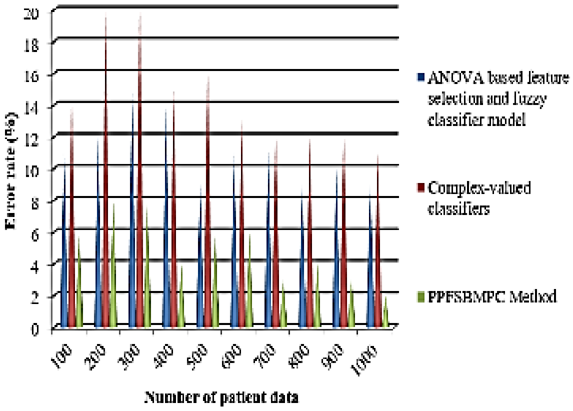


Figure.5 Performance results of Error Rate

The above figure 5 illustrates thatPPFSBMPC Methodreduces theerror rate in brain tumor disease identification. This is due to the application ofprojection pursuit feature selection and bivariate multilayer perceptred classification for performing tumor disease identification. Projection pursuit feature selection in PPFSBMPC method selects relevant features and removes the irrelevant features before performing the classification process. The bivariate multilayer perceptred classification classifies the data points as normal or tumour diseased data. This helps to reduce the error rate during brain tumor disease identification. The average of ten results illustrates that the error rate of PPFSBMPC method is minimized by 55% when compared to ANOVA based feature selection and fuzzy classifier model [1]. In addition, the error rate of PPFSBMPC method is significantly reduced by 67% when compared to Complex-valued classifiers [2].

VI. CONCLUSION

An efficient method termedPPFSBMPC method is introduced for improving the performance ofbrain tumor disease identification with minimal execution time. PPFSBMPC method collects the patient data from input dataset. The relevant feature selection and irrelevant feature elimination is carried out by applying projection pursuit. The projection matrix identifies the similarity between target and the features. The feature selection of PPFSBMPC method is to reduce theexecution time in the brain tumor disease identification. Finally, the data classification is performed using bivariate multilayer perceptred classification to classifythe data points as normal or tumour diseased data. By this way, the error rate during brain tumor disease identification gets reduced. The experimental evaluation performed with different metrics such as brain tumour identification accuracy, execution time, and error rate. The observed result demonstrates that the PPFSBMPC methodenhances the disease diagnosing accuracy and minimizes the execution time as well as error rate when compared to existing classification methods.

VII. REFERENCES

1. Varsha Harpale and Vinayak Bairagi, "An adaptive method for feature selection and extraction for classification of epileptic EEG signal in significant states", Journal of King Saud University - Computer and Information Sciences, Elsevier, Pages 1-9 2018.
2. Musa Peker, Baha Sen and Dursun Delen, "A Novel Method for Automated Diagnosis of Epilepsy Using Complex-Valued Classifiers", IEEE Journal of Biomedical and Health Informatics, Volume 20, Issue 1, Pages 108-118, 2016.
3. Danda Shashank Reddy, Chinta Naga Harshitha and Carmel Mary Belinda, "Brain tumor prediction using naïve Bayes' classifier and decision tree algorithms", International Journal of Engineering &Technology, Volume 7, Pages 137-141,2018.
4. Kostas M. Tsiouris , Vasileios C. Pezoulas , Michalis Zervakis, Spiros Konitsiotis, Dimitrios D. Koutsouris, Dimitrios I. Fotiadis, "A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals", Computers in Biology and Medicine, Elsevier ,Volume 99, Pages 24-37, 2018.
5. S Raghu, Natarajan Sriraam, Alangar Sathyaranjan Hegde, Pieter L Kubben, "A novel approach for classification of epileptic seizures using matrix determinant", Expert Systems with Applications, Elsevier, Volume 127, Pages 323-341, 2019.
6. Yuanfa Wang, Zunchao Li, Lichen Feng, Chuang Zheng, and Wenhao Zhang, "Automatic Detection of Epilepsy and Seizure using Multiclass Sparse Extreme Learning Machine Classification", Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation, Volume 2017,Pages1-10, June 2017.
7. Md. Kamrul Hasan, Md. Asif Ahamed, Mohiuddin Ahmad, and M. A. Rashid, "Prediction of Epileptic Seizure by Analysing Time Series EEG Signal Using k-NN Classifier", Applied Bionics and Biomechanics, Hindawi, Volume 2017, Pages 1-12, August 2017.
8. Md. Faizul Bari and Shaikh Anowarul Fattah, "Epileptic seizure detection in EEG signals using normalized IMFs in CEEMDAN domain and quadratic discriminant classifier", Biomedical Signal Processing and Control, Elsevier, Volume 58, Pages 1-8, April 2020.
9. Sandeep Kumar Satapathy, Satchidananda Dehuri, Alok Kumar Jagadev, "ABC optimized RBF network for classification of EEG signal for epileptic seizure identification", Egyptian Informatics Journal, Elsevier, Volume 18, Issue 1, Pages 55-66,2017.
10. Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, "A Hybrid Feature Selection with Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis", IEEE Access, Volume 4, Pages 9145- 9154, 2016.
11. Lingraj Dora, Sanjay Agrawal, Rutuparna Pand and Ajith Abraham, "Optimal breast cancer classification using Gauss-Newton representation based algorithm", Expert Systems with Applications, Elsevier, Volume 85, Pages 134-145, November 2017.
12. Chih-Jen Tseng, Chi-Jie Lu, Chi-Chang Chang and Gin-Den Chen and Chalong Cheewakriangkrai, "Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence", Artificial Intelligence in Medicine, Elsevier, Volume 78, Pages 47-54, May 2017.
13. Chu-Yu Chin, Sun-Yuan Hsieh and Vincent S. Tseng, "eDRAM: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis", PLoS ONE, Volume 13, Issue 11, Pages 1-19. 2018.
14. Varun Jain and Sunila Godara,

“Comparative Study of Data Mining Classification Methods in Brain Tumour Disease Detection”, International Journal of Computer Science & Communication, Volume 8, Issue 2, Pages 12-17, March 2017.

15. Abeg Kumar Jaiswal and Haider Banka, “Local pattern transformation based feature extraction techniques for classification of epileptic EEG signals”, Journal of Medical and Biological Engineering, Springer, Volume 38, Issue 2, Pages 222-235, April 2018.
16. Mengni Zhou, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo and Jie Xiang, “Epileptic Seizure Detection Based on EEG Signals and CNN”, Frontiers in Neuroinformatics, Pages 1-15, December 2018.
17. Diah P. Wulandari, Nomala G. P. Putriz, Yoyon K. Suprapto and Santi W. Purnami, Anda I. Juniani and Wardah R. Islamiyah, “Epileptic Seizure Detection Based on Bandwidth Features of EEG Signals”, Procedia Computer Science, Elsevier, Volume 161, Pages 568-576, 2019.
18. Zeynab Mohammadpoory, Mahda Nasrolahzadeh and Javad Haddadnia, “Epileptic seizure detection in EEGs signals based on the weighted visibility graph entropy”, Seizure, Elsevier, Volume 50, Pages 202-208 August 2017.
19. Anurag Nishad and Ram Bilas Pachori, “Classification of epileptic electroencephalogram signals using tunable-Q wavelet transform based filter-bank”, Journal of Ambient Intelligence and Humanized Computing, Springer, Pages 1-15, 2020.
20. G. Ravi Shankar Reddy and Rameshwar Rao “Automated identification system for seizure EEG signals using tunable-Q wavelet transform”, Engineering Science and Technology, an International Journal, Elsevier .Volume 20, Issue 5, Pages 1486-1493, October 2017.

ID: 09

AN EXAMINATION ON ADAPTIVE ALGORITHMS

Ms.R.DEEPA
Assistant Professor,
Department of Computer Science,
Nandha Arts and Science College,
Erode, Tamil Nadu, India

Ms.K.SAROJA
Assistant Professor,
Department of Computer Science,
Nandha Arts and Science College,
Erode, Tamil Nadu, India

ABSTRACT

The essential of adaptive filtering arises when the received signal in its course of propagation is corrupted by the noise which deviations continuously. The presence of noise affects the processing and application of the signal. Adaptive algorithms form the principal of these adaptive filters which adjusts its parameters according to the environment to get an optimal signal. This evaluation boons a survey to know the work done on diverse adaptive algorithms LMS, NLMS, RLS which are applied in the fields of signal processing, communication, signal control applications.

KEYWORDS: *Adaptive algorithm, Adaptive filter, LMS, NLMS, RLS.*

I. INTRODUCTION

Adaptive filters are used in the environment where the signal parameters or signal constraints modification continuously because of their toughness and good pursuing capabilities. Adaptive filters are time–variant and its coefficients are familiar in order to optimize the detailed objectives like mean square of the error signal. The algorithms which implement the functionality of the adaptive filters are multifaceted, hence virtually all the adaptive filters are digital filters. There are countless adaptive algorithms developed for various applications, between them Least Mean Square algorithm, Normalized Mean Square algorithm and Recursive Least Mean Square algorithm are discussed in the paper. The research work done on these algorithms which are used for optimized MSE are discussed in this paper.

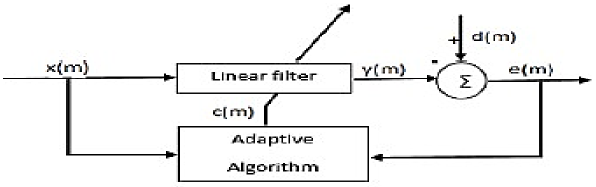


Fig. 1. General block diagram of adaptive filter.

1.1 Adaptive Algorithms

The difference between the conventional filters and adaptive filters is the skill of adaptive algorithms to adjust is filter coefficients deprived of the knowledge of previous signal or noise characteristics.

The reference signal which is consequent from the noise field is used. This reference signal is filtered and deducted from the initial input comprehending both the noise and the signal resulting in an error signal.

x(n)- input signal with noise d(n)-desired signal

e= x(n)-d(n) -error signal

Thus, the adaptive filter parameters are familiar in order to reduce the difference between x(n) and d(n) to make the error as minimum as possible. The three algorithms- LMS algorithm, NLMS algorithm and RLS algorithm have different junction speeds.

II. LITERATURE SURVEY

Different researchers used different algorithms to transmit out the process of noise cancellation. Some of the important research works are reviewed in this paper.

This paper deliberates about the need of adaptive algorithms and different types of adaptive algorithms including LMS, NLMS, RLS algorithms. These algorithms are compared on the basis of MATLAB and Verilog simulation results. The simulation results show that RLS algorithm has the highest merging rate when compared to LMS and NLMS algorithm. LMS algorithm makes use of gradient descent process to search for the optimal condition, with meant square error as the cost function. The slope requires the knowledge of authentic values of the auto-correlation matrix which will be a difficult task to know. RLS algorithm is advantages as it uses instantaneous values of auto-correlation matrix. The MSE for different SNR ratios show that the RLS is the quicker than both LMS and NLMS algorithms which the NLMS is quicker than LMS algorithm in terms of its convergence rate.[1].

In this paper an adaptive noise termination technique using NLMS algorithm in GNU is discussed. Bestowing to this paper the adaptive filters have high noise removal efficiency when related to direct filtering because of its higher noise rejection level. The advantages of NLMS algorithm are discussed in this paper. The filter belongs to the LMS family. The NLMS filter algorithm has the same principle of LMS but the weight control mechanism is different from LMS algorithm. The error estimation is done by subtracting the M-by-I tap input vector from the desired retort. The simulation is carried out in GNU radio mate. A sullied cosine signal and an audio signal are given as an input and the simulation was carried out close-fitting the advantages of noise cancellation with NLMS algorithm. [2].

In this paper drawbacks and limitation of LMS algorithm are conversed. The statistical analysis of mean-square adaptive algorithm with uncorrelated Gaussian data is presented. The analytical expression for study state means square error and performance deprivation due to weight vector mis adjustment are derived. It is found that the adaptive coefficient mu, which controls the rate of convergence of the algorithm essential be restricted to an interval significantly smaller than the domain commonly definite. The outcome of this paper, places the fundamental limitations of mean performance and rate of convergence of LMS algorithm. As a result, the LMS algorithm is very complex the noise and in order to get a momentous convergence rate other adaptive algorithms can be taken into contemplation. [3].

The objective of this paper is to study the tracking capabilities of the time-varying systems by an adaptive filter with LMS algorithm in the incidence of white, zero-mean reference input. The tracking capabilities of time- variations by the LMS algorithm is one of the characteristics while two cases of time variations of a strange system i.e., zero mean time increment and deter mistic bounded increments have been the calculated. [4].

The aim of this paper is to predict the performance of the LMS and NLMS algorithms in chasing the time-varying fields. In order to make the tracking possible a technique is implemented and simulation results show the efficiency of the proposed algorithms. A simple smoothing on the incremented weights is applied to forecast the weights for the next iteration. The trade-off between weight vector noise and lag weight vector by conventional LMS algorithm in time varying situation is modeled by Markov process of order by Windrow. The algorithm has a very short order of arithmetic complexity. Moreover, this procedure could be combined with a wide class of adaptive filters (e.g., RLS, gradient lattice algorithm, etc.) to advance their behaviors. The proposed algorithm is obtained by simplifying a Kalman filter. To this end, a Markov model of second order is considered for the weight vector. This model shows that the assessment of parameter increments inferred from the foreseen parameters improves the tracking performance. [5]

In the paper prediction of stock market movements with the help of adaptive algorithms has been discussed. The advantage of using traditional techniques such as technical analysis and signal handing out techniques such as moving averages and regression were limited due to the dynamic behavior of the markets. In this paper a very innovative approach of applying adaptive filters for the calculation of behavior of financial signals was implemented. The hybrid filters used are DCT-LMS, DCT-NLMS, DCT-RLS and kalman filters. The proposed method is used to predict the values of five of the major stock markets. The Performance of hybrid adaptive filters is equated against the conventional filters like Autoregressive (Ar), Moving Average (MA) filters and adaptive filters like lms, nlms etc. the base technique considered is the Random Walk (RW) process which acts as the yardstick technique. The results show a high degree of prediction precision for the hybrid adaptive filters, which is very high when compared to conventional filters, thus signifying that hybrid adaptive filters can be successfully used for stock market.[6].

This paper compares the adaptive algorithms like LMS algorithm, NLMS algorithm, RLS algorithm, TVLMS algorithm and FTRLS algorithm on the basis of their computational complication and SNR. The adaptive behavior of the algorithms is analyzed. The MSE, algorithm and the required filter direction are criteria on which the algorithms are analyzed. Accordingly, there are various digital signal processing applications where second order statistics cannot be applied. Therefore, adaptive algorithms can be used in those applications. The assessment of the algorithms at a constant sampling rate of 1.5KHz and at different variance values show that RLS algorithm has good SNR enhancement and can be considered as fast converging algorithm. [7].

This paper has studied and calculated the former works done on adaptive algorithms in the field of acoustic echo cancellation. The echo cancellation is considered to be one of the imperious aspects of design of modern communication system. There are primarily two types of echo that are current that are hybrid echo and acoustic echo which are the key area of concern in this paper. Hybrid echo is caused due to the mismatch of impedance in transmission lines where as Acoustic echo is a kind of noise signal in which the audio signal is vibrated in real environment due to the reflection from surrounding objects, walls, floors or surfaces etc. Here along with the original required signal the attenuated, time-delayed images of this speech signal is produced which generates disturbance. Adaptive algorithms are used to remove techno and performance of each echo is evaluated. And the discussion is concluded by proving RLS algorithm has improved echo cancellation when compared to other algorithms.[8].

This paper discusses the use of FPGA systems in apartment of Programmable Digital Signal Processor systems (PDSP) because of their greater suppleness and higher bandwidth due to its parallel architecture. In this debate FPGA is used to instrument the adaptive filters for noise cancellation. The VHDL design of adaptive filters is performed and analyzed on the basis of SNR and MSE. The FPGA platform is well-matched for the complex real time audio processing. An adaptive noise cancellation process has successfully been implemented for filter order up to 256 using Spartan -3 FPGA XC3s400pq208-5 board. When tested with dissimilar signals, the system showed an enhanced performance compared to the original signal. [9].

In this paper the advantages of using NLMS algorithm over LMS algorithm are discussed. The Normalized Least Mean Square error (NLMS) algorithm is maximum popular due to its simplicity. The conflicts of debauched convergence and low excess mean square error associated with a fixed step size NLMS are solved by using an optimal step size NLMS algorithm. The foremost objective of this paper is to derive a new nonparametric algorithm to control the step size and also the theoretical performance analysis of the steady government behavior is presented in the paper. The simulation experiments are performed in MATLAB. The simulation results show that the proposed algorithm as greater performance in Fast convergence rate, low error rate, and has superior performance in noise cancellation. [10].

The goal of this paper is to implement the adaptive digital Least Mean Square (LMS) and delayed LMS (DLMS) Finite Impulse Response (FIR) filters on Field Programmable Gate Array (FPGA) chips for typical noise termination applications and comparability the behavior of LMS and DLMS adaptive algorithms in terms of chip area utilization and the filter critical path time or filter frequency. The direct FIR architecture is measured for filter designing and the VHDL hardware explanation language is used for algorithm modeling. The obtained results by the synthesize tool QUARTUS II on a single STRATIX II chip, EP2S15F484C3, from ALTERA Inc. establish that the DLMS algorithm which has a pipeline architecture is earlier than LMS algorithm while it uses more chip part due to using further registers. [11].

This paper proposes a VHDL execution of a variable step size Least Mean Square (NLMS) adaptive algorithm. The foreseen application is the documentation of an unknown system. The good union of NLMS algorithm has made us to take it. It also has moral stability. Adaptive filtering constitutes one of the core machineries in digital signal processing and bargains numerous application areas in science as well as in industry. Adaptive filtering techniques are used in a wide range of applications, including system identification, adaptive equalization, adaptive noise cancellation, wire less communication and echo cancellation. A HDL implementation is established for a 4th order NLMS adaptive filter. As linked conventional LMS it has been proven that NLMS Algorithm has good comportment. Model simulations grades altogether with schemes obtained in MATLAB prove the same. [12].

COMPARISION OF ADAPTIVE ALGORITHMS			
Algorithm	Speed of convergence	Stability	Robustness
Lms	Slow	Stable	Less
Nlms	Fast	Stable	Less
Rls	Fast	Unstable	More

III. CONCLUSION

This paper has studied the algorithm using MATLAB and VHDL simulation. There are many algorithms modified procedures developed to meet the requirements of the application. The RLS algorithm and NLMS algorithm shows moral results in terms of conjunction and tracking capability. Whereas the LMS algorithm is the humblest and stable algorithm.

IV. REFERENCES

1. Comparative Study of Adaptive Algorithms Using Matlab and Verilog. Parvathy, A and Narayanan, G. 2018, p. 6.
2. Adaptive noise cancellation using NLMS algorithm in GNU radio,. J. Adarsh, P. Vishak and R. Gandhiraj,. Coimbatore, : s.n., 2017.
3. Feuer, E. Weinstein,. Convergence analysis of LMS filters with uncorrelated Gaussian data. s.l. : IEEE Transactions on Acoustics, Speech, and Signal Processing, 1985.
4. Tracking capability of the least mean square algorithm: Application to an asynchronous echo canceller. S. Marcos, O. Macchi., s.l.: IEEE, 1987, Vol. 35.
5. Prediction in LMS-type adaptive algorithms for smoothly time varying environments. Gazor, s.l.: in IEEE Transactions on Signal Processing, 1999, Vols. 47.
6. Application of hybrid adaptive filters for stock market prediction. Nair, Binoy.Mohandas, V.Sakthivel, N.Nagendran, S.Nareash, A.Nishanth, R. Ramkumar, S.Kumar, D. s.l.: International Conference on Communication and Computational, 2010.
7. Comparision between Adaptive filter, yothi, Dhiman, s.l international Journal of science engineeringand technology research(IJSETR),2013, Vol.2.
8. Design and implementation of Adaptive filtering algorithm for IEEE. Mrs. A.B. Diggikar, Mrs. S. Ardhapurkar. s.l.: IEEE,, 2012.
9. FPGA implementation of optimal step size NLMS algorithm and its performance. L.Bharani, P.Radhika,. s.l International Journal of research in engineering and technology (IJRET),2013, Vol.2.
10. “Performance Evaluation of LMS and DLMS Digital Adaptive FIR Filters by Realization on FPGA”. Hesam Ariyadoost, Yousef S. Kavian, Karim Ansari. s.l International Journal of science and Emerging technologies, 2010, Vol.2.
11. FPGA Implementation of NLMS Algorithm for Identification of unknown system. K. R. Rekha, Dr B. S. Nagabushan, Dr K.R. Nataraj,. s.l International Journal of Engineering science and technology, 2010, Vol.2.

ID: 10

FUZZIFIED CRISP RELATIVE CURE HIERARCHICAL CLUSTERING FOR TEMPORAL RELATIONAL DATA ANALYSIS

Dr.L.JAYA SINGH DHAS
Assistant Professor,
Department of Computer Science,
Scott Christian College (Autonomous),
Nagercoil, Tamil Nadu, India.

Mrs.P.S.RENJENI
Assistant Professor,
Department of Computer Science,
VTM College of Arts & Science,
Arumanai, Tamil Nadu, India.

ABSTRACT

A bi-temporal data clustering is a significant solution to the diverse problems for finding the intrinsic structure and compact information over temporal data. The temporal data are collected in the series of particular time periods. The various data mining methods have been developed in the temporal relational data analysis. But the accurate analysis was not performed with minimum time. An efficient technique called Fuzzy Crisp Relative Spherical CURE Hierarchy Clustering (FCRSCHC) is introduced for improving the temporal relational data analysis by partitioning the total dataset into different clusters with minimum time as well as space complexity. The CURE hierarchical structure takes the number of scattered temporal data points in the spherical surface for the clustering. After that, 'k' number of clusters and the representative points (i.e. cluster centroid) are initialized. Then the distance between the representative point and the temporal data point are calculated using spherical coordinates. The minimum distance between the data points are grouped into a particular cluster. Then the fuzzy memberships between the two cluster representative points are calculated based on the distance metric. The CURE hierarchical structure merges the two clusters based on the crisp relation between the representative points. Then, the newly obtained clusters are validated using the silhouette coefficient to identify the data points are close to its own cluster or their neighboring clusters. Finally, the optimal numbers of clusters are obtained and minimize the incorrect data clustering which improves the accuracy. The experimental evaluation is performed using a bitmeporal dataset with various parameters such as clustering accuracy, false alarm rate, clustering time and space complexity. The results show that FCRSCHC technique improves the clustering accuracy and minimize the time as well as space complexity as compared to the state-of-the-art- works.

KEYWORDS: *BI-Temporal Data Analysis, CURE Hierarchical Clustering, Fuzzy Membership, Crisp Relation, Silhouette Coefficient*

I. BACKGROUND AND RELATED WORKS

Bitemporal data analysis is a specific case of temporal database information to handle the sequential data along with two different timelines. Time series analysis is a statistical technique that describes the data in the series of particular time periods for tracking the behavior of a particular phenomenon. Clustering is an essential data mining technique that partitioning the data points into different groups. A variety of clustering algorithms has been developed for a large bitemporal dataset with very various attribute counts. Since the dimensionality of temporal data is considerably larger and more complicated. The research work is concentrated to involve the relevant data object clustering algorithm to overcome the computational problems.

A novel density-based clustering algorithm called chronoClust was developed in [1] for grouping a time-series data. The designed clustering algorithm failed to extend the performance of clustering accuracy. A bi-weighted ensemble approach was developed in [2] for grouping the temporal data using a hidden Markov model. The approach failed to minimize the temporal data clustering time as well as space complexity. An Expectation-Maximization algorithm was designed in [3] for grouping the temporal data. But the clustering accuracy was not improved and also the optimal number of clusters was not obtained.

Clustering a temporal network (ClueNet) was introduced in [4] to cluster the data into different groups. The ClueNet provides higher computational complexity while increasing the temporal network data. A hybrid clustering algorithm was developed in [5] to group the time series data based on similarity. The designed algorithm failed to use any split or merge the clustering algorithm for obtaining the optimal number of clusters. A Fuzzy clustering algorithm was designed in [6] for Spatial-Temporal data based on the autocorrelation. The clustering time was not minimized using the designed algorithm. Clustering of time series data was presented in [7] to measure variation between time series data using copula-based distance measure. But the cluster validation was not performed to achieve the higher accuracy.

A stepwise spatial and temporal clustering strategy was developed in [8] based on the similarity between the data. The designed strategy was not minimized the false alarm rate during the clustering process. A spatiotemporal (ST-OPTICS) clustering technique was developed in [9]. The clustering technique minimizes the time complexity but the performance of space complexity remained unaddressed. A Generalized k –means based clustering technique was designed in [10] for temporal data. Though the technique minimizes the time consumption, the clustering accuracy was not minimized.

A dynamic fuzzy cluster (DFC) algorithm was introduced in [11] for dynamically grouping the time series data. But the algorithm failed to perform the cluster validation. Two incremental fuzzy clustering algorithms were developed in [12] based on a Dynamic Time Warping distance. The clustering algorithms failed to minimize the incorrect clustering of time series data. An HMM-based hybrid meta-clustering algorithm was developed in [13]. The algorithm failed to minimize clustering time.

An Evolutionary Clustering based on Graph regularized Nonnegative Matrix Factorization was developed in [14] to analyzing the temporal networks. But the technique failed to solve the time-consuming problems during clustering analysis. A mixed fuzzy clustering (MFC) algorithm was designed in [15] for grouping the time series data. But the MFC algorithm failed to perform the cluster validation to minimize the false alarm rate.

A Correlation-based clustering of big spatiotemporal data was presented in [16] with less memory capacity. The clustering error was not minimized using the correlation based clustering technique. An automatic similarity measure selection was developed in [17] for clustering the time series data. But the accurate clustering results was not provided since it failed to use the distance measures for grouping the time series data into different clusters. A novel clustering technique was designed in [18] for grouping the spatiotemporal sequence data with respect to density. The designed method minimized the time complexity but the algorithm failed to describe how the total dataset was partitioned into smaller ones and the clusters were merged together.

A cross clustering algorithm was designed in [19] for grouping the time series data and improving the clustering performance. The algorithm minimizes the time complexity, but the setting of the distance parameter in the cross clustering application was not performed to further improving the clustering accuracy. A unified framework was introduced in [20] for clustering the spatiotemporal data with minimum space and time complexity. The framework failed to validate the clustering technique.

The major issues are identified from the above said literature are overcome by introducing a novel clustering technique called FCRSCHC. The proposal contribution is summarized in below subsection on the contrary to existing techniques.

1.1 PROPOSAL CONTRIBUTION

The major limitations of the proposed FCRSCHC technique on the contrary to the existing literature are summarized as follows,

- The FCRSCHC technique is designed to improve the clustering accuracy of temporal relation data and minimize the time complexity using CURE hierarchical clustering algorithm. The algorithm groups similar scattered data points in the spherical surface through the distance measure. The CURE hierarchical structure merges the two clusters by measuring the crisp relation between the two representative points. The crisp relation is defined by the fuzzy membership function.

- To minimize the false alarm rate in the clustering process, the silhouette coefficient based cluster validation is performed. The coefficient effectively measures how each data point is similar in one cluster and the neighboring clusters and thus provides a way to obtain an optimal number of clusters. This helps to correctly group all the similar data points into the particular clusters resulting in minimizes the space complexity.

1.2 Structure of the paper

This paper is organized into five different sections in the following manner. Section 2 provides a detailed description of the clustering with neat diagram. Section 3 illustrates the experimental evaluation using a temporal dataset with different parameters settings. Followed by, the statistical results analysis is presented under the different parameters using various techniques in section 4. Finally, section 5 concludes the paper.

II. METHODOLOGY

The bi-temporal time series datasets comprise a large volume of data and it leads to higher computational vulnerability. In order to solve the issues in the bi-temporal data analysis, Fuzzy Crisp Relative Spherical CURE Hierarchy Clustering (FCRSCHC) technique is developed. The time series data clustering is performed by using the following system model. Given the bi-temporal datasets D_b, the 'n' number of time series data points $DP_1, DP_2, DP_3, \dots . DP_n$ collected at any time 't'. The total dataset D_b partitioned into the number of clusters i.e. $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots \mathrm{C}_\mathrm{m}$ where $C_i \neq C_j$. The clusters are formed by grouping the similar time series data together depends on the distance similarity. The Hierarchy Clustering technique merges the two clusters based on the system similarity. Based on the system model, the FCRSCHC technique is designed.

2.1 Fuzzy Crisp Relative Spherical CURE Hierarchy clustering method

A Fuzzy Crisp Relative Spherical CURE Hierarchy clustering (FCRSCHC) method is developed for partitioning the temporal data objects into different groups. The clustering is the data mining model helps to minimize the similarity between the clusters and maximize the similarity within the cluster. The similar temporal data points are assigned into one cluster whereas the different data points are grouped into other clusters. The conventional clustering algorithm did not work well with clusters of different size and different density. Therefore, the CURE (Clustering Using REpresentatives) is a hierarchical clustering technique used for partitioning the large databases into different subsets. The block diagram of the FCRSCHC method is shown in figure 1.

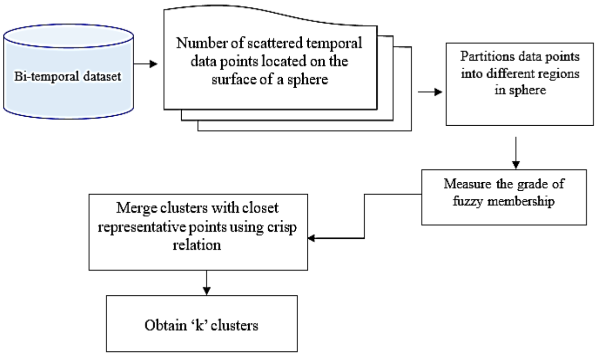

Bi-temporal dataset
Number of scattered temporal data points located on the surface of a sphere
Partitions data points into different regions in sphere
Measure the grade of fuzzy membership
Merge clusters with closet representative points using crisp relation
Obtain 'k' clusters

Figure.1 Block Diagram of the FCRSCHC Method

As shown in figure 1, a block diagram of the FCRSCHC method is illustrated to obtain the optimal 'k' clusters. The bi-temporal data points are collected from the dataset. Initially, numbers of scattered temporal data points are located on the surface of a sphere for clustering. CURE hierarchical structure identifies the well scattered bi-temporal data points $DP_1, DP_2, DP_3, \dots . DP_n$ in the sphere $'S'$. The total sphere (S) is partitioned into different regions (i.e. clusters). By applying a FCRSCHC method, initially 'k' number of clusters and representative points (i.e. centroid) are chosen randomly in surface of the sphere. Each cluster has one representative point. After that, the scattered data points are assigned to the cluster through the distance measure. The distance between the data points in the sphere and the cluster representative points are calculated using the spherical coordinates (r, θ, φ) .

Let us consider the data point $\mathrm{DP_i}$ in the spherical coordinates is $\left(\theta_{\mathrm{D_i}}, \varphi_{\mathrm{D_i}}\right)$ and the representative point R_j of the cluster in the spherical coordinates denoted as $\left(\theta_{R_j}, \varphi_{R_j}\right)$. Therefore the given two spherical coordinates, the distance between them is mathematically calculated using the below equation,

$$d \ \left(\mathrm{DP_i}, R_j\right) = arc\cos(\sin\theta_{\mathrm{D_i}}\cos\varphi_{\mathrm{D_i}}\sin\theta_{R_j}\cos\varphi_{R_j} + \sin\theta_{\mathrm{D_i}}\sin\varphi_{\mathrm{D_i}}\ \sin\theta_{R_j}\sin\theta_{R_j} + \cos\theta_{\mathrm{D_i}}\cos\theta_{R_j} \qquad (1)$$

In (1), $d\ \left(\mathrm{DP_i}, R_j\right)$ denotes a distance between the data point and the representative point of the cluster. In the above equation (1), the coordinates (θ, φ) are mathematically calculated as follows,

$$\theta = arc\cos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \qquad (2)$$

$$\varphi = arc\cos z \qquad (3)$$

In (2) (3), x, y, z denotes a Cartesian coordinates. Based on the distance measure, the FCRSCHC technique groups the data points into the particular cluster with minimum distance between them. In this way, the numbers of regions are formed on the sphere surface with similar data points over the time periods. After partitioning the data points on the sphere surface, the nearest clusters are merged for minimizing the dimensionality of the space which estimates the level of similarity of data points. The interconnections between the two clusters are identified by the fuzzy membership function. The fuzzy membership function finds the neighboring representative points for merging the two clustering set. The membership is computed by the following equation,

$$U_{ij} = \left(\sum_{j=1}^{m} \frac{R_{ij}}{R_{ik}}\right)^{-\left(\frac{2}{p-1}\right)} \qquad (4)$$

In (4), U_{ij} fuzzy membership function, R_{ij} denotes a distance between i^{th} and j^{th} cluster representative points, R_{ik} denotes a distance between i^{th} and k^{th} cluster representative points, p denotes a fuzziness parameter $p > 1$. The membership grade is generally represented by a real number in the closed interval [0, 1] and it indicates the strength of the relation between the two cluster representative points. If the distance between the two representative points are minimum, then the membership value is higher. Otherwise, the membership value is minimized. Based on the membership value, the nearest representative points are combined to form a one cluster through the crisp relation. The following operation is performed to find the crisp relation between the cluster representative points.

$$C_1 \cup C_2 = arg\max\ [U_{ij}] \qquad (5)$$

In (5), $C_1 \cup C_2$ represents the union (i.e. combine) of two clustering sets that contains all of the data points. $\arg max$ denotes a argument of the maximum function, U_{ij} denotes a membership function. The maximum membership functions between the cluster representative points are merged into one cluster in the spherical surface. After merging the clusters, the new representative point is defined i.e. updated for the newly combined cluster. The newly obtained cluster includes the all data points in the two clusters $C_1 and\ C_2$ that were merged. Finally, the well-partitioned data points are shrunken to the cluster centroid. As a result, the hierarchical clustering technique merges the nearest pairs of representative points to form a single cluster which results in minimizes the dimensionality of the cluster in sphere space.

2.2 Silhouette coefficient based cluster validation

After obtaining the 'k' number of clusters, the proposed FCRSCHC technique uses the Silhouette coefficient for validating the data points within the cluster and between the cluster to minimize the incorrect temporal data point clustering. The Silhouette coefficient is mathematically expressed as follows,

$$\rho_s = \frac{(D_n(dp) - D_w(dp))}{\max\{D_w(dp), (D_n(dp)\ \}} \qquad (6)$$

In (6), ρ_s denotes a Silhouette coefficient, $D_w(dp)$ denotes an average distance of data points with respect to all the other data points in the cluster it's assigned. $D_n(dp)$ represents the average distance of data points with respect to all the other data points to the neighboring clusters. The Silhouette coefficient provides the value in the range of -1 to +1. The value '+1' indicates that the data point is very close to the cluster it's assigned and far away from its neighboring cluster. The value '-1' indicates that the data point is close to the neighboring cluster than the cluster it's assigned. As a result, the validation effectively improves the clustering accuracy and minimizes the false positive rate. The algorithmic process of the proposed FCRSCHC technique is described as follows,

```
Input: Bi-temporal dataset $D_b$, number of time series data points $DP_1, DP_2, DP_3, \ldots. DP_n$

Output: Improves the bi-temporal data clustering accuracy

Begin
    Initialize 'k' clusters
    for each cluster
        Select one representative point $R_j$
        Measure the distance $d(DP_i, R_j)$
        Partitions dataset into 'k' clusters with minimum distance
        Calculate the degree of membership $U_{ij}$ between the cluster representative points
        Find the crisp relation $C_1 \cup C_2$
        Merge two clusters based on $arg\max\ [U_{ij}]$
        Update the representative point $R_j{}'$
        Shrink the data points $(DP)$ to the representative point $R_j{}'$
        Calculate the Silhouette coefficient' $\rho_s$'
        if ( $\rho_s = +1$) then
            Data points are correctly grouped into the cluster
        else
            Data points are close to the neighboring cluster
        end if
    end for
    Obtain the optimal k' clusters
end
```

Algorithm 1 Fuzzy Crisp Relative Spherical CURE Hierarchy Clustering

Algorithm 1 clearly describes the bi-temporal data clustering and minimizing the dimensionality as well as the clustering time. Initially, the total dataset is divided into a number of clusters on the spherical surface based on the distance between the data points and cluster representative point. After clustering, the closest representative points in the surface are merged by measuring the crisp relation based on the fuzzy membership grades. Followed by, the new clusters are formed and update the representative points. Finally, the proposed FCRSCHC technique validates the data points within and between the clusters through the Silhouette coefficient. If the coefficient returns +1, then the data points are correctly grouped into the particular cluster. Otherwise, the data points are close to the neighboring cluster. As a result, all the bi-temporal data points are grouped into the cluster resulting in improves the clustering accuracy and minimize the false positive rate.

The above-explained processes are implemented in the experimental using time series dataset in the following section.

III. Experimental evaluation and parameter settings

An experimental assessment of proposed FCRSCHC technique and existing methods ChronoClust [1] and Bi-weighted ensemble approach [2] is implemented using Java language with Activity Recognition from Single Chest-Mounted Accelerometer Data Set. The dataset is collected from the UCI machine learning repository. This dataset comprises the temporal data points from a wearable accelerometer fixed on the chest. The characteristics of the dataset are univariate, sequential and time-Series. The attributes characteristics are real and the association tasks performed by the dataset are clustering and classification. The data points are separated by each participant and each file comprises the sequential number, x acceleration, y acceleration, z acceleration, and labels. The labels are used to provide the clustering outcomes which are represented by the seven numbers. The label 1 to 7 represents the human Activity Recognition such as Working at Computer, Standing up Walking and Going updown stairs, Standing, Walking, Going UpDown Stairs, Walking and Talking with Someone and Talking while Standing.

The performance results of FCRSCHC technique and existing methods ChronoClust [1] and Bi-weighted ensemble approach [2] are evaluated with the different parameters such as,

- Clustering Accuracy
- False Alarm Rate
- Clustering Time
- Space Complexity

IV. Comparative performance analysis under different metrics

The experimental evaluation results of proposed FCRSCHC technique and existing methods ChronoClust [1], Bi-weighted ensemble approach [2] are discussed and the results are compared in this section. The different performance metrics are used in this section such as clustering accuracy, false alarm rate, clustering time and space complexity. The obtained results are discussed with the help of graphical representation. For each section, the statistical calculation is given to show the performance of the proposed technique and conventional techniques.

4.1 Impact of clustering accuracy

Clustering accuracy is referred to the ratio of a number of (no. of) time series data points are correctly clustered to the total number of data points. The mathematical formula for calculating the clustering accuracy is given below,

$$clustering\ \ accuracy = \left[\frac{No.of\ \ DP\ \ correclty\ \ clustered}{n}\right] * 100 \quad (7)$$

From (7), 'n' represents the number of data points $i.e\ DP$. The clustering accuracy is measured in percentage (%).

Sample calculation:

- **Proposed FCRSCHC technique**: No. of data points correctly clustered is 880 and the total number of data points is 1000. The clustering accuracy is calculated as,

$$clustering\ \ accuracy = \frac{880}{1000} * 100 = 88\%$$

- **Existing ChronoClust**: No. of data points correctly clustered is 810 and the total number of data points is 1000. The clustering accuracy is calculated as,

$$clustering\ \ accuracy = \frac{810}{1000} * 100 = 81\%$$

- **Existing Bi-weighted ensemble approach**: No. of data points correctly clustered is 780 and the total number of data points is 1000. The clustering accuracy is calculated as,

$$clustering\ \ accuracy = \frac{780}{1000} * 100 = 78\%$$

Case 1: Let us consider, the number of bi-temporal data taken as input from the Activity Recognition dataset. The ten runs are considered with various input data points. 1000 data points are taken in the first run to calculate the clustering accuracy. The proposed FCRSCHC technique correctly grouped 880 data points into seven different clusters among the 1000 data points. Then the classification accuracy is 88% using FCRSCHC technique. Whereas, the clustering accuracy of existing ChronoClust [1] and Bi-weighted ensemble approach [2] are 81% and 78% respectively. The statistical analysis shows that the proposed FCRSCHC technique improves the clustering accuracy than the other two existing methods. The various results of clustering accuracy are shown in the following graph.

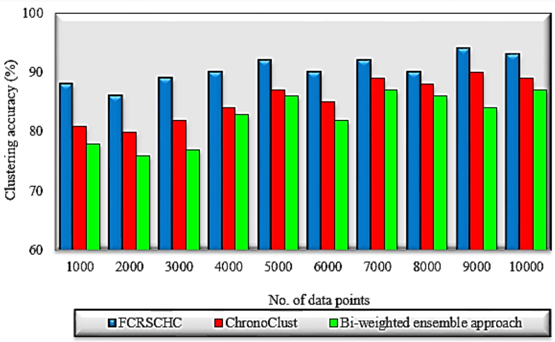


Figure.2 experimental results of clustering accuracy

Figure 2 shows the experimental results of clustering accuracy with respect to the number of temporal data points. The temporal data points are taken in the range of 1000 to 10000. The graphical results confirm that the clustering accuracy of FCRSCHC technique is said to be increased. This improvement is achieved by performing the hierarchical clustering of temporal data points. The proposed FCRSCHC technique initializes the clusters and the representative points in the sphere surface. The minimum distance between each data points and the representative point are grouped into the clusters. Similarly, all data points are grouped into the different clusters. Then the nearest representative points between the clusters are combined. The crisp relation between the representative's points is calculated by the membership grade for merging the two clusters. After merging, the optimal numbers of clusters are obtained. Based on the clustering results, human activities are correctly recognized. As a result, the hierarchical clustering technique correctly groups all the temporal data points into the different clusters.

The observed results of the FCRSCHC technique are compared to the accuracy of the ChronoClust [1] and Bi-weighted ensemble approach [2]. The comparison results confirm that the FCRSCHC technique improves the clustering accuracy by 6% and 10% than the existing methods.

4.2 Impact of false alarm rate

The false alarm rate is measured as the ratio of a number of time series data points are incorrectly clustered to the total number of data points. The mathematical formula for the false alarm rate is given below,

$$FAR = \left(\frac{No.of\ \ DPs\ \ incorrecly\ \ clusterd}{n}\right) * 100 \quad (8)$$

From (8) FAR represents the false alarm rate, DPs denotes a data points, n represents the number of data points. The false alarm rate is measured in the unit of percentage (%).

Sample calculation:

- **Proposed FCRSCHC technique**: No. of data points incorrectly clustered is 120 and the total number of data points is 1000. Then the false alarm rate is calculated as,

$$FAR = \frac{120}{1000} * 100 = 12\ \ \%$$

- **Existing ChronoClust**: No. of data points incorrectly clustered is 190 and the total number of data points is 1000. Then the false alarm rate is calculated as,

$$FAR = \frac{190}{1000} * 100 = 19\%$$

- **Existing Bi-weighted ensemble approach**: No. of data points incorrectly clustered is 220 and the total number of data points is 1000. Then the false alarm rate is calculated as,

$$FAR = \frac{220}{1000} * 100 = 22\ \ \%$$

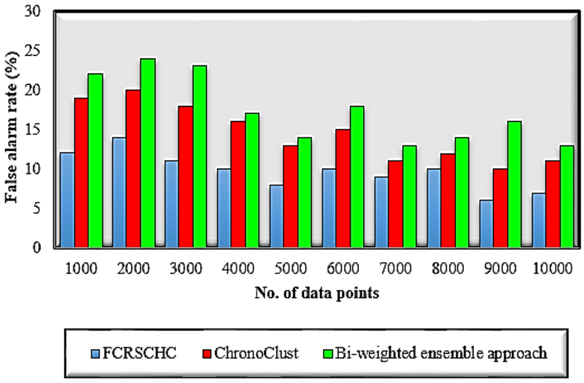


Figure.3 Experimental results of false alarm rate

Figure 3 illustrates the experimental results of false alarm rate versus a number of temporal data points taken in the range from 1000 to 10000. The data point is given to the 'x' axis and the results of the false alarm rate are obtained at the 'y' axis. The above graph shows that the proposed FCRSCHC technique outperforms well and provides accurate clustering results with minimum false alarm rate. The incorrect classification of the FCRSCHC technique is minimized by performing the cluster validation. After obtaining the 'k' clusters, the cluster centroid gets updated for the newly obtained cluster. In this case, the cluster validation is performed through the silhouette coefficient for correctly identifying the cluster members within the clusters to minimize the error rate. The silhouette coefficient measures the average distance of the data points and its cluster center along with their neighboring cluster center. Based on the coefficient values, the cluster members are correctly identified. This helps to minimize the incorrect data point clustering.

Let us consider the 1000 data points as input for grouping the temporal data points to recognize human activities. The false alarm rate of FCRSCHC technique is 12% whereas the false alarm rate of ChronoClust [1] and Bi-weighted ensemble approach [2] are 19% and 22% respectively. Similarly, the nine different results are obtained with various input data points. The observed results of the proposed clustering technique are compared to the existing results. The average is taken for the comparison results and shows that the false alarm rate is considerably minimized by 33% and 44% using FCRSCHC technique as compared to the existing clustering techniques.

4.3 Impact of clustering time

Clustering time is measured as an amount of time taken by the algorithms to group the similar data point into the cluster. The clustering time is mathematically calculated as follows,

$$C_T = No.of\ \ data\ \ points * time\ \ (grouping\ \ single\ \ DP) \quad (9)$$

From (9), C_T indicates the clustering time, DP denotes a data point. The clustering time is measured in the unit of milliseconds (ms).

Sample calculation

- **Proposed FCRSCHC technique**: No. of data points is 1000 and the time taken for grouping single data point is 0.022ms, then the overall clustering time is calculated as follows,

$$C_T = 1000 * 0.022ms \quad = 22ms$$

- ***Existing ChronoClust***: $No. of\ \ data\ \ points\ \ is$ 1000 and the time taken for grouping single data point is 0.025ms, then the overall clustering time is calculated as follows,

$$C_T = 1000 * 0.025ms \quad = 25ms$$

- **Existing Bi-weighted ensemble approach**: $No. of\ \ data\ \ points\ \ is\ \ 1000\ \ and\ \ the\ \ time\ \ taken$ for grouping single data point is 0.028ms, then the overall clustering time is calculated as follows,

$$C_T = 1000 * 0.028ms \quad = 28ms$$

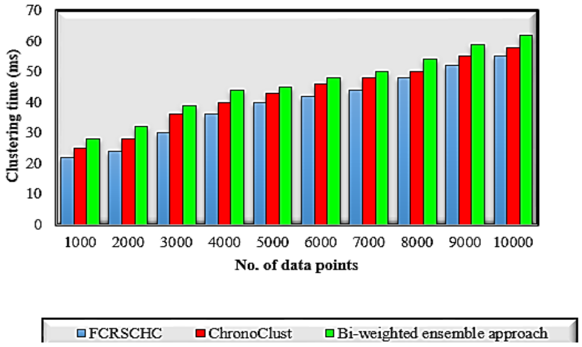


Figure.4 experimental results of clustering time

The experimental results of clustering time versus a number of temporal data points are shown in figure 4. The above figure clearly shows that with the increase in the number of data points, the clustering time is also found to be in the increasing trend. This is due to the application of Fuzzy Crisp Relative Spherical CURE Hierarchy Clustering along with the cluster validation. By applying this, with the temporal clusters, the data points are partitioned into different groups based on time series. The cure clustering technique accurately groups the temporal data points with minimum time. For example, with 1000 data points, clustering time for single data point is '0.022ms' using FCRSCHC technique, where as the clustering time for data points is ' 0.025ms ' using ChronoClust [1] and clustering time for single data point being ' 0.028ms ' using Bi-weighted ensemble approach [2]. Therefore, the overall clustering time was found to be '22ms', '25ms' and ' 28ms ' using FCRSCHC technique, ChronoClust [1] and Bi-weighted ensemble approach [2] respectively. The above statistical analysis proves that the clustering time is minimized using FCRSCHC technique. The average of ten various results shows that the FCRSCHC technique minimizes the clustering time by 9% when compared to ChronoClust [1] and 16% as compared to Bi-weighted ensemble approach [2].

4.4 Impact of Space Complexity

Space complexity is measured to the amount of memory space taken by the algorithms to store similar data points into the cluster. The space complexity is mathematically calculated as follows,

$$space\ \ complexity = No. of\ \ DP * memory\ \ space\ \ (storing\ \ single\ \ DP) \quad (10)$$

From (10), DP denotes a data point. The space complexity is measured in terms of mega bytes (MB).

Sample calculation:

- **Proposed FCRSCHC technique**: No. of data points is 1000 and the memory space taken for storing single data point is 0.01MB, then the overall space complexity is calculated as follows,

$$Space\ \ complexity = 1000 * 0.01MB \quad = 10MB$$

- **Existing ChronoClust:** No. of data points is 1000 and the memory space taken for storing single data point is 0.012MB, then the overall space complexity is calculated as follows,

$$Space\ \ complexity = 1000 * 0.012MB \quad = 12\ \ MB$$

- **Existing Bi-weighted ensemble approach:** No. of data points is 1000 and the memory space taken for storing single data point is 0.013MB, then the overall space complexity is calculated as follows,

$$Space\ \ complexity = 1000 * 0.013MB \quad = 13MB$$

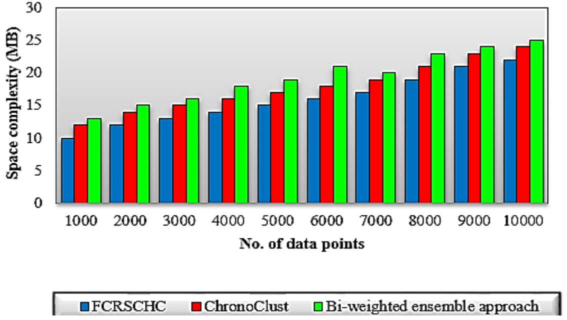


Figure.5 Experimental results of space complexity

Figure 5 illustrates the experimental

results of space complexity with respect to a number of data points using three methods namely FCRSCHC technique, ChronoClust [1] and Bi-weighted ensemble approach [2]. As shown in the above figure, the proposed FCRSCHC technique utilizes the less memory space for storing the temporal data points as compared to the existing clustering technique. This is owing to the application of the Spherical hierarchical structure based clustering for large dimensional data sets. The CURE hierarchical structure identifies the well scattered bi-temporal data points in the spherical surface. In addition, the silhouette coefficient is used to find similar data points within the cluster. This in turn minimizes space complexity. For example, 1000 data points are considered for calculating the space complexity. The FCRSCHC technique consumes 10MB for storing the 1000 temporal data points whereas the 12MB and 13MB consumed by the ChronoClust [1] and Bi-weighted ensemble approach [2]. The result shows that the FCRSCHC technique minimizes the storage space. Totally ten results are obtained for three methods and the results of the proposed technique are compared with the existing results. Then the average value is taken for ten various results. Finally, the average result confirms that FCRSCHC technique minimizes the space complexity by 12% when compared to existing ChronoClust [1]. Similarly, the performance results of space complexity also minimized by 19% when compared to existing Bi-weighted ensemble approach [2].

The above discussion clearly shows that the FCRSCHC technique improves the clustering accuracy and minimizes the time, false alarm rate as well as space complexity when compared to the state-of-the-art methods.

V. CONCLUSION

An efficient technique called FCRSCHC is designed with the aim of temporal relational data analysis through the clustering process. By applying the FCRSCHC technique, the clustering accuracy is improved and minimized the clustering time. The fuzzy crisp relation based hierarchical clustering algorithm groups the similar temporal data points into the different clusters. The hierarchical structure takes the well scattered bi-temporal data points in the spherical surface and groups the data point which is close to the representative point of that particular cluster. After that, the nearby cluster representative points are merged to minimize the dimensionality. Finally, the cluster validation is performed to identify the weather data points are well matched to its own cluster and poorly matched to neighboring clusters. This helps to minimize the false alarm rate and improve the clustering accuracy. The experimental assessment is done using the temporal dataset for recognizing human activities. The various observed results show that the FCRSCHC technique achieves better clustering accuracy and minimum clustering time, false positive rate as well as space complexity than the state-of-the-art-methods.

VI. REFERENCES

1. Givanna H. Putri, Mark N. Read, Irena Koprinska, Deeksha Singh, Uwe Röhm, Thomas M. Ashhurst, Nicholas J.C. King, “ChronoClust: Density-based clustering and cluster tracking in high-dimensional time-series data”, Knowledge-Based Systems, Elsevier, Volume 174, 15 June 2019, Pages 9-26
2. YunYang and Jianmin Jiang, “Bi-weighted ensemble via HMM-based approaches for temporal data clustering”, Pattern Recognition, Elsevier, Volume 76, April 2018, Pages 391-403
3. Hani El Assaad, Allou Samé, Gérard Govaert, Patrice Aknin, “A variational Expectation-Maximization algorithm for temporal data clustering”, Computational Statistics & Data Analysis, Volume 103, 2016, Pages 206-228
4. Joseph Crawford, Tijana Milenkovic, “ClueNet: Clustering a temporal network based on topological similarity rather than denseness”, PLoS ONE, Volume 13, Issue 5, 2018, Pages 1-25
5. Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali, “A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique” The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, March 2014, Pages 1-12
6. Pierpaolo D’Urso, Livia De Giovanni, Marta Disegna, Riccardo Massari, “Fuzzy clustering with spatial–temporal information”, Spatial Statistics, Elsevier, Volume 30, April 2019, Pages 71-102
7. Beibei Zhang and, Baiguo An,

"Clustering time series based on dependence structure", PLoS ONE, Volume 13, Issue 11, 2018, Pages 1-22

8. Xin Yao, Di Zhu, Yong Gao, Lun Wu, Pengcheng Zhang, Yu Liu, "A Stepwise Spatio-Temporal Flow Clustering Method for Discovering Mobility Trends", IEEE Access, Volume 6, 2018, Pages 44666 – 44675
9. K.P. Agrawal, Sanjay Garga, Shashikant Sharma, Pinkal Patel, "Development and validation of OPTICS based spatio-temporal clustering technique", Information Sciences, Elsevier, Volume 369, 2016, Pages 388-401
10. Saeid Soheily-Khah, Ahlame Douzal-Chouakria, Eric Gaussier, "Generalize d k -means-base d clustering for temporal data under weighted and kernel time warp", Pattern Recognition Letters, Elsevier, Volume 75, 2016, Pages 63–69
11. Min Ji, Fuding Xie, and Yu Ping, "A Dynamic Fuzzy Cluster Algorithm for Time Series", Abstract and Applied Analysis, Hindawi Publishing Corporation, Volume 2013, March 2013, Pages 1-7
12. Yongli Liu, Jingli Chen, Shuai Wu, Zhizhong Liu, Hao Chao, "Incremental fuzzy C medoids clustering of time series data using dynamic time warping distance", PLoS ONE, Volume 13, Issue 5, 2017, Pages 1-25
13. Yun Yang and Jianmin Jiang, "HMM-based hybrid meta-clustering ensemble for temporal data", Knowledge-Based Systems, Elsevier, Volume 56, 2014, Pages 299–310
14. Wei Yu, Wenjun Wanga, Pengfei Jiao, Xuewei Li," Evolutionary clustering via graph regularized nonnegative matrix factorization for exploring temporal networks", Knowledge-Based Systems, Elsevier, Volume 167, 2019, Pages 1-10
15. Cátia M. Salgado, Marta C. Ferreira, Susana M. Vieira "Mixed Fuzzy Clustering for Misaligned Time Series", IEEE Transactions on Fuzzy Systems, Volume 25, Issue 6, 2017, Pages 1777 – 1794
16. Marc Hüsch, Bruno U. Schyska, Lueder von Bremen, "CorClustST - correlation-based clustering of big spatio-temporal datasets", Future Generation Computer Systems, Elsevier, 2018, Pages 1-33.
17. Usue Mori, Alexander Mendiburu, Jose A. Lozano, "Similarity Measure Selection for Clustering Time Series Databases", IEEE Transactions on Knowledge and Data Engineering (Volume 28, Issue 1, 2016, Pages 181 – 195
18. Keon Myung Lee, Sang Yeon Lee, Kyung Mi Lee and Sang Ho Lee, "Density and Frequency-Aware Cluster Identification for Spatio-Temporal Sequence Data", Wireless Personal Communications, Springer, Volume 93, Issue No.1, March 2017, Pages 47–65.
19. Wenshuo Zhou, Kuangrong Hao, Chunli Jiang, Lei Chen, Xue-Song Tang, Xin Cai, "A New Cross Clustering Algorithm for Improving Performance of Supervised Learning", IEEE Access, Volume 7, Pages 56713 – 56723
20. Min Deng, QiLiang Liu, JiaQiu Wang, Yan Shi, "A general method of spatio-temporal clustering analysis", Science China Information Sciences, Springer, Volume 56, Issue 10, 2013, Pages 1–14

ID: 11

IN-CITY ARTIFICIAL INTELLIGENCE, ROBOTICS, AND AUTOMATION

Dr.K.SANTHOSH KUMAR
Assistant Professor,
Department of Computer Science and Applications,
Providence College for Women (Autonomous),
Coonoor, Bandishola, Tamil Nadu, India.

ABSTRACT

Globally cities are becoming experimental sites for new forms of robotic and automation technologies applied across a wide variety of sectors in multiple areas of economic and social life. As these innovations leave the laboratory and factory, this paper analyzes how robotics and automation systems are being layered upon existing urban digital networks, extending the capabilities and capacities of human agency and infrastructure networks, and reshaping the city and citizen's everyday experiences. To date, most work in this field has been speculative and isolated in nature. We set out a research agenda that goes beyond analysis of discrete applications and effects, to investigate how robotics and automation connect across urban domains and the implications for differential urban geographies, the selective enhancement of individuals and collective management of infrastructures, the socio-spatial sorting of cities and the potential for responsible urban innovation

I. INTRODUCTION

The fourth industrial revolution of robotic and automation technologies (4IR) is moving from science fiction and R&D to reality. Enabled by vast increases in computing capacity, burgeoning data harvested through powerful algorithms embedded in digital platforms, advanced material developments and urban connectivity – the capability of machines is expanding across all facets of the economy and throughout everyday life.Technological developments open up new horizons for vastly extended application of robotics and automation in production, but also beyond the factory . Building on the digital turn , and mediation of the city through corporate data platforms, the increasing presence of robotics and automation systems will uniquely (re)shape the logics, materialities, practices, processes and affects of the urban context.

There is now a growing body of literature on the implications of aspects of automation and robotics for cities specifically focused on urban surveillance, social and ethnic profiling and algorithmic governance, or the opportunities and challenges presented by driverless cars (Autonomous Vehicles – AVs) and drones (Unmanned Autonomous Vehicles – UAVs). However, the tendency has been to look at discrete applications of these technologies rather than possibilities for more systemic robotic and automated restructuring of the city. There is also scope for urban studies to engage with wider literature on applications that are not specifically city focused, but which have potential urban implications, health care for instance. Increasingly research work is focusing on wider societal implications of robotics and automation including labor, the economy, democracy, and ethics and governance. Yet, so far there has been little focus on perspectives that consider robotics and automation across different urban domains, or examine the full potential (and limitations) of these technologies to address contemporary urban issues Given these gaps, the following paper investigates why researchers, practitioners and policy-makers within the Urban Studies community need to be urgently and critically concerned with these technological developments. Our central argument is that new generation robotics and automation – by which we mean the recent intensification of *in situ* trials and implementation of automation, autonomous systems and robotics within cities – represent a potentially powerful new mode of urban restructuring, where life is shaped by extended and expanded robotic and automation possibilities, and contrasting urban contexts lead to experimentation with distinctive ensembles of technologies. Recognizing the precursor digital infrastructures and urban forms on which this mode of socio-spatial development depends, we advocate that a “whole city” perspective is needed to grasp the full significance of robotics and automation early in their co-evolution.

II. Contemporary robotics and automation: claims and concerns

The permeation of automation throughout modern society is described by four waves of technological development. Electric power enabled the first wave of industrial automation which was exemplified by sophisticated assembly line robots and mass production. Advances from analog electronic and mechanical devices to the digital technologies found in human-machine systems today emerged from the 1980s onwards. Building on growing levels of digital connectivity, new materials and advanced manufacturing, today the “Fourth Industrial Revolution” is marked by technological breakthroughs across; robotics, artificial intelligence (AI), nanotechnology, quantum computing, biotechnologies, the Internet of Things (IoT), 3D printing and autonomous vehicles. 4IR is “characterised by a fusion of technologies that [blur] the lines between the physical, digital, and biological spheres”. Collectively these developments are delivering a “Cambrian Explosion” in automation and robotics, producing technologies with diverse applications, are networked, and work closely with people. Indeed, there is already talk of an Industry 5.0 focussed on “combining human beings’ creativity and craftsmanship with the speed, productivity and consistency of robots”

The role of robotics and autonomous systems in social life is at an early stage and may initially appear restricted, in part because concerns about health, safety and security have limited possibilities for real-world experimentation outside the controlled environment of factories and laboratories. The temptation is to consider these systems as science fiction for some distant future. Yet, large-scale and real-time automated calculative processes already sit behind and inform decision-making across many facets of society, for example; trading, retail, logistics, engineering, transport management, resource extraction, hospital operations, security, etc. Additionally, some applications, such as new forms of surveillance, are undertaken hidden from view. Progress in software, hardware and materials development, pervasive digital WiFi networks, information-gathering ICTs (sensors and remote control capabilities) and global positioning systems (GPS), when coupled with advances in necessary infrastructural support systems, are enabling uniquely new and diverse “Robotics and Autonomous Systems” (RAS) applications throughout economic and social life. Besides having processing power, RAS can sense their environment, make autonomous decisions, and allow for physical activity via actuators. They therefore “enable the collection of data … analysis of data (e.g. via artificial intelligence) and … making of real time interventions in the real world”. We use the established engineering term “Robotics and Autonomous Systems” (RAS) to describe the full complement of automated and autonomous systems, robotics and AI being developed and applied to address contemporary urban challenges.

III. Robotic and automated restructuring of the city

It has been argued that contemporary cities are becoming “a distributed robot, a collection of sensors and functions linked through invisible networks of communication” . The extended and expanded abilities that this enables represent a distinctive shift from previous incarnations of technologically-mediated urbanism. However, attention has so far largely concentrated on appraising the development and anticipated benefits and concerns around; discrete technological developments, such as drone technology; van Wynsberghe, Soesilo, Thomasen, , the future of particular sectors such as mobility or social care , or concerns such as labor or ethics , often failing to connect to wider academic and policy debates or societal issues van Est. Taking a more systemic view, and examining the emerging distributed landscape and role of RAS technologies in cities, we can distinguish between:

1. ***Automated system management*** – technologies that seek to replace, reduce, supplement, enhance, extend and/or simplify human calculation and/or control in the management of systems underpinning the functionality of the city, and
2. ***Robotization of urban services*** – the deployment of mobile-controlled and partially autonomous devices (robots and automatons) that replace, augment

or extend existing approaches by enabling tasks that are too dangerous, repetitive or monotonous for humans, or that require strength or precision beyond human capabilities

Robotics and automation now necessarily form part of a critical urban research agenda. However, Urban Studies currently lack a conceptual lens to understand the rationale and processes through which this phenomenon is constituted, or to anticipate the socio-political implications of RAS implementation. Recognizing this gap, the pace of change and the increasing global (whilst often highly localized) spread of RAS developments, below we propose an analytical framework to map and understand the emergence of contrasting and overlapping modes of robotics and automation. This helps to explain how RAS technologies are (re)shaping, and themselves being (re)made, through urban processes, and to anticipate the wider consequences of this critical contemporary phenomenon (see Figure 1).

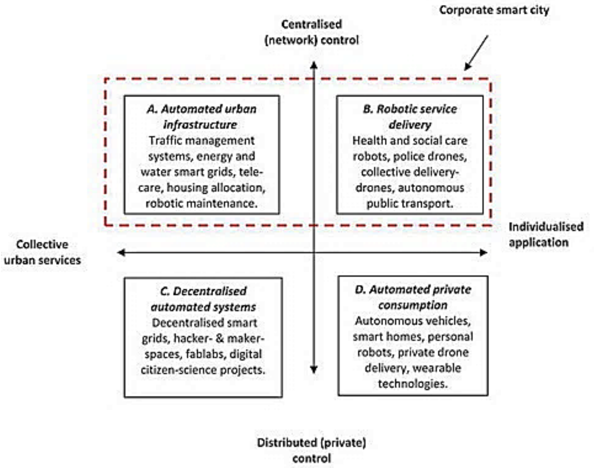

Corporate smart city
Centralised (network) control
A. Automated urban infrastructure
Traffic management systems, energy and water smart grids, tele-care, housing allocation, robotic maintenance.
B. Robotic service delivery
Health and social care robots, police drones, collective delivery-drones, autonomous public transport.
Individualised application
Collective urban services
C. Decentralised automated systems
Decentralised smart grids, hacker- & maker-spaces, fablabs, digital citizen-science projects.
D. Automated private consumption
Autonomous vehicles, smart homes, personal robots, private drone delivery, wearable technologies.
Distributed (private) control

Figure 1. Analytical Framework Mapping the Emerging Logics of Urban RAS Technologies

The analytical framework is based around the strategic positioning of RAS technologies in relation to logics of centralized (network-mediated) and distributed (or private) control, and their ability to address collective urban services (e.g. traffic flow, street lighting, policing) or provide individualized application(s). The *X*–axis contrasts the management of urban infrastructures of resource management for collective service provision (i.e. to enhance systemic efficiency) with RAS technologies that allow for individualized requirements. The *Y*–axis describes the mode of control and extent to which citizens can influence (and opt in or out of) their engagement with RAS technologies (ranging from centralized network control to distributed private control). The grid highlights the distinction between a "corporate" smart city model (top half) and the extension of RAS technologies through private consumption markets (bottom half). The corporate vision is concerned with maintenance and enhancement of city-wide systems, including elements of tailored individualized services. Individual control is strongest in private systems tailored to particular consumers.

3.1. Automated urban infrastructure

In the RAS restructured city, real-time information and predictive analytics present possibilities for radical extensions of networked digital infrastructures. RAS technologies enable complex system management in the context of vastly increased digital data, whilst algorithmic models enhance monitoring and predictive capabilities. Urban flows of people, traffic and resources are thus managed in: automated street lighting, traffic control systems, and surveillance systems. Seeking to minimize disruption, increase efficiency, and optimize network capacity, these developments enable autonomous control systems to manage urban processes with minimal human agency. The automated expert-system is exemplified by the "centralised control room. Which requires subordination of personal choice to the logic of the networked system. We position automated infrastructure system control in the corporate smart city (indicated by a dashed line) because it is concerned with collective management, and in many respects, the UA agenda fits the interests of "smart cities". However, incorporating contemporary developments in RAS, the automated management of urban infrastructure provides a specificity and focus often lacking in the smart city approach.

3.2. Robotic service delivery

Experiments are underway for remotely controlled robotic devices, drones and AVs to provide services across urban domains, for example, in customer service, logistics, policing and security, education and health and social care (see Table 1). These innovations are designed to free up time, save money, provide control, enable precision and strength and to be tailored to our personal needs.

For example, Unmanned Aerial Vehicles (UAVs) (or drones) can exploit urban airspace for the efficient delivery of goods in congested urban environments. And urban robot experimentation in Dubai (UAE) and China reveals how "Authoritarian state regimes are finding innovative ways to use robots in public services, transportation, policing and surveillance". The Dubai Police force has already deployed robot police on the street – "the friendly robot, which can salute and shake hands", can be used by the public to report crimes, complete paperwork and pay traffic fines.

3.3. Decentralized automated systems

Whilst often overlooked, RAS experiments can work through decentralized and perhaps more modest socio-technical community interventions (bottom left quadrant). Often RAS experiments in this quadrant are developed in response to perceived deficiencies in networked infrastructures, limitations in centralized solutions, or urban challenges at localized sites. Typical initiatives range from maker-spaces, hacker-spaces and fablabs, to social media crowd-sourcing campaigns for the community, digital citizen-science projects and digital platforms for urban democracy, and they commonly seek to enhance social capital and empower citizens to act. For example,describes how, as part of the "Making Sense" project, in early 2021, neighbors living around the Plaça del Sol (Barcelona) were provided with automated sensing tools by a group of activists to measure noise levels in a bustling city square, compare them with officially permissible levels with the intention of reducing noise pollution. With this data, noise was no longer a concern affecting individual residential properties, but became a collective issue. Beyond these decentralized automated systems and datasets, however, to mobilize political action to improve networked city infrastructure, old-fashioned, street-level skills in community development, lying beyond the remit of technology vendors, proved essential (*ibid.*).

3.4. Automated private consumption

Many of the technologies mapped under the bottom right-hand quadrant are being developed and marketed by smart corporate technology companies for private consumer markets. RAS experiments in this quadrant relate to user configuration, bespoke products and individualized service provision. Our everyday lives are becoming more intelligent, connected and personalized through integrated AI, voice control and other innovative RAS technologies intended for ease, comfort and convenience. Key examples are; the automated energy and environmental control of houses and consumer goods, virtual home assistants (e.g. Amazon Alexa, Google Home, Samsung's Bixby), personal robots and biometric wearable technologies, autonomous vehicles and automated domestic security systems. Whilst within this realm there will be different degrees of individual control, these technologies are about satisfying individual preferences and maintaining personal urban (social and ecological) security. Thus, Atkinson describes how automated domestic security systems (including alarms and panic buttons) are relied upon by the super-rich to ensure "private withdrawal and shielded mobility" from the unpredictability of the city, leading to segregated enclaves. Contrastingly Google Home and Google Assistant claim to help those with physical disabilities feel included and gain autonomy in their daily activities by "adapting the last mile to the special ability each one of us has".

Issues in this fourth quadrant relate to; the state regulation of privatized systems that may exclude or discriminate (e.g. security systems or smart energy technologies). For example, Strenger uses the "Resource Man" analogy to describe the goal underpinning many smart energy technologies – for householders to act as active micro-resource managers, while also passively assigning management to smart technologies. Not only is this assumption heavily gendered and rationalizing, but smart technologies also raise concerns in terms of data ownership and security. Additional concerns include possible conflicts between personal and collective networks of provision (e.g. delivery drones), and the extent to which personalized and private investments (i.e. the private smart city as opposed to the corporate smart city) reinforce social and spatial inequalities.

IV. A RESEARCH AGENDA FOR URBAN ROBOTICS AND AUTOMATION

Urban contexts are being actively targeted for RAS applications yet we minimally understand the emerging metropolitan landscape and what these developments might cumulatively and strategically mean for city life. A critical research agenda would focus on four sets of issues.

4.1. The urban geography of RAS implementation

The transitional phase of urban restructuring through RAS technologies is unlikely to affect all cities in the same way at the same time. Urban responses will be shaped by the capacity of places to invest in new technologies and the capability and willingness of urban leaders to facilitate extended applications of expensive RAS infrastructure networks in different domains. Technology firms and consultants might also prefer to operate in some cities responding to particular urban challenges, regulatory frameworks, governmental support or reputational value.

Key international sites where RAS interventions have become a strategic innovation and infrastructure priority for national governments include those striving to be the most – “smart” (Dubai & Singapore); robotized (Japan, South Korea); artificially intelligent (China); and AV-friendly. Notably, the RAS industry is dominated by corporates from developed countries, whilst wealthy nations can explore future markets; this leads to questions about how urban RAS technologies, labor and wealth will be distributed at a geopolitical level. For example, recent research examines how Kenya, Tanzania and Rwanda have allowed the use of commercial drones leading robotics companies from Europe and the US to intensively enter the (deregulated) African aerial space.

4.2. Selective RAS enhancement of citizens and infrastructure networks

It is necessary to investigate how urban RAS interventions reconfigure boundaries between human and machine decision-making and how this re (shapes) our capabilities and opportunities as individuals, part of society and urban networks, and within urban governance regimes. At the individual level, new combinations of humans, technologies and infrastructures create synthetic hybrids that are remaking ideas about what it means to be human (and the role of infrastructures). How will RAS-mediated processes and practices affect human skills and qualities such as patience, concentration, empathy, and what will this mean for our relations with one another? For example, research has found that smartphone dependency is already resulting in a diminished attention span, reduced memory and amplifying the need for instant gratification. As machines become increasingly intelligent and lifelike we need to understand; the moral, ethical and legal dilemmas this poses in terms of machine agency, how to delegate problem-solving and decision-making, whether we choose to actualize things usually beyond human capabilities, and implications in terms of human (and robot) rights (e.g. an AI humanoid robot was recently granted citizenship in Saudi Arabia.

4.3. RAS and the social sorting of cities

Third, urban research needs to examine how RAS technologies reinforce and replicate existing socio-spatial inequalities, segregations and exclusions, or create opportunities for more inclusive, socially-just and sustainable cities. Many aspects of RAS technologies have a promissory dimension in that they claim to be concerned with the making of new types of innovative future cities. Such socio-technical expectations shape urban applications of RAS, steer public and private investment, and influence the construction of markets and forms of regulation. Critical to understanding this is to examine whether urban RAS trials constitute global testbeds of living laboratories within existing infrastructural networks capable of enabling metropolitan transformation across urban geographies, or are leading to enclaves of experimentation and premium service delivery.

V. CONCLUSION

The urban context is now a critical site for the experimentation of new generation robotics and automation. The justification for this emergent trend is the claim that RAS technologies offer a superior form of decision-making and acting in the city that will help improve urban efficiency, augmenting and improving urban infrastructure, healthcare and everyday life. However, while RAS restructuring may bring significant benefits to aspects of urban life (and at least to some residents within some cities), it has the potential to reinforce and accelerate disadvantages and disparities within and between cities. It is critical that urban research subjects claims of the potential of RAS technologies to urgent scrutiny, so that informed choices can be made about which aspects of urban life should be automated or roboticized, and these developments can be managed and controlled within a framework of equitable and responsible innovation.

VI. REFERENCES

1. Acemoglu, D, & Restrepo, P (2017). *Robots and jobs: Evidence from US Labor Markets*. NBER Working Papers 23285, National Bureau of Economic Research, Inc. [Google Scholar]
2. Angin, J, Larson, J, Mattu, S, & Kirchner, L (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Google Scholar]
3. Aravindan, A, & Geddie, J (2018). Singapore to test facial recognition on lampposts, stoking privacy fears. *Reuters World News*. Retrieved from https://www.reuters.com/article/us-singapore-surveillance/singapore-to-test-facial-recognition-on-lampposts-stoking-privacy-fears-idUSKBN1HK0RV [Google Scholar]
4. Ardiny, H, Witwicki, SJ, & Mondada, F (2015) *Construction automation with autonomous mobile robots: A review*. 3rd RSI International Conference on Robotics and Mechatronics (ICROM), 7–8 October. Tehran, Iran: Tarbiat Modares University. [Crossref], [Google Scholar]
5. Ash, J, Kitchin, R, & Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, 42(1), 25–43. [Crossref], [Web of Science ®], [Google Scholar]
6. Atkinson, R. (2016). Limited exposure: Social concealment, mobility and engagement with public space by the super-rich in London. *Environment and Planning A: Economy and Space*, 48(7), 1302–1317. [Crossref], [Web of Science ®], [Google Scholar]
7. Bamburry, D. (2015, July 23). Drones: Designed for product delivery. *Design Management Institute (DMI) Review*. Retrieved from https://www.dmi.org/. [Google Scholar]

ID: 12

PERFORMANCE MEASURE OF SOCIAL MEDIA & CHANGING COMMUNICATION PATTERNS IN DIGITAL WORLD

Ms.P.THENMOZHI
Assistant Professor,
Department of Computer Science,
Nandha Arts and Science College,
Erode, Tamil Nadu, India.

Mr.P.MOHANRAJ
Assistant Professor,
Department of Computer Science,
Nandha Arts and Science College,
Erode, Tamil Nadu, India.

ABSTRACT

New media as a democratic means of communication, theprospect it being most indispensable part of our lives is not really far away. In this perspective, paper attempts to study the changing communication patterns of 21st century tech savvy generation. It has been argued that new media has brought sea changes in the Indian economy. Once upon a time traditional media was setting agenda of public discourse is lookingforward to new media for breaking news. In the absence of a proper content regulatory authority new media isdiminishing the gate keeping function in media thus making it more participatory yet less authentic in terms ofcontent. The paper explores how newmedia is redefining social roles and bringing about a change in the Indian economy. Author concludes that in the age of over communication a new kind of social order is being developed that is strengthening public and mass communication. It is also contributing towards the economy of the country through advertisements.

KEYWORDS: *Advertisements, Twitter, Social Media, Chatbot, Communication patterns.*

I. INTRODUCTION

In the age of Information and Communication Technologies whosoever is unaware of these terms isconsidered illiterate or outdated. Once obsessed with capital intensive mass media technologies developmentplanners as well as advertisers are finding new means and tools to reach out their tech savvy target audience. Educators are engrossed in new means to use the updated technology in their class rooms. In an age when there are one billion user sinter-connected through facebook this is one of the most powerful participatory medium of communication. In this scenario however is educated and is having means to use this simple technology can't ignore it.

Social media is emerging as a most vital tool of different kinds of communication which is equipped with the ability to share information, mouldopinion, connecting individuals and communities and tool of active participation. Social Media refers to the means of interactions among people in which they create, shareandexchange information and ideas in virtual communities and networks (Ahlqvist, & Halonen, 2008). And reasKaplanand Michael Haenleinde fine social mediaas "a group of Internet-based applications that are built on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content (Kaplan, &Haenlein, 2010). Furthermore, social media is available on mobile and web-based technologies to create highly interactive platforms through which individuals and communities share, co-create,discuss, and modify user-generated content. It introduces substantial and pervasive changes to communication betweenorganizations, communities and individuals (Kietzmann, & Hermkens, 2011).

In India, the number of people actively using Social Media is about 66 million but it is growing faster with theavailability of cheaper broadband connection and internet enabled handsets at very low prices. Social media istransforming the contours of social interaction. Emotions like love, friendship, family bonding, intimacy and language and are finding various platforms and forms of expression.

Interaction through SocialMedia

Social Media was created for social interaction and it is serving its purposefully. It has enabled the communication, faster, cheaper and anytime anywhere. This very feature which seems to be useful is becominga problem in interpersonal communication. People are virtually connected all the time with updated status on social media sites. But due to over obsession of up date invirtual world is leading to not communication in real world. Most of the time netizens are so much engrossed in their virtual communication they hardly find time totalk to those near and dear one who are present in their physical surroundings. In the world of over-communication we are moving away from the real world and living a virtual world of fake identities andintimacies. Those who are not really using them due to lack of knowledge, access or disinterest find themselvesalienated.

Not only it is affecting the quantity and quality of communication but itisal so affecting the kind of language weare using in our informal and formal written communication. Like on Twitter one can post messages in limitednumber of characters, netizens have innovated a variety of acronyms. This is affecting the language of present generation in various ways. Students have started using these acronyms in their examination answer sheets and many of them have forgotten the actual words for those.

Advertising through social media

In comparison to the age of mass communication which is more than four centuries old social media is ininfancy. Despite the fact, it is influencing traditional mass media in every aspect – programmed formats,content, treatment and language. Most of the big newspapers, TV channels and radio channels monitor the content of popular social networking sites Facebook and Twitter to get the latest update of happening around the world. The 24X7 update of the information on various news sites is giving immense competition to traditionalmedia. Despite the rise in competition, new media and traditional media are shaking hands in reaching the masses and reinforcing the communication messages in the interest of target audience.

Earlier journalists were dependent on the discretion of editors for the space and time they were getting innewspapers and TV channels, now new media is giving them more space in terms of blogs. A few of them areleaving their jobs and are sticking to blogging where not only they are getting more readers and audience but they are earning advertising revenue sans interference of management.

Traditional advertising vs. advertising n social media

Traditional advertising on platforms like televisions, newspapers, magazines, hoardings etc. have a limited reach. On the other hand social media platforms have a wider reach. They can be accessed even by semi-literate or literate people. Anyone who has access to a smart phone an knows how to use the internet can get access to social media advertisements and can be tapped as a potential customer. The hard-core readers of paper editions of newspapersused to say that in 21st century people will stopsubscribing newspapers as news is available free of cost on different sites. The buzz createdon social media sites sets the agenda for business through social media advertising. To some extent traditional advertising is suffering the wrath of higher costs. Latest audience research survey of the media used to collect data from opinion polls and surveys has mentioned it its findings that nowmass advertising is more effective through social networking sites that can be instantly used and used fromdifferentperspectives.

Socioeconomics- The changing economy in India through social media advertising

No other media has become so popular in such a short period of time as social media. Credit of this tremendoussuccess goes to user friendly features of Social Media. According to Nielsen, internet users continue to spendmore time with social media sites than any other type of site. At the same time, the total time spent on socialmedia in the U.S. across PC and mobile devices increased by 37 percent to 121 billion minutes in July 2012 ascompared to 88 billion minutes in July 2011(State of media report, 2012). It has been observed that Facebook isnow the primary method for communication by college students in the U.S (Harris, 2008).

According toNielsen, global consumers spend more than six hours on social networking sites. "Social Media Revolution"produced by *Socialnomics*author Erik Qualman contains numerous statistics on social media including the factthat 93% of businesses use it for marketing and that if Facebook were a country it would be the third largest (Youtube 2011). Several colleges and universities such as Harvard, Johns Hopkins, Columbia and Stanfordamong others have even introduced classes on best social media practices, preparing students for potential careers as digital strategists.

Various functions performed by different techniques and tools used by media are social networking, webpublishing, virtual reality, creating virtual relationships, Interpersonal and community interaction, sharing of text, sounds and AVs, Gaming etc. With the advent of Web3.0 technology more features and sophisticated tools are being added.

Reasons of the growth of the Indian Economy and the use of Social Media in India

No other technology has seen such an unprecedented growth, the telecommunication has seen. Now, thequestion arises, why the India, which is considered to be developing economy, is experiencing the tremendousgrowth inthissector?Thereason liesinthefollowingfactors

1. Booming economy - The economy of India is the tenth-largest in the world by nominal GDP and thethird-largest by purchasing powerparity (PPP) (Wikipedia, 2013). The country is one of the G-20majoreconomies anda member of BRICS. On aper-capita-income basis, India ranked 141st bynominal GDPand 130th by GDP (PPP) in 2012, according to the IMF (Wikipedia, 2013). India is the 19th largest exporter and the 10th largest importer in the world. The economy slowed to around5.0% for the 2012–13 fiscal year compared with 6.2% in the previous fiscal (Wikipedia, 2013). Thesefacts support the fact that Indian middle and lower class is capable of spending on consumer goods more and more at present.
2. Rapid expansion in country's middle class – With one of the fastest growing economies in the world, clock edata growth rate of 8.3% in 2010, India is faston its way to be coming a large and globally important consumer economy. The Indian middle class was estimated to be 250million people in 2007, by McKinsey & Company (Wikipedia, 2013). It will reach 600million by 2030. According to Deutsche Research the estimates are nearly 300 million people for all Middle Class (Wikipedia, 2013). If current trends continue, Indian per capita purchasing power parity will significantly increase from 4.7 to 6.1 percent of the world share by 2015 (Wikipedia, 2013).
3. Inexpensive Technology – According to AC Nielson's *The Social Media Report 2012* assesses that, "More people are using smart phones and tablets to access social media with more connectivity, consumers have more freedom to use social media wherever and whenever they want."
4. Telecom expansion Telephony introduced in India in 1882. The total number of telephones in thecountry stands at 960.9 million, while the overall tele-density has increased to 79.28% as of May 31, 2012 and the total numbers of mobile phone subscribers have reached 929.37 million as of May 2012(Wikipedia, 2013). The mobile tele-density has increased to 76.68% in May 2012. In the wirelesssegment, 8.35 million subscribers were added in May 2012 (Wikipedia, 2013). The wire line segmentsubscriberbasestoodat31.53million (Wikipedia, 2013). Indian telecom operators added a staggering 227.27 million wireless subscribers in the 12 months between Mar 2010 and Mar 2011 (Wikipedia, 2013). According to Internet and Mobile Association of India (IAMAI) report, the number of active social media user base in India is 32.5 million (82% of active mobile internet base) (IAMAI, 2012).
5. Internet expansion - The report, 'Internet in Rural India', prepared by IAMAI and the Indian Market Research Bureau, states rural India has 38 million claimed internet users and 31 million active internetusers (Business Standard, 2013). Active users are those who access internet at least once a month, while claimed users are those who have used internet atleast oncein their life time.

Current Trends in Socialmedia advertising & its growth

Despite a billion users are register edon Social networking sites yet as per its social media is still an infant. Once a popular networking site 'Orkut' it has lost its identity in virtual world. Samemay happen to other sites as well. Facebook is considered more a social burden than a loved networking site by young users. A recent study byPew research Centre on teen agers has found that Facebook was losing its crucial demographic, which longfueled to its success. According to *Huffington Post,* teens are increasingly abandoning Facebook and turning toInstagram, Twitter, MySpace where they tend to have more privacy. Teens consider that site has become over run by parents and leads toun necessary 'social drama' (Hindustan Times, 2013).

For Start-up companies, social media advertising tends to be affordable and an efficient way of reaching out to people.British-American entrepreneur and author Andrew Keen criticizes social media in his book *The Cult of theAmateur*, writing, "Out of this anarchy, it suddenly became clear that what was governing the infinite monkeysis now putting away on the Internet was the law of digital Darwinism, the survival of the loudest and most opinionated. Under these rules, the only way to intellectually prevailis by infinite filibustering" (Keen, 2007).

CONCLUSION

A big source of information, education, communication and of course entertainment social media is going to benext big challenge for the different aspects of our social and personal lives. Social media is creating a new kindofsocialorderit is strengthening social network but weakening inter-personal relations. User-generated content will lead to more plagiarism i.e. new challenge for IPR, a new form of addiction for younger generation (difficult to rehabilitate), an age of over communication (24hrs messaging and chat) and a lie nation on the other hand (leading to psychological disorder). Degradation of ethics in various sectors may provea cultural shock for the generation of 20th century. Mass media's dependency on new media may lead to new media's supremacy, leading to more representation and more participation in publics phere.

REFERENCES

1. Ahlqvist,Toni,Bäck,A.,Halonen,M.,Heinonen,S.(2008).Socialmediaroadmaps exploringthefuturestriggeredbysocial media.VTTTiedotteita-ValtionTeknillinenTutkimuskeskus(2454):13.
2. Kaplan AndreasM., Haenlein Michael. (2010). Users of the world, unite! The challenges and opportunities of social media, Business Horizons, Vol.53, Issue1 (page61)
3. Kietzmann, H., Jan, Hermkens, Kristopher. (2011). Socialmedia? Getserious! Understanding the functional building blocks of social media. Business Horizons Vol54:241–251.
4. Stateofthemedia: The social media report 2012. Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved December 9, 2012 from http://blog.nielson.com/nielsonwire/social/2012
5. Harris, Kandace (2008). Using Social Networking Sitesas Student Engagement Tools. Diverse Issues in Higher Education Vol.**25**(18).
6. Social Media Revolution Video.Youtub.Retrieved June 22, 2012 from http://www.youtube.com/watch?EconomyofIndia.RetrievedJune13,2013fromhttp://en.wikipedia.org/wiki/Economy_of_India
7. Standard of living in India. Retrieved June 13, 2013 from http://en.wikipedia.org/wiki/Standard_of_living_in_India
8. Tele communications statistics in India. Retrieved June 13, 2013 from http://en.wikipedia.org/wiki/Telecommunications_statistics_in_India
9. Internet to reach 45 mn in rural India by year end – Study. Retrieved June 13, 2013 from http://www.business-standard.com/india/technology/news/internet-to-reach-45-mn-in-rural-india-by-year-end-study/484965/
10. Teen Rage on facebook is ebbing, says new study. (2013, May 13). HindustanTimes, pp.25Keen,Andrew(2007).TheCultoftheAmateur.RandomHouse.p.15.

ID: 13

BIG DATA ANALYTICS FOR INTERNET OF THINGS

Mr.R.MURUGESAN
Head,
Department of Computer Science & Engineering,
Paavai College of Engineering,
Namakkal, Tamil Nadu, India.

Dr.M.DEEPA
Assistant Professor,
Department of Computer Science
Paavai Arts and Science College,
Namakkal, Tamil Nadu, India.

ABSTRACT

The rapid advancement of communication and technological advances made possible by the Internet has expanded connectivity between various machines and sensor-based technologies. The notion of IoT (Internet of Things) was created as a result of these equipment or gadgets being connected by way of the global web. These monitoring gadgets generate large amounts of Big Data every day. This information may be analyzed to find solutions to a variety of everyday issues. Fortunately constrained information technology, storage, and computing capabilities at the IoT device end make massive analysis of data difficult. The various Big Data methods and strategies that may be applied to IoT frameworks are discussed in this study. It also demonstrated a method for using Big Data to evaluate IoT data sets in an intelligent manner. It is clarified which of the several Big-data analytics systems is ideal for IoT data after a thorough explanation of each platform.

KEYWORDS: *Big data, Frameworks, Internet of Things (IoT), Architecture, Big Data Analytics (BDA)*

I. INTRODUCTION

Big data is a term used to describe an accumulation of extremely large and intricate data sets and data volume, which also includes current time information, social network analytics, and vast amounts of data. The practice of studying a lot of data is known as big data analytics. There are vast volumes of diverse digital data. Large data sets are measured in terms of terabytes or petabytes, and big data is all about data volume. Big data is the name for this phenomena. Big Data analytics have been introduced as a result of big data analysis. The 5Vs of big data, as well as the method and technology utilized when dealing with big data, are presented in this study. Preserving, analyzing, storing, finding, sharing, visualizing, conveying, and privacy breaches are among the difficulties.

The Internet of items (IoT) is a network of interconnected items or devices (computing, mechanical, or digital equipment) that can link these things or processes without the need for human intervention. Machine to machine (M2M) communication is taking place. IoT gathers data in many formats and from various sources, which is why it is referred to as diverse data. IoT may gather information from the healthcare sector, homes with smart appliances, intelligent traffic management, railroads, airplanes, weather forecasting systems, agricultural sensors, and many more sources.

II. V's of BIG DATA

Understanding the 5 V's enables data engineers to get more value out of their data and helps the business where they work grow more customer-centric. The five primary and fundamental qualities of big data are the five V's: velocity, volume, value, variety, and veracity.Big data was only ever discussed in terms of the three Vs, or volume, velocity, and variety, in the early years of this century. Value and veracity are two additional Vs that have been developed throughout time to aid data scientists in better expressing and conveying the key features of big data.

- **Volume:** Volume, or the initial size and quantity of data that is collected, can be thought of as the foundation of big data. Big data can be used to describe a set of data that is sufficiently massive. However, the definition of big data is subjective and subject to change depending on the market's supply of computer power.

Figure.1: 5V's of Big Data

- **Velocity:** Data generation and transmission rates are referred to as velocity. For businesses who require their data to flow fast so that it can be accessed when needed to make the best business choices, this is a crucial factor.
- **Variety:** Big data, as defined, is a big volume of diverse data. So diversity is undoubtedly a key component of large data. Today, a variety of data formats (structured, semi-structured, or unstructured) are available across data generating devices. This data may occasionally not be in the format that is anticipated. The processing of the data might be hampered by this unexpected format. Any firm should have a data storage system that can analyze and handle any type of data, regardless of its structure, in order to eliminate these issues.
- **Value:** Big Data is often produced as a result of continuous data creation. Until or unless it appears to have any worth, this data is useless. Therefore, the value of the data is undoubtedly a crucial component of big data. The useful data that various gadgets supply to the analyst or data scientist today is the foundation of big data analytics, which has now become a crucial component of society. Big data is not necessarily required to be valuable.
- **Veracity:** Veracity does not concern the volume of information. It is a part of the data that Big Data offers its customers that is easy to interpret. Any company handling a lot of data should purge "dirty information" prior to it builds up in the systems.

III. INTEGRATION OF BIG DATA AND IOT

Everything is intertwined with technology in today's lifestyle. IoT is quickly taking up in a variety of businesses. The Internet of Things (IoT) comprises of gadgets that gather data and link to the outside world using that data. We may utilize this information to aid in a variety of ways to solve our research problems. Numerous big data analytical tools and approaches can be useful for this data analysis. Big Data and the Internet of Things are viewed as two sides of the same coin. The relationship between IoT and big data analytics may be shown in the following Figure.

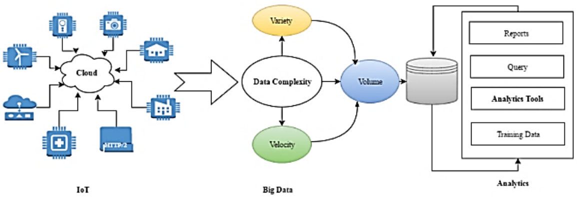


Figure 2: Relationship between Big Data and IoT

Steps for Big Data processing in IoT

The method of managing IoT Big data may be broken down into four parts, which are explained below.

- Managing various IoT data sources, i.e., IoT sensor devices, is the first stage. Monitors in a device cooperate with one another with the aid of various applications, producing highly informal, partially structured, or organized data.
- In the second stage, a big data storage system gathers and stores big IoT data, which is created by various IoT devices. This information is based on Gartner's 5V model.
- This IoT data is transformed into distributed and spread Big Data documents and stored in big data storage systems, afterwards that, it utilizes several analytical tools, like as Hadoop, MapReduce, Spark, and many more (which are further explained in the next part), to analyze data.
- The summary report created and displayed to the user corresponds to the injected data in the final stage of the process.

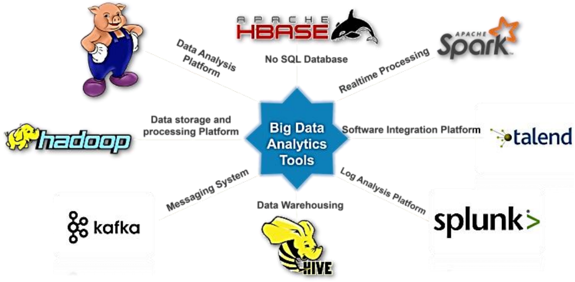
IV. BIG DATA ANALYTICS TOOLS

Big Data analytics is becoming a crucial tool for businesses of all kinds operating in a variety of sectors. Businesses are now able to learn things about their clientele, their markets, and the environment beyond them that were previously impossible by utilizing the power of big data. To convert IoT structured, semi - structured, and unregulated data into descriptive or complete form for the analysis that follows process, big data analytics needs certain tools and procedures. These technologies make use of algorithms to find trends, correlations, and patterns across different types of data. Several tools are used for displaying the results of data analysis in the form of graphs, tables, pie charts, bar charts, etc.

The list of top Big Data analytics tools with well-liked features is provided below.

- Apache Storm
- Talend
- CouchDB
- Apache Spark
- Splice Machine
- Plotly
- Azure HDInsight
- R
- Skytree
- Lumify
- Apache Hadoop
- Qubole

- **Apache Storm**: Apache Storm is an open-source and free large data processing system. Apache Storm is another Apache product that provides a real-time framework for data stream processing and may be used with any language used for programming. It provides a fault-tolerant, shared real-time processing system. With the ability to compute in real time. Storm scheduler spreads workload across several nodes based on topology and performs smoothly with Hadoop's distributed file system (HDFS).



- **Talend:** Talend is a big data technology that helps to simplify and automate large data integration. Its visual wizard creates native code. It also supports large data integration, master data management, as well as information validation. By creating native code, Talend Big Data Platform facilitates the use of MapReduce and Spark.It offers multiple connectors under one roof, allowing you to tailor the solution to your specific requirements.
- **Apache CouchDB:** Apache CouchDB is a free and open-source, cross-platform, document-oriented NoSQL database with a scalable design. It is written in the concurrency-oriented programming language Erlang. Couch DB stores data in JSON documents that may be accessed over the web or by JavaScript queries. It provides distributed scalability as well as fault-tolerant storage. It enables data access by establishing the Couch Replication Protocol.
- **Apache Spark:** Spark is another well-known open-source big data software technology. Spark contains over 80 high-level operators that make it simple to create parallel programs. It is used to process huge datasets in a variety of businesses.Spark provides in-memory data processing capabilities, which are significantly quicker than MapReduce's disk processing. Furthermore, Spark integrates with HDFS, OpenStack, and Apache Cassandra, both in the cloud and on-premises, bringing another layer of flexibility to your company's big data operations.
- **Splice Machine:** It is a tool for large data analytics. Their design is adaptable to public clouds like AWS, Azure, and Google.
- **Azure HDInsight:** It is a cloud-based Spark and Hadoop service. It offers two types of big data cloud services: Standard and Premium. It offers the company with an enterprise-scale cluster on which to perform its big data tasks. The Azure Data Engineering certification might help you get a deeper understanding.
- **R programming:** R is a computer language and free software that is used to compute statistical and visual data. For producing statistical applications and data analysis, mathematicians and data miners use technologies like the R programming language. The R programming language has a large number of statistical tests.
- **Hadoop:** The a centuries-old leader in the field of Big Data processing, well-known for its massive-scale processing of information capabilities. Because it is an open-source Big Data architecture that may run on-premises or in the public internet, it has modest hardware requirements. The following are the primary Hadoop advantages and features:
  - Hadoop Distributed File System (HDFS) is a file system designed to operate with large amounts of bandwidth.
  - (MapReduce) is a highly customizable Big Data processing methodology.

- (YARN) is a resource scheduler for Hadoop resource management.
- The glue required to allow modules from other companies to interact with Hadoop - (Hadoop Libraries)

It is intended to grow from Apache. Hadoop is a software framework used for clustered file systems and massive data processing. It uses the MapReduce programming approach to process big data collections. Hadoop is an open-source framework developed in Java that supports several platforms. Without a doubt, this is the best big data tool. Hadoop is used by more than half of the Fortune 50 enterprises. Amazon Web Services, Hortonworks, IBM, Intel, Microsoft, Facebook, and others are among the big names. Ranging from a single server to hundreds of computers

V. VISUALIZATION TOOLS

Graphic designers now have an easier approach to make visual representations of huge data sets thanks to data visualization technologies. A designer's task is greatly facilitated by streamlining the tedious task of constructing a visual representation when working with data sets that contain hundreds of thousands or millions of data points. Large data sets can be handled by the greatest technologies. The finest ones can really manage several data sets in a single graphic. The top tools can also produce a wide variety of chart, graph, and map kinds. The majority of the tools listed below can produce both dynamic charts and photos.

- **Google chart:** The ease of use of Google's offerings is well recognized and it is an apparent criterion, and Google Chart is no different. One of the simplest techniques for displaying large amounts of data is this one. You may utilize any of the chart gallery that Google Chart offers, ranging from basic line graphs to intricate hierarchical tree-like structures, depending on your needs. Additionally, customization is crucial when constructing a chart, but Google Charts' customization options are somewhat limited.
- **Tableau:** Everyone can use Tableau desktop, a fantastic data visualization application (SaaS) for working with massive data. Two further versions of it—cloud-based "Tableau Online" and "Tableau Server"—are especially made for big data-related businesses. To use this tool, you don't need to know how to code. This device is really useful and offers rapid operation. The canvas or dashboard is 'drag and drop' flexible and user-friendly, which makes it cozy in any office setting.
- **D3:** The JavaScript package D3, or Data-Driven Document, allows for nearly any kind of large data display. This is not a tool like the others, therefore in order to give the gathered data a shape, the user must have a solid understanding of JavaScript. There is no space for outdated browsers (IE 7 or 8), as they do not support SVG (Scalable Vector Graphics), which is used to present the altered data. It is not a rigid structure that must look for opportunities at every turn; instead, it tackles the issue head-on. You may apply the data-driven transformation to the data with a smooth transition and animation effect (optional) and attach any type of data with DOM (Document Object Model).
- **Fusion Chart:** A Javascript charting package called Fusion chart XT is used in 120 different countries by companies like Google, Intel, Microsoft, and many more. To implement it, though, you'll need some understanding of Javascript. Technically, it gathers information in XML or JSON format and presents it as charts in Javascript (HTML5), SVG, and VML format. With a variety of capabilities including scrolling, panning, and animation effects, it offers more than 90 different chart types in both 2D and 3D graphic formats. Additionally, it offers 950+ maps of various locations throughout the world. Here, exporting charts is simple; you may send any chart in PNG, JPG, or PDF format to any location. The platforms Android, iPhone, iPad, MAC, and Windows all support Fusion Charts.
- **Canvas:** Javascript graphing toolkit Canvas.js has a straightforward API and a variety of attractive styles. It loads much quicker than standard SVG or Flash charts. Additionally, it has an adaptive layout that enables it to function on a variety of platforms, including Android, iPhone, tablets, Windows, and Mac. There are 24 various sorts of charts in the chart gallery, but its quickness is its USP. In under 100 milliseconds, it can produce one million data points.
- **Microsoft Power BI:** Microsoft Power BI is a collection of corporate analytics tools designed primarily for data analysis and sharing. Your data may be explored and mined for insights using whatever platform you want, including PCs, tablets, and

smartphones. It assists you in getting rapid answers from the data and offers real-time mapping and analysis capabilities by connecting to on-premises data sources. Power BI allows you to connect to hundreds of sources, including streaming data, data stored on cloud services, excel spreadsheets, and the majority of other data file types, so your data source or format won't be a constraint.

- **Oracle Visual Analyser:** This web-based solution from the Oracle Enterprise Intelligence Cloud Service, which was introduced in 2015, earned a place in Gartner's Magic Quadrant for Business Intelligence and Analytics Platform. Some of Oracle Visual Analyzer's primary features include interactive visualizations, very sophisticated analysis, and a customized dashboard. Due to its tremendous scalability, this data visualization tool is ideal for deployments on a large scale in businesses where in-depth analysis and well selected reports are crucial. Every piece of data has a narrative to tell, and these tools for data visualization are the key to understanding what tale each piece of data is trying to tell us. It aids in our comprehension of the market's existing data and prospective tendencies.

VI. CONCLUSION

A special period in data analysis history has resulted from the accessibility of Big Data, low-cost commodity technology, and innovative information management and analytical tools. We now have the skills necessary to evaluate astounding data volumes fast and affordably for the first time in history thanks to the convergence of these tendencies. These skills are neither only hypothetical nor unimportant. They offer a true advancement and a great chance to achieve significant increases in effectiveness, production, income, and profitability.

IoT is currently becoming a substantial source of Big Data, which is meaningless unless properly examined. This study concentrates on the Big Data setting in the framework of the Internet of Things. It explains the fundamental ideas of IoT and its architecture. It provides an expanded framework of Gartner's 3 V's Big Data model in the form of a 5 V's model. This article improves the viewer's awareness of the relationship between IoT, Big Data, and analytics. It introduces the reader to several Big Data Analytics tools that can handle diverse IoT datasets. After reading his article, the reader will be aware of several platforms and will be able to choose one for their specific difficulties.

Big data analytical tools offer more precise results and enhance the process even with complicated data. Based on the comparison and other qualities added to the tools created for future empowerment, the user and data scientist can choose any of the tools. Due to the ease of use and effectiveness of the big data analytics tool, the supply and demand of the tools are growing daily among consumers.

VII. REFERENCES:

1. Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analyticsenabled transformation model: Application to health care. Inf. Manage, 55(1), 64–79.
2. Wang, Yichuan, Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13. doi:10.1016/j.techfore.2015.12.019.
3. Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. IEEE Access, 5, 5247–5261.
4. Ahmed, E., Yaqoob, I., Hashem, I. A. T., & Khan, I. (2017). Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, Athanasios V. Vasilakos, "The role of big data analytics in Internet of Things. Computer Networks, 129, 459–471.
5. Tanwar, S., Tyagi, S., & Kumar, N. (2020). Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms, and Solutions. Singapore: Springer.
6. Dhankhar, A., & University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak. (2019). Predicting Student's Performance by using Classification Methods. International Journal of Advanced Trends in Computer Science and Engineering, 1532–1536. doi:10.30534/ijatcse/2019/75842019.
7. Shibli, M. A., Head of Artificial Intelligent and Autonomous Systems Engineering Technology Program, Project Manager of Joint Aviation Command (JAC) Program, Abu Dhabi Polytechnic, IAT, Abu Dhabi,

UAE, & Mathew, B. (2019). Artificial intelligent machine learning and big data mining of desert geothermal heat pump: Analysis, design and control. International Journal of Intelligent Systems and Applications, 11(4), 1–13. doi:10.5815/ijisa.2019.04.01

8. Nicolalde, F., Silva, F., Herrera, B., & Pereira, A. (n.d.). Big Data Analytics in IoT.
9. Rocha, A., Adeli, H., Reis, L. P., & Costanzo, S. (Eds.). (2018). Trends and advances in information systems and technologies (1st ed.) [PDF]. doi:10.1007/978-3-319-77712-2
10. Integrating IoT with Big Data, a Revolutionary Step," Experfy Insights. (n.d.).
11. (N.d.). Retrieved 26 July 2023, from https://www.experfy.com/blog/integrating-iot-with-big-data-a-revolutionary-step.

ID: 14

APPLICATIONS OF MACHINE LEARNING AND DEEP LEARNING

Mr.E.TAMIZHAN
Student,
Department of Computer Science and Engineering,
Paavai College of Engineering,
Pachal, Namakkal, Tamil Nadu, India.

Mr.M.P.VENKAT VIJAY
Student,
Department of Computer Science and Engineering,
Paavai College of Engineering,
Pachal, Namakkal, Tamil Nadu, India.

ABSTRACT:

Machine learning (ML) and deep learning (DL) have revolutionized the world of artificial intelligence, providing powerful tools for data analysis, pattern recognition, and decision-making. ML algorithms learn from data to make predictions, while DL models emulate the human brain's neural networks to process vast amounts of information. This paper explores the diverse applications of ML and DL across various domains, showcasing their impact on industries and society. From healthcare and finance to autonomous systems and natural language processing, the versatility of ML and DL has led to transformative advancements, making our lives more efficient, safer, and enjoyable.

KEYWORDS: *Machine Learning, Deep Learning, Natural Language Processing.*

I. INTRODUCTION:

Machine Learning (ML) is a transformative field within the realm of artificial intelligence that has gained immense traction in recent years. As a subset of AI, ML focuses on developing algorithms and models that enable computers to learn from data, recognize patterns, and make informed decisions without being explicitly programmed. The power of ML lies in its ability to uncover hidden insights and knowledge from vast datasets, leading to groundbreaking advancements and applications across various industries.

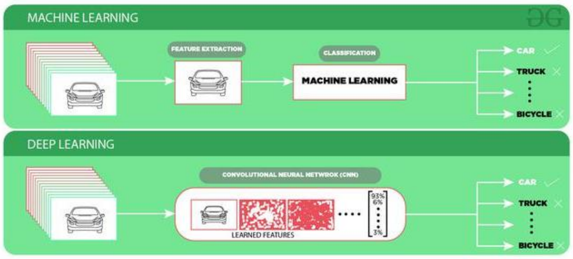
Deep Learning (DL) has emerged as a transformative technology within the field of artificial intelligence, revolutionizing the way machines learn, process information, and make decisions. Machine Learning and Deep Learning have become the driving forces behind numerous breakthroughs in artificial intelligence. The ability of these models to automatically learn from data and improve their performance with experience makes them indispensable tools in modern technology. As a subset of machine learning, DL employs complex artificial neural networks with multiple layers to process vast amounts of data, enabling computers to perform sophisticated tasks that were once considered beyond their capabilities. The power of ML lies in its ability to uncover hidden insights and knowledge from vast datasets, leading to groundbreaking advancements and applications across various industries.

Machine Learning (ML) and Deep Learning (DL) have emerged as transformative technologies within the realm of artificial intelligence, revolutionizing how we approach complex problems and data-driven decision-making. ML enables computers to learn patterns and relationships from data, while DL leverages sophisticated artificial neural networks to process vast amounts of information, mimicking the workings of the human brain. These powerful techniques have found applications across various domains, leading to groundbreaking advancements and reshaping industries worldwide.

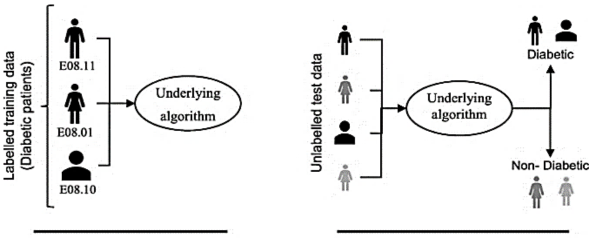
This research aims to explore and analyze the diverse applications of ML and DL, highlighting their significant impact on different fields. By delving into real-world case studies and success stories, we seek to uncover how these technologies have enhanced efficiency, accuracy, and decision-making processes, driving innovation and shaping the future of technology.

II. APPLICATIONS OF MACHINE LEARNING

- **Image and Object Recognition:** ML is extensively used for image recognition, object detection, and classification tasks. Applications include facial recognition, self-driving cars, security surveillance, and medical image analysis.



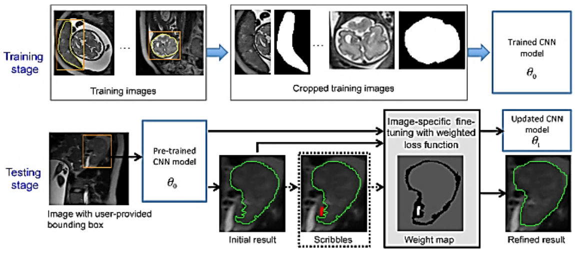
- **Natural Language Processing (NLP):** ML algorithms are employed in language translation, sentiment analysis, chatbots, and virtual assistants, enabling machines to understand and interact with human language.
- **Healthcare:** ML is utilized for disease diagnosis, medical image analysis, drug discovery, and personalized treatment plans. It can predict patient outcomes, identify patterns in medical data, and assist in medical research.



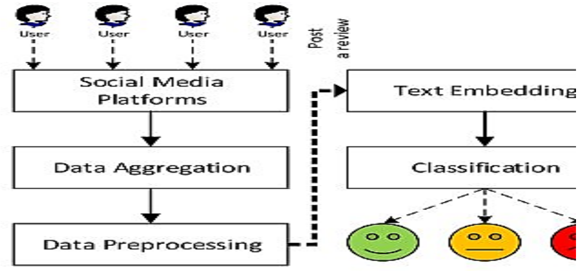
Step1: Train the algorithm using labelled training data

Step2: Feed the trained algorithm on unlabelled data

- **Medical Image Segmentation and Reconstruction:** ML algorithms help segment and reconstruct medical images, facilitating surgical planning and treatment evaluation.

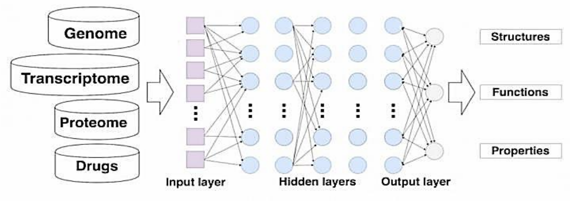


- **Radiology and Pathology Automation:** ML assists radiologists and pathologists in automating routine tasks, such as image analysis and tissue sample classification, improving efficiency and accuracy.
- **Electronic Health Record (EHR) Analysis:** ML algorithms process electronic health records to extract valuable insights, identify trends, and support clinical decision-making.
- **Recommendation Systems:** ML powers recommendation engines used by platforms like Netflix, Amazon, and Spotify to suggest personalized content, products, and services to users based on their preferences and behavior.
- **Fraud Detection and Cybersecurity:** ML is applied in detecting fraudulent activities, credit card fraud, and cybersecurity threats by analyzing patterns and anomalies in data to identify potential risks.
- **Financial Analysis:** ML models are used in stock market prediction, credit risk assessment, algorithmic trading, and portfolio optimization, aiding in making data-driven financial decisions.
- **Autonomous Systems:** ML algorithms are critical for the development of autonomous systems like self-driving cars, drones, and robots, enabling them to perceive their environment and make decisions.
- **Social Media and Sentiment Analysis:** ML helps analyze social media data to understand public opinions, sentiment, and trends. It is used for market research, brand reputation management, and political analysis.

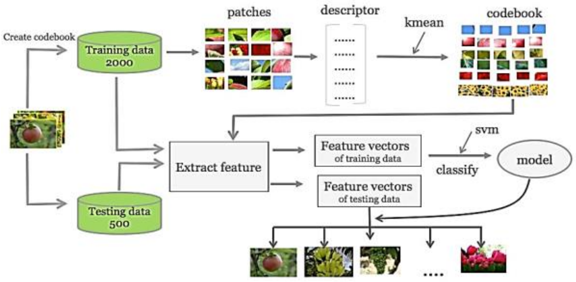


RE 1. Basic steps of sentiment analysis on social media.

- **Manufacturing and Industrial Processes:** ML optimizes industrial operations by predicting equipment failures, optimizing production processes, and ensuring quality control.
- **Drug Discovery and Genomics:** ML is used to analyze biological data, identify potential drug candidates, and understand genetic interactions, accelerating drug discovery processes.



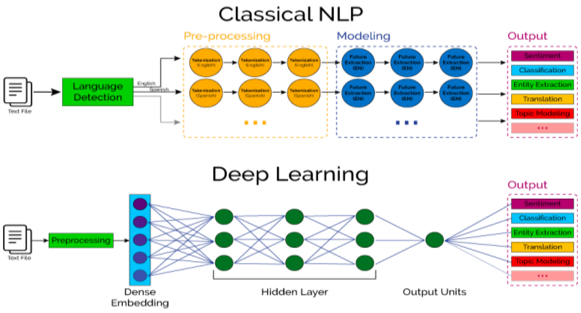
- **Customer Service:** ML is used in chatbots and virtual assistants to provide automated customer support, answer frequently asked questions, and handle basic inquiries, improving response time and efficiency.
- **Climate Modeling and Environmental Monitoring:** ML is applied to analyze climate data, predict weather patterns, and monitor environmental changes to aid in climate research and natural disaster prediction.
- **Speech Recognition:** ML is used for speech recognition systems in applications like voice assistants, voice-controlled devices, and transcription services.
- **Predictive Maintenance:** ML helps predict equipment failures in industries like manufacturing, aviation, and transportation, allowing for timely maintenance and reduced downtime.
- **Marketing and Advertising:** ML enables targeted advertising, customer segmentation, and personalized marketing campaigns based on consumer behavior and preferences.
- **Medical Imaging:** ML is used in medical imaging for tasks like identifying tumors, lesions, and anomalies in X-rays, CT scans, and MRIs. It aids in diagnosing diseases and assists medical professionals in making more accurate assessments.
- **Quality Control in Manufacturing:** ML-based image recognition systems can inspect products on assembly lines, identifying defects and ensuring quality control in real-time.
- **Visual Search Engines:** ML enables visual search engines, where users can upload images to search for similar products, artworks, or related content.
- **Agriculture and Crop Monitoring:** ML-powered image recognition is used for monitoring crop health, identifying pests and diseases, and optimizing agricultural practices.



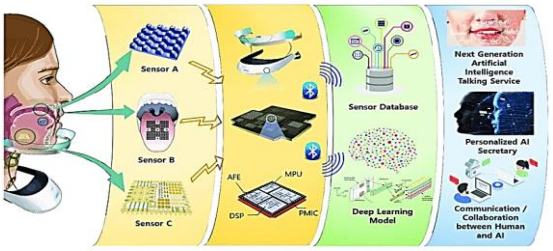
- **Automated Optical Inspection (AOI):** ML algorithms perform AOI in electronic manufacturing, checking circuit boards and electronic components for defects.
- **Augmented Reality (AR):** ML plays a crucial role in AR applications by recognizing real-world objects and overlaying digital information onto them.
- **Security and Surveillance:** ML enables intelligent security systems that can detect suspicious activities, identify intruders, and track objects in surveillance videos.
- **Document Analysis:** ML is used for text recognition and document classification, making it easier to extract and categorize information from scanned documents.
- **Art and Culture:** ML powers applications that analyze artworks, recognize artistic styles, and assist in art restoration and preservation efforts.
- **Wildlife Monitoring:** ML-based camera traps and drones help identify and track wildlife species, aiding in wildlife conservation efforts.
- **Gesture Recognition:** ML algorithms can recognize and interpret hand gestures in images and videos, enabling intuitive human-computer interaction.

III. APPLICATIONS OF DEEP LEARNING

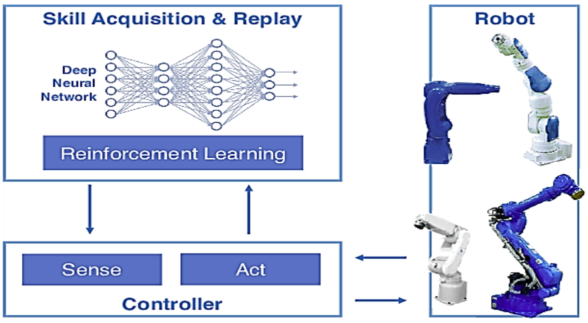
- **Computer Vision:** Deep learning powers advanced computer vision tasks, such as image classification, object detection, semantic segmentation, and image generation. These applications find use in autonomous vehicles, medical imaging, surveillance, and facial recognition systems.
- **Natural Language Processing (NLP):** DL has revolutionized NLP, enabling machines to understand, generate, and interact with human language. Applications include language translation, sentiment analysis, chatbots, and voice assistants like Siri and Alexa.



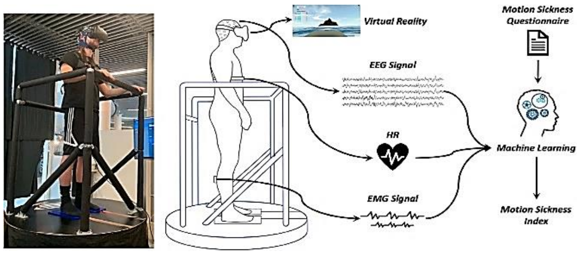
- **Speech Recognition:** DL algorithms excel in speech recognition tasks, enabling accurate transcription, voice-controlled devices, and interactive voice response systems.


Sensor A
Sensor B
Sensor C
AFE
MPU
DSP
PMIC
Sensor Database
Deep Learning Model
Next Generation Artificial Intelligence Talking Service
Personalized AI Secretary
Communication / Collaboration between Human and AI

- **Recommendation Systems:** DL models drive recommendation engines on platforms like Netflix, Amazon, and Spotify, providing personalized content and product suggestions to users.
- **Healthcare:** DL is utilized for medical image analysis, disease diagnosis, drug discovery, and predicting patient outcomes, improving healthcare research and patient care.
- **Autonomous Systems:** DL is at the core of autonomous systems, such as self-driving cars, drones, and robots, enabling perception, decision-making, and navigation.
- **Generative Models:** DL-based generative models, like Generative Adversarial Networks (GANs), can create realistic images, videos, and audio, with applications in art, entertainment, and content creation.
- **Financial Analysis:** DL is employed in stock market prediction, fraud detection, credit risk assessment, and algorithmic trading, optimizing financial decision-making processes.
- **Drug Discovery and Molecular Design:** DL accelerates drug discovery by predicting drug-target interactions and optimizing molecular structures.
- **Social Media and Sentiment Analysis:** DL is used to analyze social media data for sentiment analysis, trend detection, and understanding public opinions.
- **Robotics:** DL plays a crucial role in robotics, enabling robots to navigate complex environments, perform object manipulation, and interact with humans.



- **Gaming and Entertainment:** DL enhances gaming experiences by creating realistic characters, dynamic game environments, and providing adaptive gameplay.
- **Climate Modeling:** DL is used for climate modeling and analysis, predicting weather patterns, and understanding climate change impacts.
- **Virtual Reality (VR):** DL powers VR applications by generating realistic virtual environments and enhancing user interactions.



- **Language Generation:** DL models are capable of generating human-like text, leading to applications like automated content creation, storytelling, and natural language generation.

IV. CONCLUSION:

Machine learning and deep learning have become indispensable technologies, reshaping industries and enhancing the quality of our lives. From healthcare to finance, transportation to entertainment, these applications have demonstrated their ability to solve complex problems, optimize processes, and drive innovation. ML and DL applications have touched every aspect of our lives, from the way we interact with technology to how industries operate. They have the potential to address complex challenges and create solutions that were once thought impossible. However, their responsible and ethical use is crucial to ensure that they continue to benefit humanity while mitigating potential risks. As ML and DL continue to advance, their impact will only grow, presenting novel opportunities and addressing the challenges of a rapidly evolving world. Embracing these technologies responsibly and ethically will be crucial in maximizing their benefits while ensuring a positive and inclusive future for society.

V. REFERENCE:

1. Yinglong Li, “Research and application of deep learning in image recognition”, 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), 994-999, 2022.
2. Neha Sharma, Reecha Sharma, Neeru Jindal, “Machine learning and deep learning applications-a vision”, Global Transitions Proceedings 2 (1), 24-28, 2021.
3. Ata Jahangir Moshayedi, Atanu Shuvam Roy, Amin Kolahdooz, Yang Shuxin, “Deep learning application pros and cons over algorithm”, EAI Endorsed Transactions on AI and Robotics 1 (1), 2022
4. Li Deng, Dong Yu, “Deep learning: methods and applications”, Foundations and trends® in signal processing 7 (3–4), 197-387, 2014
5. Iqbal H Sarker, “Machine learning: Algorithms, real-world applications and research directions”, SN computer science 2 (3), 160, 2021
6. Raffaele Cioffi, Marta Travaglioni, Giuseppina Piscitelli, Antonella Petrillo, Fabio De Felice, “Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions”, Sustainability 12 (2), 492, 2020
7. Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, Amir H Gandomi, “Machine learning in medical applications: A review of state-of-the-art methods”, Computers in Biology and Medicine 145, 105458, 2022
8. Abdelmalek Bouguettaya, Hafed Zarzour, Ahmed Kechida, Amine Mohammed Taberkit, “Deep learning techniques to classify agricultural crops through UAV imagery: A review”, Neural Computing and Applications 34 (12), 9511-9536, 2022
9. Jeevith Hegde, Børge Rokseth, “Applications of machine learning methods for engineering risk assessment–A review”, Safety science 122, 104492, 2020
10. Yaguo Lei, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, Asoke K Nandi, “Applications of machine learning to machine fault diagnosis: A review and roadmap”, Mechanical Systems and Signal Processing 138, 106587, 2020.

ID: 15

PRIVACY AND CRYPTOLOGY

Ms.V.SOWMIYA
Student,
Department of Computer Science and Engineering,
Paavai College of Engineering,
Pachal, Namakkal, Tamil Nadu, India.

Ms.S.SRI RANJANI
Student,
Department of Computer Science and Engineering,
Paavai College of Engineering,
Pachal Namakkal, Tamil Nadu, India.

ABSTRACT:

In an interconnected world where personal data is constantly exchanged and stored electronically, preserving privacy has become an essential concern. Cryptology, the science of secure communication, plays a pivotal role in protecting sensitive information from unauthorized access. This chapter explores the critical intersection of privacy and cryptology, emphasizing the significance of encryption techniques in safeguarding digital assets. From understanding the fundamental principles of cryptology to exploring various cryptographic methods for data protection, this chapter provides insights into how privacy and security are maintained in the digital realm.

KEYWORDS: *Cryptology, Encryption, Decryption.*

I. INTRODUCTION:

The digital age has revolutionized how we communicate and access information. While this interconnectedness brings unprecedented convenience, it also raises concerns about the privacy and security of personal data. Cryptology, the art and science of secure communication, has emerged as a vital tool to address these challenges. This chapter begins by introducing the concept of cryptology and its historical development, from ancient ciphers to modern encryption techniques. We will then explore the role of cryptography and cryptanalysis in ensuring confidentiality, integrity, and authenticity of data. As we dive deeper into the realm of cryptology, the chapter aims to shed light on how privacy can be preserved in a data-driven world.

Definition of Cryptology:

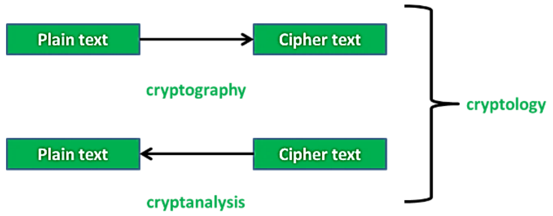
Cryptology is an interdisciplinary field that encompasses both cryptography and cryptanalysis. It involves the study and application of techniques to ensure secure communication, confidentiality, and data integrity. Cryptology plays a crucial role in protecting sensitive information from unauthorized access and interception in an interconnected world.

Historical Development of Cryptology

The history of cryptology can be traced back to ancient civilizations, where encryption methods were employed to encode secret messages during wars and diplomatic exchanges. Notable historical examples include the Caesar cipher and the use of the Enigma machine during World War II. The advent of computers and the digital revolution led to the development of modern encryption algorithms, strengthening the field of cryptology.

The Role of Cryptography and Cryptanalysis:

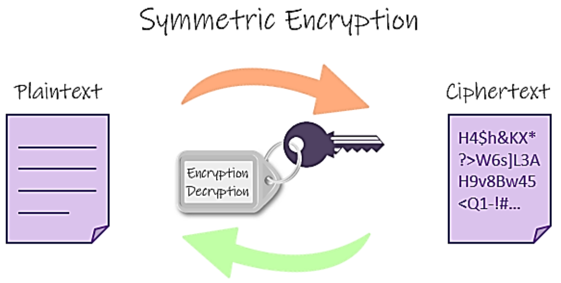
Cryptography involves the creation of secure communication by converting plaintext data into ciphertext using encryption algorithms. It relies on cryptographic keys to encode and decode information, ensuring that only authorized parties can access the original message. Cryptanalysis, on the other hand, focuses on breaking encrypted messages without possessing the corresponding decryption key. Cryptanalysts use various techniques, including mathematical analysis and computational power, to decipher encrypted data and uncover the original information.



II. FUNDAMENTALS OF CRYPTOGRAPHY:

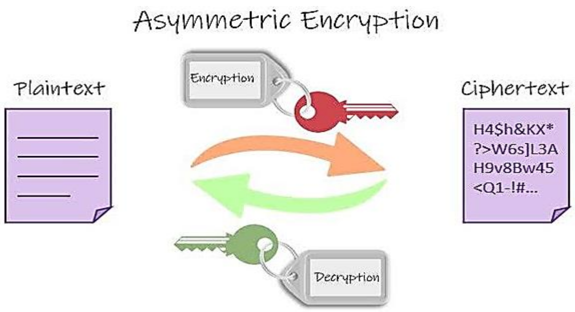
Symmetric Encryption:

Symmetric encryption employs a single secret key for both encryption and decryption. This section explores the principles of symmetric encryption, where the same key is used by both the sender and the recipient. We delve into the Data Encryption Standard (DES), an early symmetric encryption algorithm, and the Advanced Encryption Standard (AES), the widely used successor to DES.



Asymmetric Encryption:

Asymmetric encryption, also known as public-key cryptography, employs a pair of keys: a public key used for encryption and a private key kept secret for decryption. This section examines the RSA algorithm, a popular asymmetric encryption technique. RSA provides a secure means of exchanging keys and authenticating communication between parties without the need for prior key exchange.



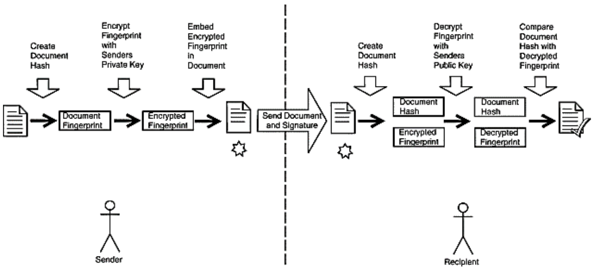
III. ENSURING DATA INTEGRITY AND AUTHENTICITY:

Cryptographic Hash Functions:

Cryptographic hash functions generate fixed-length hash values (digests) based on the input data. These functions possess properties such as collision resistance and pre-image resistance, ensuring the integrity and authenticity of data. We explore the applications of hash functions in data integrity, including verifying data integrity during transmission and securing passwords.

Digital Signatures:

Digital signatures provide a means to verify the authenticity and integrity of a message or document. This section delves into the process of generating digital signatures using public-key cryptography, where the signature is encrypted with the private key and verified with the corresponding public key. Digital signatures ensure that the message remains unaltered and is from the claimed sender.


Create Document Hash
Encrypt Fingerprint with Senders Private Key
Embed Encrypted Fingerprint in Document
Document Fingerprint
Encrypted Fingerprint
Send Document and Signature
Create Document Hash
Decrypt Fingerprint with Senders Public Key
Compare Document Hash with Decrypted Fingerprint
Document Hash
Encrypted Fingerprint
Document Hash
Decrypted Fingerprint
Sender
Recipient

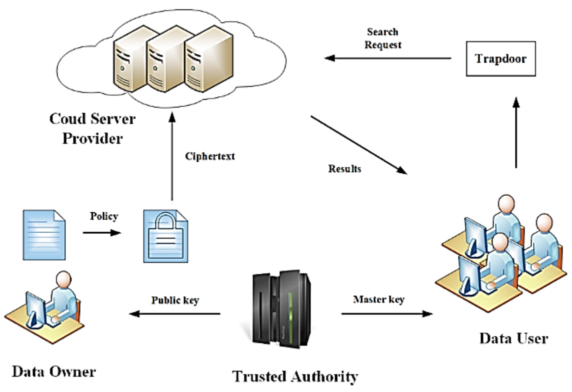
Cryptology for Privacy Preservation:

Securing Communication:

Securing communication is crucial to safeguard sensitive information from unauthorized access. End-to-end encryption ensures that messages and data are encrypted at the sender's end and decrypted at the recipient's end, preventing interception by unauthorized parties. Virtual Private Networks (VPNs) are also discussed as a means of creating secure tunnels for data transmission over the internet, ensuring confidentiality and privacy.

Data Protection and Storage:

Data protection at rest is equally important, as data stored in various repositories can be vulnerable to theft or unauthorized access. Full-disk encryption encrypts the entire disk to protect data on storage devices, while database encryption secures sensitive information within databases, ensuring that only authorized users with the decryption key can access the data.



IV. PRIVACY IN THE AGE OF BIG DATA AND AI:

Privacy-Preserving Data Analytics:

As the volume of data generated continues to grow, privacy-preserving data analytics techniques become essential to analyze data without compromising individual privacy. Homomorphic encryption allows computations to be performed on encrypted data directly without decrypting it, ensuring privacy. Differential privacy adds controlled noise to the data to provide statistical guarantees while preserving individual privacy.

AI and Privacy Concerns:

With the increasing use of artificial intelligence (AI) applications, privacy concerns arise due to the vast amounts of personal data processed. Federated learning allows AI models to be trained on decentralized data without sharing raw data, preserving privacy. Encrypted machine learning techniques protect the privacy of trained models and data during the learning process.

Challenges and Future Directions:

Quantum Threats to Cryptology:

Quantum computing poses a significant threat to traditional cryptographic systems, as quantum algorithms can efficiently break certain encryption schemes. This section explores the impact of quantum computing on cryptanalysis and the need for quantum-resistant cryptography to ensure the continued security of encrypted data.

Privacy and Ethical Considerations:

As cryptology becomes more pervasive in various applications, striking a balance between privacy and security becomes paramount. Ethical considerations in the use of cryptology are discussed, addressing issues related to data collection, transparency, and the responsible use

Advantages of Privacy and Cryptology:

1. Data Protection: Privacy and cryptology techniques play a crucial role in protecting sensitive data from unauthorized access and malicious actors. Encryption ensures that data remains confidential, even if it is intercepted during transmission or stored in unsecured environments.

2. Confidentiality: Cryptology ensures that only authorized individuals or entities can access encrypted information, maintaining confidentiality and preventing unauthorized parties from understanding the content.

3. Authentication: Cryptography provides a means of verifying the authenticity of messages or documents through digital signatures. This helps ensure that messages come from genuine senders and have not been tampered with during transmission.

4. Secure Communication: Cryptology enables secure communication over networks and the internet. End-to-end encryption and Virtual Private Networks (VPNs) ensure that communication remains private and protected from eavesdropping.

5. Privacy Preservation: Privacy-preserving data analytics techniques, such as homomorphic encryption and differential privacy, allow data to be analyzed without revealing sensitive information, striking a balance between data utility and individual privacy.

6. Compliance and Trust: Implementing privacy and cryptology measures helps organizations comply with data protection regulations and builds trust with customers, partners, and stakeholders, as it demonstrates a commitment to safeguarding sensitive information.

7. Protection against Quantum Threats: Quantum-resistant cryptography research ensures that cryptosystems remain secure in the face of quantum computing advances, safeguarding data against future threats.

Disadvantages of Privacy and Cryptology:

1. Performance Overhead: Implementing cryptology measures, especially strong encryption algorithms, can impose a performance overhead on computational processes, leading to slower data processing and communication.

2. Key Management: Proper key management is essential for effective encryption. Securely generating, distributing, and managing encryption keys can be challenging, especially in large-scale systems.

3. Potential for Misuse: While cryptology is intended to protect privacy and security, it can also be misused for illegal activities or to conceal illicit communications.

4. Complexity: Cryptology concepts and implementations can be complex, requiring specialized knowledge and expertise. Misconfigurations or improper use can lead to vulnerabilities and compromise security.

5. Data Recovery Challenges: Strong encryption can make data recovery difficult or impossible if the encryption keys are lost or unavailable.

6. Quantum Computing Risks: While quantum-resistant cryptography is being developed, current cryptographic systems may be vulnerable to future quantum computing threats.

7. False Sense of Security: Overreliance on cryptology without proper cybersecurity measures may lead to a false sense of security. Security breaches can still occur through other means, such as social engineering or software vulnerabilities.

V. APPLICATIONS OF PRIVACY AND CRYPTOLOGY:

1. Secure Communication: Cryptology is widely used to ensure secure communication over networks and the internet. Encryption techniques, such as end-to-end encryption and SSL/TLS (Secure Sockets Layer/Transport Layer Security), protect data during transmission, preventing eavesdropping and interception.

2. Data Protection and Storage: Cryptology is crucial for safeguarding data at rest. Full-disk encryption ensures that data stored on devices, such as laptops and smartphones, remains secure even if the device is lost or stolen. Database encryption protects sensitive information within databases, restricting access to authorized users with the decryption keys.

3. Online Banking and E-Commerce: Privacy and cryptology play a pivotal role in securing online transactions. SSL/TLS encryption secures online banking sessions and e-commerce transactions, protecting users' financial data and ensuring safe online purchases.

4. Digital Signatures: Cryptographic digital signatures are used to verify the authenticity and integrity of electronic documents, contracts, and transactions. They provide non-repudiation, ensuring that the signer cannot later deny having signed the document.

5. Virtual Private Networks (VPNs): VPNs use cryptology to create secure tunnels for data transmission over public networks. They enable secure remote access to corporate networks and protect users' online activities from prying eyes.

6. Password Protection: Cryptography is used to securely store and transmit passwords. Hashing algorithms ensure that passwords are not stored in plaintext, reducing the risk of unauthorized access to user accounts.

7. Secure Messaging and Email Encryption: Privacy and cryptology are applied in secure messaging applications that provide end-to-end encryption, ensuring that only the intended recipients can read the messages. Email encryption techniques, such as Pretty Good Privacy (PGP) and Secure/Multipurpose Internet Mail Extensions (S/MIME), protect the confidentiality of email communication.

8. Privacy-Preserving Data Analytics: Cryptology techniques like homomorphic encryption and differential privacy enable data analysis while protecting individual privacy. Privacy-preserving data analytics allows organizations to gain insights from data without revealing sensitive information.

9. Secure Voice and Video Calls: Cryptography is used in secure communication applications for voice and video calls, ensuring that the conversations remain confidential and protected from interception.

10. Internet of Things (IoT) Security: Cryptography is vital in securing IoT devices and communications, protecting sensitive data transmitted between IoT devices and backend servers.

11. Secure Cloud Computing: Cryptology ensures the security of data stored and processed in the cloud. Encrypted cloud storage and secure cloud computing environments protect user data from unauthorized access by cloud service providers or malicious actors.

12. Blockchain Technology: Cryptography forms the backbone of blockchain technology, providing the cryptographic primitives necessary for secure transaction validation, data immutability, and user identity protection in decentralized networks.

VI. CONCLUSION:

Cryptology remains an indispensable tool for preserving privacy and security in the digital age. This chapter has explored the fundamental concepts of cryptology, from encryption methods to data integrity and digital signatures. By understanding the principles of cryptology, individuals and organizations can better protect their sensitive data and communications.

As technology continues to advance, challenges such as quantum threats and privacy in the age of big data and AI will require innovative solutions. Quantum-resistant cryptography will play a critical role in ensuring that data remains secure even in the face of powerful quantum computers. Additionally, the responsible and ethical use of cryptology will be essential to strike a balance between privacy and security.

In conclusion, privacy and cryptology are inseparable in the modern digital landscape. As individuals and societies become increasingly reliant on digital technologies, safeguarding privacy through robust cryptology becomes paramount. By embracing the principles of cryptology and staying vigilant against emerging threats, we can foster trust, protect our digital assets, and preserve the sanctity of privacy for generations to come.

VII. REFERENCES:

1. Menasce, "Security performance," in IEEE Internet Computing, vol. 7, no. 3, pp. 84-87, May-June 2003, doi: 10.1109/MIC.2003.1200305.
2. W. -B. Lee and C. -D. Lee, "A Cryptographic Key Management Solution for HIPAA Privacy/Security Regulations," in IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp. 34-41, Jan. 2008, doi: 10.1109/TITB.2007.906101.
3. W. Diffie and M. E. Hellman, "Privacy and authentication: An introduction to cryptography," in Proceedings of the IEEE, vol. 67, no. 3, pp. 397-427, March 1979, doi: 10.1109/PROC.1979.11256.
4. M. A. Al-Shareeda, M. Anbar, S. Manickam, A. Khalil and I. H. Hasbullah, "Security and Privacy Schemes in Vehicular Ad-Hoc Network With Identity-Based Cryptography Approach: A Survey," in IEEE Access, vol. 9, pp. 121522-121531, 2021, doi: 10.1109/ACCESS.2021.3109264.
5. Menezes and D. Stebila, "Challenges in Cryptography," in IEEE Security & Privacy, vol. 19, no. 2, pp. 70-73, March-April 2021, doi: 10.1109/MSEC.2021.3049730.
6. Khaliq, A. Anjum, A. B. Ajmal, J. L. Webber, A. Mehbodniya and S. Khan, "A Secure and Privacy Preserved Parking Recommender System Using Elliptic Curve Cryptography and Local Differential Privacy," in IEEE Access, vol. 10, pp. 56410-56426, 2022, doi:10.1109/ACCESS.2022.3175829.
7. N. Jaber and Mohamad Fadli Bin Zolkipli, "Use of cryptography in cloud computing," 2013 IEEE International Conference on Control System, Computing and Engineering, Penang, Malaysia, 2013, pp. 179-184, doi: 10.1109/ICCSCE.2013.6719955.
8. M. A. Pathak, B. Raj, S. D. Rane and P. Smaragdis, "Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise," in IEEE Signal Processing Magazine, vol. 30, no. 2, pp. 62-74, March 2013, doi: 10.1109/MSP.2012.2230222.

ID: 16

BIG DATA AND BIG DATA ANALYTICS IN HEALTHCARE

Ms.M.SOUNDARYA
II M.Sc,
Department of Information Technology,
Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

Ms.V.PRABHA
II M.Sc,
Department of Information Technology,
Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT:

In recent times, huge quantities of structured, unshaped and semi-structured data have been generated by colourful institutions around the world and, inclusively, this miscellaneous data is appertained to as big data. The health assiduity sector has been brazened by the need to manage the big data being produced by colourful sources, which are well known for producing high volumes of miscellaneous data. Colourful big-data analytics tools and ways have been developed for handling these massive quantities of data, in the healthcare sector. In this paper, we bandy the impact of big data in healthcare, and colourful tools available in the Hadoop ecosystem for handling it. We also explore the abstract armature of big data analytics for healthcare which involves the data gathering history of different branches, the genome database, electronic health records, textbook/ imagery, and clinical opinions support system.

KEYWORDS: *Big Data; Healthcare; Hadoop; MapReduce*

I. INTRODUCTION:

Every day, data is generated by a range of different operations, bias, and geographical exploration conditioning for the purposes of rainfall soothsaying, rainfall vaticination, disaster evaluation, crime discovery, and the heath assiduity, to name a many. In current scripts, bigdata is associated with core technologies and colourful enterprises including Google, Facebook, and IBM, which prize precious information from the huge volumes of data collected [1-3]. A period of open information in healthcare is now under way. Big data is being generated fleetly in every field including healthcare, with respect to patient care, compliance, and colourful nonsupervisory conditions. As the global population continues to increase along with the mortal lifetime, treatment delivery models are evolving snappily, and some of the opinions underpinning these fast changes must be grounded on data [4]. Healthcare shareholders are promised new knowledge from big data, so called both for its volume as well as its complexity and range. Pharmaceutical- assiduity experts and shareholders have begun to routinely dissect big data to gain sapience, but these conditioning are still in the early stages and must be coordinated to address healthcare delivery problems and ameliorate healthcare quality.

Beforehand systems for big- data analytics of healthcare informatics have been established across numerous scripts., the disquisition of patient characteristics and determination of treatment cost and results to pinpoint the stylish and most cost-effective treatments [4]. Health informatics is described as the assimilation of healthcare science, computing science and information science in the study of healthcare information. Health informatics involves data accession, storehouse, and reclamation to give better results by healthcare providers. In the healthcare system, data is characterized by its diversity and variety as a result of the linking of a different range of biomedical data sources including, for illustration, detector data, imagery, gene arrays, laboratory tests, free textbook, and demographics [5]. Utmost data in healthcare system (e.g., doctor’s notes, lab test results, and clinical data) is unshaped and isn't stored electronically, i.e., it exists only in hard clones and its volume is adding veritably fleetly. Presently, there's a major focus on the digitization of these vast stores of hard dupe data. The revolutions of data size are actually creating a problem in order to achieve this thing [6]. The colourful languages and models that have been developed to resolve the problems associated with big data concentrate on working four issues known as the four V's, volume, variety, velocity, and veracity. The colourful classes of data in healthcare operations include Electronic Health Records (EHR), machine generated/ detector data, health information exchanges, patient registries, doors, inheritable databases, and public records. Public records are major sources of big- data in the healthcare assiduity and bear effective data analytics to resolve their associated healthcare problems. According to a check conducted in 2012, healthcare data totalled nearly 550 petabytes and will reach nearly 26 000 petabytes in 2020 [5]. In light of the miscellaneous data formats, huge volume, and related misgivings in the big- data sources, the task of realizing the metamorphosis of raw data into practicable information is dispiriting.

Being so complex, the identification of health features in medical data and the selection of class attributes for health analytics demands largely sophisticated and architecturally specific ways and tools.

II. BIG DATA ANALYTICS IN HEALTH INFORMATICS

The main difference between traditional health analysis and big- data health analytics is the prosecution of computer programming. In the traditional system, the healthcare assiduity depended on other diligence for big data analysis. Numerous healthcare shareholders trust information technology because of its meaningful issues their operating systems are functional and they can reuse the data into standardized forms. Moment, the healthcare assiduity is faced with the challenge of handling fleetly developing big healthcare data. The field of big data analytics is growing and has the implicit to give useful perceptivity for the healthcare system. As noted over, utmost of the massive quantities of data generated by this system is saved in hard clones, which must also be digitized [7]. Big data can ameliorate healthcare delivery and reduce its cost, while supporting advanced patient care, perfecting patient issues, and avoiding gratuitous costs [8]. Big data analytics is presently used to prognosticate the issues of opinions made by physicians, the outgrowth of a heart operation for a condition grounded on case's age, current condition, and health status. Basically, we can say that the part of big data in the health sector is to manage data sets related to healthcare, which are complex and delicate to manage using current tackle, software, and operation tools. In addition to the burgeoning volume of healthcare data, payment styles are also changing [9]. Thus, purposeful use and pay grounded on performance have surfaced as important factors in the healthcare sector. In 2011, associations working in the field of healthcare had produced further than 150 exabytes of data [10], all of which must be efficiently anatomized to be at all useful to the healthcare system [11]. The storehouse of healthcare related data in EHRs occurs in a variety of forms. An unforeseen increase in data related to healthcare informatics has also been observed in the field of bioinformatics, where numerous terabytes of data are generated by genomic sequencing [11]. There are a variety of logical ways available for interpreting medical, which can also be used for patient care [12]. The different origins and forms of big data are challenging the healthcare informatics community to develop styles for data processing. There's a big demand for fashion that combines different data sources [13]. A number of abstract approaches can be employed to fete irregularities in vast quantities of data from different datasets. The fabrics available for the analysis of healthcare data are as follows Predictive Analytics in Healthcare For the once two times, prophetic analysis has been honoured as one of the major business intelligence approaches, but its real-world operations extend far beyond the business environment. Big data analytics includes colourful styles, including textbook analytics and multimedia analytics [14]. Still, one of the most pivotal orders is prophetic analytics which includes statistical styles like data mining and machine literacy that examine current and literal data to prognosticate the future. Prophetic styles which are being used moment in the hospital environment to determine if case may be at threat for readmission [15]. This data can help doctors to make important case care opinions. Prophetic analysis requires an understanding and use of machine literacy, which is extensively applied in this approach.

- Machine literacy in Healthcare: The conception of machine literacy is veritably analogous to that of data mining [4], both of which check-up data to identify patterns. Rather than rooting data grounded on mortal understanding, as in data mining operations, machine literacy uses that data to ameliorate the program's understanding. Machine literacy identifies data patterns and also alters the program function consequently [16].
- Electronic Health Records: EHR represents the widest health operation of big data in healthcare. Each case has his/ her own medical records, with details that include their medical history, disinclination's opinion, symptoms, and lab test results. Case records are participated in both public and private sectors with healthcare providers via a secure information system. These lines are adjustable, in those croakers can make changes over time and add new medical test results, without the need for paper work or duplication of data.

III. FOUR VS OF BIG DATA IN HEALTHCARE

Four primary attributes that are associated with big data: volume, velocity, variety, and veracity.

o **Volume:** Big data is a term to pertaining to huge volumes of collected data. There's no fixed threshold for the volume of this data. generally, the term is used with respect to massive- scale data which must be managed, stored, and anatomized using traditional databases and data recycling armature [14]. The volume of data generated by ultramodern IT and the healthcare system has been growing and is driven by the reduced costs of data storehouse and processing infrastructures and the need to prize precious perceptivity from data to ameliorate business processes, edge, and services to consumers [4].

o **Velocity:** Velocity which represents primary reason for the exponential growth of data, refers to how fast data is collected [14]. Healthcare systems are generating data at decreasingly advanced pets. In the volume and variety of the structured or unshaped data collected, the haste of the generation of this data after recycling requires a decision grounded on its affair.

o **Variety:** Variety refers to the form of the data, i.e., unshaped or structured, textbook, medical imagery, audio, videotape, and detector data. Structured data information includes clinical data (case record data), which must simply be collected, stored, and reused by a particular device. Structured data comprises just 5% to 10% of healthcare data. Unshaped or semi-structured data includes e-mails, prints, vids, audios, and other health related data similar as hospital medical reports, physician's notes, paper conventions, and radiograph flicks [13].

o **Veracity:** The veracity of data is the degree of assurance that the meaning of data is harmonious. Different data sources vary in their situations of data credibility and trust ability [9]. The issues of big data analytics must be believable and error-free, but in healthcare, unsupervised machine literacy algorithms make opinions that are used by automated machines grounded on data that may be empty or deceiving [4]. Healthcare analytics are assigned with rooting useful perceptivity from this data to treat cases and make the stylish possible opinions.

IV. IMPACT OF BIG DATA ON THE HEALTHCARE SYSTEM

The eventuality of big data is that it could revise issues regarding the most suitable or accurate case opinion and the delicacy information used in the health informatics system [15].

As similar, the disquisition of huge quantities of information will have an important effect on medicinal services frame in five felicitations, or pathways. Perfecting issues for cases with respect to these pathways, as described below, will be the focus of the healthcare system and will directly impact the case. Right Living Right living refers to the case living a better and healthier life [15]. By right living, cases could manage themselves by making the stylish opinions for themselves, grounded on the application of information mining more choices and enhancing their good. By choosing the right path for their diurnal health, regarding their diet, preventative care, exercise, and other conditioning of diurnal living, cases can play an active part in realizing a healthy life [16].

- Right Care: This pathway ensures that cases admit the most applicable treatment available and that all providers gain the same data and has the same objects to avoid redundancy of planning and trouble [17]. This aspect has come more feasible in the period of big data.

- Right Provider: Healthcare providers in this pathway can gain an overall view of their cases by combining data from colourful sources similar as medical outfit, public health statistics, and socioeconomic data [15]. The availability of this information enables mortal service providers to conduct targeted examinations and develop the chops and capacities to identify and give better treatment options to cases [18].
- Right Innovation: This pathway recognizes that new complaint conditions, new treatments, and new medical will continue to evolve [15]. Likewise, advancements in the provision of case services, for illustration, elevation specifics and the effectiveness of exploration and development sweats, will enable new ways to promote good and patient health via public social insurance system [17]. The vacuity of early trial data is important for stakeholders. This data can be used to explore high-implicit targets and identify ways for perfecting traditional clinical treatment styles.
- Right Value: To ameliorate the quality and value of health- related services, providers must pay careful and ongoing attention to their cases. Cases must gain the most salutary results linked by their social insurance system [18]. Measures that could be taken to ensure the intelligent use of data includes, for illustration, relating and destroying data misrepresentation, manipulations, and waste, and perfecting coffers [19].

V. HADOOP- GROUNDED OPERATIONS FOR HEALTH INDUSTRY

In light of the fact that healthcare data exists primarily in published form, there's a need for the active digitization of print form data. The maturity of this data is also unshaped, so it's a major challenge for this assiduity to prize meaningful information regarding patient care, clinical operations, and exploration. The collection of software serviceability known as the Hadoop ecosystem can help the healthcare sector to manage this vast quantum of data. The colourful operations of the Hadoop ecosystem in the healthcare sector are as follows

- Treatment of Cancer and Genomics: We know that mortal DNA contains three billion base dyads. To fight cancer, it's vital that large quantities of data are efficiently organized. The patterns of cancer mutations and their responses vary grounded on individual genetics, which explains then on-curability of some cancer. Oncologists have determined that in feting the patterns of cancer, it's important to give specific treatment for specific cancers, grounded on the case's inheritable makeup. The Hadoop technology MapReduce facilitates the mapping of three billion DNA base dyads to determine the applicable cancer treatment for each particular case. Arizona State University is working on design to develop a healthcare model that takes individual genomic data and selects a treatment grounded on identification of the case's cancer gene. This model provides base for treatment through big data analysis to ameliorate the chances of saving cases lives.
- Monitoring of Case Vitals: Hospital staff throughout the world connect their workaffair using big- data technology. Colourful hospitals around the globe use Hadoopgrounded factors in the Hadoop Distributed train System (HDFS), including the Impala, HBase, Hive, Spark, and Flume fabrics, to convert the huge quantum of unshaped data generated by detectors that take patient vital signs, jiffs per nanosecond, blood pressure, blood sugar position, and respiratory rate. Without Hadoop, these healthcare staff couldn't dissect this unshaped data being generated by patient healthcare systems. In Atlanta, Georgia, there are 6200 Intensive Care Units (ICUs) for paediatric healthcare, where children can stay for further than one month depending on their problem.
- These ICUs are equipped with a detector technology that tracks the child's health status with respect to twinkle, blood pressure, and other vital signs. However, an alert is automatically generatedto medical staff to ensure the child's safety, if any problem occurs.
- Hospital Network: Several hospitals use the Hadoop ecosystem's NoSQL database to collect and manage their huge amounts of real- time data from different sources related to patient care, finances, and a payroll, which helps

them identify high- trouble cases while also reducing day- to- day expenditures.

- Healthcare Intelligence: Hadoop technology also supports the healthcare intelligence operations used by hospitals and insurance companies. Hadoop ecosystem's Pig, Hive, and MapReduce technologies process large datasets related to medicines, conditions, symptoms, opinions, geographic regions, and other factors to prize meaningful information (e.g., asked age) for insurance companies.
- Prevention and Detection of Frauds: In the early faces of big data analytics, healthpredicated insurance groups use multiple paths to identify fraud exertion and establish styles to help medical fraud. With Hadoop, companies use operations predicated on a prophecy model to identify those committing fraud via data regarding their former health claims, voice recordings, pay check, and demographics. Hadoop's NoSQL database is also helpful in preventing fraud related to medical claims at an early stage by the use of realtime Hadoop predicated health operations, authentic medical claim bills, downfall auguring data, voice data recordings, and other data sources.

BIG DATA ANALYTICS ARCHITECTURE FOR HEALTH INFORMATICS:

Presently, the main focus in big- data analytics is to gain an in- depth sapience and understanding of big data rather than to collect it [20]. Data analytics involves the development and operation of algorithms for assaying colourful complex data sets to prize meaningful knowledge, patterns, and information. In recent times, experimenters have begun to consider the applicable architectural frame for healthcare systems that use big- data analytics, one of which uses a four- subcaste armature that comprises a metamorphosis subcaste, data-source subcaste, big data platform subcaste, and logical subcaste [14]. In this layered system, data originates from different sources and has colourful formats and storehouse systems.

Each subcaste has a specific data-processing functionality for performing specific tasks on the HDFS, using the MapReduce processing model. The other layers perform other tasks, i.e., report generation, query end, data mining processing, and online logical processing. The main demand in big- data logical processing is to rush the data at high speed to minimize the speeding time. The coming precedence in big- data logical processing is to efficiently modernize and transfigure queries at a constant time [21]. The third demand in the big- data logical processing is to use and efficiently manage the storehouse area space.

The last specification of big- data analytics is to efficiently come familiar with the fleetly progressing workload memos. Big- data analytics fabrics differ from traditional healthcare processing systems with respect to how they reuse big data [22]. In the current health care system, data is reused using traditional tools installed in a single stage-alone system like a desktop computer. In discrepancy, big data is reused by clustering and reviews multiple bumps of clusters in the network [23]. This processing is grounded on the conception of community to handle large medical data sets [24].

Freely available fabrics, similar as Hadoop, MapReduce, Pig, Sqoop, Hive, and HBase Avro, all have capability to reuse the health-related data sets for healthcare systems. Big- data technologies astronomically relate to scientific inventions that mimic those used for large datasets [25]. In the first element is the demand for big data sources for processing. In the alternate element clusters with a centralized big- data processing structure are at the peak of high performance [24].

It has been observed that the tools substantially available for big- data analytics recycling give data security, scalability, and manageability with the help of the MapReduce paradigm. In the third element, big data analytics operations have a storehouse sphere to integrate penetrated databases that use different operations [26]. In the fourth element, are the most popular big- data analytics operations in healthcare systems, which include reports,

Online Analytical Processing (OLAP), queries, and data mining. Healthcare data come from a range of sources including EHRs, genome databases, genome data lines, textbook and imagery (unshaped data sources), clinical decision support systems, government affiliated sources, medical test labs and apothecaries, and health insurance companies. These data are constantly available in different scheme tables, and are in ASCII/ textbook and stored at colourful locales. In the coming section, we describe the colourful bigdata Hadoop- grounded processing tools that support the development of health- grounded operations for the health assiduity.

VI. REFERENCES

1. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods and analytics, International Journal of Information Management, vol. 35, no. 2, pp. 137 144, 2015.
2. A. O Driscoll, J. Daugelaite, and R. D. Sleator, Big Data , Hadoop and cloud computing in genomics, Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774 781, 2013.
3. C. L. P. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, vol. 275, pp. 314 347, 2014.
4. M. Herland, T. M. Khoshgoftaar, and R.Wald, A review of data mining using big data in health informatics, Journal of Big Data, vol. 1, no. 1, p. 2, 2014.
5. D. H. Shin and M. J. Choi, Ecological views of big data: Perspective and issues, Telematics and Informatics, vol. 32, no. 2, pp. 311 320, 2015.
6. B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. Basha, and P. Dhavachelvan, Big data and Hadoop-A study in security perspective, Procedia Computer Science, vol. 50, pp. 596 601, 2015.
7. X. Wu, X. Zhu, G. Q. Wu, and W. Ding, Data mining with big data, IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97 107, 2014.
8. S. Sharma and V. Mangat, Technology and trends to handle big data: Survey, in Proc. 5th International Conference on Advanced Computing & Communication Technologies, 2015, pp. 266 271.
9. R. Mehmood and G. Graham, Big data logistics: A healthcare transport capacity sharing model, Procedia Computer Science, vol. 64, pp. 1107 1114, 2015. [10] D. P. Augustine, Leveraging big data analytics and Hadoop in developing India healthcare services, International Journal of Computer Applications, vol. 89, no. 16, pp. 44 50, 2014.
10. J. A. Patel and P. Sharma, Big data for better health planning, in Proc. International Conference on Advances in Engineering and Technology Research, 2014, pp. 1 5.
11. A. E. Youssef, A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments, International Journal of Ambient Systems and Applications, vol. 2, no. 2, pp. 1 11, 2014.
12. MAPR, Healthcare and life scienceUsecases,https://mapr.com/sol utions/indu try/healthcare-and-lifescienceuse-cases/, 2018.
13. W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: Promise and potential, Health Information Science and Systems, vol. 2, no. 1, p. 3, 2014.
14. J. Sun and C. K. Reddy, Big data analytics for healthcare, in Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1525 1525.
15. C. Mike, W. Hoover, T. Strome, and S. Kanwal. Transforming health care through big data strategies for leveraging big data in the health care industry, http://ihealthtran.com/iHT2 BigData 2013.pdf, 2013.
16. J. Anuradha, A brief introduction on big data 5Vs characteristics and Hadoop technology, Procedia Computer Science, vol. 48, pp. 319 324, 2015.
17. M. Viceconti, P. J. Hunter, and R. D. Hose, Big data, big knowledge: Big data for personalized healthcare, IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209 1215, 2015.
18. Y. Sun, H. Song, A. J. Jara, and R. Bie, Internet of things and big data analytics for smart and connected communities, IEEE Access, vol. 4, pp. 766 773, 2016.
19. A. Jain and V. Bhatnagar, Crime data analysis using Pig with Hadoop, Procedia
20. Computer Science, vol. 78, pp. 571 578, 2016.
21. T. Jach, E. Magiera, and W. Froelich, Application of Hadoop to store and process big data gathered from an urban water distribution system, Procedia Engineering, vol. 119, pp. 1375 1380, 2015.
22. C. Uzunkaya, T. Ensari, and Y. Kavurucu, Hadoop ecosystem and its analysis on tweets, Procedia-Social and Behavioral Sciences, vol. 195, pp. 1890 1897, 2015.
23. S. G. Manikandan and S. Ravi, Big data analysis using Apache Hadoop, in Proc. International Conference on IT Convergence and Security, 2014, pp. 1 4.
24. V. Ubarhande, A. M. Popescu, and H. GonzalezVelez, Novel data-distribution

technique for Hadoop in heterogeneous cloud environment, in Proc. 9th International Conference on Complex, Intelligent, and Software Intensive Systems, 2015, pp. 217 224.

25. S. Maitrey and C. K. Jha, Handling big data efficiently by using map reduce technique, in Proc. International Conference on Computational Intelligence & Communication Technology, 2015, pp. 703 708.
26. J. Dean and S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM, vol. 51, no. 1, pp. 107 113, 2008.
27. Cloudera, Whole genome research drives healthcare to Hadoop, https://www.cloudera.com/content/dam/www/ marketing/resources/solutionbriefs/ whole-genomeresearch-inhealthcare.pdf.landing.html., 2018.
28. R. Misra, B. Panda, and M. Tiwary, Big data and ICT applications: A study, in Proc. 2nd International Conference on Information and Communication Technology for Competitive Strategies, 2016, p. 41.
29. A. G. Picciano, The evolution of big data and learning analytics in american higher education, Journal of Asynchronous Learning Networks, vol. 16, no. 3, pp. 9 20, 2012.
30. Apache Hadoop, http://hadoop.apache.org/, 2018.
31. A. Katal, M. Wazid, R. H. Goudar, and T. Noel, Big data: Issues, challenges, tools and good practices, in Proc. 6th International Conference on Contemporary Computing, 2013, pp. 404 409.
32. Apache Hive, https://hive.apache.org/, 2018.
33. K. K. Y. Lee, W. C. Tang, and K. S. Choi, Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage, Computer Methods and Programs in Biomedicine, vol. 110, no. 1, pp. 99 109, 2013.
34. Apache Pig, https://pig.apache.org/, 2018. [35] E. Dede, B. Sendir, P. Kuzlu, J.Weachock, M. Govindaraju, and L. Ramakrishnan, Processing Cassandra datasets with Hadoop-streaming based approaches, IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 46 58, 2016.
35. Apache HBase, http://hbase.apache.org/, 2018.
36. Apache Oozie, https://oozie.apache.org/, 2018.
37. Apache Avro, https://avro.apache.org/, 2018.
38. Apache Zookeeper, https://zookeeper.apache.org/, 2018.
39. Apache Zookeeper, https://www.ibm.com/analytics/ hadoop/zookeeper, 2018.
40. Apache Yarn, https://yarn.apache.org/, 2018.
41. Apache Sqoop, https://sqoop.apache.org/, 2018.
42. Apache Flume, https://flume.apache.org/, 2018

ID: 17

AI IN ROBOTICS

Mr.S.KALAICHELVAN
Student,
Department of Digital & Cyber Forensic Science,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

Ms.S.MADHUSRI
Student,
Department of Information Technnology,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

ABSTRACT:

The use of artificial intelligence (AI) is increasing significantly day by day, resulting in a reduction in human workload due to its ability to adapt to various situations and individuals. Furthermore, integrating AI with robotics provides specific traits necessary for using robots in environments where humans might not be able to survive, disregarding personal preferences. This chapter will provide a detailed explanation of how AI and robotics combine to create an ideal model. The integration process is still in the development stage, but it can potentially revolutionize various workspaces, including industries and households. AI acts as the system's brain, controlling all other components. Machine Learning, Natural Language Processing, Computer Vision, Deep Learning, Simulations, and Virtual Environments are some of the popular methods used to implement AI in robotics. However, these algorithms are still being researched for improved outcomes. This discussion focuses on the use of specific algorithms.In summary, this paper explores the utilization of AI-enabled robotics in various scenarios and the necessary system requirements for their implementation. Additionally, the paper highlights the challenges and advancements needed in the development process, while also discussing the drawbacks associated with using AI in robotics.

KEYWORDS: *Artificial Intelligence, Robotics, Machine Learning.*

I. INTRODUCTION

Artificial intelligence has emerged as a rapidly growing field in computer technology and is essential to the advancements of Industrial Revolution 4.0. It is intricately connected to other innovative technologies like the Internet of Things and Robotics. By integrating AI into Robotics, robots are enabled to collaborate and work independently without relying solely on user instructions. However, a major challenge in implementing AI in Robotics lies in the extensive time required for training datasets, labelling data, and training the AI system from scratch. Fortunately, there are prebuilt datasets available that can be utilized while developing your own system. The input data for AI in Robotics is gathered through sensors that detect and analyze the surrounding environment, catering to specific use cases. Various platforms are dedicated to practical analysis and learning processes, with the capability to incorporate AI even in edge devices. This amalgamation of IoT and AI is considered a groundbreaking approach. Our specific application of AI with Robotics leads to the development of self-operating machines equipped with vision, efficient workloads, motion control, and other essential functionalities.

II. PROBLEM IDENTIFICATION

Problem identification involves determining how to solve a problem using the latest technologies. In this case, developing new methods of solving problems more efficiently than traditional approaches adds value to your solutions. Before generating solutions, it is important to focus on the specific problem you are facing and find a solution that is suitable in all aspects. By utilizing analytics and statistical data, analyze the true nature of the problem and determine what you need to address. Think creatively and outside the box when identifying the problem. For instance, if your problem revolves around object detection or specific tasks, there are numerous existing algorithms and techniques as well as hardware options available in the industry. Assess your actual requirements and then attempt to solve the problem by adapting existing solutions or creating your own from scratch.

III. DATA COLLECTION

Data collection is a crucial and essential step in training AI models and creating AI-enabled robots. It provides the necessary information needed to train the models effectively. Therefore, it is important to gather clean and accurate data in order to achieve optimal output results. The collected data can consist of various types, such as images, videos, lidar scans, audio recordings, annotations, and labels. These data are considered raw and must be labeled and organized to train the AI using this dataset. When creating AI robots for real-time detection, it is necessary to employ high-quality sensors and high-resolution cameras to ensure the best results in optimal conditions. Additionally, data storage and management are important for continuous adaptation and improvement of the system. Effective data collection ensures that the AI models are exposed to diverse and representative data, allowing them to accurately learn from the data and perform their intended tasks.

IV. PREPROCESSING AND FEATURE ENGINEERING

The process of preprocessing involves organizing and correcting errors in raw data before using it in a dataset. This includes addressing noise, inconsistencies, missing values, and outliers. Once these corrections are made, the raw data can be labelled and used with algorithms. Scaling the data is necessary after error correction, and addressing missing values is crucial to prevent issues with the pretraining model. Each data point must be carefully analyzed before adding it to the algorithms. The preprocessing steps include Data Cleaning, Data Normalization, Handling Missing Data, and Dimensionality Reduction.

Feature Engineering, on the other hand, involves extracting meaningful data from existing ones to enhance the performance of algorithms. The extracted features are more useful than the original data, enabling faster detection and easier execution of algorithms. The steps involved in feature engineering are Feature Extraction, Feature Encoding, Feature Scaling, Feature Transformation, and Feature Combination.

V. ALGORITHM SELECTION

In order to solve a problem, it is usually necessary to use a suitable AI algorithm. You can either create your own algorithm or utilize open-source algorithms by adding your own datasets to them. However, creating your own algorithm is generally better. Additionally, you can even combine multiple algorithms to address your specific problems. It is crucial to prioritize execution speed when developing algorithms, as the entire system of a robot should operate simultaneously with minimal delay. Therefore, it is recommended to create algorithms with fast execution. The process of selecting an algorithm typically takes into account factors such as the nature of the data, complexity of the problem, available resources, and desired performance. This process involves problem understanding, data analysis, algorithm familiarization, algorithm evaluation, trade-offs and considerations, and iterating through the steps. Among these considerations, the selection of an appropriate algorithm is particularly significant. There are major categories of algorithms including supervised learning, unsupervised learning, reinforcement learning, and natural language processing. It is important to choose an algorithm that suits the specific problem at hand.

It is crucial to evaluate the problem using various algorithms and datasets. This allows for an assessment of execution speed, performance, dataset design, and identifies the advantages and disadvantages of the dataset. Testing the model repeatedly is essential to ensure that the AI model can train with all errors. Correcting errors during the preparation stage is important to prevent any damage to the hardware due to robot malfunctions. Optimal computation is necessary to achieve the desired output.

VI. MODEL TRAINING AND INTEGRATION

Model training involves training algorithms to detect and perform the specified tasks embedded in the model. During this process, the algorithm receives complete responses, and all system components work together to achieve the desired output. The embedding algorithm uses labelled or unlabelled data to learn patterns, relationships, and dependencies within the data, and optimize its internal parameters. The training process includes data preparation, model initialization, forward propagation, loss calculation, backpropagation, parameter update, iterative training, evaluation, hyperparameter tuning, early stopping, and model persistence. These factors need to be considered when creating an AI model. Errors can potentially cause severe damage, especially in the case of robots performing physical work, so an override function and emergency methods should be trained along with the preprocessing model. Training should be conducted and surveys taken to ensure maximum results are achieved.

Model integration involves the seamless integration of a trained AI model into the software or hardware architecture of a robot system. This encompasses integrating the model with the existing codebase, creating suitable interfaces, and guaranteeing that the robot can efficiently utilize the model for decision-making, perception, or any other pertinent tasks.

VII. TESTING AND ITERATION

Testing and iteration are essential in the development of AI systems for robotics. These steps involve evaluating the integrated AI models' performance, identifying areas for improvement, and refining the system through iterative cycles. Furthermore, it is crucial to analyze the entire model for bugs and issues in both hardware and software after testing. Some software may not perfectly engage with all hardware, so benchmarking can help identify system flaws and determine the need for further tests. Through continued testing and iteration, the system can achieve maximum results. In the context of AI in robotics, this involves testing setup, test plan design, test execution, performance evaluation, identifying weaknesses and areas for improvement, incorporating feedback, iterative refinement, real-world testing and validation, documentation, and reporting. By identifying weaknesses, gathering feedback, and making iterative updates, the system can evolve and adapt to various challenges, ultimately leading to improved performance.

VIII. DEPLOYMENT AND MONITORING

Real-time deployment of the robot system is crucial for effective analysis of required changes and evaluating hardware and software responses in real-world scenarios. Therefore, it is important to continually monitor the system until no potential issues arise in the future. By continuously monitoring the system, we can determine what modifications are needed and anticipate potential errors. It is essential to configure system parameters, settings, and calibration in order to achieve optimal performance in the deployment environment. This may involve adjusting sensor configurations, network connections, or communication protocols.

Performing these calibrations in the user environment can significantly improve the model's performance. Additionally, comprehensive safety and reliability checks should be conducted to ensure that the system operates within acceptable limits and complies with relevant regulations and standards. It is crucial to verify that all components, including AI models, hardware, and software, function as intended and meet safety requirements.

IX. CONTINOUS ADAPTATIONS AND IMPROVEMENT

Continuous adaptation and improvement are crucial in the development and deployment of AI-enabled robot systems. This process involves actively updating and enhancing the system based on feedback, new data, changing requirements, and evolving technologies.

To achieve long-term improvement, it is necessary to analyze the statistical data of the system over time. This analysis allows us to understand the system's progress and identify areas where enhancements can be made. Feedback from the system is particularly essential in this process, as it provides valuable insights into its performance, usability, and opportunities for improvement.

Collecting feedback can be done through various methods, such as surveys, interviews, user testing, or monitoring user interactions with the system. This feedback acts as a reflection of the system's past performance and gives us information on how the assigned tasks were executed.

By continuously adapting and improving the AI-enabled robot system based on feedback and data analysis, we can ensure its effectiveness and adaptability in response to evolving requirements and technologies.

X. CONCLUSION

The Creation AI-enabled robots is technically challenging due to the high cost of materials and the complexity involved in creating AI models that are suitable for various situations. We are currently in the early stages of developing AI-enabled robots, but they hold the key to a future revolution. The integration of AI into robotics has immense potential to enable intelligent, autonomous, and adaptable robot systems. By following steps such as problem identification, data collection, preprocessing, algorithm selection, model training, deployment, monitoring, and continuous adaptations and improvements, ultimately, the integration of AI into robotics has the potential to revolutionize industries such as manufacturing, healthcare, agriculture, and transportation. AI-enabled robots can perform complex tasks, assist humans in various domains, and operate in environments that were previously challenging or unsafe.AI can greatly improve the capabilities and performance of robots. As AI and robotics progress, it is crucial for researchers, engineers, and end-users to collaborate in order to unlock the full potential of this technology. By combining expertise, sharing knowledge, and embracing continuous improvement, AI-enabled robot systems will continue to bring about transformative changes.

REFERENCE

1. R. C. Arkin. Just what is a robot architecture anyway? Turing equivalency versus organizing principles. In AAAI Spring Symposium on Lessons Learned from Implemented Software Architectures for Physical Agents, 1995.
2. A. Bicchi, and G. Tonietti. Fast and soft arm tactics: Dealing with the safety-performance tradeoff in robot arms design and control. IEEE Robotics and Automation Magazine 11(2), 2004.
3. A. Bonarini, M. Matteucci, and M. Restelli. Filling the gap among coordination, planning, and reaction using a fuzzy cognitive model. In RoboCup 2003: Robot Soccer World Cup VII, pages 662–669, Berlin, Heidelberg, 2003. Springer-Verlag.
4. R. A. Brooks. A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation, 2(1), 1986.
5. C. Castelpietra, A. Guidotti, L. Iocchi, D. Nardi, and R. Rosati. Design and implementation of cognitive soccer robots. In RoboCup 2001: Robot Soccer World Cup V, pages 312–318, Berlin, Heidelberg, 2002. Springer-Verlag.
6. A. Chella, M. Frixione, and S. Gaglio. Understanding dynamic scenes. ArtiÞcial Intelligence, 123:89–132, 2000.
7. A. Chella, S. Gaglio, and R. Pirrone. Conceptual representations of actions for autonomous robots. Robotics and Autonomous Systems, 34:251–263, 2001.
8. L. Chittaro and A. Montanari. Efficient temporal reasoning in the cached event calculus. Computational Intelligence Journal, 12(3):359–382, 1996.
9. S. Coradeschi and A. Saffiotti. An introduction to the anchoring problem. Robotics and Autonomous Systems, 43(2-3):85–96, 2003.
10. P. I. Corke. Visual Control of Robots: HighPerformance Visual Servoing. Wiley, New York, 1996.
11. G. De Giacomo, L. Iocchi, D. Nardi, and R. Rosati. A theory and implementation of cognitive mobile robots. Journal of Logic and Computation, 5(9):759– 785, 1999.
12. R. Fikes and N. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. ArtiÞcial Intelligence, 2, 1971.
13. A. Finzi and F. Pirri. Combining probabilities, failures and safety in robot control. In Proceedings of IJCAI-01, pages 1331–1336, 2001.
14. E. Giunchiglia, G. N. Kartha, and V. Lifschitz. Representing action: Indeterminacy and ramifications. ArtiÞcial Intelligence, 95(2):409–438, 1997.
15. J. Hallam and H. Bruyninckx. An ontology of robotics science. In H.I. Christensen, editor, European Robotics Symposium 2006, pages 1–14, Berlin, Heidelberg, 2006. Springer-Verlag.
16. L.Iocchi, D. Nardi, M. Piaggio, and A. Sgorbissa. Distributed coordination in heterogeneous multi-robot systems. Autonomous Robots, 15:155– 168, 2003.

ID: 18

MACHINE LEARNING AND DEEP LEARNING APPLICATIONS

Ms.Dr.K.KAVITHA
Associate Professor,
Department of Computational Sciences,
Brainware University, Kolkata, West Bengal.

Dr.T.SARATHAMANI
Associate Professor,
Department of Computational Sciences,
Brainware University, Kolkata, West Bengal.

ABSTRACT

Machine Learning (ML) and Deep Learning (DL) have revolutionized various industries, offering a wide range of transformative applications. This chapter provides a comprehensive overview of the fundamental concepts, working methodologies, techniques, and algorithms utilized in both machine learning and deep learning. In the context of machine learning, the chapter covers the essential principles and approaches to model training, evaluation, and optimization. It delves into various types of machine learning algorithms, such as supervised, unsupervised, and reinforcement learning, elucidating their specific use cases and advantages.

Furthermore, the chapter delves into the working mechanism of deep learning, architecture of deep neural networks, types of deep learning models including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for sequential data analysis.

In addition to the theoretical aspects, the chapter extensively examines practical applications of machine learning and deep learning across different industries. It highlights successful use cases in computer vision, natural language processing (NLP), and healthcare, finance, and recommendation systems, showcasing the effectiveness of these technologies in solving real-world challenges.

I. An Introduction to Machine Learning

Machine learning is an evolving field that empowers computers to learn autonomously from historical data. By employing diverse algorithms, machine learning constructs mathematical models capable of making predictions based on past information. Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed.

The core idea behind machine learning is to develop algorithms that can automatically identify patterns, relationships, and insights from data, and then use this knowledge to make informed decisions or predictions about new, unseen data. This process is analogous to how humans learn from experience, generalize from past knowledge, and apply that understanding to new situations.

1.1. Features of Machine Learning

Machine learning encompasses a wide range of techniques and algorithms, each with its own unique set of features. Here are some key features of machine learning:

- **Learning from Data:** Machine learning algorithms are designed to learn, process and analyze large datasets to identify patterns, relationships, and insights from past data.
- **Automation:** Once trained, machine learning models can automate decision-making processes, reducing the need for manual intervention and enabling real-time predictions or responses.
- **Pattern Recognition:** Machine learning algorithms excel at recognizing complex patterns and structures in data that might not be easily identifiable through traditional rule-based programming.
- **Scalability:** ML algorithms are scalable to handle large datasets and complex problems efficiently.
- **Flexibility:** ML can adapt and learn from new data, making them more flexible and capable of handling dynamic and evolving scenarios without requiring code modifications.

1.2. Need of Machine Learning:

Machine learning plays a crucial role in addressing various challenges and leveraging opportunities in today's data-driven and technology-centric world. Here are some key reasons why we need machine learning:

- Machine learning enables efficient and effective analysis of large datasets to extract valuable insights and patterns.
- Machine learning excels at solving complex problems with numerous variables and interactions.
- Machine learning allows us to make accurate predictions and forecasts based on historical data. This capability is valuable in various applications, such as weather forecasting, stock market predictions, and customer behavior analysis.
- Machine learning helps optimize processes and decision-making in diverse domains, including supply chain management, transportation, and resource allocation.
- Machine learning models can continuously learn from new data, making them adaptive and responsive to evolving circumstances and changing conditions.

II. How Machine Learning Works?

Machine learning works by enabling computers to learn from past data and improve their performance on a specific task over time. The process typically involves the following steps:

- **Data Collection and Pre-processing:** The first step in machine learning is to collect relevant data for the task at hand. This raw data may contain noise, missing values, or inconsistencies that need to be cleaned and processed before feeding it into the machine learning algorithm.
- **Feature Extraction:** For certain types of data, feature extraction is performed to convert the raw data into a suitable format that can be used as input for the machine learning model. Feature extraction helps identify the most relevant information from the data.
- **Model Selection:** Choosing an appropriate machine learning algorithm or model is crucial. The selection depends on the type of problem, the nature of the data, and the desired output.
- **Model Training:** During the training phase, the selected model is fed with the preprocessed data. The model adjusts its internal parameters based on the data to learn patterns and relationships that relate inputs to outputs. During training, the model's performance is evaluated using a loss function, which measures how well the model's predictions match the actual outputs
- **Model Evaluation:** After training, the model's performance is evaluated on a separate dataset, called the test set. This evaluation provides an indication of how well the model will perform on new, unseen data.
- **Model Deployment:** Once the model demonstrates satisfactory performance, it can be deployed to make predictions or decisions on new, real-world data.

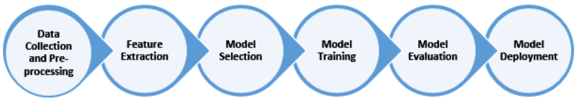


Fig.1 Working Methodology of Machine Learning

2. Classification of Machine Learning

Machine learning techniques can be broadly categorized into three main types based on the learning process and the availability of data. These types are supervised learning, unsupervised learning and reinforcement learning.

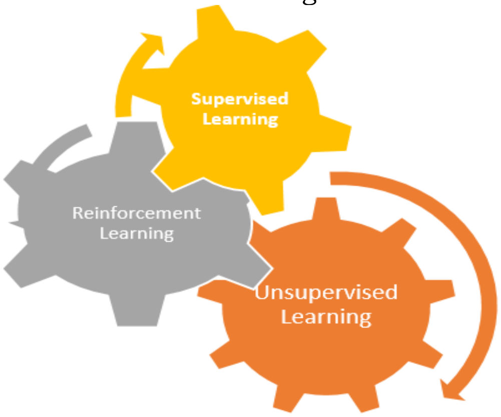


Fig.2 Classification of Machine Learning

2.1. Supervised Learning:

Supervised learning algorithms are trained on labeled data, where each input is associated with a corresponding output or target value. The goal of supervised learning is to learn a mapping between inputs and outputs so that the algorithm can make accurate predictions on new, unseen data. Classification and regression are two fundamental types of supervised learning tasks in machine learning.

Classification involves identifying or uncovering a model or function that aids in the division of data into distinct categories or discrete values. This method classifies data into various labels based on specific input parameters and subsequently predicts the appropriate labels for the given data. Common classification algorithms are Logistic Regression, Support Vector Machines, K-Nearest Neighbours, Naïve Bayes, Decision Tree Classification and Random Forest Classification

Machine Learning Regression is a method used to examine the association between independent variables or features and a dependent variable or outcome. It serves as a predictive modeling technique in machine learning, where an algorithm is leveraged to make predictions for continuous outcomes. Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression and Random Forest Regression are commonly used regression algorithms.

2.2. Unsupervised Learning:

Unsupervised learning algorithms are trained on unlabeled data, where the algorithm tries to identify patterns, structures, or relationships within the data without explicit guidance.

The primary goal of unsupervised learning is to explore the inherent structure of the data and group similar data points together. Common unsupervised learning tasks include clustering and regression.

Clustering is a technique used to group objects into clusters, where objects that share the most similarities are placed together within a cluster, while having minimal or no similarities with objects from other clusters. Cluster analysis identifies common traits among data objects and organizes them based on the presence or absence of those shared characteristics. Some of the commonly used algorithms are Common algorithms: K-Means Clustering, Hierarchical Clustering

An association rule is used to discover relationships between variables within extensive databases. It identifies sets of items that frequently occur together in the dataset and utilize association rules for the enhanced predictions. Commonly used algorithms are Apriori and Eclat.

2.3. Reinforcement Learning:

Reinforcement learning is a type of machine learning where an agent interacts with an environment and learns to take actions that maximize cumulative rewards. The agent receives feedback in the form of rewards or penalties based on its actions, allowing it to learn through trial and error to achieve specific goals.

The primary goal of reinforcement learning is to learn the best strategy or policy to maximize the cumulative reward over time.

Common algorithms used in reinforcement learning are Q-Learning, Deep Q Networks (DQNs), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG).

III. Applications of Machine Learning

Machine learning is a powerful technology that has a wide range of application areas across various industries. Some of the key application areas of machine learning include:

- **Natural Language Processing (NLP):** Machine learning enables machines to understand, interpret, and generate human language. NLP is used in applications such as language translation, sentiment analysis, chatbots, voice recognition, and text summarization.
- **Computer Vision:** Machine learning is applied to process and interpret visual information from the world. Computer vision is used in facial recognition, object detection, image classification, autonomous vehicles, medical imaging, and surveillance systems.
- **Financial Services:** Machine learning is used in fraud detection, credit risk assessment, algorithmic trading, and customer service in the financial industry.
- **Healthcare:** Machine learning is applied in medical diagnosis, drug discovery, personalized medicine, disease prediction, and patient monitoring.
- **Manufacturing and Industrial Processes:** Machine learning is used for predictive maintenance, quality control, process optimization, and supply chain management.
- **Marketing and Advertising:** Machine learning helps optimize advertising campaigns, target specific audiences, and analyze customer behaviour to improve marketing strategies.

- **Gaming:** Machine learning is used to create intelligent NPCs (Non-Playable Characters), adaptive game mechanics, and dynamic game environments.
- **Internet of Things (IoT):** Machine learning is applied in smart homes, smart cities, and industrial IoT applications for data analysis and decision-making.
- **Autonomous Systems:** Machine learning plays a crucial role in autonomous vehicles, drones, and robotics to perceive the environment and make decisions.
- **Energy Management:** Machine learning is used in energy consumption prediction, energy efficiency, and smart grid optimization.
- **Climate Prediction and Environmental Monitoring:** Machine learning is used to analyze large-scale environmental data for climate prediction, weather forecasting, and environmental monitoring.
- **Education:** Machine learning is applied to adaptive learning platforms, intelligent tutoring systems, and educational analytics to personalize learning experiences.
- **Human Resources:** Machine learning is used in candidate screening, workforce planning, and employee sentiment analysis.
- **Social Media and Content Filtering:** Machine learning algorithms are used to analyze user behavior and preferences to deliver relevant content and filter inappropriate content.
- **Recommendation Systems:** Machine learning algorithms are used to analyze user preferences and provide personalized recommendations in areas such as movie or product recommendations, music playlists, and online shopping.

IV. An Introduction to Deep Learning

Deep learning is a subfield of machine learning that uses artificial neural networks to learn complex patterns and representations from data. It is particularly effective for tasks involving large-scale data and has achieved remarkable success in areas like computer vision, natural language processing, and speech recognition. Deep learning involves training artificial neural networks with multiple layers to learn hierarchical representations of data. The term "deep" in deep learning refers to the depth of these neural networks, which typically comprise multiple hidden layers.

The primary goal of deep learning is to enable machines to automatically learn hierarchical representations of data from raw input. By learning these hierarchical representations, deep learning models can handle large and complex datasets, extract intricate patterns, and make accurate predictions or classifications in various domains.

V. How Deep Learning Works?

The working principles of deep learning are based on the structure and functioning of artificial neural networks. Deep learning works by using artificial neural networks with multiple layers (deep neural networks) to automatically learn hierarchical representations of data. The process involves the following steps:

- **Data Collection and Preprocessing:** Gathering relevant data for the task at hand is the first step. This data may include images, text, audio, or any other type of data depending on the application. The data is then preprocessed to ensure it is in a suitable format for training the deep learning model.
- **Neural Network Architecture:** A deep learning model consists of an input layer, one or more hidden layers, and an output layer. Each layer contains multiple artificial neurons (nodes) that process and transform the data. The hidden layers are responsible for learning intermediate representations of the data.
- **Activation Functions:** Activation functions introduce non-linearity to the neural network, allowing it to learn complex patterns and relationships in the data. Common activation functions include ReLU (Rectified Linear Unit), Sigmoid, Tanh, and Softmax.
- **Forward Propagation:** During the training process, data is fed into the input layer, and it propagates through the neural network layer by layer. At each layer, the neurons perform a computation on the data, combining inputs with their associated weights and biases, and passing the result through the activation function.
- **Loss Function:** The output of the neural network is compared to the ground-truth labels (target values) using a loss

function, which quantifies the difference between the predicted outputs and the actual targets. The choice of the loss function depends on the nature of the task, such as mean squared error for regression tasks or categorical cross-entropy for classification tasks.

- **Backpropagation:** Backpropagation is the core training algorithm in deep learning. It calculates the gradients of the loss function with respect to the model's parameters (weights and biases). These gradients indicate how much each parameter should be adjusted to minimize the error between predicted and actual values.
- **Optimization:** Optimization algorithms like stochastic gradient descent (SGD) or its variants use the calculated gradients to update the model's parameters, iteratively reducing the error during training.
- **Iterative Training:** The training process is repeated for multiple epochs (iterations over the entire dataset) to fine-tune the model's parameters and improve its performance.
- **Generalization:** The goal of deep learning is to achieve good generalization, where the trained model can make accurate predictions on new, unseen data.

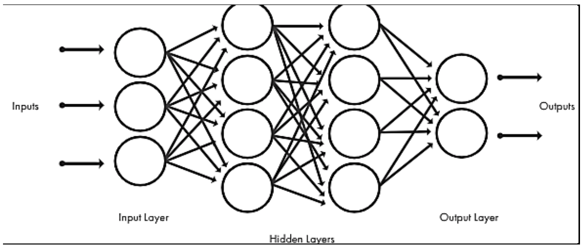


Figure: Structure of Neural Layers in Deep Learning Models.

VI. Types of Deep Learning Models

Deep learning encompasses various types of models, each tailored to specific tasks and data types. Some of the commonly used types of deep learning models include:

Convolutional Neural Networks (CNNs)

CNNs are primarily used for computer vision tasks, such as image classification, object detection, and image segmentation. They leverage convolutional layers to automatically learn spatial hierarchies of features from images.

Recurrent Neural Networks (RNNs)

RNNs are designed for sequential data, like natural language processing, time-series analysis, and speech recognition. They have connections that form loops, allowing them to maintain hidden states and process sequences of data.

Long Short-Term Memory (LSTM) Networks:

LSTM is a specialized type of RNN that addresses the vanishing gradient problem and can handle long-range dependencies in sequential data.

Gated Recurrent Units (GRUs)

GRUs are another type of RNN that can handle sequential data and are computationally more efficient than LSTMs.

Transformers:

Transformers have revolutionized natural language processing and are widely used for tasks like machine translation, text generation, and language understanding. They leverage self-attention mechanisms to capture long-range dependencies in input sequences.

Autoencoders

Autoencoders are used for unsupervised learning and dimensionality reduction tasks. They aim to learn compressed representations of the input data and then reconstruct the original data from these compressed representations.

Generative Adversarial Networks (GANs):

GANs consist of two neural networks, a generator, and a discriminator, that compete against each other during training. They are used for generating realistic data, such as images, music, or text.

Variational Autoencoders (VAEs):

VAEs combine elements of autoencoders and probabilistic models, enabling them to generate new data points similar to the training data.

Siamese Networks:

Siamese networks are used for similarity-based tasks, such as facial recognition and signature verification. They learn representations of input pairs and measure their similarity in the learned feature space.

Deep Reinforcement Learning:

Deep reinforcement learning combines deep learning with reinforcement learning to train agents that can make decisions in environments based on rewards and punishments.

3. Deep Learning Applications

Deep learning has revolutionized various industries and applications due to its ability to process large amounts of data and automatically learn patterns and features from it. Here is a detailed note on some of the key applications of deep learning:

3.1. Large Scale Deep Learning

- Large-scale deep learning is utilized to tackle intricate tasks that demand extensive computational power and efficient management of vast datasets. These models are capable of handling massive datasets, complex architectures, and high-performance computing resources, making them well-suited for addressing these challenges effectively.
- DL is essential to handle the large scale applications such as parallel computing, distributed training and so on.

3.2. Computer Vision

- Deep learning models can classify images into predefined categories, making them useful in areas like medical diagnosis, facial recognition, and autonomous vehicles.
- It also enables real-time detection of multiple objects within images or videos, leading to applications in surveillance, robotics, and augmented reality.
- Deep learning can separate different objects or regions within an image, assisting in medical imaging, image editing, and satellite imagery analysis.

3.3. Speech Recognition

- Deep learning models can convert spoken language into text, enabling voice assistants, transcription services, and voice-controlled devices.
- DL can generate human-like speech from text, facilitating accessibility services, audiobook production, and voice-over applications.

3.4. Natural Language Processing

- Deep learning models can determine the sentiment or emotion expressed in text, aiding businesses in understanding customer feedback and social media analysis.
- DL has greatly improved machine translation systems, making services like Google Translate more accurate and efficient.
- It is also used to extract entities such as names, locations, and dates from unstructured text, which is essential for information retrieval and question answering systems.

3.5. Recommended System

- Deep learning-based recommendation engines can analyze user preferences and behaviours to suggest personalized products, movies, music, and content, enhancing user experience and increasing engagement for businesses.

3.6. Knowledge Representation

- Deep learning techniques are used to encode and process information in a manner that captures the underlying relationships and structure within the data. Knowledge representation is a fundamental aspect of AI systems that aims to organize and model knowledge in a format that is suitable for reasoning, learning, and decision-making.

3.7. Other Applications

- **Autonomous Vehicles:** Deep learning plays a crucial role in enabling self-driving cars by processing sensor data, detecting obstacles, and making real-time driving decisions.
- **Robotics:** Deep learning empowers robots to perceive their environment, manipulate objects, and adapt to changing conditions, enabling tasks in manufacturing, healthcare, and logistics.
- **Medical Imaging Analysis:** Deep learning models excel at detecting abnormalities in medical images such as X-rays, MRIs, and CT scans, aiding in disease diagnosis and early detection.
- **Drug Discovery:** Deep learning is utilized in drug design and discovery,

predicting molecular interactions and identifying potential drug candidates.

- **Fraud Detection:** Deep learning helps in identifying fraudulent transactions by analyzing patterns and anomalies in financial data.
- **Stock Market Prediction:** Deep learning models are used to forecast stock prices and market trends.
- **Environmental Monitoring:** Deep learning is applied to analyze satellite images and sensor data to monitor environmental changes, track deforestation, manage natural disasters, and study climate patterns.

4. Future of Machine Learningv and Deep Learning

The future of machine learning and deep learning is full of immense promise and potential, as ongoing progress and revolutionary effects are anticipated across diverse domains. This technological advancement is poised to profoundly influence numerous aspects of our lives and various industries. Here are some key trends and possibilities that may shape the future of machine learning:

- **Advanced Deep Learning Techniques:** Future developments in Machine Learning are likely to lead to even more powerful deep learning models that can tackle complex problems with greater accuracy and efficiency.
- **Explainable AI and Interpretable Models:** Researchers are focusing on developing methods to make machine learning and deep learning models more transparent and understandable, allowing humans to comprehend and trust the decisions made by AI systems.
- **Transfer Learning:** Transfer learning techniques, which allow models to leverage knowledge from one domain to perform better in related domains, will continue to play a vital role in training AI systems with limited data.
- **Edge AI and Federated Learning:** Edge AI, which involves running AI algorithms directly on edge devices like smartphones and IoT devices, will become more prevalent, enabling real-time and privacy-preserving applications. Federated learning, a decentralized learning approach, will allow models to be trained across distributed devices without sharing raw data, enhancing data privacy.
- **Autonomous Systems and Robotics:** Machine learning and deep learning model will be instrumental in the development of autonomous systems and robots, enabling them to perceive, learn, and adapt to dynamic environments and perform complex tasks in various industries, such as transportation, logistics, and healthcare.
- **AI in Healthcare and Drug Discovery:** Both ML and DL will continue to revolutionize healthcare by assisting in medical diagnosis, personalized treatment plans, and drug discovery. AI-driven medical imaging analysis and patient care systems will become more prevalent.
- **Quantum Machine Learning:** With advancements in quantum computing, quantum machine learning and deep learning models holds the potential to solve complex problems exponentially faster, significantly expanding the capabilities of machine learning models.
- **AI for Sustainability and Climate Change Solutions:** Machine learning can play a vital role in addressing environmental challenges and developing solutions for climate change, such as optimizing energy usage, managing resources, and monitoring environmental impacts.
- **Human-AI Collaboration:** The future of machine learning and deep learning involves seamless integration between humans and AI systems, enabling more effective and natural interactions. Human-AI collaboration will lead to the coexistence of AI systems as tools that augment human capabilities.

CONCLUSION

Ultimately, this chapter serves as a comprehensive guide for readers to understand the foundations of machine learning and deep learning, their practical implementation, and their transformative impact on various industries. By exploring the recent application areas and discussing the effectiveness of these technologies, the chapter offers valuable insights into harnessing the full potential of machine learning and deep learning for solving complex problems in the modern AI-driven world. This chapter will also provides insights into the ongoing research and future directions of machine learning and deep learning applications.

REFERENCE

1. Machine Learning and Deep Learning using Python and Tensor Flow Paperback, Venkata Reddy Konasani, McGraw Hill, First Edition, 2021.
2. Machine Learning, Zhi-Hua Zhou, Springer, 2021.
3. The Hundred-Page Machine Learning Book, Andriy Burkov, Ingram Short Title, 2019.
4. An Introduction to Machine Learning, Miroslav KubatSpringer, Second Edition, 2017.
5. Machine Learning for Absolute Beginners, Oliver Theobald, Second Edition, 2017.
6. Deep Learning Paperback, Ian Goodfellow, Yoshua Bengio, Aaron Courvillc, The MIT Press, Illustrated Edition 2019.
7. A Practical Approach for Machine Learning and Deep Learning Algorithms, Abhishek Kumar Pandey, Pramod Singh Rathore, Dr. S. Balamurugan, BPB Publications, 2019.
8. Deep Learning, John D. Kelleher, MIT Press, 2019
9. Learning Deep Learning, Ekman, Pearson Education, First Edition, June 2022
10. Deep Learning: Algorithms and Applications, Pedrycz, Springer, First Edition Kindle Edition, 2020.

ID: 19

AI TRANSPORT MANAGEMENT SYSTEM

Ms.V.M.SAHIRA
Research Scholar,
Department of Computer Science,
Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

Ms.K.P.SREELAKSHMI
Research Scholar,
Department of Computer Science,
VEL's University,
Chennai, Tamil Nadu, India.

I. TRANSPORT MANAGEMENT SYSTEM

The greatest technological boom and advancement of modern times are currently taking place. These new technologies bring progress to every aspect of business, including customer facing technologies, IoT in Transport Management. IoT is Internet of Things, It is a network of interconnected devices, generally via wi-fi, software and sensors, which exchange data. These devices can include, but are not limited to, vehicles, electronics, home appliance and robots.

KEYWORDS: *Transport Management, IoT*

II. LITERATURE REVIEW

A Literature Review on the Intelligent Transportation System using Deep Learning and Machine Learning Techniques 1 Dr. K.S.Ramakrishnan, 2 Dr. C. Jothi Venkateswaran, 1 Assistant Professor, 2 Joint Director (P & D) 1 School of Education,2 Directorate of Collegiate Education 1 Tamil Nadu Open University, Chennai, India proposed that Road traffic accidents are very common. An estimated 1.2 million deaths and 50 million injuries happen all over the world every year. In this emerging world, the road accidents are among the principal reason of fatality and injury. The current research represents that the Machine Learning techniques and Big Data Analytics that are employed in the field of transport have been investigated. Through this comprehensive investigation, the techniques of Machine Learning and Big Data Analytics in the traffic study can enhance the administration level of road traffic safety productively.

III. PROPOSED SYSTEM

1. Computer vision & AI enable pedestrian and vehicle safety

Ensuring public road safety is a crucial responsibility of transportation management systems. The complex environment created by vehicles and pedestrians needs to be kept under close surveillance, in order to ensure the safety of all traffic participants. Luckily, technology is available that makes it possible to automate such surveillance tasks and delegate them to software and algorithms. Computer vision and video analytics can be implemented both on roadside cameras, but also on cars. Algorithms can perform computation on the edge and can detect situational and behavioral abnormalities at the moment when they happen. From the automated reading of license plates to detecting walking patterns, a variety of applications become possible thanks to computer vision. When implemented as part of traffic management systems, they can minimize the high risks associated with careless driving and ensure the safety of public pedestrian areas.

2. IoT Sensors enable accurate traffic monitoring in smart cities

Understanding traffic is a task that needs to be done in real-time, in order to be able to optimize the traffic flow, both within and outside of urban areas. This involves the identification and communication of accidents, congestion, and temporary roadside obstacles, among other traffic events. Sensor technologies and advanced wireless communication protocols make it possible for all kinds of vehicles to communicate direction, speed, and travel times. There is no limit to the amount of information that they can communicate, given the increased customizability of IoT Devices.

Not only can they be attached to any moving object, but they also make it possible to collect and communicate contextual information from the environment. Sensor-collected data makes it possible to run real-time analytics that power immediate traffic management decisions. Such an example application is that of adaptive traffic signals, which are not simply programmed, but take into account live traffic information.

IV. CONCLUSION

In this study, we primarily examine the challenges that Transportation System currently faces before looking toward the use of Artificial Intelligence in the Transportation System. Afterward, we proposed a method that uses Artificial Intelligence to govern the problems of Transportation Management System. We use Sensors to track the speed of the system in order to avoid traffic congestion.

V. REFERENCES

1. Al Qundus, J., Dabbour, K., Gupta, S., Meissonier, R., &Paschke, A. (2020). Wireless sensor network for AI-based flood disaster detection. *Annals of Operations Research*, 1-23.
2. Mitra, A., Basu, D., &Ghosh, A. (2022). Swarm Intelligence-Based Smart City Applications: A Review for Transformative Technology with Artificial Intelligence. *Data Science and Security*, 81-92.
3. Sharma, H., Haque, A., &Blaabjerg, F. (2021). Machine learning in wireless sensor networks for smart cities: a survey. *Electronics*, *10*(9), 1012.
4. Kalinin, M., &Zegzhda, P. (2020, November). AI-based Security for the Smart Networks.In *13th International Conference on Security of Information and Networks* (pp. 1-4).
5. Ovasapyan, T. D., Danilov, V. D., &Moskvin, D. A. (2021). Application of synthetic data generation methods to the detection of network attacks on the internet of things devices. *Automatic Control and Computer Sciences*, *55*(8), 991-998.
6. Ognev, R. A., Zhukovskii, E. V., &Zegzhda, D. P. (2021). Detection of Malicious Executable Files Based on Clustering of Activities. *Automatic Control and Computer Sciences*, *55*(8), 1092-1098.
7. Ovasapyan, T. D., Knyazev, P. V., & Moskvin, D. A. (2021). Automated search for vulnerabilities in arm software using dynamic symbolic execution. Automatic control and computer sciences, 55(8), 932-940.

ID: 20

PARKINSONS DISEASE DETECTION USING CLASSIFICATION ALGORITHMS

Dr.R.SABIN BEGUM
Assistant Professor,
Department of Computer Applications,
B.S.Abdur Rahman Crescent Institute of Science and Technology,
Chennai, Tamil Nadu, India.

Dr.A.HAJA ALAUDEEN
Assistant Professor,
Department of Computer Applications,
B.S.Abdur Rahman Crescent Institute of Science and Technology,
Chennai, Tamil Nadu, India.

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative movement disorder that affects a significant number of individuals globally. As time passes, the symptoms of the disease gradually worsen, making it difficult for non-specialist clinicians to diagnose it accurately, particularly in its early stages. The initial symptoms include a slight tremor in one hand and a feeling of stiffness in the body, which can progress to more severe manifestations. The complex and diverse nature of PD symptoms, such as depression, anxiety, sleep disturbances, memory problems, and loss of smell, poses challenges for proper identification and treatment. Currently, manual examinations result in a considerable misdiagnosis rate, leading to delayed therapeutic interventions. To tackle these diagnostic issues, our proposed solution involves a comparative analysis of machine learning algorithms such as Xgbooster, KNN, and SVM for early detection of PD. This innovative approach aims to enhance the patient's lifespan and improve their overall well-being by facilitating timely medical interventions. The effectiveness of our system was validated using a Kaggle dataset, which showcased the potential of machine learning techniques to revolutionize PD diagnosis and subsequently improve patient care outcomes.

KEYWORDS: *Xgbooster, KNN, SVM.*

I. INTRODUCTION

Parkinson's disease is a neurological disorder that affects the brain, resulting in body tremors and stiffness. Despite significant research, there is currently no definitive cure or treatment available for the advanced stages of the disease. Early detection and intervention are crucial for effective management, reducing the overall cost burden and potentially improving patient outcomes.

The disease's symptoms may include depression, sleep disturbances, memory issues, anxiety, and a loss of the sense of smell. These signs become more prominent as nerve cells in the basal ganglia, responsible for controlling movement, become impaired or die. These cells typically produce dopamine, a critical brain chemical that regulates movement. However, in Parkinson's disease, the reduction of dopamine production leads to movement problems, though the exact cause of neuron death remains unknown.

In the existing system, PD is detected at the secondary stage only which leads to medical challenges. Also, doctor must manually examine and suggest medical diagnosis inwhich the symptoms might vary from person to person so suggesting medicine is also a challenge. Thus, the mental disorders are been poorly characterized and have many health complications.The result in a high misdiagnosis rate and many years before diagnosis, people can have the disease. Thus, the existing system is not effective in early prediction and accurate medicinal diagnosis to the affected person.

Due to these limitations, the existing system exhibits a high misdiagnosis rate, with the disease often going undetected for many years. Consequently, early prediction and accurate medicinal diagnosis for affected individuals are not effectively achieved using the current approach.

This research paper aims to explore the potential of classification algorithms for Parkinson's disease detection. Specifically, we will compare and evaluate the performance of popular algorithms such as Xgbooster (extreme gradient boosting), KNN (k-nearest neighbor), and SVM (support vector machine) using a dataset comprising voice features. Previous studies have demonstrated the correlation between voice characteristics and Parkinson's disease, making it a promising avenue for early detection.

The objectives of this research paper are threefold: first, to investigate the effectiveness of classification algorithms in distinguishing Parkinson's disease from healthy individuals; second, to compare the performance of different algorithms to identify the most accurate and efficient approach.

II. METHODOLOGY

Methodology pertains to the systematic approach or technique employed for organizing and analyzing information. The system adopts a top-down approach, which prioritizes thorough planning and comprehensive of the system. Python is the programming language utilized for developing this research work .The below figure 2.1 shows that System Architecture Diagram.

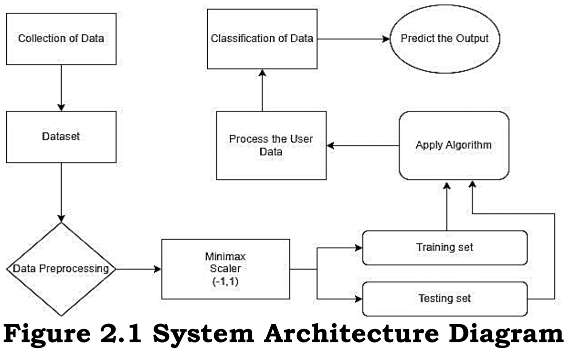


Figure 2.1 System Architecture Diagram

2.1.1. Machine learning

Machine learning involves the scientific examination of algorithms and statistical models utilized by computer systems to accomplish tasks effectively without explicit instructions, relying on patterns and inference instead. It falls under the umbrella of artificial intelligence. Machine learning algorithms construct mathematical models from sample data, known as "training data," enabling them to make predictions or decisions without explicit programming for the task. These algorithms find applications in various domains, such as email filtering and computer vision, where creating task-specific algorithms is impractical. Machine learning is closely linked to computational statistics, which revolves around making predictions using computers. The study of mathematical optimization contributes methods, theory, and application domains to the field of machine learning. Additionally, data mining is a subfield of machine learning that emphasizes exploratory data analysis through unsupervised learning techniques.

2.1.2. Xgbooster

To tackle the Parkinson's disease detection challenge, we employed the Xgbooster algorithm, renowned for its effectiveness with large datasets. This algorithm, introduced by data scientist Tianqi Chen in 2014, combines software and hardware optimization techniques, delivering superior results while utilizing fewer computing resources in record time. Compared to other machine learning techniques, Xgbooster stands out for its simplicity, accessibility, and advanced features, making it versatile across Windows, Linux, and OS X platforms.

Xgbooster is based on the gradient boosting method and excels in regression, classification, and ranking tasks. Utilizing a loss function, weak learner, and additive model, it effectively minimizes errors in decision trees. The algorithm's scalability allows it to rapidly generate billions of outcomes, and its versatility extends beyond tree learning to encompass linear models. Given its exceptional speed, accuracy, and accessibility, Xgbooster has gained popularity across various industries, with companies like Google and start-ups incorporating it for diverse applications. As an open-source software library, Xgbooster proves to be a powerful tool in addressing regression, classification, and ranking challenges.

2.1.3. K- Nearest Neighbor

The K Nearest Neighbors (KNN) classification method is an effective approach for solving non-linearly classified data points. When a new data point is introduced in the classification process, a value of k is chosen, representing the number of nearest neighbors to consider based on distance. Euclidean Distance is commonly used to calculate the distance from the new data point to other data points.

For two points, P1 ($x1$, $x2$) and P2 ($y1$, $y2$), the Euclidean Distance (d) is computed as $d = \sqrt{}(x2 - x1)^2 + (y2 - y1)^2$.

2.1.4 Support Vector Machine

SVM proves valuable in addressing both classification and regression problem statements. In classification tasks, SVM utilizes a hyperplane to separate or classify two classes. This hyperplane is accompanied by two marginal lines at a specific distance, ensuring linear separability for the classification points. Additionally, two parallel planes are created, passing through the nearest points of the two classes (support vectors). The distance between these planes, known as the marginal distance, acts as a cushion, facilitating a more accurate point classification. The best hyperplane is chosen based on its maximum marginal distance.

III. Results and Discussion

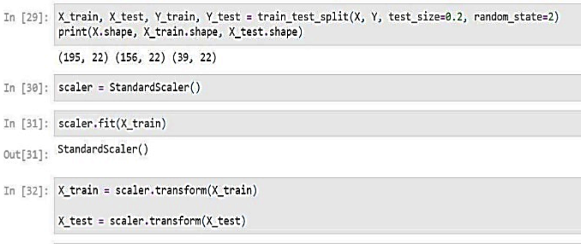
The results of proposed approach are discussed in following sections.

3.1.1. Data Collection and Pre-Processing

The text-based classification dataset was obtained from the Kaggle website. Before using the dataset for analysis, preprocessing was performed to eliminate any noise data, specifically by removing the Null values.

3.1.2. Training Data and Test Data

Machine learning involves extracting information or properties from a dataset to classify and test properties of another set. Typically, data is split into two categories: a training set used for learning properties, and a testing set for assessing properties, which may differ from those in the training set. Using various classifiers with the training set can yield different results and accuracies.Among these classifiers, deep learning stands out for its highest classification accuracy, making it suitable for further analysis and consideration.



```
In [29]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
         print(X.shape, X_train.shape, X_test.shape)

         (195, 22) (156, 22) (39, 22)

In [30]: scaler = StandardScaler()

In [31]: scaler.fit(X_train)

Out[31]: StandardScaler()

In [32]: X_train = scaler.transform(X_train)

         X_test = scaler.transform(X_test)
```

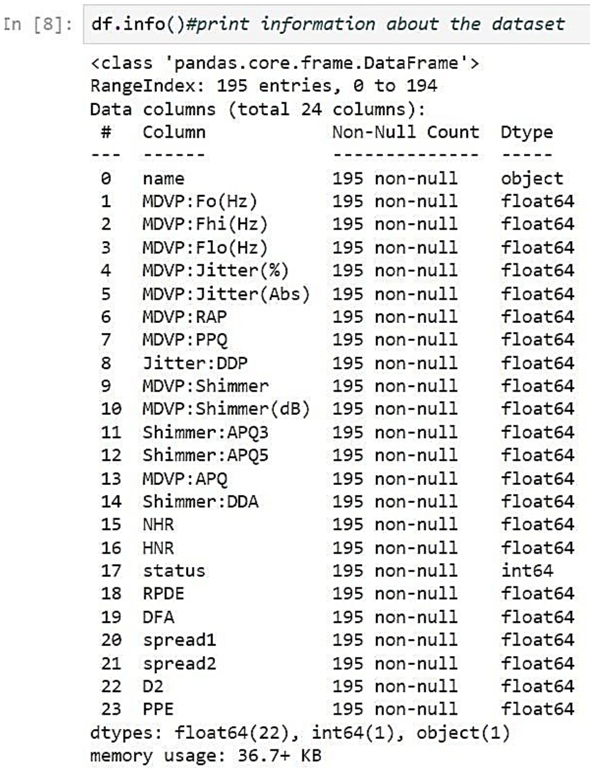
Figure 3.1. Training and Testing Data

3.1.3. Data Analysis

Data analysis involves the systematic cleaning, transformation, and processing of raw data to derive meaningful and relevant insights that support informed decision-making for businesses. By presenting valuable information in charts, images, tables, and graphs, the analysis helps mitigate decision-making risks, providing actionable statistics and insights.

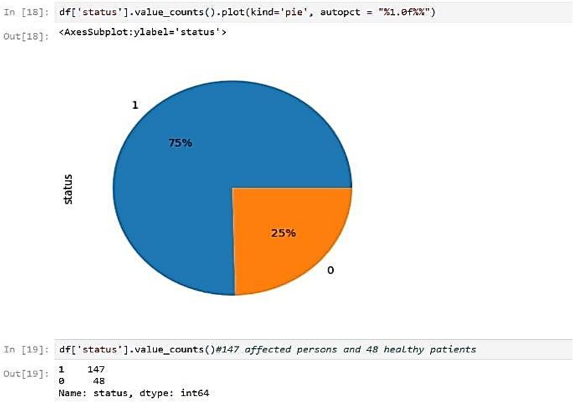
3.1.4. Descriptive Analytics

Data analysis is a methodical approach that encompasses the cleansing, transformation, and processing of raw data to extract valuable and pertinent insights. These insights play a pivotal role in facilitating informed decision-making for businesses. By presenting the information in visually appealing formats such as charts, images, tables, and graphs, data analysis aids in reducing decision-making risks. It provides actionable statistics and valuable insights that empower businesses to make well-informed choices.



```
In [8]: df.info()#print information about the dataset

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 195 entries, 0 to 194
        Data columns (total 24 columns):
         #   Column            Non-Null Count  Dtype  
        ---  ------            --------------  -----  
         0   name              195 non-null    object 
         1   MDVP:Fo(Hz)       195 non-null    float64
         2   MDVP:Fhi(Hz)      195 non-null    float64
         3   MDVP:Flo(Hz)      195 non-null    float64
         4   MDVP:Jitter(%)    195 non-null    float64
         5   MDVP:Jitter(Abs)  195 non-null    float64
         6   MDVP:RAP          195 non-null    float64
         7   MDVP:PPQ          195 non-null    float64
         8   Jitter:DDP        195 non-null    float64
         9   MDVP:Shimmer      195 non-null    float64
         10  MDVP:Shimmer(dB)  195 non-null    float64
         11  Shimmer:APQ3      195 non-null    float64
         12  Shimmer:APQ5      195 non-null    float64
         13  MDVP:APQ          195 non-null    float64
         14  Shimmer:DDA       195 non-null    float64
         15  NHR               195 non-null    float64
         16  HNR               195 non-null    float64
         17  status            195 non-null    int64  
         18  RPDE              195 non-null    float64
         19  DFA               195 non-null    float64
         20  spread1           195 non-null    float64
         21  spread2           195 non-null    float64
         22  D2                195 non-null    float64
         23  PPE               195 non-null    float64
        dtypes: float64(22), int64(1), object(1)
        memory usage: 36.7+ KB
```

Figure 3.2. Data Analysis



```
In [18]: df['status'].value_counts().plot(kind='pie', autopct = "%1.0f%%")
Out[18]: <AxesSubplot:ylabel='status'>
```

```
In [19]: df['status'].value_counts()#147 affected persons and 48 healthy patients
Out[19]: 1    147
         0     48
         Name: status, dtype: int64
```

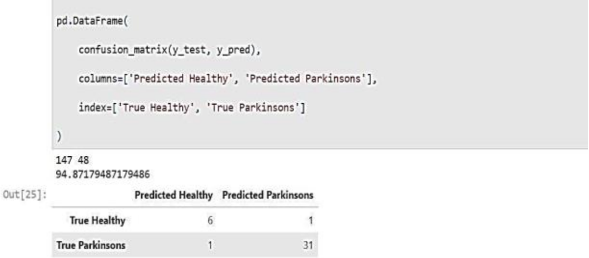
Figure 3.3 Descriptive Analytics – Data Visualization

3.1.5. Predictive Analytics:

Predictive analytics is a powerful tool that leverages historical data, statistical modeling, data mining techniques, and machine learning to make predictions about future outcomes. Companies utilize predictive analytics to identify patterns in the data, enabling them to uncover risks and opportunities.

Xgbooster:

To address the challenge of Parkinson's disease detection, we employed the Xgbooster algorithm. This algorithm is particularly beneficial when dealing with large datasets. It is an exceptional amalgamation of software and hardware optimization techniques, delivering superior results while utilizing fewer computing resources and minimizing processing time. Xgbooster is an extreme gradient boosting method introduced by data scientist Tianqi Chen in 2014.

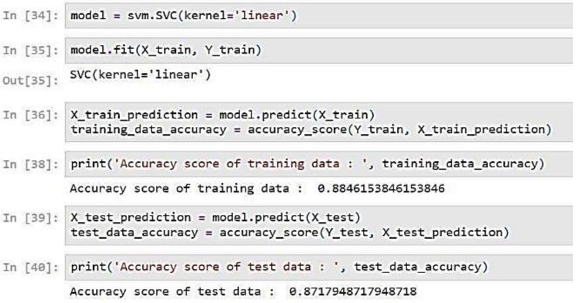


```
pd.DataFrame(

    confusion_matrix(y_test, y_pred),

    columns=['Predicted Healthy', 'Predicted Parkinsons'],

    index=['True Healthy', 'True Parkinsons']

)
147 48
94.87179487179486
Out[25]:                Predicted Healthy  Predicted Parkinsons
            True Healthy                6                     1
         True Parkinsons                1                    31
```

Figure 3.4 a) Predictive Analytics – Xgbooster

3.1.6. Support Vector Machine:

SVM is a versatile tool for addressing both classification and regression problems. In classification tasks, SVM separates two classes by creating a hyperplane. To ensure easy linear separability for the classification points, SVM generates two marginal lines along with the hyperplane at a certain distance. Additionally, two parallel planes are formed, passing through the nearest points of the two classes (support vectors). The distance between these planes, known as the marginal distance, serves as a cushion that enhances the effective division of points into classes

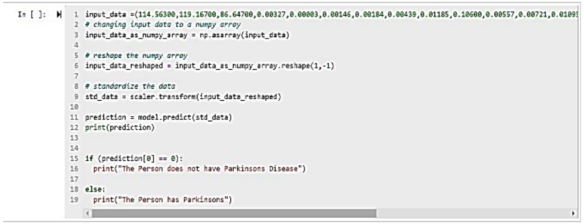


```
In [34]: model = svm.SVC(kernel='linear')

In [35]: model.fit(X_train, Y_train)
Out[35]: SVC(kernel='linear')

In [36]: X_train_prediction = model.predict(X_train)
         training_data_accuracy = accuracy_score(Y_train, X_train_prediction)

In [38]: print('Accuracy score of training data : ', training_data_accuracy)
         Accuracy score of training data :  0.8846153846153846

In [39]: X_test_prediction = model.predict(X_test)
         test_data_accuracy = accuracy_score(Y_test, X_test_prediction)

In [40]: print('Accuracy score of test data : ', test_data_accuracy)
         Accuracy score of test data :  0.8717948717948718
```

Figure 3.4 b) Predictive Analytics – SVM

3.1.7. K-Nearest Neighbor:

K Nearest Neighbors (KNN) classification is a highly effective approach for handling non-linearly classified data points. In this method, when a new data point is introduced in the classification process, a value of k is chosen to determine the number of nearest neighbors considered based on distance. The distance from the new data point to others can be calculated using Euclidean Distance. For two points, P1 $(x1, x2)$ & P2 $(y1, y2)$, the Euclidean Distance (d) is calculated as $d = \sqrt{}(x2 - x1)$^2 + $(y2 - y1)$^2. Alternatively, another distance metric known as Manhattan Distance can also be used.



```
In [ ]: input_data =(114.56300,119.16700,86.64700,0.00327,0.00003,0.00146,0.00184,0.00439,0.01185,0.10600,0.00557,0.00721,0.0109
        # changing input data to a numpy array
        input_data_as_numpy_array = np.asarray(input_data)

        # reshape the numpy array
        input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

        # standardize the data
        std_data = scaler.transform(input_data_reshaped)

        prediction = model.predict(std_data)
        print(prediction)


        if (prediction[0] == 0):
          print("The Person does not have Parkinsons Disease")

        else:
          print("The Person has Parkinsons")
```

Figure 4.8 Predictive System Comparison of Algorithms

It concluded that the Xgbooster gives more accuracy than others. Hence using the Xgboosteralgorithm

Table 3.1 Comparison of accuracy

Algorithms	Accuracy
Xgbooster	94.871
K-NN	89.74
SVM	88.46

IV. CONCLUSION

This study offers a comprehensive survey of recent neuroimaging modalities and associated analysis techniques used in diagnosing Parkinson's disease. Unlike previous review papers that focused on specific imaging modalities or dementia types, this research covers a broader spectrum of machine learning technologies for mental illness diagnostics. By doing so, researchers in the field can readily identify the latest advancements in this domain.

The research work also highlights the significance of early detection and prediction of Parkinson's disease. Early identification allows for timely treatment and support for patients, enhancing their overall management and well-being.

V. REFERENCES

1. Abos, A., Baggio, H. C., Segura, B., Campabadal, A., Uribe, C., Giraldo, D. M., et al. (2019). Differentiation of multiple system atrophy from Parkinson's disease by structural connectivity derived from probabilistic tractography. Sci. Rep. 9:16488. doi: 10.1038/s41598-019-52829-8.
2. Abujrida, H., Agu, E., and Pahlavan, K. (2017). "Smartphone-based gait assessment to infer Parkinson's disease severity using crowdsourced data," in 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT) (Bethesda, MD), 208–211. doi: 10.1109/HIC.2017.8227621.
3. Benba, A., Jilbab, A., and Hammouch, A. (2016a). Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis. IEEE Trans. Neural Syst. Rehab.
4. Benba, A., Jilbab, A., Hammouch, A., and Sandabad, S. (2016b). "Using RASTA-PLP for discriminating between different neurological diseases," in 2016 International Conference on Electrical and Information Technologies (ICEIT) (Tangiers), 406–409.
5. Choi, H., Ha, S., Im, H. J., Paek, S. H., and Lee, D. S. (2017). Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. NeuroImage.
6. Chou, K. L. (2017). Diagnosis and Differential Diagnosis of Parkinson Disease. Waltham, MA: UpToDate.
7. Haller, S., Badoud, S., Nguyen, D., Barnaure, I., Montandon, M. L., Lovblad, K. O., et al. (2013). Differentiation between Parkinson disease and other forms of Parkinsonism using support vector machine analysis of susceptibility-weighted imaging (SWI): initial results. Eur. Radiol. 23, 12–19.

ID: 21

SOCIAL MEDIA DATA MINING – NEED, TECHNIQUES AND CHALLENGES

Dr.T.SARATHAMANI
Associate Professor,
Department of Computational Sciences,
Brainware University, Kolkata, West Bengal.

Dr.K.KAVITHA
Associate Professor
Department of Computational Sciences
Brainware University, Kolkata, West Bengal.

ABSTRACT

Given the sheer volume and complexity of social media data, organizations and researchers employ advanced data analytics techniques, machine learning algorithms, and natural language processing tools to make sense of this wealth of information and extract valuable insights. The analysis of social media data has become an integral part of decision-making processes for businesses, governments, and various other sectors in today's data-driven world. The purpose of this chapter is to provide an insight of how data mining methods are applied to analyse and derive valuable insights from the vast and complex datasets generated on social media platforms and to showcase the significance of data mining in understanding social media trends, user behaviour, and sentiments, while also discussing the challenges associated with mining social media data.

KEYWORDS: *Social Media, Data Analysis, Data Mining, Big Data, Social Media Mining*

I. INTRODUCTION

1.1 Social Media

Social media refers to a collection of online platforms and technologies that enable users to create, share, and exchange information, ideas, and content in virtual communities and networks. It allows individuals and groups to interact, connect, and communicate with each other through various digital mediums, such as text, images, videos, and audio. The importance of social media in today's world cannot be overstated, as it has transformed the way people communicate, access information, and engage with the world around them. Here are some key aspects of its significance:

Communication and Connectivity: Social media has revolutionized communication by breaking down geographical barriers. It enables people from different parts of the world to connect instantly, fostering friendships, relationships, and professional networks. People can stay in touch with family and friends, collaborate with colleagues, and even form global communities based on shared interests.

Information Sharing: Social media platforms serve as a massive repository of information. News, updates, opinions, and educational content are shared rapidly, allowing users to be well-informed about current events, trends, and developments in real-time.

Marketing and Business: Businesses, both large and small, utilize social media as a powerful marketing tool. These platforms provide a cost-effective way to reach a broader audience, target specific demographics, and engage directly with customers. Social media marketing has become an integral part of modern advertising strategies.

Social and Political Movements: Social media has played a significant role in empowering social and political movements. It has facilitated the organization of protests, awareness campaigns, and grassroots activism, giving a voice to marginalized groups and facilitating social change.

Entertainment and Content Creation: Social media offers a wide array of entertainment options. Users can consume and share entertaining content like videos, memes, and creative artwork. Additionally, it has given rise to a new generation of content creators, allowing individuals to showcase their talents and build dedicated audiences.

Education and Learning: Social media platforms have become valuable educational resources. They provide access to educational content, tutorials, and online courses, enabling continuous learning and skill development.

Cultural Exchange and Diversity: Social media connects people from diverse cultures, promoting cross-cultural understanding and appreciation. It allows for the exchange of ideas, customs, and traditions, fostering a global community that celebrates diversity.

Real-Time Communication during Emergencies: During disasters or emergencies, social media can serve as a vital communication channel to disseminate critical information, coordinate relief efforts, and connect affected individuals with resources and aid.

While social media brings many benefits, it also comes with challenges, such as privacy concerns, online harassment, misinformation spread, and the potential for addiction. As its influence continues to grow, society must strive to strike a balance between leveraging its advantages and addressing its drawbacks to ensure a positive and responsible digital environment.

1.2 Social Media Data and Its Characteristics

Social media data refers to the vast amount of information generated and collected on social media platforms through user interactions, posts, and engagements. This data is a valuable resource for various purposes, such as market research, sentiment analysis, personalized advertising, and understanding user behaviour. The characteristics of social media data include:

1. **Volume**: Social media platforms generate an enormous volume of data every second. Millions of users post text, images, videos, and other content, resulting in a continuous stream of data that requires powerful data storage and processing capabilities.
2. **Velocity**: Social media data is produced at an incredible speed. The real-time nature of social media interactions means that new data is constantly being generated and shared, making it essential to capture and analyze information quickly.
3. **Variety**: Social media data comes in various formats, including text, images, videos, hashtags, likes, shares, comments, and more. This diversity of data types makes it challenging to process and analyse, requiring specialized tools and techniques.
4. **Veracity**: Social media data can be characterized by a lack of quality control and potential for misinformation. Not all content shared on social media platforms is accurate or reliable, and distinguishing between credible and false information can be difficult.
5. **Volatility**: Social media data has a relatively short lifespan. Trends and topics can become popular and fade away quickly, making it necessary to analyse data promptly to derive meaningful insights.
6. **User-Generated**: Social media data is primarily generated by users themselves. People willingly share their thoughts, opinions, and experiences on these platforms, providing valuable insights into their preferences and behaviours.
7. **Privacy Concerns**: Social media data raises significant privacy concerns, as it involves personal information shared by users. Ensuring the responsible handling and protection of user data is crucial to maintaining trust and ethical practices.

1.3 Data Mining

Data mining is a process of discovering patterns, relationships, and insights from large volumes of data. It involves using various techniques, algorithms, and tools to analyse data from different sources and extract valuable information for decision-making and knowledge discovery. In the context of social media analysis, data mining plays a crucial role in extracting meaningful insights from the vast amounts of user-generated content, interactions, and behaviours on social media platforms.

Analysing social media using data mining techniques is essential for several reasons:

Understanding Customer Sentiment: Social media platforms are a goldmine of user-generated content, including opinions, reviews, and feedback. Data mining enables businesses to analyse this vast amount of data to understand customer sentiment towards their products, services, and brand. Identifying positive and negative sentiments helps businesses gauge customer satisfaction and sentiment trends over time.

Market Research and Competitive Analysis: Social media data provides valuable insights into consumer preferences, behaviour, and trends. Data mining allows businesses to perform market research, track competitors' performance, and identify emerging market opportunities.

Enhancing Customer Engagement: Analysing social media interactions, businesses can gain insights into what content resonates best with their audience. Data mining helps identify patterns and preferences, enabling businesses to tailor their content and engagement strategies to increase user engagement and build a stronger online community.

Influencer Marketing: Identifying influential users and social media influencers is crucial for influencer marketing campaigns. Data mining techniques can help businesses find the right influencers whose audience aligns with their target market, leading to more effective marketing efforts.

Crisis Management and Brand Reputation: Social media can be a powerful tool for disseminating information, both positive and negative. Data mining aids in early detection of potential crises, allowing businesses to respond promptly and mitigate negative impacts on their brand reputation.

Personalized Recommendations: Data mining enables social media platforms to provide personalized content and recommendations to users. By analysing users' interactions, interests, and behaviour, platforms can suggest relevant content, products, and connections, enhancing user experience.

Identifying Trends and Viral Content: Social media data mining helps identifying emerging trends and viral content, enabling businesses to capitalize on opportunities or participate in trending conversations.

Fake News Detection and Misinformation: Data mining can be employed to detect and combat the spread of fake news and misinformation on social media platforms, promoting a more informed online environment.

Social Network Analysis: Data mining techniques can be used to analyse social networks, identifying key influencers, understanding user relationships, and studying information diffusion patterns.

Measuring Campaign Effectiveness: Businesses can use data mining to measure the effectiveness of their social media marketing campaigns, track performance metrics, and optimize their strategies accordingly.

II. DATA MINING TECHNIQUES IN ANALYSING SOCIAL MEDIA DATA

Data mining techniques play a crucial role in analysing social media data due to its sheer volume and complexity. Here are some commonly used data mining techniques for social media analysis these data mining techniques enable businesses, researchers, and organizations to gain valuable insights from social media data, helping them make informed decisions, understand user behaviour, and tailor their strategies to better engage with their target audience. Here are some commonly used data mining techniques for social media analysis:

Text Mining and Natural Language Processing (NLP): Text mining and NLP techniques are used to process and analyse the textual content of social media posts, comments, and messages. These techniques involve tasks such as sentiment analysis, named entity recognition, topic modelling, and keyword extraction. NLP algorithms help in understanding the meaning, sentiment, and themes present in the text data.

Social Network Analysis (SNA): Social Network Analysis (SNA) is a powerful methodology that examines the relationships and interactions among individuals, groups, or entities within a social network. It provides a structured approach to understanding the dynamics of social connections and the patterns that emerge from these relationships. At the core of SNA lies the concept of nodes (representing entities) and edges (representing connections). By visualizing the network and analysing its properties, SNA offers valuable insights into information flow, influence, and network centrality. This analysis can help identify influential nodes (individuals or entities with a high degree of connectivity), detect key communities, and understand how information or influence propagates through the network. Social Network Analysis finds applications in diverse fields, including sociology, anthropology, marketing, and cyber security. In sociology, it helps study social structures and their impact on individual behaviours. In marketing, SNA is utilized to identify potential brand ambassadors and target influential users for word-of-mouth marketing. Additionally, in cyber security, it aids in detecting anomalies and understanding the spread of threats within a network. With the increasing prominence of online social platforms, SNA continues to play a vital role in unveiling the intricate web of relationships that shape our digital and real-world interactions.

Clustering and Segmentation: Clustering and segmentation are indispensable techniques in analysing social media data due to its massive volume and diversity. Clustering groups similar users or content together based on shared characteristics, interests, or behaviours. This helps in identifying distinct user segments, such as frequent shoppers, tech enthusiasts, or sports fans, enabling businesses to personalize marketing strategies and content recommendations. Segmentation allows marketers to target specific groups more effectively and tailor their messaging accordingly. Additionally, clustering helps uncover patterns and trends within social media data, revealing popular topics, hashtags, or trends that can guide content creation and marketing campaigns. By employing clustering and segmentation, analysts can make sense of the vast and complex social media landscape, gaining deeper insights into user behaviour, preferences, and sentiment, and ultimately, making more informed decisions to engage with their audience successfully.

Association Rule Mining: Association rule mining identifies patterns of co-occurring items, such as frequently mentioned topics or hashtags in social media posts. These patterns can be used to understand user interests, recommend related content, and even uncover hidden relationships between different topics.

Geospatial Analysis: Geospatial analysis is a valuable tool in analysing social media data as it adds a geographical dimension to the insights gained. By mapping social media posts, comments, and interactions to specific locations, analysts can identify regional trends, popular hotspots, and location-based user behaviour. This information is essential for businesses to target localized marketing campaigns, understand regional preferences, and enhance location-based services. Geospatial analysis also aids in tracking the spread of information during events, disasters, or campaigns, helping authorities and organizations respond effectively. By integrating geospatial data with social media analysis, businesses and researchers gain a comprehensive understanding of how location influences user behavior and sentiment, enabling them to make data-driven decisions in a spatial context.

Time Series Analysis: Time series analysis is a crucial method for analysing social media data as it provides insights into temporal patterns and trends. By studying the data over time, analysts can identify recurring patterns, seasonal trends, and spikes in social media activity related to specific events or topics. Time series analysis helps understand the dynamics of user engagement, track the impact of marketing campaigns, and monitor the popularity of products or services over different periods. It also aids in predicting future trends and anticipating user behaviour based on historical data. Leveraging time series analysis empowers businesses and researchers to make informed decisions and develop strategies that align with the ever-changing landscape of social media.

Anomaly Detection: Anomaly detection is a vital technique in analysing social media data as it helps identify unusual or abnormal patterns that deviate from the norm. Social media platforms generate an enormous amount of data, making it challenging to manually spot anomalies. Anomaly detection algorithms can automatically identify irregular spikes in user activity, sudden changes in sentiment, or unexpected trends, which may indicate viral content, spam, or potential security threats. By promptly detecting anomalies, businesses can respond proactively to emerging issues, protect users from harmful content, and ensure the integrity of their social media platforms. Anomaly detection in social media data plays a crucial role in maintaining a safe and engaging environment for users, while also enabling organizations to make data-driven decisions based on accurate and reliable insights.

Image and Video Analysis: Image and video analysis are essential techniques in analysing social media data, where visual content plays a significant role. With the vast amount of images and videos shared daily on social platforms, these methods provide valuable insights. Image analysis involves tasks such as object detection, image classification, and facial recognition, enabling businesses to identify brand logos, products, or gauge user sentiment from emojis. Video analysis, on the other hand, allows tracking of brand mentions, sentiment expressed through gestures, and even understanding the context behind viral videos. Leveraging image and video analysis empowers marketers to measure the impact of visual content, detect trends, and tailor campaigns based on user engagement with compelling visuals. These techniques offer a comprehensive approach to understanding user behaviour and preferences, making social media data analysis more effective and informative.

Sentiment and Emotion Analysis: Sentiment and emotion analysis are powerful tools for analysing social media data, as they provide valuable insights into user reactions and feelings towards various topics, products, or events. Sentiment analysis involves determining the overall sentiment (positive, negative, or neutral) expressed in social media posts and comments, helping businesses understand public opinion and gauge customer satisfaction. Emotion analysis goes a step further by identifying specific emotions conveyed through text and emojis, such as happiness, anger, or sadness. This deeper understanding of user sentiment and emotions aids in tailoring marketing strategies, improving customer service, and responding to issues promptly. By leveraging sentiment and emotion analysis, businesses can make data-driven decisions, engage with their audience more effectively, and build stronger connections with their customers on social media platforms.

Influencer Identification: Influencer identification is a crucial aspect of analysing social media data as it helps businesses and marketers identify key individuals with a significant impact on their audience. By leveraging data mining and social network analysis techniques, businesses can pinpoint influential users based on factors such as follower count, engagement rate, and reach. These influencers can serve as valuable brand advocates, driving brand awareness, and promoting products to their dedicated followers. Analysing social media data for influencer identification allows businesses to form authentic partnerships with influencers who align with their brand values and target audience. This approach enables more effective influencer marketing campaigns, ultimately leading to increased brand credibility, customer loyalty, and expanded reach across social media platforms.

III. CHALLENGES ASSOCIATED WITH SOCIAL MEDIA DATA MINING

Social media data mining comes with several significant challenges that researchers and analysts must address to extract meaningful insights effectively. Some of the key challenges include:

Volume and Velocity: Social media platforms generate an enormous volume of data at a rapid pace. Handling and processing such vast amounts of data require robust computational resources and real-time analytics capabilities.

Data Noise and Quality: Social media data is often noisy and unstructured, containing spelling errors, slang, abbreviations, and irrelevant content. Ensuring data quality and accuracy is crucial for reliable analysis.

Privacy and Ethical Concerns: Social media data may contain sensitive and personal information, raising ethical concerns about data privacy and consent. Respecting user's privacy while conducting data mining is essential.

User Engagement Bias: Social media users may not represent the entire population, leading to bias in the data. The analysis must account for this bias to draw unbiased conclusions.

Misinformation and Fake News: Social media is susceptible to the spread of misinformation and fake news. Data mining techniques should address the challenge of detecting and mitigating the impact of false information.

Contextual Understanding: Understanding the context and sarcasm in social media posts can be challenging for data mining algorithms. Proper context analysis is essential for accurate insights.

Dynamic Nature of Social Media: Social media trends, topics, and user behavior constantly evolve. Data mining models should be adaptable to these changes to provide up-to-date insights.

Multilingual Data: Social media data can be multilingual, making language processing and analysis more complex.

Identifying Influencers: Identifying genuine influencers amidst a sea of bots and fake accounts is challenging. Data mining techniques should accurately differentiate authentic influencers from fraudulent ones.

IV. CONCLUSION

By extracting valuable insights from social media platforms, businesses can make informed decisions, improve customer engagement, strengthen their brand reputation, and gain a competitive edge in the market. Addressing challenges requires a combination of advanced data mining techniques, ethical considerations, and continuous adaptation to the dynamic nature of social media platforms. By overcoming these obstacles, researchers and businesses can harness the full potential of social media data for valuable insights and informed decision-making.

V. REFERENCES

1. Rahman MS, Reza H. A Systematic Review Towards Big Data Analytics in Social Media. *Big Data Mining and Analytics*, 2022, 5(3): 228-244.
2. Mustafa Abdalrassual Jassim and Sarah N. Abdulwahid 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1090** 01205
3. Qixuan Hou, Meng Han, Zhipeng Cai. Survey on Data Analysis in Social Media: A Practical Application Aspect. Big Data Mining and Anyalytics 2020, 3(4): 259-279.
4. N. A. Ghani, S. Hamid, I. A. Targio Hashem and E. Ahmed, "Social media big data analytics: A survey", *Comput. Human Behav.*, vol. 101, pp. 417-428, 2019.
5. L. Cao, "Data science: Challenges and directions", *Communication of the ACM*, vol. 60, no. 8, pp. 59-68, 2017.
6. Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining: An Introduction. Cambridge University Press**.**
7. Danneman, N., & Heimann, R. (2014). Social Media Mining with R: Deploy Cutting-edge Sentiment Analysis Techniques to Real-world Social Media Data using R. Packt Publishing.

ID: 22

REMOTE SURVEILLANCE VIDEO ANALYTICS MODELS TO IMPROVE OBJECT ACCURACY

Dr.S.SELVI
Associate Professor,
Department of Computer Science,
PSG College of Arts& Science,
Coimbatore, Tamil Nadu, India,

Dr.D.MOHANAPRIYA
Assistant Professor,
Department of Computer Science,
PSG College of Arts & Science,
Coimbatore, Tamil Nadu, India,

I. INTRODUCTION

Video surveillance is an essential tool for ensuring public safety and security in various environments, such as airports, train stations, shopping centers, and residential areas. This chapter dealt with various existing video analytics methods, its significant limitations and challenges, such as low accuracy, high computational complexity, and limited adaptability to changing environments.

Remote Surveillance Video Analytics models

Video surveillance systems have become an essential tool for ensuring public safety and security in various settings, including airports, public transportation, and critical infrastructure facilities (Khan & Han, 2018; Li et al., 2020). With the advent of digital video cameras and high-speed networks, remote surveillance has become an increasingly popular approach to monitor and secure large areas without the need for on-site personnel (Wang et al., 2018).

However, the effectiveness of remote surveillance systems depends largely on the accuracy and efficiency of their video analytics algorithms. Object detection, tracking, and behavior recognition are crucial tasks in remote surveillance, as they allow security personnel to identify potential threats and respond quickly and appropriately(Kumar et al., 2020; Wang et al., 2021).

Current video analytics algorithms and machine learning techniques have made significant advances in recent years, but there is still room for improvement. Some ofthe challenges that need to be addressed include occlusion, low lighting conditions, complex scenes with multiple objects, and the need for real-time processing (Gao etal., 2019; Li et al., 2021; Zhang et al., 2020).

Therefore, this chapter aims to explore availed ideas on advanced video analytics algorithms and machine learning techniques to enhance the accuracy and efficiency of object detection, tracking, and behavior recognition in remote surveillance systems.

One of the main challenges is the accurate recognition of specific behaviors and the timely detection of abnormal activities. Existing behavior recognition and anomaly detection algorithms face challenges in accurately identifying complex behaviors and distinguishing abnormal activities from normal ones. These challenges can lead to false alarms, missed events, and reduced overall performance of remote surveillancesystems.

Another challenge is the need for real-time processing, as remote surveillance systems often operate in dynamic and fast-changing environments. Delayed or inaccurate recognition of behaviors and anomalies can result in missed opportunities to prevent or mitigate security incidents.

Furthermore, with the increasing availability of high-resolution cameras, remote surveillance systems are generating vast amounts of data that need to be processed and analyzed. This presents a significant challenge for traditional analytics approaches, which may not be able to handle the large volume and complexity of thedata. Therefore, there is a need to explore advanced machine learning and computer vision techniques that can handle large-scale data processing and analysis, enabling more effective remote surveillance capabilities.

II. HISTORICAL OVERVIEW OF THE THEORY AND RESEARCH LITERATURE

Remote surveillance systems have been in use for several decades, with closed-circuit television (CCTV) cameras being the most common form of surveillance technology. However, the effectiveness of these systems has been limited by several factors, including the inability to accurately recognize specific behaviors and detect abnormalactivities.

The development of video analytics algorithms and machine learning techniques in recent years has led to significant advances in behavior recognition and anomaly detection in remote surveillance. These algorithms and techniques can process andanalyze large volumes of data, allowing security personnel to identify potentialthreats and respond quickly and appropriately.

Early research in the field of video analytics focused on object detection and tracking, which are essential tasks in behavior recognition and anomaly detection. One of the earliest approaches was the Viola-Jones algorithm, which used Haar-like features and an AdaBoost classifier to detect faces in images (Viola & Jones, 2001). This algorithm was later extended to detect other objects, such as cars and pedestrians (Bahlmann,Haasdonk, & Burkhardt, 2002). However, these approaches were limited by their inability to handle complex and dynamic environments and their reliance on handcrafted features.

In recent years, there has been a significant shift towards using deep learning approaches for video analytics in remote surveillance. Deep learning algorithms, such as convolutional neural networks (CNNs), have shown remarkable performance in object detection, tracking, and behavior recognition (Redmon & Farhadi, 2018). Transfer learning, which involves fine-tuning pre-trained CNN models on new datasets, has also been shown to be effective in reducing the amount of labeled training data required for behavior recognition and anomaly detection (Zhou, Huang,& Wang, 2019).

Despite the significant advances made in video analytics algorithms and machine learning techniques, there are still many challenges that need to be addressed to improve the overall performance of remote surveillance systems. These challenges include the accurate recognition of complex behaviors, the ability to distinguish abnormal activities from normal ones, and the need for real-time processing in dynamic and fast-changing environments.

The theory and research literature specific to the topic

Video analytics algorithms and machine learning techniques have been the focus ofsignificant research in recent years for improving behavior recognition and anomaly detection in remote surveillance. This section will discuss the key theories and research literature on video analytics and machine learning for remote surveillance.

III. OBJECT DETECTION AND TRACKING

Object detection and tracking are fundamental tasks in video analytics for remote surveillance. In the early years, object detection was achieved using handcrafted features and classifiers, such as the Viola-Jones algorithm (Viola & Jones, 2001). However, these methods were limited by their inability to handle complex and dynamic environments. In recent years, deep learning-based object detection approaches, such as You Only Look Once (YOLO) (Redmon & Farhadi, 2018), have shown significant improvements in detection accuracy and speed.

Behavior Recognition

Behavior recognition involves identifying and categorizing different actions and activities in video streams. There are several approaches to behavior recognition, including rule-based systems, feature-based methods, and deep learning-based methods. Rule-based systems use predefined rules to identify specific behaviors, such as loitering or running. Feature-based methods use handcrafted features to represent the behavior, such as motion features or shape features. Deep learning-based methods, such as convolutional neural networks (CNNs), have shown remarkable performance in behavior recognition (Wu et al., 2019).

Anomaly Detection

Anomaly detection involves identifying abnormal activities that deviate from normal behavior patterns. Anomaly detection is a challenging task because it requires modeling normal behavior patterns accurately. There are several approaches to anomaly detection, including statistical methods, rule-based methods, and machine learning-based methods. Machine learning-based methods, such as unsupervised deep learning models, have shown promising results in anomaly detection (Sabokrouet al., 2018).

Real-time Processing

Real-time processing is a crucial requirement for remote surveillance systems, as it enables security personnel to respond quickly to potential threats. Real-time processing involves analyzing video streams in real-time and providing alerts when abnormal activities are detected. There are several challenges in achieving real-time processing, including the need for high computational power and the ability to handle large volumes of data. To address these challenges, researchers have proposed several approaches, such as parallel processing and hardware acceleration (Zhao et al.,2018).

Overall, the literature shows that video analytics algorithms and machine learning techniques have the potential to significantly improve behavior recognition and anomaly detection in remote surveillance. However, there are still several challenges that need to be addressed, such as handling complex and dynamic environments, accurate recognition of complex behaviors, and real-time processing. The proposedstudy aims to address these challenges by exploring advanced algorithms and techniques that can enhance behavior recognition and anomaly detection in remote surveillance.

Research in cognate areas relevant to the topic

Video analytics for remote surveillance is an interdisciplinary field that draws on theories and research from several cognate areas, including computer vision, machine learning, and signal processing. This section will discuss the key theories and research literature in these areas that are relevant to enhancing video analytics for remotesurveillance.

Computer Vision

Computer vision is a field of study that focuses on enabling machines to interpret andunderstand visual data from the world around us. In recent years, significant progress has been made towards developing video analytics algorithms that use computer vision techniques to analyze video data in real-time and recognize objects, behaviors, and activities. These algorithms are crucial for remote surveillance, as they enable the detection and tracking of objects and the recognition of behaviors and activities, improving the effectiveness and efficiency of surveillance systems.

To achieve this goal, several computer vision algorithms have been developed, including YOLO (You Only Look Once), which is a real-time object detection algorithm that predicts object locations and categories in a single pass (Redmon et al., 2015). Faster R-CNN is a popular region-based object detection algorithm that uses proposals to identify object locations and categories in a video frame (Ren et al., 2015). Mask R-CNN, a variant of Faster R-CNN, can predict object masks in addition to object locations and categories (He et al., 2017). SSD (Single Shot Detector) is a one-stage object detection algorithm that uses anchor boxes to predict object locations and categories (Liu et al., 2016).

Object tracking is another crucial task in video analytics, and several algorithms have been developed for this purpose. DeepSORT (Deep Learning for Object Tracking and Re-identification) is a deep learning-based tracker that uses a combination of convolutional neural networks (CNNs) and Kalman filtering to track objects in a video sequence (Wojke et al., 2018). Siamese Trackers are another type of deep learning-based tracker that use a Siamese network to learn object appearance and motion models (Bertinetto et al., 2016).

Overall, computer vision techniques have provided the foundation for many video analytics algorithms used in remote surveillance and have led to significant progress in the development of object detection, tracking, and recognition algorithms. These algorithms have the potential to enhance public safety and improve the effectiveness of remote surveillance systems

IV. COMPUTER VISION ALGORITHMS

YOLO (You Only Look Once)

YOLO is a real-time object detection algorithm that directly predicts object locations and categories in a single pass. The algorithm divides the input image into a grid of cells and predicts objectness scores and bounding boxes for each cell. The predictions are made using a single neural network that simultaneously predicts objectness scores and bounding boxes for each cell. YOLO is fast and accurate, making it a popular algorithm for real-time object detection in video analytics.

Mask R-CNN

Mask R-CNN is a variant of Faster R-CNN that can also predict object masks in addition to object locations and categories. The algorithm uses a third neural network to predict object masks from the bounding box proposals generated by the RPN. The mask network is a fully convolutional network that takes the object proposal as input and outputs a binary mask that indicates the objectpixels. Mask R-CNN is accurate and flexible, making it a popular algorithm for instance segmentation in video analytics.

SSD (Single Shot Detector)

SSD is a one-stage object detection algorithm that uses anchor boxes to predict object locations and categories. The algorithm predicts object ness scores and bounding boxes for each anchor box using a single neural network.The anchor boxes are fixed shapes and sizes that are designed to cover a range of object scales and aspect ratios. The algorithm is fast and accurate, making it popular algorithm for object detection in video analytics.

V. DEEPSORT (DEEP LEARNING FOR OBJECT TRACKING AND RE-IDENTIfiCATION)

DeepSORT is a deep learning-based tracker that uses a combination of CNNs and Kalman filtering to track objects in a video sequence. The algorithm uses aCNN to extract features from the object appearance and a Kalman filter tomodel the object motion. The CNN features and Kalman filter predictions arecombined to update the object position and size in the next frame. DeepSORT is robust to occlusion and changes in object appearance, making it a popularalgorithm for object tracking in video analytics.

Siamese Trackers

Siamese Trackers are a type of deep learning-based tracker that use a Siamese network to learn object appearance and motion models. The algorithm consists of two identical CNNs that share weights and learn a similarity metric betweentwo input images. The output of the network is a similarity score that indicates the similarity between the current object patch and the target object patch.

The tracker uses the similarity score to update the object position and size in the next frame. Siamese Trackers are robust to occlusion and changes in object appearance, making them a popular algorithm for object tracking in video analytics.

Machine Learning

Machine learning is a subfield of artificial intelligence that focuses on developingalgorithms that can learn from data without being explicitly programmed. In recentyears, significant progress has been made in the development of machine learning algorithms that can be used in video analytics for remote surveillance.

These algorithms enable the development of models that can recognize patterns and anomalies in large volumes of data, improving the effectiveness of surveillance systems.

To achieve this goal, several machine learning algorithms have been developed, including LSTM (Long Short-Term Memory) Networks, which are a type of recurrent neural network that can model sequential data and are used for tasks such as action recognition (Hochreiter & Schmidhuber, 1997). GANs (Generative Adversarial Networks) are another type of deep learning algorithm used for image and video generation, and can be used for tasks such as generating synthetic data for training (Goodfellow et al., 2014). Autoencoders are another type of deep learning algorithm that can be used for tasks such as feature extraction and anomaly detection (Hinton & Salakhutdinov, 2006). HMMs (Hidden Markov Models) are a probabilistic graphical model used for modeling temporal data and can be used for tasks such as activity recognition (Rabiner, 1989).

Overall, the research in machine learning has led to significant progress in the development of deep learning, reinforcement learning, and transfer learning algorithms, which are all essential for enhancing behavior recognition and anomaly detection in remote surveillance. These algorithms have the potential to improve the effectiveness and efficiency of surveillance systems and enhance public safety.

VI. MACHINE LEARNING ALGORITHMS APPLIED IN VIDEO PROCESSING

LSTM (Long Short-Term Memory) Networks

LSTM Networks are a type of RNN that can learn long-term dependencies in sequential data, making them well-suited for modeling temporal patterns in video data. The algorithm consists of a memory cell and three gates: input, forget, and output gates. The input gate controls which information is stored inthe memory cell, the forget gate controls which information is discarded from the memory cell, and the output gate controls which information is outputted from the memory cell. LSTM Networks have achieved state-of-the-art performance on several benchmark datasets for action recognition and video prediction.

GANs (Generative Adversarial Networks)

GANs are a type of deep learning architecture that can generate synthetic video data that is similar to real-world data, which can be used to augment training datasets. The algorithm consists of two neural networks: a generator that generates synthetic data and a discriminator that distinguishes between real and synthetic data. The generator is trained to generate synthetic data that is similar to real data, while the discriminator is trained to distinguish between real and synthetic data. GANs have been used in various video analytics tasks, such as video synthesis, video prediction, and video anomalydetection.

Autoencoders

Autoencoders are a type of neural network that can learn a low-dimensional representation of normal patterns and detect deviations from it, making them useful for anomaly detection tasks. The algorithm consists of an encoder that maps the input data to a lower-dimensional latent space and a decoder that maps the latent space back to the input data. The encoder and decoder are trained to minimize the reconstruction error between the input data and the reconstructed data. Autoencoders can be used to learn normal patterns in video data and detect anomalies, such as unexpected events or behaviors.

VII. HMM's (HIDDEN MARKOV MODELS)

HMMs are a statistical model that can model temporal dependencies between actions, making them useful for behavior recognition tasks. The algorithm consists of a set of states and a set of observations. The transitions between states are modeled using transition probabilities, while the observations are modeled using emission probabilities. HMMs can be used to model temporal patterns in video data and recognize behaviors, such as walking, running, or loitering. HMMs have been used in various video analytics tasks, such as activityrecognition and behavior analysis.

Signal Processing

Signal processing is a field of study that focuses on the analysis, manipulation, and synthesis of signals, such as audio and video signals. Signal processing techniques are essential for video analytics in remote surveillance, as they enable the extraction of features and the reduction of noise and interference in video signals. The research in signal processing has led to significant advances in feature extraction, signal denoising, and compression, which are all essential for enhancing video analytics in remote surveillance (Gómez et al., 2017).Overall, the research in computer vision, machine learning, and signal processing has been instrumental in advancing the field of video analytics for remote surveillance.

VIII. REFERENCES

1. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
2. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition

(pp. 779-788).

3. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp.21-37). Springer, Cham.

ID: 23

COMPARATIVE STUDY OF ETHEREUM AND BITCOIN USING BLOCKCHAIN TECHNOLOGY

Mr.K.RAJESHKUMAR
Assistant Professor
Department of Computer Science
Arignar Anna Government Arts College,
Namakkal, Tamil Nadu, India.

Mr.R.RAMESH
Assistant Professor,
Department of Computer Applications,
Government Arts and Science College for Women,
Karimangalam, Dharmapuri, Tamil Nadu, India.

ABSTRACT:

Bitcoin was introduced in the year 2009, after the launch several new cryptocurrencies have been seen with variations to Bitcoin's original creation. Bitcoin still remains the most prominent actor in the digital world. The ideal of the exploration paper is to determine whether the newer cryptocurrencies solves the problem arises in the digital world, rather of assessing several cryptocurrencies for this comparison, the crypto currency Ethereum has been chosen as a deputy for the others. Ethereum was started in 2014, is extensively backed in the community and is alternate in line to Bitcoin when it comes to market capitalization. As a base for the relative analysis a rigorous study of the Bitcoin and Ethereum protocols have been performed, and resemblant descriptions of the systems have been considered. Three major problems have shaped the focus of the study such as computational waste, attention of power and nebulosity of deals.

KEYWORDS: *Cryptocurrency, Security, Blockchain, Bitcoin, Internet, Ethereum, Digital Signature, etc.*

I. INTRODUCTION

Traditionally, our financial systems depends on physical exchange of money and digital credit. In the world of online transactions the only way of exchanging value is by utilizing trusted third parties, such as banks or intermediate payment systems – for instance,

PayPal to relay the transaction. A weakness of these kinds of online systems is that you have to trust the mediating third party to act in your interest. Even fiat currencies, i.e. the physical money system, have their weaknesses – trust needs to be placed in the institution issuing such currencies, that they will not act in ways that may cause unreasonable levels of inflation or financial crises.

Cryptocurrencies gives an alternate solution to the various existing systems. By utilizing a Peer-to-Peer transaction system – where identities and ownership can be validated using cryptographic means – the users of the system do not have to rely on trust in third parties to exchange value online. Most cryptocurrencies mint their own coin to create value within the system. This is an essential part of their functionality that decouples the value in the cryptocurrency from any fiat currency that relies on governments or organizations for issuance. It also means that crypto-currencies can be exchanged globally - independently of what currency is native to a region.

When Bitcoin emerged in 2009 it was just a small group of early adopters that saw the potential in the technology and made use of it. Since then a growing ecosystem of different cryptocurrencies has emerged. It becomes increasingly interesting to notice the differences and challenges of the several systems as they grow in popularity and value.

1.1 Problem Description

The Bitcoin is extensively honored as the first successful attempt at a distributed cryptocurrency, with bitcoins being accepted as payment in a growing number of cases. Despite its popularity, central problems remain with the system design. These issues can be categorized into three major categories such as: wastefulness of computational resources, tendency to centralization over time and ambiguity of transaction finalization. After the launch of bitcoin many alternatives to it have been designed, with many trying to combat these issues. First among the alternatives, in terms of market value and popularity, is the Ethereum system. The developers claim that it provides a wider scope of functionality and higher levels of effectiveness compared to Bitcoin, all the while maintaining the same levels of security. This paper will put a light on a comparative analysis of Bitcoin and Ethereum by considering on the above mention parameters. Our goal is to determine if Ethereum suffers from the same problems, and attempts to offer some insight into the future of this technology in general. The study will use the published technical descriptions of both systems as well as statistics from the live blockchains of each where appropriate.

1.2 Research Objective

Over the past few years, Bitcoin has been thoroughly reviewed, and technical issues for the system have been identified:

Computational Waste

Large amounts of computation and energy is wasted in the validation process of Bitcoin. This is because nodes – i.e. Bitcoin system participants – are investing in expensive hardware to get an advantage when competing to receive rewards for validating the transactions of peers. Combined with the dynamic adjustment of the difficulty of the validation process to fit the expected time of ten minutes between each set of transactions, it creates a situation where unnecessarily large amounts of computation are expended.

Concentration of Power

The validation is concentrated in a few centrally organized groups, and not spread out in the distributed manner as it was intended. For the validation process to be profitable for nodes in the peer-to-peer network regarding costs related to the process, nodes join together and split rewards received from successful validation. In a group like this, the cost of specialized validation equipment and power expenses is split between the nodes in the group, and any reward for a successful validation is shared with the others.

Ambiguity

Because of the underlying architecture of Bitcoin, transactions take time to process. Knowing exactly when a transaction can be trusted or not is the matter of variance, causing the validity of the transactions to remain ambiguous for a period of time after they have been relayed to the network.

The main objective of this research work is to determine if the later cryptocurrencies have avoided the known problems of Bitcoin, by performing a comparative analysis of Bitcoin and Ethereum.

Newer cryptocurrencies have the advantage of the knowledge of the problems of Bitcoin before development; comparing these cryptocurrencies to Bitcoin therefore seems like the natural choice when reviewing how well these challenges are being met and handled by the newcomers. Rather than try to answer this by examining multiple instances of cryptocurrencies the main focus has been put on Ethereum.

There are many reasons for selecting Ethereum as a proxy for the other cryptocurrencies. Firstly, the currency is relatively recent, Ethereum was launched in the summer of 2015, the developers of the system will likely be aware of the current research in the field. Secondly, the system is explicitly aimed at fixing the shortcomings of Bitcoin, which means that we can assume they have had these problems in mind when designing the protocol. Lastly, although Bitcoin is still the biggest of the cryptocurrencies in terms of adoption and value, Ethereum comes next in line as the second most popular alternative and is widely backed in the community. To give a quantifiable measure of the two systems, the website coinmarketcap.com values the current Bitcoin market capitalization as being over 561 billion US dollars. After this, Ethereum comes next with more than 229 billion US dollars. However, it should be noted that these values are highly volatile and subject to change on a daily basis.

1.3 Methodology

As a methodology for this research paper consists of deriving parallel technical descriptions for Bitcoin and Ethereum to detail both systems in a scientifically rigorous manner to accurately compare and contrast them. Especially, while Bitcoin has recently been described well in the literature, information on Ethereum is dispersed over online platforms which is subject to edits and changes by the different authors sporadically – making the task of detailing Ethereum significantly.

Furthermore, the technical problems of Bitcoin have been elaborated in terms of the protocol detail, and Ethereum has been analyzed to evaluate whether it exhibits the same problems. The two descriptions are placed in contrast to each other, highlighting their similarities and differences. The comparative study will also consider data available from the live systems when this is appropriate. This data has been gathered and synthesized to enlighten the discussion and to determine if the findings from the comparison align with the data.

II. WHAT IS BITCOIN?

An entirely P-2-P interpretation of electronic cash would allow online deals to be transferred directly from one party to another without going through a fiscal association. Digital signs are used as a part of the overall result, but the key benefits are lost if a third party is still need to help double- spending.

The network instants deals by mincing them into an ongoing chain of hash-grounded evidence- of- work, forming a record that can not be changed without redoing the evidence- of- work. The longest chain not only serves as evidence of the sequence of events witnessed, but evidence that it came from the largest pool of CPU power. As long as a maturity of CPU power is controlled by bumps that aren't cooperating to attack the network, they'll induce the longest chain and outpace bushwhackers. The network itself requires minimum structure. Dispatches are broadcast on a stylish trouble base, and bumps can leave and return the network at will, accepting the longest evidence- of-work chain as evidence of what happed while they were gone.[6]

III. THE BLOCKCHAIN

Blockchain is a new technology that has surfaced with the appearance of the Bitcoin, which has added a new way of dealing financially. Grounded on the success of this fashion with the idea of Bitcoin, the fashion has been reckoned upon and applied gradationally in colorful conditioning, whether governmental or private and entered the confidence and satisfaction of guests. The paper highlights the challenges ahead and openings in this ultramodern technology that's all set to develop our digital world.[4]

Blockchain technology is one of the approaches that has the possibility to enhance decentralization, translucency, equivalency, and responsibility on the internet [1]. Blockchain is a distributed database of records that can be either public tally of digital issues or deals that got achieved and have been participated among sharing parties across a large network of untrusted actors. It stores data in blocks that can corroborate information which are veritably delicate to hack. It avoids the demand of a third- party verification and therefore deactivates any sector that leverages it traditionally. [2]. Using blockchain can give advanced security compared to storing all data in a central database. The use of these technologies in Bitcoin “mining” was ground-breaking in the data storehouse and operation side, detriment from attacks on a database can be averted. Further, since the blockchain has an openness trait, it can give translucency in data when applied to an area taking the exposure of data [3].

IV. What is Ethereum?

Ethereum is an open source project first introduced in 2013, initially described as a “Next-Generation Smart Contract and Decentralized Application Platform”. At first glance Ethereum is a peer-to-peer network and an exchangeable cryptocurrency that allows nodes to share computing resources for the execution of programmable smart contracts on the blockchain. There are however multiple different ways to describe Ethereum depending on ones point of view. In the official guides Ethereum is also described as a ‘World Computer’, in the sense that it can be seen as a single computing platform which anyone in the world is able to use. In this computer world any number of programs can be encoded and executed, and any participating code can interact and have access to the state of each one of these programs.

In other words, with Ethereum any user can have access to a cheap, zero-infrastructure, global platform that provides a very interesting set of features:

- User authentication will be validated by the use of cryptographic signatures.
- Easily deployable payment logic. A payment system can be setup on

Ethereum very quickly with no third party reliance.

- Total DDoS resistance. Each application on Ethereum is not executed on any single node; rather it is executed on each and every node on the system. As long as there is one node maintaining the blockchain the application will run perpetually and will be able to be interfaced by any joining node.
- Limitless interoperability. Each Ethereum contract can seamlessly interact with any other contract instance via the provided interfaces in the Ethereum ecosystem
- No server infrastructure. As mentioned before Ethereum is completely built on top of a Peer-to-Peer network with no central server infrastructure involved. Thus, the deployment of an application on the blockchain does not require the setup and the costs of setting and maintaining servers. Having said this, we can understand that Ethereum strives to provide a platform where anyone can easily deploy and run Internet services.[5]

V. CONCLUSION:

Comparative Analysis of Bitcoin and Ethereum

Comparison of Bitcoin and Ethereum is shown in the following Table.

Sr. No.	Parameters	Bitcoin	Ethereum
01	Maximum block size	1 MB	Flexible Limit
02	Target block time	10 min	13 sec
03	State	Stateless	Stateful
04	Consensus mechanism	Proof of work	Proof of work
05	Consensus Protocol	Nakamoto consensus	GHOST
06	Mining algorithm	Hashcash	Ethash

Table 1: Comparative Analysis of Bitcoin and Ethereum.

Among the cryptocurrencies available moment, Bitcoin is the more well- known, but Ethereum has been gaining ground in recent times. Both have their unique benefits and downsides. Bitcoin is the original bone and remains the most well-known. It's also the most precious, with a high market cap. Bitcoin is frequently used as a store of value, and it's seen as a more dependable investment than Ethereum.

Ethereum, on the other hand, is a platform that allows for the creation of decentralized operations. These operations can be deployed on top of the Ethereum blockchain, and they run on Ethereum's native currency, Ether. Ethereum also has a much lower transaction charges as compared to Bitcoin.

Both Bitcoin and Ethereum have their strengths and drawbacks. Bitcoin is more precious and is seen as a more dependable investment, but Ethereum has lower transaction charges and allows for the creation of decentralized operations.

VI. References

1. Walid A., Nicolas S.,2017,"Blockchain technology for social impact: opportunities and challenges ahead", available: https://doi.org/10.1080/23738871.2017.1400084
2. Arijit C., Ashesh K., 2017," Blockchain and its Scope in Retail"available: https://irjet.net/archives/V4/i7/IRJET-V4I7616.pdf
3. Ketki R., Sheetal Y., 2018," Blockchain Technology in Cloud Computing : A Systematic Review", available : https://www.irjet.net/archives/V5/i4/IRJET-V5I4428.pdf
4. https://www.irjet.net/archives/V5/i5/IRJET-V5I5659.pdf
5. https://delaat.net/rp/2015-2016/p53/report.pdf
6. https://bitcoin.org/bitcoin.pdf

ID: 24

UNVEILING NEW HORIZONS IN RESEARCH: BIG DATA AND ITS ANALYTICAL ODYSSEY

Mr.M.VENKATESH SARAVANAKUMAR
Assistant Professor,
SRM Trichy Arts & Science College,
[Affiliated to Bharathidasan University, Tiruchchirappalli,
Tiruchchirappalli, Tamil Nadu, India.

Dr.M.SABIBULLAH
Associate Professor,
Department of Computer Science,
PG & Research Dept. of Computer Science,
Jamal Mohamed College (Autonomous),
[Affiliated to Bharathidasan University, Tiruchchirappalli],
Tiruchchirappalli, Tamilnadu, India.

ABSTRACT:

Rapid technological advancements have ushered in the era of big data, characterized by the exponential growth of diverse data sources like the web, social networks, business records, and medical databases. The convergence of extensive data generation and cloud computing power has propelled the field of "Big Data Analytics," revolutionizing the way information is extracted from these vast datasets. However, as the private sector increasingly embraces big data analytics, it brings forth new challenges for privacy advocates. Big data encompasses voluminous, complex, and highly unstructured data from various origins. Extracting insights from this data treasure trove can significantly benefit organizations across diverse sectors, influencing decisions that shape the future of businesses and social platforms. The historical evolution of big data, encapsulated chronologically from the birth of relational databases to modern cloud-based solutions, reflects the paradigm shift in technology development and data access methodologies. The "Three Vs" of big data - Volume, Velocity, and Variety - describe its unique characteristics, emphasizing the need for scalable storage, real-time processing, and the ability to handle various data types. These new dimensions amplify the complexity of managing and analysing big data, fuelling further innovations in the field. Big data is supported by a rich ecosystem of platforms, technologies, and algorithms. Platforms like Hadoop and Spark provide the infrastructure for data storage and processing, while technologies encompass tools for storage, computation, analytics, and visualization. Machine learning and deep learning algorithms extract meaningful patterns from the data, enabling predictive and prescriptive analytics.

KEYWORDS: *Big Data, Big Data Analytics, Hadoop, Spark, Python, Machine Learning*

1. INTRODUCTION:

We now inhabit the era of big data, where the generation of vast economic and social value takes place. Progress in data collection, coupled with the substantial growth in cloud computing power, has made technologies capable of effectively analysing extensive sets of heterogeneous data increasingly prevalent. Despite the manifold advantages of big data, the emergence of "Big Data Analytics" in the private sector presents novel challenges for privacy advocates. Big data pertains to voluminous, intricate, and highly unstructured data stemming from myriad sources. Its expansion is occurring at an astounding rate, underscoring the imperative to adeptly handle and manage data to facilitate the analysis of requisite information, thereby conserving both time and cost. The gleaned insights can substantially benefit organizations across multiple dimensions. A multitude of decisions that profoundly influence diverse entities, including businesses and social networking platforms, hinge upon this extensive data repository, aiding in the anticipation of futures yet to unfold."

1.1 Data:

Data refers to raw facts, figures, and observations that are collected, recorded, or generated. Data can be in various forms, such as numbers, text, images, audio, and more. It is the foundation of any analytical process and serves as the input for analysis.

1.2 Data Analysis:

Data analysis involves the process of inspecting, cleaning, transforming, and modelling data to discover useful information, draw conclusions, and support decision-making. It aims to extract meaningful insights from data to reveal patterns, trends, correlations, and relationships.

1.3 Data Analytics:

Data analytics is a broader concept that encompasses the entire lifecycle of data, from collection to interpretation. It involves the application of various techniques and methods to explore data, uncover insights, and make predictions or recommendations. Data analytics often includes descriptive analytics (summarizing historical data), diagnostic analytics (identifying reasons for past outcomes), predictive analytics (forecasting future trends), and prescriptive analytics (suggesting actions based on analysis).

1.4 Big Data:

Big data refers to extremely large and complex datasets that are beyond the capacity of traditional data processing methods to efficiently handle, store, or analyze. These datasets typically include a wide variety of data types such as structured, semi-structured, and unstructured data. Big data is characterized by the three "Vs": volume (large amount of data), velocity (high speed at which data is generated and processed), and variety (different types of data sources).

1.5 Big Data Analysis:

Big data analysis involves the process of examining and interpreting large and intricate datasets to uncover patterns, trends, correlations, and other valuable insights. This analysis often requires advanced computational and statistical techniques, as well as specialized tools and platforms that can handle the volume and complexity of the data. The goal of big data analysis is to extract meaningful information from the data to inform decision-making and strategy.

1.5.1 Types of Data Analytics:

Analytics technology refers to the systematic computational analysis of transforming data into information; it is described as data-driven decision-making. The main goal of analytics technology is to capture data collected from different sources and analyse these data to reach an optimal decision. DA is a science to examine the raw data and drawing conclusions about the information, especially used in healthcare sector to cope-up, cost-effective management, instant and to find treatment effectiveness[1]. Machine Learning (ML), is a one area of computer science and ML is Data Mining (DM), defined to execute and model the huge datasets from which unknown patterns/relationships would be discovered. [2]. BD Analytics [3][4] helps to identify new opportunities by harnessing the data.

Varieties of analytics classifications exist, offering valuable outcomes for businesses, organizations, and healthcare sectors through analytics acquisitions [5][6]. These solutions are grouped into distinct categories (Table 1):

Table 1: Categories of Analytics

Descriptive Analytics: This category models past behaviours, encompassing model estimation, validation, and scoring. Diagnostics Analytics: It unveils the reasons behind occurrences, proving beneficial for tasks like evaluating social media campaigns through metrics like post count, page views, and reviews.
Predictive (Understanding the future): Forecasts based on available data (Attempts to predict the future). Tools used for this analytics are; Rapid Miner, WEKA and Waffles.
Prescriptive (Possible Outcomes): Assess actions, assist decision-making (Solutions are to assist analysts in making decisions). The term prescriptive analytics introduced by IBM and later trademarked by Ayata, Australia based Software Company.

1.6 Big Data Analytics:

Big data analytics is a broader concept that encompasses the entire process of collecting, processing, analysing, and interpreting large datasets to gain insights and support decision-making. It involves using various techniques, including statistical analysis, data mining, machine learning, and predictive modelling, to derive actionable insights from big data. Big data analytics often focuses on discovering hidden patterns, predicting future trends, and optimizing processes based on the analysis results.

II. LITERATURE REVIEW:

Here explored the [7] characteristics, and necessity of Big Data, as well as the array of options available in the market for exploring vast, unstructured datasets. The rapid surge in technological advancements has resulted in an inundation of data originating from diverse sources such as the web, social networks, business records, and medical archives, over the preceding years [8]. Contrasted with traditional data, Big Data manifests a distinctive attribute through its three V's, signifying its unstructured nature. Effectively managing and handling this data has become a critical challenge, essential for analysing requisite insights and conserving both time and cost. The harnessed insights can greatly benefit organizations across multifaceted dimensions [9].

Conventional data storage methods fall short in accommodating and analysing this immense data volume. Numerous researchers are dedicatedly exploring dimensionality reduction techniques for Big Data to enhance effective analytics reporting and data visualization [10]. This paper's focal point is the inception of extensive information and the imperative of scrutinizing such data. Additionally, it offers a concise glimpse into real-world Big Data analytics applications within various fields, coupled with the associated challenges and advantages. Furthermore, an examination of diverse tools for processing massive data volumes in various real-world domains is conducted [11].

Notably, Big Data exhibits unique traits that conventional methods struggle to handle and process, consequently posing a substantial industry challenge, this work provides a comprehensive outline of Big Data's characteristics and also delves into the prospective future scope, particularly in terms of the direction Big Data is headed [12]. Analytics companies are cultivating the ability to fortify their decisions through analytical reasoning, employing a range of statistical and mathematical techniques. Recent research has unveiled that over 59% of organizations lack the requisite information for informed decision-making.

Embracing "Data Analysis with R" not only supplements existing analytical knowledge and methodologies but also exposes individuals to contemporary analytics techniques encompassing forecasting, social media analytics, text mining, and more. And further facilitates an opportunity to engage with real-time data sourced from platforms like Twitter, Facebook, and other social networks [13].

III. EVOLUTION OF BIG DATA

With the advancement and transformative impact of computer science, a new paradigm has emerged, signifying a shift from traditional databases of the past to the modernized databases prevalent today.

This evolution encompasses not only the methods of accessing these databases but also the dramatic changes witnessed in technology development over time. These shifts are presented in a chronological tabulation (Table 2).

Table 2: History of Big data

S.No.	Year	Advancements
1	1970	The Birth of Relational Databases
2	1980	The Rise of Data Warehousing
3	1990	The Internet and Digital Revolution
4	1997	The term "data mining" becomes more widely recognized as a process of discovering patterns in large datasets
5	2000	Web usage data starts to accumulate rapidly due to the increasing popularity of the Internet.
6	2003	Google publishes the paper on the Google File System (GFS), introducing a scalable way to store and manage large datasets
7	2004	Doug Cutting and Mike Cafarella develop Hadoop, an open-source implementation of Google's GFS and MapReduce
7	2005	Hadoop Platform form for handling extremely large datasets were introduced
8	2006	Amazon Web Services (AWS) launches its Elastic Compute Cloud (EC2) and Simple Storage Service (S3), paving the way for cloud-based Big Data solutions.
9	2008	Hadoop becomes an Apache open-source project. Yahoo! starts using Hadoop for large-scale data processing.
10	2009	Hadoop, an open-source framework for distributed storage and processing of big data, gained

		prominence, enabling scalable data processing for machine learning tasks.
11	2010	Microsoft, Amazon and Google introduced low cost database and computers in the cloud. With the growth of social media, IoT devices, and other data sources, the need for advanced analytics on big data led to increased interest in combining machine learning with big data technologies. The term "Big Data" gains popularity as it's used to describe the challenges and opportunities of managing and analysing massive datasets.
12	2011	IBM's Watson supercomputer wins Jeopardy!, showcasing the potential of artificial intelligence and data analytics, Apache Storm was launched for a distributed computation framework for data stream
13	2012	Google introduced its machine learning framework, Tensor Flow, making it easier for researchers and developers to implement machine learning algorithms on large datasets.
14	2013	Apache Hadoop 2.0 is released, bringing improvements in scalability and resource management. Spark [14], an open-source cluster-computing framework, gains attention for its speed and ease of use.
15	2014	The emergence of the Internet of Things (IoT) leads to a surge in data generated from connected devices. Apache Spark emerged as a powerful tool for big data processing and analytics, providing machine learning libraries that enabled distributed and parallel processing.
16	2015	Microsoft, Amazon and Google incrementally release new technologies to adopt machine learning and AI to have deeper insights and better outcome prediction.
17	2016	Machine learning and AI techniques are increasingly applied to Big Data for predictive and prescriptive analytics .Deep learning, a subset of machine learning focused on neural networks, started gaining attention for its ability to analyse and derive insights from massive datasets.
18	2017	Rise of open source tools like R, Python to ease data analytics and better insights
19	2018	Cloud-based Big Data solutions become more prevalent, offering scalable and cost-effective ways to manage large datasets
20	2019	Edge computing gains momentum, allowing data processing and analysis to occur closer to the data source.
21	2020	The use of big data and machine learning became more mainstream across industries, with applications in healthcare, finance, marketing, and more.
22	2021	Big Data continues to evolve with advancements in AI, machine learning, and analytics tools, shaping industries and research

3.1. Characteristics of Big Data:

In the context of big data, the "Vs" refer to the three main characteristics that are often used to describe the unique properties of big data. These characteristics are known as the "Three Vs" of big data:

Volume, Velocity, and Variety. Over the time, additional characteristics are emerging and being recognized by domain experts, contributing to the expansion of the big data paradigm (Table 3).

Table 3: Big Data -V's

Volume: This refers to the sheer size of the data generated and collected from digital devices, sensors, social media, and more
Velocity: This refers to the speed at which data is generated, processed, and made available for analysis.
Variety: Big data comes in various formats and types, including structured, semi-structured, and unstructured data.
Veracity: This refers to the quality and reliability of data.
Validity: Ensuring that the data is accurate, valid, and trustworthy for analysis.
Vulnerability: Addressing data security and privacy concerns to protect sensitive information.
Volatility: Dealing with the dynamic nature of data, including how quickly data changes over time.
Visibility: Gaining insights into the flow, usage, and access of data across an organization.
Vexing: Navigating the challenges and complexities of working with Big Data.
Vitality: Recognizing the critical role of data in the success and competitiveness of businesses

Vagueness: Managing uncertainty and incomplete information present in unstructured data
Vigilance: Being alert to potential biases and ethical considerations when analyzing data.
Vitalization: Transforming raw data into valuable information and actionable insights
Variance: Dealing with the variability and fluctuations in data patterns.
Voracity: The scale and impact of data processing on storage, computing power, and resources.
Variance in Data Models: Addressing differences in data models and schemas across different data sources.

3.2 Big Data Platforms:

A big data platform typically encompasses a broader ecosystem of tools, technologies, and frameworks that collectively enable the storage, processing, management, and analysis of large and complex datasets. This includes not only the core technologies like Hadoop and Spark but also other components such as data storage systems, data integration tools, analytics libraries, visualization tools, and more. Big data platforms provide a comprehensive environment for handling various aspects of big data challenges.

3.3Big Data Technologies:

Big data technologies, on the other hand, specifically refer to the individual software tools and frameworks that address different aspects of big data processing. Hadoop and Spark are prominent examples of big data technologies. They are foundational components that play a crucial role in enabling distributed computing and data processing for large datasets.

In practical usage, the terms "big data platforms" and "big data technologies" are often used interchangeably, especially when discussing specific tools like Hadoop and Spark. However, it's important to recognize that these tools are part of a larger ecosystem that collectively constitutes a big data platform. Other components like data storage systems (HDFS), data querying (Hive, Impala), data streaming (Kafka), and machine learning libraries contribute to the overall big data platform.

3.4 Navigating Big Data Realms: Unmasking the Divide between Platforms and Technology

Frequently, there arises a perplexity surrounding the distinction between Big Data (BD) technology and Big Data platforms. In order to dispel such uncertainties, let's consider the following illustration.

Consider a healthcare establishment with ambitions to harness big data in enhancing patient care, streamlining operations, and advancing medical research. To attain these objectives, they establish an all-encompassing big data infrastructure. This highlights the clear distinction between technology and platform, akin to the disparity illustrated earlier.

3.4.1 Big Data Platform Components:

1. **Data Storage:** The platform includes Hadoop Distributed File System (HDFS) for storing vast amounts of patient records, medical images, sensor data from wearable devices, and research papers.
2. **Data Processing:** Apache Spark is integrated into the platform to process and analyse medical data. It enables real-time analysis of patient vital signs and historical data to detect anomalies or trends.
3. **Data Integration:** Data from different sources, such as electronic health records, laboratory results, and patient demographics, are integrated using tools like Apache Kafka, ensuring all relevant data is available for analysis.
4. **Analytics Libraries:** The platform includes machine learning libraries like Tensor Flow and Scikit-learn, enabling researchers to develop predictive models for disease diagnosis and treatment recommendations.
5. **Visualization:** Tableau and D3.js are used to create interactive visualizations and dashboards. Medical professionals can explore patient data, track treatment outcomes, and identify patterns.
6. **Cloud Integration:** The platform leverages cloud services to scale resources based on demand, ensuring efficient handling of data processing during peak times.

3.4.2 Big Data Technologies:

1. **Hadoop (Big Data Technology):** Hadoop, specifically HDFS, stores massive amounts of patient data securely and ensures fault tolerance. It also supports batch processing of data for long-term analytics.
2. **Spark (Big Data Technology):** Spark processes real-time data streams, enabling instant monitoring of patient vitals and generating alerts in case of critical changes.
3. **Kafka (Big Data Technology):** Kafka facilitates the integration of diverse data sources, allowing continuous data ingestion from various hospital departments.
4. **Tensor Flow (Big Data Technology):** Researchers use Tensor Flow to build deep learning models that analyse medical images and help identify early signs of diseases like cancer.
5. **Tableau (Big Data Technology):** Tableau provides interactive visualizations of patient data, making it easier for doctors to identify trends and patterns in treatment outcomes.

In this example, the healthcare organization's big data platform comprises various big data technologies, including Hadoop and Spark, along with other components that collectively create an environment for storing, processing, analysing, and visualizing healthcare data. The platform enables the organization to leverage the power of big data to enhance patient care and advance medical research.

IV. BIG DATA SUPPORTED STORAGE

Big Data storage solutions are crucial for efficiently managing and handling the vast volumes of data generated and processed in the Big Data ecosystem. Here are some key storage technologies (Table 4) and concepts related to Big Data [15].

Table 4: Big Data Supported Storage

S.No.	Storage System	Features
1	Distributed File Systems	Distributed file systems like Hadoop Distributed File System (HDFS) are designed to store large datasets across a cluster of commodity hardware. Data is distributed and replicated across nodes for fault tolerance and high availability. Allows parallel processing of data using Map Reduce or other processing frameworks.
2	Object Storage:	Object storage systems, such as Amazon S3, Google Cloud Storage, and Azure Blob Storage, provide scalable and cost-effective storage for unstructured data.Data is stored as objects with unique identifiers and can be accessed via APIs.
3	NoSQL Databases:	NoSQL databases, like MongoDB, Cassandra, and Couch base, are designed to handle large volumes of unstructured or semi-structured data. They offer high scalability, flexibility, and support for horizontal data distribution.
4	Columnar Databases:	Columnar databases like Apache Cassandra and Apache HBase organize data by columns rather than rows, enabling efficient storage and retrieval of specific attributes. Suited for analytical workloads and data warehousing.
5	In-Memory Databases:	In-memory databases, such as Apache Ignite, store data in memory for fast data access and low-latency processing. Ideal for real-time analytics and applications that require rapid data retrieval.
6	Graph databases	Like Neo4j and Amazon Neptune store data in nodes and edges, making them suitable for analysing complex relationships and networks. Used in applications like social networks, recommendation systems, and fraud detection
7	Cloud storage services	Like Amazon S3, Azure Blob Storage, and Google Cloud Storage, provide scalable and on-demand storage resources.

		Organizations can store and access Big Data in the cloud, reducing the need for on-premises infrastructure.

V. Big Data Supported Languages:

In the realm of Big Data analytics, several programming languages are used to process, analyse, and derive insights from large and complex datasets. These languages (Table 5) provide various libraries, frameworks, and tools that facilitate handling massive amounts of data efficiently. Here are some of the key programming languages used in the context of Big Data illustrated in Figure-1:

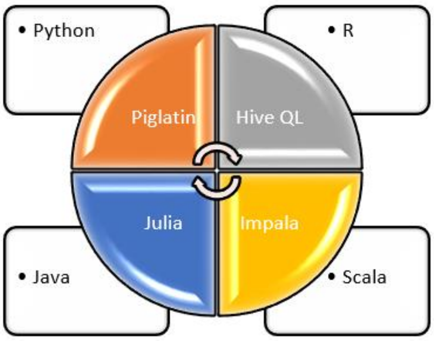


Figure 1: Big Data Supported language

Table 5: Big Data supported Language

Python: Python is widely used for data analysis and machine learning tasks. PySpark, a Python library, enables Big Data processing and analytics using the Apache Spark framework.	**R**: A popular language for statistical computing and graphics, provides a wide range of packages for data analysis, visualization, and modelling, includes dplyr and ggplot2 packages used for data manipulation and visualization.
Java: Java is used in Big Data frameworks like Apache Hadoop and Apache Flink for distributed processing and analytics.	**Scala**: Scala is a language that runs on the Java Virtual Machine (JVM) and is used in Apache Spark.
HiveQL: HiveQL is a query language used with Apache Hive, a data warehouse infrastructure built on top of Hadoop. It allows SQL-like querying of large datasets stored in Hadoop.	**Pig Latin**: Pig Latin is a scripting language used with Apache Pig, another data processing tool on Hadoop. It simplifies data processing tasks by providing a higher-level abstraction.
Julia: Julia is a programming language designed for high-performance scientific computing. It is gaining traction in data analysis and numerical computation due to its speed and versatility.	**Impala**: is an open-source, high-performance SQL query engine, provides interactive, low-latency SQL queries on large datasets, making it well-suited for ad-hoc querying, data exploration, and business intelligence (BI) analytics in big data environments.

6. Big Data Supported Algorithms:

Big Data analytics involves a variety of algorithms and techniques to process and analyse large and complex datasets. These algorithms help uncover insights, patterns, and relationships within the data. Here are some key types of algorithms used in Big Data analytics:

6.1 Machine Learning Algorithms:

6.1.1 Regression Algorithms:

Linear Regression, Logistic Regression, Polynomial Regression.

6.1.2 Classification Algorithms:

Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naive Bayes.

6.1.3 Clustering Algorithms:

K - Means, Hierarchical Clustering, DBSCAN.

6.1.4 Dimensionality Reduction Algorithms:

Principal Component Analysis (PCA), t-Distributed Stochastic.

CONCLUSION:

The realm of big data, along with its associated intricacies such as storage methodologies, supporting languages, tools, technologies, and platforms, continues to evolve, offering abundant opportunities for exploration. The captivating nuances of big data consistently yield novel insights, igniting the curiosity of researchers with each endeavour.

This expansive landscape has transformed big data into a virtual goldmine for researchers, ceaselessly providing uncharted territories for discovery. Moreover, its influence extends beyond research, permeating into education, drug development, healthcare, and beyond. The impact of big data technology resonates widely, leaving its marks across the expanse of the universe. In the present era, its influence is penetrating ever deeper, leaving indelible imprints as it continues to make profound strides.

The ongoing research endeavours centre around the synergistic fusion of big data, machine learning algorithms, Spark, and Python. This amalgamation aims to harness resources to their fullest potential and expedite the attainment of optimal solutions. The trajectory of this research is unequivocally leading towards the zenith of excellence, marked by efficiency, speed, and breakthrough outcomes.

REFERENCES:

1. S.S. Kumaran Nair, Nandita Ganesh, “An exploratory study on big data processing: a case study from bio medical informatics”, 3rd MEC International conference on Big Data and Smart city, 2016.
2. W.R. Herse, “Healthcare Data Analytics”, Health Informatics: Practical Guide for Healthcare& information Technology Professionals, 6th edition, 2014.
3. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, Athanasios V. Vasilakos, “Big data analytics: a Survey”,Journal of Big, vol. 2(21), Springer International Publishing, 2015.
4. Pekka Paakkonen, Daniel Pakkala, “Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems”, Elsevier Big Data Research, vol. 2(3), pp.166-186, December 2015.
5. M. V. Saravanakumar and S. M. Hanifa, "BIGDATA: Harnessing insights to healthier analytics — A survey," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICAMMAET.2017.8186648.
6. H. Shuijing, "Big Data Analytics: Key Technologies and Challenges," 2016 International Conference on Robots & Intelligent System (ICRIS), ZhangJiaJie, China, 2016, pp. 141-145, doi: 10.1109/ICRIS.2016.30
7. S. Singh and N. Singh, "Big Data analytics," 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2012, pp. 1-4, doi: 10.1109/ICCICT.2012.6398180.
8. D. Gaurav, J. K. P. Singh Yadav, R. K. Kaliyar and A. Goyal, "An Outline on Big Data and Big Data Analytics," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 74-79, doi: 10.1109/ICACCCN.2018.8748683.
9. T. Garg and S. Khullar, "Big Data Analytics: Applications, Challenges & Future Directions," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 923-928, doi: 10.1109/ICRITO48877.2020.9197797.
10. M. D. A. Praveena and B. Bharathi, "A survey paper on big data analytics," 2017 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 2017, pp. 1-9, doi: 10.1109/ICICES.2017.8070723.
11. N. Pavithra. and C. M. Manasa., "Big Data Analytics Tools: A Comparative Study," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/CSITSS54238.2021.9683711.
12. A. Ashabi, S. B. Sahibuddin and M. S. Haghighi, "Big Data: Current Challenges and Future Scope," 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Malaysia, 2020, pp. 131-134, doi: 10.1109/ISCAIE47305.2020.9108826.
13. S. Kohli and H. Singal, "Data Analysis with R," 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, UK, 2014, pp. 537-538, doi: 10.1109/UCC.2014.81.
14. Venkatesh Saravanakumar, M., Hanifa, S.M. (2019). Processing Using Spark—A

Potent of BD Technology. In: Mittal, M., Balas, V., Goyal, L., Kumar, R. (eds) Big Data Processing Using Spark in Cloud. Studies in Big Data, vol 43. Springer, Singapore. https://doi.org/10.1007/978-981-13-0550-4_9.

15. Aasha, D., Saravanakumar Venkatesh, & Sabibullah, M."Emerging Big Data Storage Architectures: A New Paradigm", i-manager's Journal on Pattern Recognition, Vol;4, Issue:2, ISSN-2349-7912,2017.

ID: 25

SURVEY ON DATA MINING TECHNIQUES AND ITS APPLICATIONS

Ms.K.JAYADEVI
III B.Sc. Information Technology,
Department of Information Technology,
Rathinam College of Arts and Science, Eachanari,
Coimbatore, Tamil Nadu, India.

Ms.M.SWETHA
III B.Sc. Information Technology,
Department of Information Technology,
Rathinam College of Arts and Science, Eachanari,
Coimbatore, Tamil Nadu, India.

ABSTRACT

Data mining is a process which finds useful patterns from large amount of data. The paper discusses few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results. Data mining is a rapidly growing field which has wide applications in variety of fields. It is a multi-disciplinary field which integrates statistics, neural networks, machine learning, visualization etc. This paper is an attempt to briefly review the various tools and techniques used in data mining. The paper also reviews some of the important applications of data mining in various areas. Data mining is the process of extracting hidden and useful patterns and information from data. Data mining is a new technology that helps businesses to predict future trends and behaviours, allowing them to make proactive, knowledge driven decisions. The aim of this paper is to show the process of data mining and how it can help decision makers to make better decisions. Practically, data mining is really useful for any organization which has huge amount of data. Data mining help regular databases to perform faster. They also help to increase the profit, because of the correct decisions made with the help of data mining. This paper shows the various steps performed during the process of data mining and how it can be used by various industries to get better answers from huge amount of data.

KEYWORDS: *Data Mining Techniques, Data Mining Algorithms, Data Mining Applications.*

I. INTRODUCTION

Data mining can be defined as the process of extracting valid, previously unknown and actionable information from large data sets. The purpose of the data mining is to use the extracted information to make crucial business decisions. So, Data Mining helps end users extract useful business information from large volume of data. This is a commonly used word for any kind of large-scale data processing. The mined results should be valid, novel, useful, and understandable. Data Mining is related to the subarea of statistics called exploratory data analysis and subarea of artificial intelligence called knowledge discovery and machine learning. This paper presents a brief introduction about data mining in section one. The second section illustrates the process of data mining while the third section reviews different data mining techniques. The fourth section is committed to various application areas of Data Mining and fifth section discusses conclusion and future scope.

II. RELATED WORKS

Anshu [1], Process of Data Mining: Data mining process is a step-by-step procedure that cannot be completed in a single step. In other words, you cannot get the required information from the large volumes of data as simple as that. It is not specific to any industry. Basically, the process has evolved from the knowledge discover processes used widely in industry. The major aim of Data Mining process is to make large data projects to run more efficiently. The processes including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation are to be completed in the given order.

Business Understanding:

Business understanding phase focuses on understanding the project objectives and requirements, assessing the current situation, establishing data mining goals from the business point of view. In this phase we prepare the preliminary plan for the project.

Data Understanding:

This phase includes activities like initial data collection, data description, data exploration, and the verification of Review Paper on Data Mining Techniques and Applications Anshu Data understanding deployment Business understanding evaluation modelling Data preparation Review Paper on Data Mining Techniques and Applications It is basically concerned with establishing the main characteristics of data which includes the data structures, data quality and identifying any interesting subsets of the data. The major tasks performed under this phase are collecting initial data, describing data and exploring data and verify data.

- First, data is collected from multiple data sources available in the organization.
- Next, the step is to search for properties of acquired data.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

Data Preparation:

This phase involves all the activities for constructing the final data set into the desired form. The main activities performed during this phase are select data, cleaning data, data integration and data transformation. In this phase, data is made production ready. The output of this phase is data set that can be used in modelling.

Modelling:

In Data Modelling step, we select modelling techniques, modelling parameters and assess the model created based on the business objectives. Once greater data understanding is gained (often through pattern recognition triggered by viewing model output), more detailed models suitable to the data can be applied. The various activities performed during this phase are selecting modelling technique, generate test design, build model and assess model. For creating suitable model following steps are taken:

- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Evaluation:

This phase validates the model from the data analysis point of view. In this step the model and the steps in modelling are verified within the context of achieving the business objectives. The various activities performed during this phase include evaluating results, reviewing process. Evaluation results should be evaluated according to the business objectives. A go or no-go decision is taken to move the model in the deployment phase.

Deployment:

In this phase the knowledge obtained in the form of model is to be organized and presented in such a form that can be used by the business users. This process can be as simple like generating report or it may be complex as implementing the repeatable data mining process again and again. This is the execution phase. The various tasks involved in this phase are plan deployment, plan monitoring and maintenance, produce and review the final report. So, in the deployment phase, Patterns are deployed for desired outcome.

R. Tamilselvi, S. Kalaiselvi e [2], Data mining has attracted a great attention in the information industry and in society as a whole in recent years, due to wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, to production control, disaster management and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of various functionalities: data collection and database creation, database management (including data storage and retrieval, and database transaction processing and advance data analysis Knowledge discovery as a process consists of an iterative sequence of following steps:

1. Data cleaning, that is, to remove noise and inconsistent data.

2. Data integration, that is, where multiple data sources are combined.
3. Data selection, that is, where data relevant to the analysis task are retrieved from the database.
4. Data transformation, that is, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining, that is, an essential process where intelligent methods are applied in order to extract the data patterns.
6. Knowledge presentation, that is, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Data Mining Applications:

Data Mining for Financial Data Analysis

In Banking Industry data mining is used:

1. Predicting Credit fraud
2. Evaluation Risk
3. Performing trend analysis
4. Analysing profitability
5. Helping with direct marketing campaigns

In financial markets and neural networks data mining is used:

1. Forecasting stock prices
2. Forecasting commodity-price prediction
3. Forecasting financial disasters

Mrs. Bharati M. Ramageri, Lecturer [3], Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases.

Classification

It is the most applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and creditrisk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analysed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations.

Clustering:

It can be said as identification of similar classes of objects. By using clustering techniques, we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of Clustering Methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods.
- Grid-based methods
- Model-based methods

Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of Regression Methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Association Rule:

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing, and customer shopping behaviour analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However, the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of Association Rule:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule.

Neural Networks:

It is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example, handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of Neural Networks

- Back Propagation

Dr. Malla Reddy Jogannagari, Mrs. Maheshwari Manchala[4], Data Mining Tools: Data mining tools predicts future trends and make business into proactive. The various data mining tools such as Rapid Miner, Mahout, Orange, Weka and Data Melt can tackle the business problems effectively.

Rapid Miner: It is open source software for data and text mining. Rapid analytics is a server product.

Mahout: Mahout offers clustering and based collaborative filtering that run on top of Hadoop.

Orange: This project make data mining effective for both novices and experts. It provides a wide variety of visualization plus a tools boxes with widgets.

Weka: Weka provides set of data mining algorithms for knowledge analysis. User can apply data directly or use with java application.

Data Melt: Data Melt provides mathematical computation data mining statistical analysis and data visualization.

Fathimath Zuha Maksood, Geetha Achuthan[5], The above definitions and descriptions of Knowledge Discovery and Data Mining are theoretically formulated from the need for it in today‟s world. The development of technology has gradually resulted in the replacement of manual logs by machine. Data have since been accumulated, using the traditional file processing systems in an unordered way. This led to the mismanagement of data which was later replaced by databases. The ease of use of related databases have allowed most organizations to adopt the technology for storing their transactions and related information. The number of databases was approximated to be around five million in early 90s; 20 years later, one can just imagine the volume of data accumulated in various technological sites. Hence, data acquisition can be recognized as a double-edged sword.

Even though it implements an ease-of-use data storage strategy for organizations, data can pile up at an immense rate creating raw and unprocessed records. Despite this perception, the large amount of accumulated records can be utilized in an advantageous manner if it can be processed using appropriate means. It is logical to predict the existence of interesting relationships or hidden patterns in records aggregated over the years. Often, this information can be used to describe the records stored, find patterns in a users" transaction which was previously unknown, predict forthcoming data and most importantly use these details to create an information rich smart system. If a user's behaviour can be predicted by an organization using machine learning methods on existing dataset, it can prove to be an advantage as they now have to culminate only a known required amount of resources for the particular customer. This is one example of how data mining can be used to create a smarter environment. The next section further illustrates certain real time applications of data mining.

Prachi Deshmukh-Chaudhari, Abhijeet Deshmukh [6], DATA MINING TOOLS: Most of the data mining tools can be classified into three major categories:

1. Traditional data mining tools
2. Dashboards
3. Text-mining tools

Traditional data mining tools:

Traditional data mining tools such as OLAP (On Line Analytical Processing) help companies to establish data patterns and trends by using a number of complex algorithms and techniques.

Dashboards:

Dashboards are installed in computers to monitor information in a database. They reflect data changes and updates onscreen. It can be a chart, graph or table which enables the user to observe the business performance. Dashboards are easy to use and preferred because they can give an overview of the company's performance at a glance.

Text-mining tools:

These tools scan content of a document and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Some of the examples of text mining software's are RapidMiner, Fair Isaac, OpenNLP, IBM Language Ware, SAS etc.

Shital H. Bhojani, Dr. Nirav Bhatt[7], Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge. In general, there are three main steps in DM: preparing the data, reducing the data and, finally, looking for valuable information. The specific approaches, however, differ from companies to companies and researchers to researchers. Fayyad et al. (1996) proposed the following steps:

1. Retrieving the data from a large database.
2. Selecting the relevant subset to work with.
3. Deciding on the appropriate sampling system, cleaning the data and dealing with missing fields and records.
4. Applying the appropriate transformations, dimensionality reduction, and projections.
5. Fitting models to the pre-processed data.

K.Suguna, Dr.K.Nandhini[8], Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. In "Literature review on data mining research"Given a set of examples (training data) described by some set of attributes (ex. Sex, rank, background) the goal of the algorithm is to learn the decision function stored in the data and then use it to classify new inputs. The concept of information gain or Gini index.

In 2013, Ms. Pooja Agrawal et al. [9] This review of sequential pattern-mining algorithms in shows that the important heuristics employed includes the optimally sized data structure representations of the sequence database; early pruning of candidate sequences; mechanisms to reduce support counting; and maintaining a narrow search space.

In 2014, Vishal S. Motegaonkar, Prof. Madhav V. Vaidya et al. [10] Initial work on this topic is concentrated on improvement of the performance of algorithms by using different data structure or different representation. So, on the basis of these problems the sequential pattern mining is categorized into two types, Apriori approach based algorithms and pattern growth approach based algorithms. This survey and previous some studies by various researchers on sequential pattern mining algorithms it is found that the algorithm which are based on the approach of pattern growth are better in terms of scalability, time-complexity and space-complexity.

In 2013, P. IndiraPriya, Dr. D.K.Ghosh et al. [11] describes about the Cluster analysis, the group of data objects based only on the information found in the data that describes the objects and their relationships. The aim is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in the other groups. The greater similarity of clustering is within a group and the greater difference between groups, the more distinct the clustering. The cluster analysis splits the space into regions, characteristic of the clusters found in the data. The main benefit of a clustered solution is automatic recovery from failure. The difficulties of clustering are complication and inability to recover from database corruption.

IV. Research Methodology

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate. Many other terms carry a similar or slightly different meaning to data mining such as knowledge mining from data, knowledge extraction, data/pattern analysis data dredging.

Data mining treats as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. In others view data mining as simply an essential step in the process of knowledge discovery, in which intelligent methods are applied in order to extract data patterns.

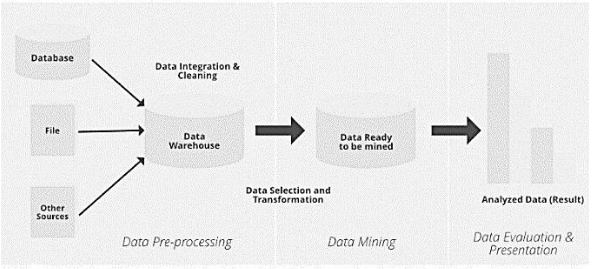


Fig 1. Data mining

Knowledge Discovery From Data Consists of the Following Steps:

Knowledge discovery from data (KDD) is a multi-step process that involves extracting useful knowledge from data. The following are the steps involved in the KDD process:

Data Selection:

The first step in the KDD process is to select the relevant data for analysis. This involves identifying the data sources and selecting the data that is necessary for the analysis.

Data Preprocessing:

The data obtained from different sources may be in different formats and may have errors and inconsistencies. The data preprocessing step involves cleaning and transforming the data to make it suitable for analysis.

Data Transformation:

Once the data has been cleaned, it may need to be transformed to make it more meaningful for analysis. This involves converting the data into a form that is suitable for data mining algorithms.

Data Mining:

The data mining step involves applying various data mining techniques to identify patterns and relationships in the data. This involves selecting the appropriate algorithms and models that are suitable for the data and the problem being addressed.

Pattern Evaluation:

After the data mining step, the patterns and relationships identified in the data need to be evaluated to determine their usefulness. This involves examining the patterns to determine whether they are meaningful and can be used to make predictions or decisions.

Knowledge Representation:

The patterns and relationships identified in the data need to be represented in a form that is understandable and useful to the end-user. This involves presenting the results in a way that is meaningful and can be used to make decisions.

Knowledge Refinement:

The knowledge obtained from the data mining process may need to be refined further to improve its usefulness. This involves using feedback from the end-users to improve the accuracy and usefulness of the results.

Knowledge Dissemination:

The final step in the KDD process involves disseminating the knowledge obtained from the analysis to the end-users. This involves presenting the results in a way that is easy to understand and can be used to make decisions.

V. Reference

1. Hsu, J. 2002. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002), ISSN: 1542-7382. Available Online: http://colton.byuh.edu/isecon/2002/224b/ Hsu.pdf.
2. Data mining techniques and applications – A decade review from 2000 to 2011.
3. LITERATURE REVIEW ON DATA MINING TECHNIQUES K. Suguna Asst.Professor, Dr.K.Nandhini Professor
4. Anshu, NET Qualified, M.Tech Professional, with a rich Experience of 9 Years in the field of Teaching and Education, and current working as an Asst. Professor, Computer Science at GVM Girls college, Sonipat.
5. Jeffrey Voasand JiaZhang, ―Cloud Computing: New Wine or Just a New Bottle? ‖, Database Systems Journal vol. III, no. 3/2012 71IEEEInternet Computing Magazine.
6. Tayel, Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141149, 2014.
7. Han. J, Kamber. M, Pei. J, "Data Mining Concepts and Techniques", Third edition The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011.
8. Kabra. R, Bichkar. R, "Performance Prediction of Engineering Students using Decision Tree", International Journal of computer Applications, December, 2011.
9. Nolofar Rehman,“ Data Mining Techniques Methods, Algorithms and Tools", International Journal of Computer Science & Mobile Computing, Vol. 6(7), July, 2017.
10. Priyanka Gautam, "Impact of Data Mining on Big Data Analytics: Challenges and Opportunities", Vol. 57(1), March, 2018.
11. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman.
12. D. Borthakur, "HDFS Architecture Guide," 4 August 2013. Available Online: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf. [Accessed 30 November 2015]

ID: 26

HYDRO CHEMICAL PREDICTION USING CLASSIFICATION ALGORITHMS

Ms.U.SARALA DEVI
Assistant Professor,
Department of Software Application,
Agurchand Manmull Jain College,
Meenambakkam, Chennai, Tamil Nadu, India.

Ms.G.JOTHIPRIYA
Assistant Professor,
Department of Software Application,
Agurchand Manmull Jain College,
Meenambakkam, Chennai, Tamil Nadu, India.

ABSTRACT

Freshwater is a critical resource for agriculture and industry's survival. Examination of water quality is a fundamental stage in the administration of freshwater assets. As indicated by the World Health Organization's yearly report, many individuals are getting sick or some are dead due to the lack of safe drinking water. It is critical to test the quality of water prior to involving it for any reason, whether it is for animal watering, chemical spraying (Pesticides etc..), or drinking water. Evaluation, monitoring, and prediction of water quality as become significant and popular study topics. Many experts and researches have done the research on quality of water in the past years using variety of methodologies. Most of the researchers used pH, Biological Oxygen Demand, Electric Conductivity, Chemical Oxygen Demand in their research to examine the water quality. The Hydro-chemical is the study that reveals the quality of water which describes the condition of water, including chemical, physical and biological characteristics that is suitable for drinking, agriculture and industries purposes. The hydro-chemical is measured by several factors, such as the concentration of dissolved oxygen, bacterial levels, the amount of salts, or the amount of material suspended in thewater.

PH, Hardness, Solids, Chloromines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity are the parameters used as a feature vector to depict the quality of water. It is observed that machine learning models such as Random Forest Classifier, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree Classifier have shown the best accuracy for predicting water quality. The proposed model is a comparison of machine learning algorithms (KNN, SVM, Logistic Regression, Xgboost, etc.,) . The result from the proposed model can accurately predict the quality of water based on the features.

I. INTRODUCTION

Water, a substance composed of the chemical elements hydrogen and oxygen and existing in gaseous, liquid, and solid states. It is one of the most plentiful and essentialof compounds. A tasteless and odourless liquid at room temperature, it has the important ability to dissolve many other substances. Indeed, the versatility of water as a solvent is essential to living organisms. The water on the surface of Earth is found mainly in its oceans (97.25 percent) and polar ice caps and glaciers (2.05 percent), withthe balance in freshwater lakes, rivers, and groundwater. As Earth's population grows and the demand for fresh water increases, water purification and recycling become increasingly important. Interestingly, the purity requirements of water for industrial useoften exceed those for human consumption.

Water quality refers to the suitability of water for different uses according to its physical, chemical, biological, and organoleptic (taste-related) properties. It is especially important to understand and measure water quality as it directly impacts human consumption and health, industrial and domestic use, and the natural environment. Regulations such as the EU Drinking Water Directive and regulatory agencies such as the US Environmental Protection Agency (EPA) set standards for enforcement of water quality, with local governments around the world usually acting as the front-line enforcers. Water quality is measured using laboratory techniques or home kits. Laboratory testing measures multiple parameters and provides the most accurate results but takes the longest time. Home test kits, including test strips, provide rapid results but are less accurate. Water suppliers including municipalities and bottledwater companies often make their water quality reports publicly available on their websites. The tested water quality parameters must meet standards set by their local governments which are often influenced by international standards set by industry or water quality organizations such as the World Health Organization (WHO).

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions. The parameters used to predict the quality of water are:

- pH,
- Solids,
- Hardness,
- Chloromines,
- Sulfate,
- Conductivity,
- Organic Carbon,
- Trihalomethanes,
- Turbidity,
- Potability

2. METHODOLOGY

Methodology refers to the procedure or technique used for analyzing informationin an orderly way. This system follows the top down approach. This emphasizes planning and complete understanding of the system. This project is developed by usingPython.

2.1 Machine Learning

It is the scientific study of algorithms and statistical models that computer systemsuse to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning.

Random Forests

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individualtree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Random forest is a supervised learning algorithm which is used for both classification as well as regression. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selectsthe best solution by means of voting.

Decision Tree

There are many types of machine learning methods. One of them is decision tree. A decision tree is similar looking to a tree diagram. A tree diagram begins with a single node and from that node, branches will reach out to new nodes that represent mutually exclusive decisions or events, meaning that they cannot occur simultaneously and is not influenced by the other.

The diagram starts by the first node and there after decisions and events will bring it onto the next node. So in other words, it is a sequence of events and this is useful in probability since it can record all the possible outcomes by adding more branches. Ifit is used to calculate probabilities, the probability is then put onto the separate branches and the outcome is the next-coming node.

A decision tree is a set of questions connected in a tree, so that answering one question leads to another and eventually to a final answer. It can be used for either regression or classification problems. When it is used to solve regression problems it is called regression tree. A decision tree is a directed tree and has one root node without any incoming edges. Other nodes have exactly one incoming edge and they are grown by a conditional split that divides the result into two new branches where the tree can continue to grow. There can be consecutive questions with different conditions by each new branch, creating additional branches that form the tree in a downward matter, or else the branch will end there. The ends of a branch are called a leaf or a terminal node and it displays the observation that falls into that specific branch.

The splits are done with a greedy top-down approach; the best split is done in eachstep even if another split could have created a better tree in a future step.

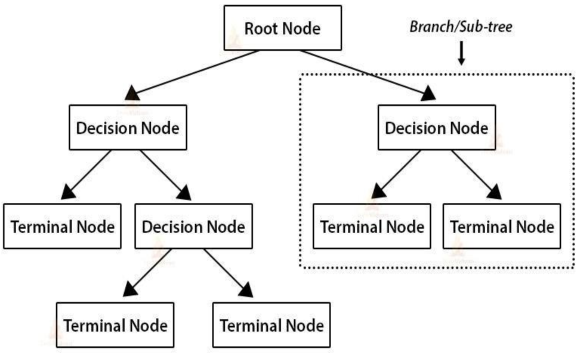


Figure 2.1 Decision Tree

K-Nearest Neighbor

In pattern recognition, the k-nearest neighbors algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classifiedby a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In K-NN regression, the output is the property value for the object. This value isthe average of the values of k nearest neighbors. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).Logistic regression is used to describe data and to explain the relationship between onedependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

AdaBoost Classifier

AdaBoost makes 'n' number of decision trees during the data training period. As the first decision tree/model is made, the incorrectly classified record in the first model is given priority. Only these records are sent as input for the second model. The processgoes on until we specify a number of base learners we want to create. Remember, repetition of records is allowed with all boosting techniques.

The record which is incorrectly classified is used as input for the next model. This process is repeated until the specified condition is met. there are 'n' number of models
made by taking the errors from the previous model. This is how boosting works. The models 1,2, 3,…, N are individual models that can be known as decision trees. All typesof boosting models work on the same principle.

XGBoost Classifier

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for "Extreme Gradient Boosting" and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-in support for parallel processing, making it possible to train models on large datasets in a reasonable amount of time.

III. IMPLEMENTATION

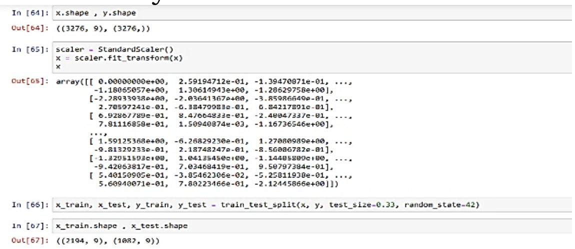
Implementation includes all those activities that take place to convert from old system to the new. The new system may be totally new, replacing an existing system orit may be major modification to the system currently put into use.

3.1.1. Data collection and Pre-processing

Text based classification dataset is collected from the Kaggle website. Dataset is preprocessed to clean the noise data by removing the Null values.

Training Data and Test Data

Machine learning concept includes learning some information or we can say properties from dataset and testing properties of another data can be classified. Mostly we are splitting data in two categories as training set and testing set. Training set (67%) is the set of learning properties from given data. Testing set (33%) is to check properties of the data given for query which may include some different properties. Depending on Training set with different classifier we may get different results with different accuracy. The classifier of type deep learning which is having highest classification accuracy can beconsidered for further analysis.

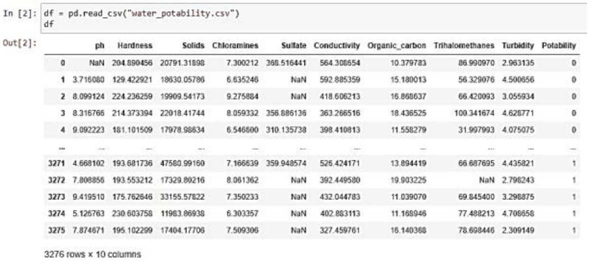


```
In [64]: x.shape , y.shape

Out[64]: ((3276, 9), (3276,))

In [65]: scaler = StandardScaler()
         x = scaler.fit_transform(x)
         x

Out[65]: array([[ 0.00000000e+00,  2.59194712e-01, -1.39470871e-01, ...,
                 -1.18065057e+00,  1.30614943e+00, -1.28629758e+00],
                [-2.28933938e+00, -2.03641367e+00, -3.85986649e-01, ...,
                  2.70597241e-01, -6.38479983e-01,  6.84217891e-01],
                [ 6.92867789e-01,  8.47664833e-01, -2.40047337e-01, ...,
                  7.81116858e-01,  1.50940874e-03, -1.16736546e+00],
                ...,
                [ 1.59125368e+00, -6.26829230e-01,  1.27080989e+00, ...,
                 -9.81329233e-01,  2.18748247e-01, -8.56086782e-01],
                [-1.32951593e+00,  1.04135450e+00, -1.14485889e+00, ...,
                 -9.42063817e-01,  7.03468419e-01,  9.50797384e-01],
                [ 5.40150905e-01, -3.85462306e-02, -5.25811938e-01, ...,
                  5.60940071e-01,  7.80223466e-01, -2.12445866e+00]])

In [66]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)

In [67]: x_train.shape , x_test.shape

Out[67]: ((2194, 9), (1082, 9))
```

Figure 3.1 Training Data and Test Data

3.1.2 Data Analysis:

Data analysis is the process of cleaning, changing, and processing raw data and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graphs.



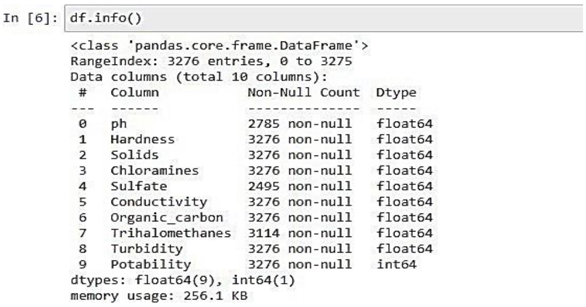
```
In [2]: df = pd.read_csv("water_potability.csv")
        df
```

Out[2]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890456	20791.31898	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.05786	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.54173	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771	0
4	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681736	47580.99160	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.80216	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.57822	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.86938	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.17706	7.509306	NaN	327.459761	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

Figure 3.2 Data Analysis – Importing Dataset

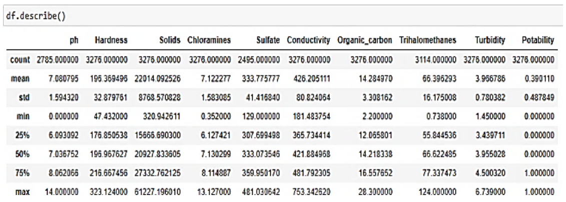


```
In [6]: df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 3276 entries, 0 to 3275
        Data columns (total 10 columns):
         #   Column           Non-Null Count  Dtype
        ---  ------           --------------  -----
         0   ph               2785 non-null   float64
         1   Hardness         3276 non-null   float64
         2   Solids           3276 non-null   float64
         3   Chloramines      3276 non-null   float64
         4   Sulfate          2495 non-null   float64
         5   Conductivity     3276 non-null   float64
         6   Organic_carbon   3276 non-null   float64
         7   Trihalomethanes  3114 non-null   float64
         8   Turbidity        3276 non-null   float64
         9   Potability       3276 non-null   int64
        dtypes: float64(9), int64(1)
        memory usage: 256.1 KB
```

Figure 3.3 Data Analysis – Info( )

3.1.3. Descriptive Analytics

Descriptive Analytics help describe and understand the features of a specific dataset by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing bythe number of figures within the set.

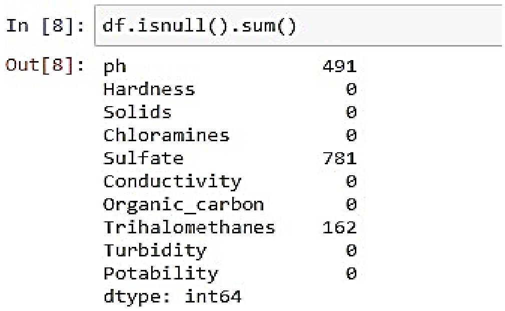


```
df.describe()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690300	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833605	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762125	8.114887	359.950170	481.792305	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196010	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

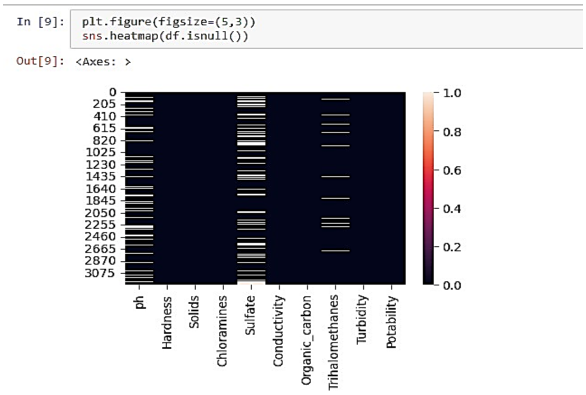
Figure 3.4 Descriptive Analytics – Describe( )

Finding the Null values



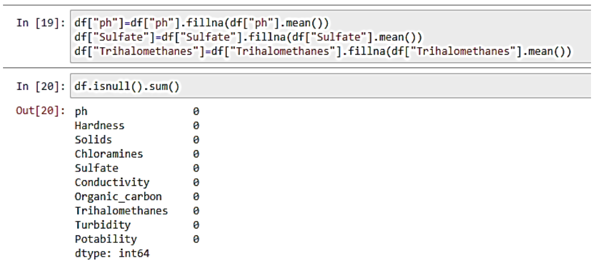
```
In [8]: df.isnull().sum()

Out[8]: ph                 491
        Hardness             0
        Solids               0
        Chloramines          0
        Sulfate            781
        Conductivity         0
        Organic_carbon       0
        Trihalomethanes    162
        Turbidity            0
        Potability           0
        dtype: int64
```

Figure 3.5 Descriptive Analytics – Null Values



```
In [9]: plt.figure(figsize=(5,3))
        sns.heatmap(df.isnull())

Out[9]: <Axes: >
```

Figure 3.6 Descriptive Analytics – Null Value Visualization



```
In [19]: df["ph"]=df["ph"].fillna(df["ph"].mean())
         df["Sulfate"]=df["Sulfate"].fillna(df["Sulfate"].mean())
         df["Trihalomethanes"]=df["Trihalomethanes"].fillna(df["Trihalomethanes"].mean())

In [20]: df.isnull().sum()

Out[20]: ph                 0
         Hardness           0
         Solids             0
         Chloramines        0
         Sulfate            0
         Conductivity       0
         Organic_carbon     0
         Trihalomethanes    0
         Turbidity          0
         Potability         0
         dtype: int64
```

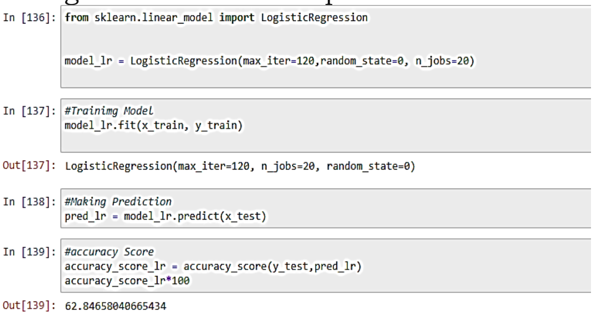
Figure 3.7 Descriptive Analytics – Fixing Null Value

Predictive Analytics

Predictive analysis that makes predictions about future outcomes using historicaldata combined with statistical modeling, data mining techniques and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities.

Logistic Regression:

This type of statistical model (also known as *logit model*) is often used forclassification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independentvariables.

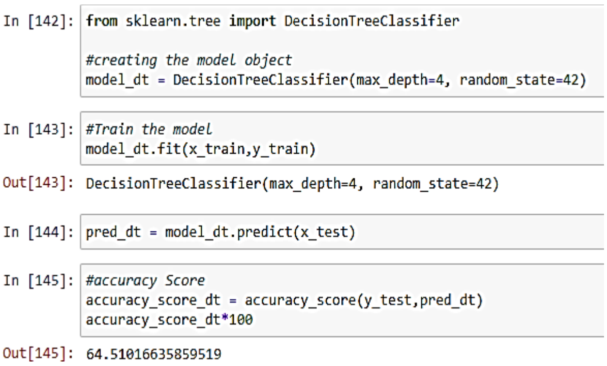


```
In [136]: from sklearn.linear_model import LogisticRegression


          model_lr = LogisticRegression(max_iter=120,random_state=0, n_jobs=20)

In [137]: #Trainimg Model
          model_lr.fit(x_train, y_train)

Out[137]: LogisticRegression(max_iter=120, n_jobs=20, random_state=0)

In [138]: #Making Prediction
          pred_lr = model_lr.predict(x_test)

In [139]: #accuracy Score
          accuracy_score_lr = accuracy_score(y_test,pred_lr)
          accuracy_score_lr*100

Out[139]: 62.84658040665434
```

Figure 3.8 Predictive Analytics – Logistic Regression

Decision Tree Classifier:

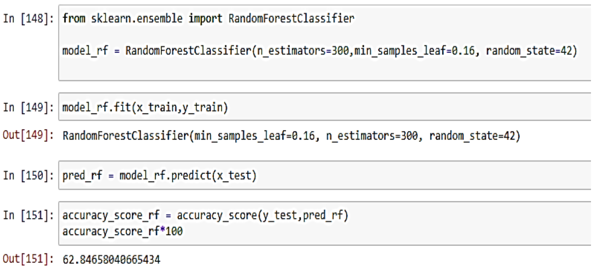
A decision tree is a non-parametric supervised learning algorithm, which is utilizedfor both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.



```
In [142]: from sklearn.tree import DecisionTreeClassifier

          #creating the model object
          model_dt = DecisionTreeClassifier(max_depth=4, random_state=42)

In [143]: #Train the model
          model_dt.fit(x_train,y_train)

Out[143]: DecisionTreeClassifier(max_depth=4, random_state=42)

In [144]: pred_dt = model_dt.predict(x_test)

In [145]: #accuracy Score
          accuracy_score_dt = accuracy_score(y_test,pred_dt)
          accuracy_score_dt*100

Out[145]: 64.51016635859519
```

Figure 3.9 Predictive Analytics – Decision Tree ClassifierRandom Forest Classifier

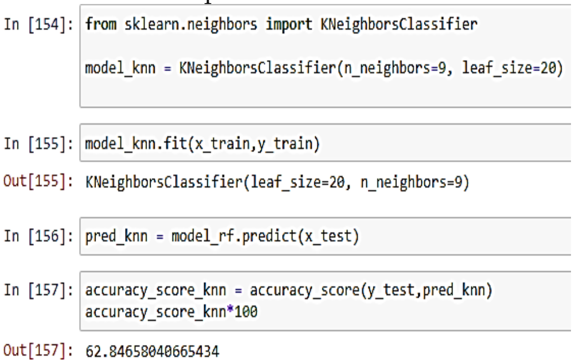
Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.



```
In [148]: from sklearn.ensemble import RandomForestClassifier

          model_rf = RandomForestClassifier(n_estimators=300,min_samples_leaf=0.16, random_state=42)

In [149]: model_rf.fit(x_train,y_train)

Out[149]: RandomForestClassifier(min_samples_leaf=0.16, n_estimators=300, random_state=42)

In [150]: pred_rf = model_rf.predict(x_test)

In [151]: accuracy_score_rf = accuracy_score(y_test,pred_rf)
          accuracy_score_rf*100

Out[151]: 62.84658040665434
```

Figure 3.10 Predictive Analytics – Random Forest ClassifierK-Nearest Neighbors:

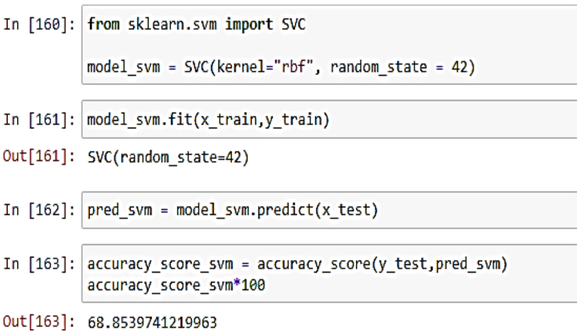
The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non- parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.



```
In [154]: from sklearn.neighbors import KNeighborsClassifier

          model_knn = KNeighborsClassifier(n_neighbors=9, leaf_size=20)

In [155]: model_knn.fit(x_train,y_train)

Out[155]: KNeighborsClassifier(leaf_size=20, n_neighbors=9)

In [156]: pred_knn = model_rf.predict(x_test)

In [157]: accuracy_score_knn = accuracy_score(y_test,pred_knn)
          accuracy_score_knn*100

Out[157]: 62.84658040665434
```

Figure 3.11 Predictive Analytics – K-Nearest NeighborSupport Vector Machine

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

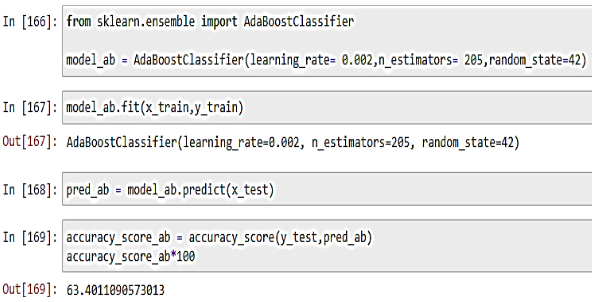


```
In [160]: from sklearn.svm import SVC

          model_svm = SVC(kernel="rbf", random_state = 42)

In [161]: model_svm.fit(x_train,y_train)

Out[161]: SVC(random_state=42)

In [162]: pred_svm = model_svm.predict(x_test)

In [163]: accuracy_score_svm = accuracy_score(y_test,pred_svm)
          accuracy_score_svm*100

Out[163]: 68.8539741219963
```

Figure 3.12 Predictive Analytics – Support Vector Machine

Adaboost Classifier:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

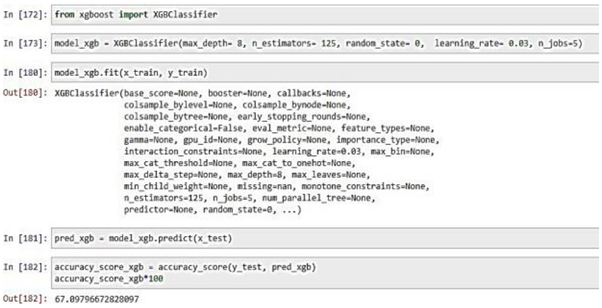


```
In [166]: from sklearn.ensemble import AdaBoostClassifier

          model_ab = AdaBoostClassifier(learning_rate= 0.002,n_estimators= 205,random_state=42)

In [167]: model_ab.fit(x_train,y_train)

Out[167]: AdaBoostClassifier(learning_rate=0.002, n_estimators=205, random_state=42)

In [168]: pred_ab = model_ab.predict(x_test)

In [169]: accuracy_score_ab = accuracy_score(y_test,pred_ab)
          accuracy_score_ab*100

Out[169]: 63.4011090573013
```

Figure 3.13 Predictive Analysis – Adaboost Classifier

Xgboost Classifier:

XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

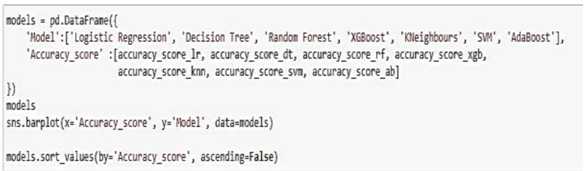


```
In [172]: from xgboost import XGBClassifier

In [173]: model_xgb = XGBClassifier(max_depth= 8, n_estimators= 125, random_state= 0,  learning_rate= 0.03, n_jobs=5)

In [180]: model_xgb.fit(x_train, y_train)

Out[180]: XGBClassifier(base_score=None, booster=None, callbacks=None,
                        colsample_bylevel=None, colsample_bynode=None,
                        colsample_bytree=None, early_stopping_rounds=None,
                        enable_categorical=False, eval_metric=None, feature_types=None,
                        gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
                        interaction_constraints=None, learning_rate=0.03, max_bin=None,
                        max_cat_threshold=None, max_cat_to_onehot=None,
                        max_delta_step=None, max_depth=8, max_leaves=None,
                        min_child_weight=None, missing=nan, monotone_constraints=None,
                        n_estimators=125, n_jobs=5, num_parallel_tree=None,
                        predictor=None, random_state=0, ...)

In [181]: pred_xgb = model_xgb.predict(x_test)

In [182]: accuracy_score_xgb = accuracy_score(y_test, pred_xgb)
          accuracy_score_xgb*100

Out[182]: 67.09796672828097
```

Figure 3.14 Predictive Analytics – XGboost Classifier

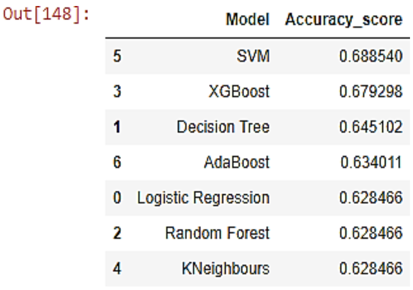
a) Comparison of Algorithms

In order to decide which algorithm to choose over another, they are compared in terms of their efficiency: the time it takes to find the solution and the resources which are consumed in the process.



```
models = pd.DataFrame({
    'Model':['Logistic Regression', 'Decision Tree', 'Random Forest', 'XGBoost', 'KNeighbours', 'SVM', 'AdaBoost'],
    'Accuracy_score' :[accuracy_score_lr, accuracy_score_dt, accuracy_score_rf, accuracy_score_xgb,
                       accuracy_score_knn, accuracy_score_svm, accuracy_score_ab]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

models.sort_values(by='Accuracy_score', ascending=False)
```

Figure 3.15 Predictive Analysis – Comparing Algorithm



Out[148]:

	Model	Accuracy_score
5	SVM	0.688540
3	XGBoost	0.679298
1	Decision Tree	0.645102
6	AdaBoost	0.634011
0	Logistic Regression	0.628466
2	Random Forest	0.628466
4	KNeighbours	0.628466

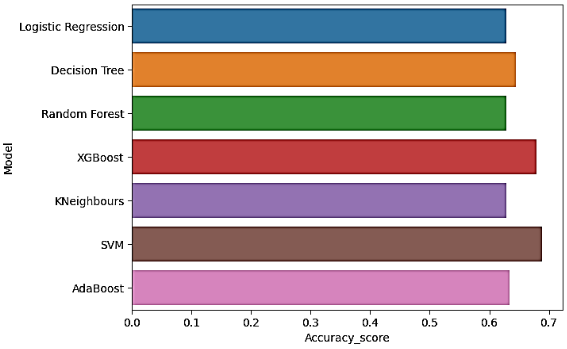
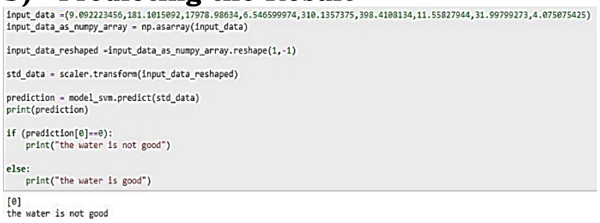


Figure 3.16 Predictive Analysis – Comparing Algorithm Visualization

Here, we concluded that the support vector machine algorithm gives more accuracy then others. Hence using the SVM algorithm we are creating the web application.

b) Predicting the Result



```
input_data =(9.092223456,181.1015092,17978.98634,6.546599974,310.1357375,398.4108134,11.55827944,31.99799273,4.075075425)
input_data_as_numpy_array = np.asarray(input_data)

input_data_reshaped =input_data_as_numpy_array.reshape(1,-1)

std_data = scaler.transform(input_data_reshaped)

prediction = model_svm.predict(std_data)
print(prediction)

if (prediction[0]==0):
    print("the water is not good")

else:
    print("the water is good")
```

```
[0]
the water is not good
```

Figure 3.17 Predictive Analysis – Prediction

IV. CONCLUSION

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize waterresource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality: (1) Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations. (2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hindersthe wide application of machine learning approaches. (3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

V. REFERENCES

1. A. N. Prasad, K. Al Mamun, F. R. Islam, and H. Haqva, “Smart water quality monitoring system,” in *Proceedings of the 2nd IEEE Asia Pacific World Congress on Computer Science and Engineering*, IEEE, Fiji Islands, December 2015. View at: Publisher Site | Google Scholar
2. P. Li and J. Wu, “Drinking water quality and public health,” *Exposure and Health*, vol. 11, no. 2, pp. 73–79, 2019. View at: Publisher Site | Google Scholar
3. Y. Khan and C. S. See, “Predicting and analyzing water quality using machine learning: a comprehensive model,” in *Proceedings of the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, IEEE, Farmingdale, NY, USA, April 2016. View at: Publisher Site | Google Scholar
4. D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, “Using machine learning models for predicting the water quality index in the La buong river, Vietnam,” *Water*, vol. 14, no. 10, p. 1552, 2022. View at: Publisher Site | Google Scholar
5. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, “Efficient water quality prediction using supervised machine learning,” *Water*, vol. 11,p. 2210, 2019. View at: Publisher Site | Google Scholar
6. S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S Kateb, “Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast),” *Applied Water Science*, vol. 11, no. 12, p. 190, 2021.
7. J. P. Nair and M. S. Vijaya, “Predictive models for river water quality using machine learning and big data techniques - a Survey,” in *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, Coimbatore, India, March 2021.

ID: 27

PROGNOSTICATION OF COVID-19 CASES UTILIZING AN ENHANCED K-NEAREST NEIGHBOR (KNN) APPROACH BY SMOTE MACHINE LEARNING

Ms.K.SINDHU
Research Scholar,
PG & Research Department of Computer Science,
Government Arts College (Autonomous), Karur,
Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

Ms.S.VANISRI
Assistant Professor,
Department of Master of Computer Applications,
M.Kumarasamy College of Engineering (Autonomous), Karur,
Anna University, Chennai, Tamil Nadu, India.

ABSTRACT:

In the twilight of 2019, an unprecedented variant of the SARS-CoV-2 coronavirus emerged, unleashing a pernicious respiratory affliction known as COVID-19. This insidious malady swiftly propagated across the globe, stealthily infiltrating its hosts, many of whom bore no outward signs of infection. Alas, these asymptomatic or pre-symptomatic carriers remained unwitting vectors of transmission, thereby confounding efforts to contain this pernicious scourge. Consequently, an urgent imperative arose to unveil novel biomarkers spanning the vast terrain of omics, empowering large-scale screening and diagnostic endeavors targeting COVID-19. Although preliminary investigations have yielded potential genetic signposts for COVID-19, their clinical utility has proven wanting, particularly in the realm of discriminating this malady from other infectious afflictions. In the present study, we harnessed the formidable prowess of three distinct classification algorithms, namely logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB), to meticulously scrutinize the comprehensive dataset at hand. As diligent custodians of scientific rigor, we employed various preprocessing techniques to ensure the data's fidelity and subsequently subjected it to the rigors of a 10-fold cross-validation, deftly partitioning it for rigorous assessment. Furthermore, we employed the esteemed EKSMOTE method to redress the inherent imbalances residing within the dataset. Ultimately, we discerned the predictive potency of twenty salient clinical features, discerned to be pivotal in forecasting the divergent fortunes of COVID-19 patients—those destined for survival and those tragically doomed. Our findings unveiled the supremacy of the random forest classifier, transcending the prowess of its algorithmic counterparts. Exhibiting an accuracy rating of 0.9 and a peerless area under the curve (AUC) value of 0.9, this discerning model emerges as a beacon of hope in the labyrinthine realm of COVID-19. Propitious it is, for this groundbreaking endeavor harbors the potential to bestow invaluable aid upon healthcare professionals and decision-makers alike, allowing them to swiftly identify the imperiled souls imperiled by COVID-19's clandestine clutches, when temporal intervention can still rewrite their fates.

KEYWORDS: *KNN, Machine Learning, COVID-19, Smote, Cross - Validation*

I. INTRODUCTION

The highly contagious nature of COVID-19, coupled with its ever-evolving characteristics and the absence of an effective vaccine or treatments, presents an urgent and widespread dilemma. In order to tackle this pressing issue, it becomes imperative to establish a framework that can assess and impede the relentless transmission of this disease, while concurrently providing intelligent healthcare interventions. This necessitates harnessing state-of-the-art technologies such as artificial intelligence, adaptive algorithms, deep learning, cognitive analysis, cloud-based solutions, mist computing, and peripheral computing.

The exponential surge of COVID-19 has compelled researchers to swiftly devise immediate measures aimed at curtailing its spread. Scientists have proposed and implemented a variety of innovative approaches to mitigate the adverse impacts of this global crisis and expedite the recovery process. Within the realm of statistics and computer science, time-series prediction algorithms utilize computational analyses based on gathered data. Scikit-Learn, a collaborative Python library, serves as a robust tool set for regression modelling, enabling the identification of trends within datasets encompassing both probable and resolved cases. Machine learning (ML) techniques can play a vital role in obtaining comprehensive insights into the pandemic dynamics and predicting the trajectory of infections.

II.METHODOLOGY

The subsequent segment elucidates the comprehensive elucidation of the dataset, along with the employed methodology. Given the presence of a pronounced class imbalance within the dataset, we utilized the Enhanced K-Nearest Neighbours (KNN) algorithm in conjunction with the synthetic minority oversampling technique known as EKSMOTE.

III. ENHANCED K-NEAREST NEIGHBOR METHODOLOGY

The present research endeavor focuses on elucidating the limitations of the prevailing system in order to surmount the challenges posed by the Decision Tree and Random Forest algorithms in the domain of Breast Cancer diagnosis. In an earnest pursuit of enhancing both precision and expeditiousness, this study puts forth an innovative approach known as Enhanced K-Nearest Neighbors (KNN). The rationale behind opting for the KNN algorithm resides in its resilience to noise-ridden data, as well as its efficacy in handling vast datasets, while concurrently exhibiting superior computational efficiency when compared to the Decision Tree and Random Forest methodologies.

Algorithm 1: Enhanced KNN

The breast cancer dataset has undergone meticulous preprocessing in order to meet the necessary prerequisites of the advanced machine learning technique known as Enhanced K-Nearest Neighbours (KNN). This preprocessing stage ensures that the dataset is refined and prepared to optimize the efficacy of the subsequent predictive model. The outcome of this rigorous data preprocessing is a sophisticated and refined Enhanced KNN predictive model. This model harnesses the power of the KNN algorithm, enhanced with additional features and optimizations to deliver accurate and reliable predictions regarding the nature of breast tumors—whether they are benign or malignant.

3.1 PROCEDURE

Step 1: Acquire a comprehensive dataset on Breast Cancer from the renowned Kaggle repository, renowned for its vast collection of high-quality data sources.

Step 2: Employ the sophisticated data preprocessing techniques of All Filter and Class Balancer to effectively cleanse and refine the acquired dataset. These methods ensure that the data is appropriately prepared for subsequent analysis, eliminating any potential biases or anomalies.

Step 3: Employ the advanced dimensionality reduction approach known as CfsSubsetEval in conjunction with the Attribute Evaluator framework. By leveraging the CfsSubsetEval technique, we can effectively select the most relevant and informative attributes, thereby reducing the overall dimensionality of the dataset. To enhance the search process, we employ the Greedy Stepwise Search method, which facilitates an efficient exploration of the attribute space, enabling us to identify the most influential features.

Step 4: Implement a robust k-fold cross-validation strategy, where the value of k ranges from 5 to 10. This approach ensures a comprehensive and unbiased assessment of the model's performance, as it rigorously tests the model's ability to generalize across various folds and subsets of the data.

Step 5: Employ the renowned Euclidean Distance metric to calculate the dissimilarity between each test data point and every row of the training data. By quantifying the distance, we gain valuable insights into the similarity or dissimilarity between the test and training instances, aiding in the subsequent evaluation process.

Step 6: Conduct a meticulous evaluation of the model's performance. This assessment involves a comprehensive analysis of various performance metrics, such as accuracy, precision, recall, and F1-score, among others. By meticulously scrutinizing these measures, we can effectively gauge the model's predictive capabilities and its ability to classify instances accurately.

IV. PERFORMANCE EVALUATION

The efficacy of the model was assessed by means of established evaluation metrics, including accuracy, precision, sensitivity, specificity, and F-score. Additionally, the classifiers were compared using the discerning measures of area under the curve and receiver operating characteristic (ROC). This procedure stands as a prominent examination to investigate the delicate balance between the detection of true-positive instances (sensitivity) and the occurrence of false-positive instances (specificity) within the diagnostic test.

4.1 Accuracy

In the realm of scientific inquiry, accuracy manifests as a profound metric delineated by the ratio of properly classified instances over the aggregate count of instances within a given dataset. We denote this measure as accuracy, and it is mathematically expressed as the summation of True Positives (TP) and True Negatives (TN) divided by the summation of True Positives, False Positives (FP), True Negatives, and False Negatives (FN):

Accuracy = (TP + TN) / (TP + FP + TN + FN)

True Positives signify instances that are correctly identified as positive, while False Positives denote instances that are incorrectly classified as positive. True Negatives, on the other hand, represent instances correctly recognized as negative, and False Negatives encompass instances falsely categorized as negative. By combining these quantities, we achieve a holistic assessment of accuracy, unveiling the proportion of accurately classified instances to the entirety of instances under scrutiny.

True Positive Rate (TPR), also known as sensitivity, represents the capability to accurately identify a high number of true positive instances. It signifies the proportion of true positives among the sum of true positives and false positives.

TPR = TP / (TP + FP)

Precision signifies the relationship between the count of correctly classified modules and the total count of modules classified as fault-prone. It measures the accuracy of identifying faulty units.

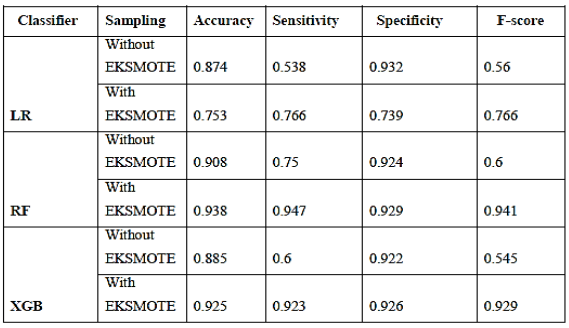
Precision = TP / (TP + FP)

In the context of imbalanced classification scenarios involving two classes, recall quantifies the number of true positives in relation to the total number of true positives and false negatives.

Recall = TP / (TP + FN)

The F-Measure serves as a composite metric that combines both precision and recall, providing a comprehensive evaluation score. It utilizes the harmonic mean of precision and recall to compute the final measure.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)



Classifier	Sampling	Accuracy	Sensitivity	Specificity	F-score
LR	Without EKSMOTE	0.874	0.538	0.932	0.56
	With EKSMOTE	0.753	0.766	0.739	0.766
RF	Without EKSMOTE	0.908	0.75	0.924	0.6
	With EKSMOTE	0.938	0.947	0.929	0.941
XGB	Without EKSMOTE	0.885	0.6	0.922	0.545
	With EKSMOTE	0.925	0.923	0.926	0.929

V. EXPERIMENTAL SETUP AND RESULTS

The Synthetic Minority Over-sampling Technique (SMOTE), originally developed by Chawla et al. [11], has emerged as a notable algorithm specifically tailored to mitigate the concerns stemming from imbalanced datasets within the realm of machine learning. This ingenious algorithm leverages the k-nearest neighbour (KNN) methodology to calculate the Euclidean distance between instances belonging to the minority class. Subsequently, it employs this information to generate fresh synthetic samples within the vicinity of the minority class.

To illustrate this process further, let us consider a hypothetical minority class denoted as "A," which is composed of "x" instances, expressed as A = {x1, x2, ..., xn}. When examining the k-nearest neighbors of x1, which can be denoted as {x6, x7, ..., xk}, we observe that A1 of x1 can be represented as {x7, x4, ..., xn}, where xk€A1 (with k ranging from 1 to N).

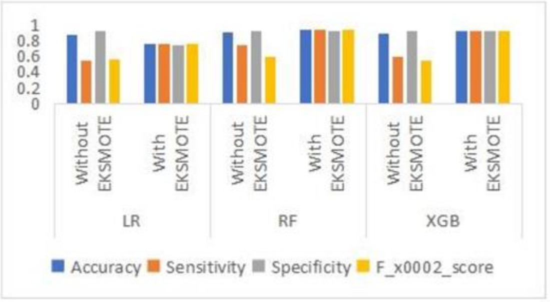


Figure 1: Performance comparison of classifiers using all features using original and EKSMOTE data

The implemented representations were developed utilizing the Python programming language within the Jupyter notebook environment (version 6.1.4), along with the sklearn library (version 0.23.2). To partition the data, a robust 10-fold cross-validation technique was employed. A series of experiments were conducted on both the original dataset and the dataset transformed with EKSMOTE. Multiple feature sets were generated using Extra tree classifiers combined with the feature importance technique.

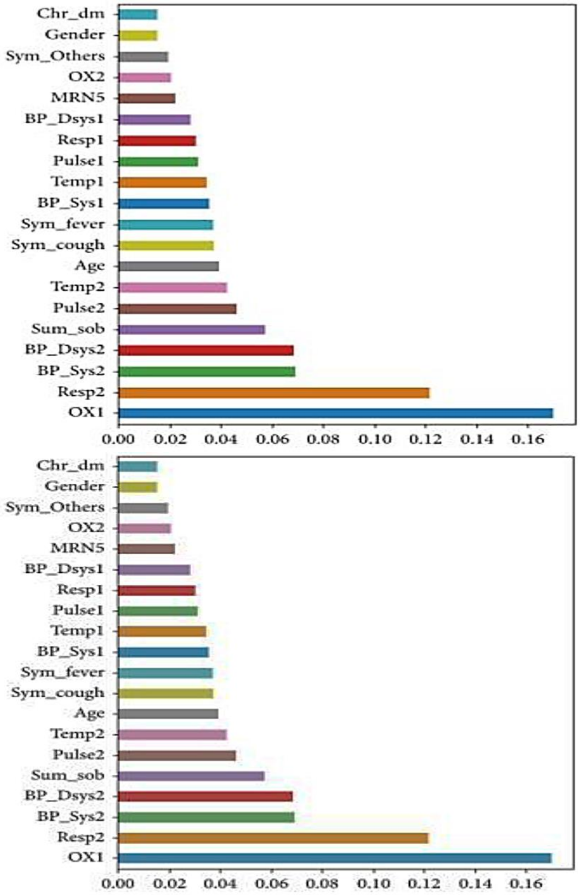


Figure 2: Correlation of top 20 features in the dataset

Within our own study, age emerged as one of the preeminent ten characteristics among the comprehensive set of 25 features employed in our predictive model. Furthermore, our study surpassed the performance of previously examined studies detailed in the literature review, boasting an exceptional accuracy of 0.952 and an outstanding AUC-ROC curve of 0.99.

VI.CONCLUSION

The objective was to discern their capacity to prognosticate the mortality and survival rates of individuals afflicted with COVID-19, employing the potent tools of supervised machine learning algorithms. The findings divulged that the random forest algorithm exhibited superior performance vis-à-vis its counterparts when subjected to rigorous evaluation using the 10-fold cross-validation technique. The grid search methodology was judiciously employed to optimize the parameters governing the algorithm's efficacy. This diligent study yielded remarkable outcomes, boasting an accuracy metric of 0.952 and an area under the curve (AUC) statistic of 0.99. Despite the auspicious nature of these achievements, it is imperative to acknowledge the necessity for further refinement and validation of these models, by employing diverse datasets for comprehensive scrutiny.

VII. REFERENCES

1. K. Yuki, M. Fujiogi, and S. Koutsogiannaki, “COVID-19 pathophysiology: a review,” Clinical Immunology, vol. 215, p. 108427, 2020.
2. T. P. Velavan and C. G. Meyer, “The COVID-19 epidemic,” Tropical Medicine & International Health, vol. 25, no. 3, pp. 278–280, 2020.
3. K. Dhama, S. Khan, R. Tiwari et al., “Coronavirus disease 2019-COVID-19,” Clinical Microbiology Reviews, vol. 33, no. 4, 2020.
4. L. C. Tindale, M. Coombe, J. E. Stockdale et al., “Transmission interval estimates suggest pre-symptomatic spread of COVID-19,” MedRxiv, 2020.
5. H. Han, Z. Xu, X. Cheng et al., “Descriptive, retrospective study of the clinical characteristics of asymptomatic COVID-19 patients,” mSphere, vol. 5, no. 5, 2020.
6. M. Chiara, D. S. Horner, C. Gissi, and G. Pesole, “Comparative genomics suggests limited variability and similar

evolutionary patterns between major clades of SARS-Cov-2," BioRxiv, 2020.

7. L. Yan, H.-T. Zhang, Y. Xiao et al., "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan," medRxiv, 2020.
8. K. C. Y. Wong and H.-C. So, "Uncovering clinical risk factors and prediction of severe COVID-19: a machine learning approach based on UK biobank data," medRxiv, 2020.
9. L. Sun, F. Song, N. Shi et al., "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19," Journal of Clinical Virology, vol. 128, p. 104431, 2020.
10. H. Yao, N. Zhang, R. Zhang et al., "Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests," Frontiers in Cell and Developmental Biology, vol. 8, pp. 1–10, 2020.
11. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

ID: 28

BIG DATA ANALYTICS IN HEALTH CARE

Dr.V.JANE VARAMANI SULEKHA
Assistant Professor,
Department of Information Technology,
Fatima College (Autonomous),
Madurai, Tamilnadu, India.

ABSTRACT

The application of big data in healthcare is a rapidly growing field. Big data refers to datasets that are not only large in size but also have a high level of variety and velocity, making it difficult to manage them with conventional tools and techniques. Medical data is also one of the fastest-growing data types because information can be collected from patients' own records or Electronic Health Records (EHRs). Due to the tremendous growth of such medical data, there is a need to provide appropriate tools and techniques in order to handle and extract information and knowledge from these datasets to improve the quality of patient's healthcare and cost of health care. The application of advanced analytics techniques on big data can be provided using big data analytics. This chapter presents an overview of big data, big data analytics, opportunities and challenges in health care.

KEYWORDS: *Big Data, Healthcare, Big Data Analytics, Data-driven healthcare.*

1. INTRODUCTION

Big Data has already drawn a remarkable attention for current researchers. More attention is given to the big data because data collection has become much cheaper now. Data is growing tremendously as it is generated by low-priced several information-sensors like mobile devices, wireless sensor networks, cameras, etc [1]. It has been reported that more than 2.5 Exabyte of data is being generated every day and technological per-capita capacity of the world to store information gets doubled after every forty months [2]. Every domain is becoming data rich and hence, are more excited to utilize the complex and massive data. The problem is that big data concepts should be understood well. There is no threshold value that defines the size of the big data. It is not only about the size but big data is an opportunity to get valuable insights from the massive available data [3]. Due to massive growth of data in last ten years, every business organization requires business insights. It helps them to improve their strategic as well as operational decisions.

Since data sources, types, and structures are not similar, integrating them into a shared data pool and processing is critical. Data generation in the healthcare industry is increasing day by day, where the data volume is higher. So, it is considered healthcare big data and needs to be analyzed. Unstructured or un-analyzed data cannot assure mining accuracy. Hence, it is necessary to create a tool or approach for big data analytics. Some reports calculated that 2.5 quintillion bytes of data are collected in a day. It is imperative to provide a big-data analytics method to manage volume, velocity, variety, veracity, and value. Most of the existing research work has been proposed using various data-analytical methods to enhance data processing, decision-making, and optimize the mining process. This chapter analyse handling big data analytics in the healthcare industry.

The medical industry with medical data or clinical data is increasing day by day. Because of the improved diagnostics systems it is easy to diagnose and treat various kinds of diseases effectively. Nowadays Cognitive technology is used in Health Care Industry. Cognitive technology combine individual medical information with larger-scale statistics and scientific data, these applications allow doctors to identify targeted treatment by immediately accessing all of the available information about similar cases. The method of cognitive technology in healthcare analyzes and understands all of the unstructured and structured information that is related to the patient's condition. It helps diagnose and identify the treatment from the available medical records. Various research methods, patient medical records, success stories of treatment methods, and literature are the potential resources for medical experts and doctors to do effective diagnosis and treatment only when it is easier to access the existing medical information. Simultaneously, manual methods can't scale the medical information of a new patient with acute diseases.

2. BIG DATA ANALYTICS

This section presents the concept of big data, data analytics and the business intelligence.

2.1. BIG DATA

"Big Data: it's not the data" [3]. In mid 1990s, the word "big data" is emanated in the lunch-table conversation at Silicon Graphics Inc. It became widespread in 2011 [4]. Over the years, the definitions of big data are evolving leading to the confusion in the mind of researchers. The numerous definitions of big data are: "Big data is high-velocity, high-volume and high-variety information resources that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [5].

"Big data is a term that defines huge volumes of high velocity, complex and variable data that need advanced methods and technologies to enable the capture, storage, distribution, management, and analysis of the information" [6]. Big Data cannot define be defined in single line as there are numerous explanations defining different characteristics of Big Data from 3Vs, 4Vs to 7Vs as shown in Figure 2.1 and are explained below [7]:

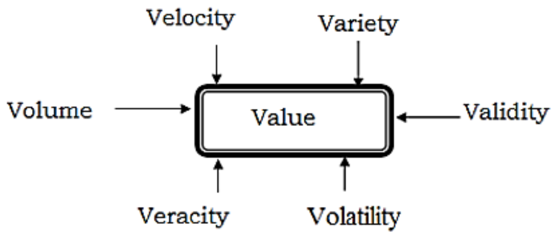


Figure 2.1: 7 Vs of Big Data

i. **Volume:** It defines the big size of data-set.

ii. **Velocity:** It defines the dynamic aspects of data i.e. the speed with which data in coming from different sources and going out after processing.

iii. **Variety:** It defines multimodal nature of big data i.e. the diverse nature or types of data set i.e. structured, unstructured and semi- structured. Example: images, text, chats, tables, posts, tweets, videos, etc. Although 3Vs defines the basic characteristics of big data, more dimensions are contributed by researchers in upcoming years to better understand and define the big data characteristics as explained below [8]:

iv. **Veracity:** It specifies the truthfulness of the data. It includes the certainty in the facts defined in the data as well as the meaningfulness of results. In the absence of veracity, incorrect inferences will be drawn.

v. **Validity:** It is similar concept to veracity but validity focuses more on accuracy and correctness of data with regards to its usage and requirements.

vi. **Volatility:** It defines that the specific data has some kind of retention period and removing such data after expiry is also needed in real time data storage. Example: An electrical business company destroys the history of a customer after 1 year warranty period is completed.

vii. **Value:** It defines the required outcome for big data processing. It indicates the worth of data for those who are consuming it.

Big Data = Data + Value

2.2. BIG DATA ANALYTICS (BDA)

Big data is a massive information asset that demands cost-effective, innovative forms of information processing techniques for enhanced insight and decision-making [9]. Big data was considered as an extremely profound problem in early 2000. As the data was skyrocketing, the organizations were facing scalability issues. Nowadays, business organizations are exploring big data as an opportunity to discover unknown facts about their data. It is about examining huge datasets to identify hidden patterns, real time insights, unknown correlations and hidden novel patterns [10]. BDA is an emerging area of research which offers systematic and advanced methods. Hence, BDA can be considered as a combination of big data and modern analytics. The rationale is to promote more profound business intelligence (BI) trends today [10]. There are numerous benefits of big data analytics for business organizations like drastic improvement in their operational efficiency, better support for customer services, identifying new and wide range of products in the competitive market, etc. as presented in the Figure 2.2 [11].

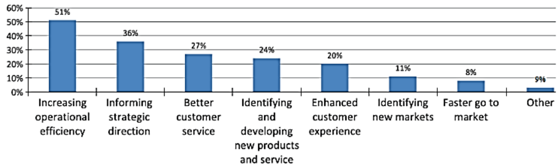


Figure 2.2: Big Data Analytics Opportunities

2.3. BUSINESS INTELLIGENCE

Business Intelligence is an umbrella term that can be considered as the combinations of all the required advanced skills, technology, techniques, methods, people and practices that are helpful for business organization in their decision making process. Business Intelligence converts raw data into useful knowledge that becomes the asset for every business organization. Nowadays, companies don't go for gut instinct for decision making. Highly competitive market demands spending huge amount of money on business intelligence (BI) tools and business analytics. As every business organization is data driven so the need to shift them to business analytics is highly desirable to survive in the today's market. Business organization collects useful information for interpreting it for better decision making which will ultimately optimize their business process. Traditional Business Intelligence used to focus on descriptive analytics by analyzing their historical data but now due to the competitive business environment, organizations are working more on predictive analytics which can help them to avoid making costly errors in future [12].

3. HEALTH CARE BIG DATA ANALYTICS

For the healthy life, everyone should more cognizant for their lifestyle; it can be achieved with awareness of healthcare, knowledge to keep safe health, referring different suggestions, proper diet plan etc. In the current digital era the health industry generates giant of electronic data per seconds, and it becomes very inconvenient to store, manage the data with conventional software's and technologies. Also the imminent data gets generated in the form of variable types like structured, unstructured and or semi structured manner. The speed of evolution of data and due to multiplicity; it must be overwhelmed about health care data. Hence the impression of big data can be used to handle the health related data. Today, it can be viewed that as a route for performing expansive scale and ease medicinal services, Big Data can be utilized to formulate proposal which can reduce cost and improve medicinal services quality. Then again, expanding gap between public health expenses and results is a standout amongst the most essential issues, and numerous endeavors to fill this gap are in progress in many nations. The gap between expenses of health related care and results was broken down to be the aftereffect of poor administration of bits of knowledge from research, very few utilization of existing knowledge and few skill in the field of healthcare, from those reasons many available chances were lost which could be useful for the betterment of patient Advantages of health related Big Data suggests us different solutions namely: Prevention of disease, identification of risk factors, mediations to improve ailment chance component, expenses for cure, decision making etc., Big data could be boon for the patients suffering with different disease as it could help to prevent the disease. Research performed on big data may provide patterns that could help to identify risk factors well in advance. Big data can also be useful to come up with treatment models that reduce the expenses drastically. Also, better decision making can be achieved through big data analysis. Big data provides wide opportunities in order to store and index previously unused, unstructured healthcare data for further use. Many sources available to collect big data [13] are shown in Figure 3.1.

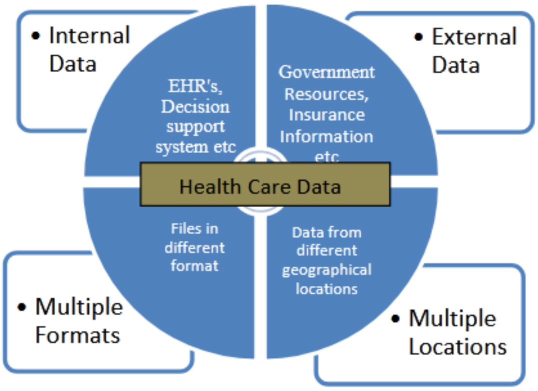

• Internal Data
EHR's, Decision support system etc
Government Resources, Insurance Information etc
• External Data
Health Care Data
Files in different format
Data from different geographical locations
• Multiple Formats
• Multiple Locations

Figure 3.1 2.2: Sources of big data

Internal data: electronic health records, clinical decision support systems, etc. External Data: Government sources, laboratories, pharmacies, insurance companies. Multiple formats: Files in different formats, Relational tables. Multiple locations: Geographic and in different Healthcare providers sites. Clinical Data:

Health data could be documents, reports, clinical and transcribed notes etc. Real concern of health data is that approximately 80percentof data is in unstructured form. Lab Reports: Different readings of lab reports can be helpful performing research on healthcare domain. Publications: Health research data and medical references material provides useful facts of health area. Clinical references: Practice guidelines available in text format and health products like drug information etc. Genomic Data: It represents gene sequencing data. Streamed Data: Data gathered from home monitoring, tele-health, wireless or smart devices which are new data sources. Web and Social Networking: Data taken from internet like search engines, social media sites etc. Business, Organization and External Data: Good amount of data can be received from administrative data of an organization [13]. Insurance: Data available with insurance companies can help making healthcare decisions. Consequent outcomes of big data analytics will help to improve clinicians' decisions, advance personalized healthcare and avoid unnecessary costs.

4. OPPORTUNITIES FOR BDA IN HEALTH CARE

Big data analytics signifies a new approach to analytics. The continuing digitization of health records together with the interoperable electronic health record (EHR), presents new opportunities to investigate a numerous of clinical and administrative questions. There is potential to layer BDA-type applications, in a privacy-protective manner, on top of the foundational health IT infrastructure to derive value that might not otherwise be found. What follows are some innovative ideas and solutions. BDA technologies used in Clinical decision support examine through huge amounts of data, recognize, categorize and learn from it, and then predict outcomes or suggest alternative treatments to doctors and patients. BDA is used in personalized care. They are used in Predictive data mining or analytic solutions that can influence personalized care (Example: genomic DNA sequence and gene expression for cancer care) in real time to highlight best practice treatments to patients. These solutions may suggest early detection and diagnosis before a patient develops disease symptoms. BDA solutions are also used in Public and population health that can mine web-based and social media data to predict flu outbreaks or other disease outbreaks based on user's search, social content and query activity. BDA solutions can also support doctors and epidemiologists to perform analysis on patient populations and care spots to help identify disease trends. BDA can give precise solutions for clinical operations without waiting for lengthier time to take fast decisions. BDA has supported the decision makers to integrate and analyze data related to key performance indicators on policy and financial aspects.

5. CHALLENGES FOR BDA IN HEALTH CARE

Big data analytics pose challenges and problems caused by exponential growth of medical data. The constant change of big data poses many challenges in analyzing, storing, and restoring large amounts of data. Traditional or standard database systems cannot be used to process, store and collect information due to its sheer volume [14]. Big clinical data contains large amounts of unstructured data such as handwritten data and natural language. Analyzing, integrating, and storing clinical big data presents significant challenges. It is not enough for authorities to exchange structured data, the exchange of unstructured data between organizations is even more complex. Efficiently searching large amounts of unstructured data is a major challenge. Big data has many characteristics. One is the diversity of data sources. The medical industry places particular demands on data processing speed, especially when patient conditions are deteriorating rapidly. Patients and their privacy and safety are impacted by the challenges and struggles associated with real-time applications such as cloud computing for data analysis. Recently, cloud computing has opened up new possibilities for mining and sharing medical big data. Several challenges must be overcome to make cloud computing more practical. [15]. First, cloud computing offers a simple and flexible way to reduce resources. However, it increases the risk of privacy leakage. It is a clinically evident fact in clinical informatics. The second is importing or exporting vast amounts of medical data to the cloud. Network bandwidth increases data costs and limits speed [16].

6. CONCLUSION

In recent years, big data has impacted

almost every industry. Medicine is no exception. In fact, the healthcare industry is the largest producer of digital data. Big data technology has brought new opportunities to healthcare. This has benefited not only patients, but medical facilities and hospitals as well. Adopting the right big data tools and underlying architecture should be carefully considered to maximize the benefits. Big data technology is having a major impact on modern healthcare, but it needs to be pushed further to reach its full potential. The traditional healthcare system still struggles to adjust to "big" change. Nevertheless, big data will put the healthcare industry on the right track of rapid transformation, which will undoubtedly bring amazing benefits to mankind.

REFERENCES

1. Segaran, T., & Hammerbacher, J. (2009). *Beautiful data: the stories behind elegant data solutions.* " O'Reilly Media, Inc.".
2. Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *science*, *332*(6025), 60-65.
3. Jagadish, H. V. (2015). Big data and science: Myths and reality. *Big Data Research*, *2*(2), 49-52.
4. Diebold, F. X., Cheng, X., Diebold, S., Foster, D., Halperin, M., Lohr, S., ... & Pospiech, M. (2012). A personal perspective on the origin (s) and development of 'big data': The phenomenon, the term, and the discipline. *Second Version.*
5. Gartner, I. T. Glossary (nd). Big Data.
6. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.
7. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.
8. MA-u-d Khan, M. F., & Uddin, N. G. (2014). Seven V's of big data understanding big data to extract value. *Proceedings of the 2014 Zone*, *1*, 1-5.
9. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.
10. Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1-34.
11. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347.
12. Rud, O. P. (2009). *Business intelligence success factors: tools for aligning your business in the global economy* (Vol. 18). John Wiley & Sons.
13. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*, 1-10.
14. Reddy, A. R., & Kumar, P. S. (2016, February). Predictive big data analytics in healthcare. In *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)* (pp. 623-626). IEEE.
15. Chen, J., Qian, F., Yan, W., & Shen, B. (2013). Translational biomedical informatics in the cloud: present and future. *BioMed research international*, *2013*.
16. Hong, L., Luo, M., Wang, R., Lu, P., Lu, W., & Lu, L. (2018). Big Data in Health Care: Applications and Challenges. Data and Information Management, 2 (3), 175–197.

ID: 29

BLOCKCHAIN TECHNOLOGY AND CRYPTOCURRENCIES

Mr.S.ANANTHARAJ
Department of Information Technology,
Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

Mr.T.KARTHIKEYAN
Department of Information Technology,
Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT:

A blockchain can be referred to as a collection of house, phone phone and rarity or it should be non-physical, for instance, offers of Associate in Nursing affiliation. It have to records or open record that gets shared amongst participating parties. Every transaction that gets incorporated is first verified by all the participants of that transaction. Once the data gets recorded by the blockchain, can never be rewritten or changed. Thus the blockchain can be termed as a record book of all the transactions held. Cryptocurrencies, the decentralized bitcoin or say ethereum which can be termed as peertopeer computerized cash also uses the blockchain technology. This paper includes history of bitcoin, a few literary reviews, working of the blockchain and its application.

KEYWORDS: Black Chain; CriptoCurrencies; Bitcoin; Bitcoin Technology.

I. INTRODUCTION

A blockchain is on a really basic level a scattered information of records or open record everything thought of or modernised occasions that are dead and shared among sharing parties. Every exchange the excellent network record is genuine by accord of Associate in nursing an excellent deal of the individuals within the structure. Likewise, once entered, data will ne'er be eradicated. The blockchain contains an explicit and clear record of every and each exchange whenever created. To utilize a foremost equivalence, it's not at all troublesome to require a treat from a treat thump, unbroken in an exceedingly confined place than taking the treat from a treat knock unbroken in an exceedingly business center, being seen by a monster range of people. Bitcoin is that the most recommended perspective that's remarkably related to blockchain progression. It's likewise the foremost off from being clearly true one since it empowers a multibillion-dollar normally market of unclear exchanges with no body management. On these lines it must administer distinctive body problems together with national governments and fund affiliations.

The benefits of Blockchain advancement trounce the executive problems and centred inconveniences. One key creating use event of blockchain headway consolidates "splendid contracts". Sharp contracts area unit primarily computer programs that may during this manner execute the terms of a comprehension. Sharp Property is another connected plan that is regarding dominant the requirement with relevance property or resource by methods for blockchain utilizing sensible Contracts. The property will be physical, for instance, auto, be compelled to be noted here that even Bitcoin is not usually a money - Bitcoin is tied in with dominant the commitment with reference to.

II. LITERATURE REVIEW

The writing survey performed preceding this examination offers a comprehension of the thought of blockchain innovation, and also the dispersion of productions within recognized points. The review demonstrates that the topics discussed further lacks in depth coverage; blockchain as administration innovation, savvy contracts, plans of action, enterprising probabilities and challenges, and blockchain as a universally helpful innovation. The creators afterwards observe the blockchain writing, for the foremost half, being of a discerning kind, wherever the potential probabilities of the innovation area unit usually secured, but the discussion on how blockchain will augment build an incentive within organizations is thus far deficient. Major focus is on what may occur if blockchain is received by the bulk, and basic potential utilize cases, while not going into esteem creating procedures of blockchain.

We are going to rather explore the thinking for utilizing blockchain innovation to require care of a problem and what esteem the innovation includes for the organizations utilising it. The enterprising issue of the blockchain is an improvement issue, closely resembling that in new advancement monetary aspects, requiring non-value coordination over the complementarity of uses and openings.

III. HISTORY OF BITCOIN

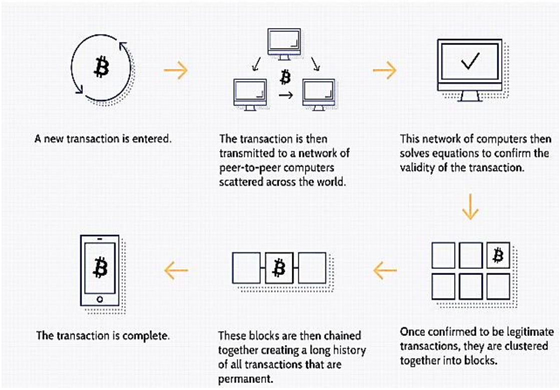
In year 2008, a private or event creating underneath the name out of Satoshi Nakamoto distributed a paper entitled "Bitcoin: A Peer-To-Peer Electronic money System". This paper delineate a disseminated version of the electronic money that will empower on-line parts to be sent significantly starting with one collection then onto the subsequent while not encountering a fund association. Bitcoin was the essential affirmation of this thought. Directly word processed financial standards is that the make certain is employed to depict all frameworks and mediums of exchange that usages cryptography to grapple trades as against those structures wherever the trades area unit redirected through a gathered sure in part.

IV. HOW A BLOCKCHAIN WORKS

The blockchain innovation has relevancy to any advanced resource exchange listed on the net. Internet business is completely fixing to the fund foundations filling in because the sure third party UN agency method and intervene any electronic exchange. The work of sure third party is to approve, defend and defend exchanges. a selected level of deceit is ineluctable in on-line exchanges which requirements intercession by cash connected exchanges. This outcomes in high exchange prices.

Bitcoin utilizes scientific discipline proof instead of the trust within the outsider for 2 willing partakers to execute a web exchange over the web. Each exchange is secured through a processed signature. Each exchange is distributed to "general society key" of the collector rigorously marked utilizing the "private key" of the sender. Keeping in mind the tip goal to burn through money, businessman of the digital cash has to demonstrate the responsibility for "private key". The part acceptive the advanced money confirms the processed signature – on these lines responsibility for "private key"- - on the exchange utilizing "the general population key" of the sender. Each exchange is communicated to every hub within the Bitcoin prepare and is then recorded in an open record when check.

The Bitcoin tackled this issue by a framework that's as of currently loosely referred to as Blockchain development. The Bitcoin system orders trades by putting them in social occasions known as squares and a brief time later interfacing these squares through what's known as Blockchain. The trades in an exceedingly solitary square square measure thought of to own happened meantime. These squares square measure related to one {another} (like a chain) in an authentic immediate, ordered demand with every square containing the hash of the past square.


A new transaction is entered.
The transaction is then transmitted to a network of peer-to-peer computers scattered across the world.
This network of computers then solves equations to confirm the validity of the transaction.
Once confirmed to be legitimate transactions, they are clustered together into blocks.
These blocks are then chained together creating a long history of all transactions that are permanent.
The transaction is complete.

There still remains one issue. Any center within the framework will accumulate unproven trades and build a square and at that time conveys it to remainder of the framework as a suggestion relating to that square ought to be the concomitant one within the blockchain. However will the framework choose that square ought to be next within the blockchain? There may be totally different | completely different} squares created by different center points meantime. One cannot depend upon the demand since squares will converge at totally different | completely different} solicitations at different concentrations within the framework.

V. CORPORATE FINANCING AND INTEREST THROUGH BITCOIN

Corporate finance into Bitcoin and Blockchain structure is making and creating excitement for a few of segments. Information system is sound blockchain development to create a safer, profitable system to trade stocks. DocuSign, associate degree association that invests large energy in electronic contracts, primarily discovered a joint plan with Visa to use blockchain to trace motorcar rentals and diminish written material. Microsoft can reveal bits of information regarding its enterprise into “shrewd gets” that use blockchain advancement. Within the in the meantime, this new obsession with blockchain development has accomplished some extent that associations area unit despite investigation distinctive roads with reference to creating a lot of diminutive, “private blockchains” within their own one in every of a sort operating environments.

VI. BLOCKCHAIN IN IOT

The IOT is dynamically obtaining the chance to be customary advancement in each the consumer and also the endeavor area. This specific essential has provoke makes an attempt towards localized IoT stages. The blockchain development energizes the execution of localized IoT stages, for example, moored and trustworthy in knowledge exchange and moreover record keeping. In such an overview, the blockchain fills in because the general record, keeping a trustworthy in record of the in depth range of messages changed between splendid devices in a much localized IoT topology.

IBM in relationship with Samsung has developed a part ADEPT that usages components of the bitcoin’s hid diagram to amass a flowed arrangement of contraptions a localized net of Things. Skilful uses 3 traditions BitTorrent (record sharing), Ethereum (sensible Contracts) and TeleHash (Peer-To-Peer Messaging)- within the stage.

VII. THE ACES AND CONS OF BITCOIN

ACES

- With a decentralized game plan of money, government or banks don’t have any associations with the cash. This can be helpful if a nation is in hardship or experiences a broad money related downturn (like the “Unique Recession” in the United States).
- Exchanges are normally assess absolved and modest
- Cash is definitely not hard to trade to zones the world over. As a matter of fact, it takes in every practical sense no time.
- Banks can’t use a man’s saved bitcoins for their own one of a kind hypotheses. Afresh, this suggests government related monetary torments won’t influence the estimation of a bitcoin.
- The square chain development is greatly powerful at removing the requirement for go betweens whose purpose behind existing is to platform the esteem based trust gap.

CONS:

- Bitcoin and other computerized monetary standards are exceedingly capricious. This suggests the estimation of a bitcoin can sway unquestionably–-and regularly there is no genuine method to foresee a change or clear up why one may have occurred.
- Since bitcoins are not settling to a fused establishment, government, or bank their expenses may rise and fall fundamentally.
- Clients may pick bitcoins to pay for unlawful items and endeavors (illegal substances, firearms, etc) by methods for the online dull web, as bitcoins can be harder to pursue.
- Bitcoins are starting at now saved in virtual, online wallets. While it would take the capacity and inclination of a fit software engineer to get to these virtual wallets, it has a tendency to be done, and hacking has happened already.
- Numerous clients encounter genuine troubles understanding bitcoin or its convoluted square chain.

VIII. UTILIZATION OF BLOCKCHAIN BEYOND CRYPTOCURRENCY

Bitcoin is simply a wonderful usage of the Blockchain. Blockchain is believed to be a completely unique miracle within the area of enrolling sanctionative unfathomable applications, for example, securing and checking definitive reports together with deeds and distinctive validations, therapeutic administrations information, IoT, Cloud so on. Tapscott BEFITTINGLY indicated Blockchain to be the “General

Ledger", partaking numerous new applications past checking trades, for example, in: wise deeds, suburbanized and additionally self-administering affiliations/citizen driven associations et cetera.

In the cloud condition, the chronicled background of arrangement of any cloud information challenge and its ensuing assignments performed quickly square measure recorded by the information structure a part of 'Data Provenance', or, in different words of cloud information. Henceforward this is often basic to allow the foremost outrageous security to {the information| the info| the information} birthplace for making certain its data insurance, sociology and obligation. Liang propels a Blockchain based mostly sure in cloud information birthplace define, 'ProvChain', or, in different words. Such appointment of the Blockchain in an exceedingly cloud circumstance will provide sturdy protection against records being modified afterward partaking a redesigned straightforwardness and moreover further information obligation. This furthermore grows the provision, steadfastness, assurance and at last the estimation of the birthplace information itself.

IX. CONCLUSION

To close, Blockchain is the development spine of Bitcoin. The passed on record value joined with security of BlockChain, makes it to a great degree charming advancement to understand the current Financial and furthermore non-cash related business issues. To the degree the advancement cares, the computerised money based mostly technical school is either within the slippy inclination of vainglorious desires or in trough of disappointment. The efforts laid on making blockchain even more advanced has allowed us to use it for trades. Properties that shields its security, assurance, traceability, trademark knowledge birthplace and timestamping has seen its assignation past its basic application zones.

The Blockchain itself and its varieties square measure by and by wont to grapple any reasonably trades, paying very little relation to whether or not or not it's human-to-human correspondences or machine-to-machine. Its gathering emits an effect of being secure particularly with the general ascent of the Internet-of-Things. The Blockchain has been particularly appeared to be correct in creating nations wherever making certain trust is of a vital concern.

X. REFERENCES

1. M. Marchesi, "Why blockchain is important for programming designers and why programming building is crucial for blockchain programming (Keynote)", *2018 International Workshop on Blockchain orientating software system Engineering (IWBOSE)*, pp. 1-1, 2018.
2. T. N. Dinh and M. T. Thai, AI and Blockchain: A turbulent Integration, vol. 51, no. 9, pp. 48-53, Gregorian calendar month 2018.
3. L. Kan, Y. Wei, A. Hafiz Muhammad, W. Siyuan, G. Linchao and H. Kai, "A Multiple Blockchains design on Inter-Blockchain Communication", *2018 IEEE International Conference on software system Quality responsibility and Security Companion (QRS-C)*, pp. 139-145, 2018.
4. D. Mill operator, "Blockchain and therefore the net of Things within the Industrial Sector", *IT skilled*, vol. 20, no. 3, pp. 15-18, May. 2018.
5. R. Henry and A. Kate, "Blockchain Access Privacy: Challenges and Directions", *IEEE Security and Privacy*, vol. 16, no. 4, pp. 38-45, July 2018.
6. N. Kshetri and J. Voas, "Blockchain in Developing Countries", *IT skilled*, vol. 20, no. 2, pp. 11-14, Mar. 2018.
7. T. Aste, P. Tasca and T. Di Matteo, Blockchain Technologies: The predictable Impact on Society and trade, vol. 50, no. 9, pp. 18-28, 2017.
8. J. Fiaidhi, S. Mahomet and S. Mohammed, "EDI with Blockchain as associate Enabler for Extreme Automation", *IT skilled*, vol. 20, no. 4, pp. 66-72, Jul. 2018.
9. V. Gatteschi, F. Lamberti, C. Demartini, C. Pranteda and V. Santamaría, "To Blockchain or to not Blockchain: that's the Question", *IT skilled*, vol. 20, no. 2, pp. 62-74, Mar./Apr. 2018.
10. T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi and J. Wang, Unraveling Blockchain: a knowledge process read of Blockchain Systems, IEEE.
11. Transactions on information and information Engineering, vol. 30, no. 7, pp. 1366-1385, July 2018.
12. N. Kshetri, "Can Blockchain Strengthen the net of Things?", *IT skilled*, vol. 19, no. 4, pp. 68-72, 2017.
13. K. R. 'Özyilmaz and A. Yurdakul, "Work-

in-advance: incorporating low-control IoT gadgets to a blockchain-based foundation", *2017 International Conference on Embedded software system (EMSOFT)*, pp. 1-2, 2017.

14. Y. Gupta, R. Shorey, D. Kulkarni and J. Tew, "The connectedness of blockchain within the net of Things", *2018 tenth International Conference on Communication Systems and Networks (COMSNETS)*, pp. 561-564, 2018.
15. R. Han, V. Gramoli and X. Xu, "Assessing Blockchains for IoT", *2018 ninth IFIP International Conference on New Technologies quality and Security (NTMS)*, pp. 1-5, 2018.
16. J. Ellul and G. J. Pace, "AlkylVM: A Virtual Machine for good Contract Blockchain Connected net of Things", *2018 ninth IFIP International Conference on New Technologies quality and Security (NTMS)*, pp. 1-4, 2018.

ID: 30

AN OVERVIEW OF GENE EXPRESSION USING MICROARRAY

Ms.P.NANCY VINCENTINA MARY
Assistant Professor,
PG Department of Computer Applications,
Fatima College, Madurai, Tamil Nadu, India. &
Research Scholar,
Department of Computer and Information Science, Faculty of Science,
Annamalai University, Chidambaram, Tamil Nadu, India.

Dr.R.NAGARAJAN
Assistant Professor,
Department of Computer and Information Science,
Faculty of Science,
Annamalai University, Chidambaram, Tamil Nadu, India.

ABSTRACT

Analysing and interpreting gene expression data is an important task in bioinformatics. Microarray technology is one of the most important tools that can be used to monitor the expression levels of thousands of genes in a particular organism. The insights gained through microarray data analysis are becoming increasingly important as they help to classify disease phenotypes. And these microarray gene expression datasets are used for various diagnostic purposes. Gene expression profiles have recently been used to classify cancers. Cancer classification based on gene expression data is an advance in DNA microarray technology and genomic sequencing. Extensive research has been done to develop better solutions to improve the prediction accuracy of cancer gene prediction. This paper explains about gene expression micro array, the importance of gene expression analysis, the importance of microarray technologies in cancer and then list few methods used in gene expression for classification purpose.

KEYWORDS: *Gene Expression, Micro Array, Classification, ML, Machine Learning*

I. INTRODUCTION

The human genome is also called as the "blueprint of life." This can be is realized through the process of gene expression. Recent advances have shown that many stages on the path from gene sequences to active proteins are interrelated and suggesting a unified theory of gene expression.

A. *Gene*

Genes are considered to be the basic physical and functional units of heredity and these Genes are made up of DNA. Some genes act as instructions for making molecules called proteins. Though, many genes do not encode proteins. In humans, genes vary in size from hundreds of DNA bases to over 2 million bases. An international research project called the Human Genome Project, which worked to sequence the human genome and identify the genes it contains, estimated that humans carry between 20,000 and 25,000 genes [1].

Each person is born with two copies of each gene, one from each parent. Most genes are the same for all, but a few genes (less than 1% of the total) vary slightly from person to person. Variant forms of a gene (Alleles) are the same gene morphology, with slight differences in the sequence of DNA bases. These slight differences play a vital role in each person's unique physical characteristics.

Scientists track genes by giving them unique names. Genes are also associated with symbols because gene names can be long. Here, a symbol is a short combination of letters (and in some cases numbers) that represents an abbreviation for a gene name. For example, the gene on chromosome 7 that is associated with cystic fibrosis is called the transmembrane conductance regulator of cystic fibrosis. Its symbol is CFTR. Figure 2 describes Chromosome and Gene.

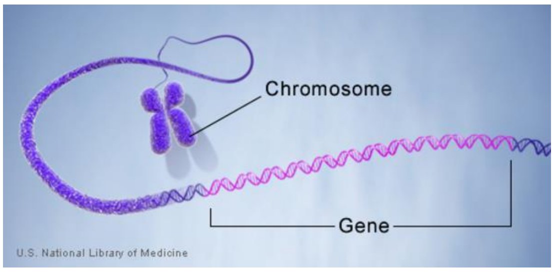


Fig. 1 Chromosome and Gene
Credit: U.S. National Library of Medicine

B. HOW PROTEIN PRODUCTION IS DIRECTED BY GENES?

Most genes contain the information needed to make functional molecules called proteins (Some genes produce regulatory molecules that help cells assemble proteins). The gene-to-protein pathway is complex and tightly regulated within each cell. It consists of two main steps, they are transcription and translation. Transcription and translation are together called gene expression.

During the transcription process, the information stored in the DNA of a gene is passed to a similar molecule called RNA (ribonucleic acid) in the cell nucleus. Although both DNA and RNA are made up of chains of building blocks called nucleotides, but their chemistries are slightly different. The type of RNA that carries the information to make a protein is called messenger RNA (mRNA) because it carries information and messages from the DNA to the cytoplasm. In the second step, the translation from gene to protein, takes place in the cytoplasm. In this second step, mRNA communicates with the ribosome that "decodes" or “reads” the sequence of mRNA nucleotides.

Each sequence of three nucleotides, called a codon, usually encodes a particular amino acid (Amino acids are constituents of proteins). A type of RNA called transfer RNA (tRNA) assembles the amino acids of a protein into amino acids. The assembling of Protein continues until the ribosome encounters a "stop" codon (A codon is a RNA or DNA sequence of three nucleotide sequence that does not encode an amino acid).

The flow of information from DNA to proteins via RNA is one of the basic principles of molecular biology. This step is very important and it is called as "central dogma".

C. GENE EXPRESSION

Gene expression is the process by which the information coded in a gene is used to direct the assembly of a protein molecule. The cell identifies the gene's sequence in three-base groups. A codon is a DNA or RNA sequence of three nucleotides relates to one of the 20 amino acids that create the protein. In other words, Gene Expression is the phenotypic manifestation (physical manifestation of genotype) of a gene or genes as a result processes of genetic transcription and genetic translation.

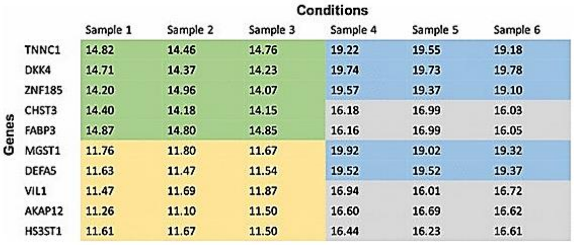
When a gene is expressed, the genetic information (base sequence) of the DNA is first copied to the mRNA molecule (transcription). The mRNA molecule then exits the nucleus and enters the cytoplasm. In the cytoplasm, the mRNA take part in protein synthesis by delegating the specific amino acids that make up individual proteins (translation).

D. GENE EXPRESSION MATRIX (MICROARRAY)

Microarrays are one of the latest breakthroughs in experimental molecular biology, enabling parallel monitoring of gene expression in tens of thousands of genes and already generating vast amounts of valuable data. Processing and analysing these data is becoming one of the main bottlenecks in using this technology. Raw microarray data is an image that needs to be converted into a gene expression matrix table. In gene expression, rows denote genes, columns denote different samples such as tissues and experimental conditions, and numbers in each cell indicate the expression level of a particular gene in a particular sample (or vice versa). We can also say that, Gene Expression matrix has hundreds of rows and thousands of columns, with each row and column representing a sample and gene. A gene expression level is associated to the generated protein value. Gene expression serves as a criterion for assessing gene activity in a certain biochemical environment. Gene expression is a dynamic process that can be either temporary or continuous. As a result, it can resound transient and irreversible changes in the biologic state of cells, tissues, and organisms. Microarray technology is very beneficial for analysing the pattern and degree of gene expression in many types of cells or tissues.

Gene expression microarrays give a clear description of all the transcriptional activity (the process by which the information in a part of DNA is copied into a new molecule of (messenger RNA) mRNA in a biological sample. Microarray technology is different from the traditional molecular biology tools, which usually allow the study of a single gene or a small set of genes. Microarrays are used in the discovery of totally novel and unexpected functional roles of genes. Microarray technologies have been used for a variety of purposes, including discovering new disease subtypes, measuring gene expression, detecting specific DNA sequences, developing new diagnostic tools, and identifying underlying mechanisms of disease or drug response. Though, this technology necessarily produces a large amount of data, challenging us to interpret it by exploiting modern computational and statistical tools.

Microarray is used to discover the gene expressions in prostate cancer, oral cancer, breast cancer, ovarian cancer, and other types of cancers. Figure 2 describes sample Micro array data.



	Conditions					
Genes	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
TNNC1	14.82	14.46	14.76	19.22	19.55	19.18
DKK4	14.71	14.37	14.23	19.74	19.73	19.78
ZNF185	14.20	14.96	14.07	19.57	19.37	19.10
CHST3	14.40	14.18	14.15	16.18	16.99	16.03
FABP3	14.87	14.80	14.85	16.16	16.99	16.05
MGST1	11.76	11.80	11.67	19.92	19.02	19.32
DEFA5	11.63	11.47	11.54	19.52	19.52	19.37
VIL1	11.47	11.69	11.87	16.94	16.01	16.72
AKAP12	11.26	11.10	11.50	16.60	16.69	16.62
HS3ST1	11.61	11.67	11.50	16.44	16.23	16.61

Fig. 2 SAMPLE MICRO ARRAY DATA

I. IMPORTANCE OF MICROARRAY ANALYSIS

Cancer, also known as malignant neoplasm, is a complex disease that develops when some cells exhibit particular characteristics of unrestrained proliferation and then spreads to other parts of the body (American Cancer Society, 2017). According to the WHO Fact Sheet, over 14 million new cases of cancer were diagnosed in 2012, making cancer one of the leading causes of illness and mortality worldwide (WHO, 2018). In the next two decades, new cases are expected to increase by around 70%. Cancer causes nearly one out of every six deaths worldwide. In 2015, 8.8 million people died from cancer, making it the world's second leading cause of death (WHO, 2018).

However, if cancer is detected and treated early, the mortality rate can be reduced. Medical data mining is a subset of data mining that assists in the analysis, extraction, transformation, interpretation, and visualization of medical records stored in repositories. Medical data mining is both interesting and challenging, because disease diagnosis and prediction are not only issues of accuracy, but also a matter of life and death. A misclassification or forecast can have devastating consequences for patients and their families. Be a result, medical data mining is frequently referred to as an expert system that uses machine learning to make decisions that assist specialists in diagnosing and forecasting diseases more readily and rapidly [2].

Microarray data is commonly used in cancer research, where an early detection of cancer is critical in defining the type of treatment and survivability [3]. Microarray Technology (MT) enables researchers to examine the activity of tens of thousands of genes in a single experiment, providing crucial information on the cell's function. This exact information can be used to diagnose a variety of disorders, including Alzheimer's disease [4], diabetes, and cancer [5]. This method may be used to generate gene expression data, which is extremely useful for cancer classification and prediction. Researchers have recently used computational intelligence algorithms to look at the most informative genes that help with cancer detection [6].

In clinical research, gene expression profiling is transforming the way biomarkers are discovered. Rather than relying on traditional clinical prognostic assessment, gene expression profiling of disease advises relying on unique genomic 'signatures' (groups of genes that can identify disease samples from healthy samples) with prognostic and predictive consequences in clinics.

The categorization of tissue samples into normal or malignant tissues is one of the most important applications of microarray data. Microarrays provide real-time data for thousands of genes, and a single experiment can readily generate many results. Microarray provides quick and simple results. It has the potential to find cures for diseases and cancer. Gene expression can be studied using various regions of DNA.

II. MICROARRAY TECHNOLOGIES IN HUMAN CANCER

Microarray can be a boon to researchers as it provides a platform for simultaneous testing of a large set of genetic samples. It helps particularly in the identification of single nucleotide polymorphisms (SNPs) and mutations, identification of target genes of tumor suppressors, classification of tumors, and identification of genes associated with chemo resistance, identification of cancer biomarkers, and drug discovery. For example, we can analyse different patterns of gene expression levels between a group of cancer patients and a group of normal patients and identify the gene associated with that particular cancer.

The microarray analysis is a powerful molecular technology that allows the simultaneous study of the expression of thousands of genes or their RNA products. Microarray analysis provides an accurate picture of gene expression in the cell or the sample at the time of the study. For example, the expression of all the genes for drug resistance and metabolism or all the known oncogenes in a cell can be recognized and evaluated in the same timeframe [7-9]. The microarray can be defined as an ordered collection of micro spots (the probes), the genes of interest are represented by each spot, which contains a single species of nucleic acid. This Method is based on hybridization between labelled free targets derived from a biological sample and an array of many DNA probes that are immobilized on a matrix [10]. The targets are created by reverse transcription and the concurrent labelling of RNA extracts from a biological sample hybridized with DNA fragment probes. The hybridization sign produced on each probe is the mRNA expression level of the corresponding gene in the sample at the time of the examination. The signals are quantified, detected, integrated and normalized with efficient software and reflect the 'gene expression profile' or 'molecular portrait' for each biological sample.

On a silicon or glass slide or a nylon solid state base, many thousands or tens of thousands of distinct spots can be printed. There are mainly two variants of microarrays: cDNA and oligonucleotide microarrays [11, 12].

Although both types of microarray are used to analyse gene expression patterns, these variants are fundamentally different [13].

In cDNA microarrays, relatively long DNA molecules are fixed on a solid surface. This type of microarray is frequently used for expression studies and large-scale screening. The oligonucleotide microarray is fabricated by in situ light directed chemical synthesis or by conventional synthesis followed by immobilization on a glass matrix. This microarray is used for recognition of gene mapping, mutations and expression studies and allows for the differential detection of gene family members or alternative transcripts that are not distinct by cDNA microarrays.

The chemistry of the microarray in itself is not new, since hybridization technology has been well established for decades. However, the simultaneous study of thousands of genes transforms the microarray technique into a powerful whole system analytical tool. This technology is still improving and advancing.

A. *THE IMPORTANCE OF MEASURING GLOBAL GENE EXPRESSION IN HUMAN CANCERS*

The characterization of the transcribed gene population led to the creation of a new term, the transcriptome [14]. This concept defines the complete set of transcribed genes that are expressed as messenger RNA of a particular species. As a result, the transcript represents the world of RNA messengers that can encode proteins. Only about 5% of the genes are active in any cell at any given time.

Most genes are suppressed and this regulation can be done either at the transcriptional level or at the translational level. Most regulation is done at this level, as regulation of protein expression is more efficient at the transcriptional level. A cell's gene expression profile determines its function, phenotype, and response to external stimuli. Therefore, gene expression profiles help interpret cell function, biochemical pathways, and regulatory mechanisms. In addition, the gene expression profile of diseased cells / tissues compared to normal controls can enable understanding of the pathology of the disease, identify new therapeutic intervention points, improve diagnosis, and reveal prognosis.

Several techniques for gene expression profiling have emerged in recent years and have been used successfully in cancer research. These include differential display, continuous analysis of gene expression, and microarrays [15].

Microarrays are important because they are easy to use, do not require large DNA sequences, and can quantify thousands of genes from multiple samples in parallel. Cancer gene expression profiling is the largest category of research using microarray technology and appears to be the most comprehensive approach to molecular property assessment of cancer. Although the phenotype of cancer is only partially determined by its transcriptome, it still provides a clear picture of the physiological state of the cell. The power of this approach has been demonstrated in studies conducted on a variety of malignancies, including breast, head and neck, liver, lung, ovarian, pancreatic, prostate, and gastric cancers.

Several cancer profiling studies using microarray analysis have used different methodologies, such as tumour versus control, in which the tumour gene expression profile is compared to its equivalent control sample in order to measure the differences and similarities between the two phenotypes. In cancer stratification, the gene expression profiles from different samples of the same cancer type are compared to reveal distinct subgroups to better define molecular classification of a common histological type of cancer. In temporal evaluation of the tumor, the gene expression patterns from tumor samples derived from different stages of progression are compared to elucidate the differences between the early and advanced stages of the disease.

III. RELATED WORK

A. *AN INTELLIGENT ARTIFICIAL BEE COLONY AND ENHANCED MONARCHY BUTTERFLY OPTIMIZATION TECHNIQUE (IABC-EMBOT)*

In this method, IABC-EMBOT (An Intelligent Artificial Bee Colony and Enhanced Monarchy Butterfly Optimization Technique) scheme [16] is developed as an automatic MBO and ABC based hidden node size optimization approach of the incorporated MLP (Multilayer perceptron) in the process of the breast cancer diagnosis.

Advantage: Improved Accuracy

Future Work: Authors planned to develop an integrated ABC- BFA (Bacterial Foraging Algorithm) that introduces swarming operator in the onlooker and employee bee phase of ABC for improving the degree of main exploitation and study that results in improved classification accuracy with increased speed and precision during the process of intelligent breast cancer detection.

B. *ENHANCED FIREFLY ALGORITHM (EFA) BASED GENE SELECTION AND ADAPTIVE NEURO NEUTROSOPHIC INFERENCE SYSTEM (ANNIS) PREDICTION MODEL FOR DETECTION OF CIRCULATING TUMOR CELLS (CTCS)*

The enhanced firefly algorithm (EFA) and the adaptive neuro neutrosophic inference system (ANNIS) [17], methods use chosen attributes for Circulating Tumor Cell (CTC) detection.

Advantage: Better classification result.

Future Work: Planned to include more gene samples.

C. *SEQUENCE ALIGNMENT AND BACK PROPAGATION NEURAL NETWORK (BPNN)*

Sequence alignment and back propagation neural network (BPNN) method [18] use two stages; first whether the affected person (patient) has mutations or not by used sequence alignment for two DNA sequences in BRCA1 and BRCA2 genes (mutant and non- mutant sequences), the second stage classify these mutations that causes the disease by using BPNN algorithm to training and testing with 5-fold cross validation.

Advantage: Better classification Result.

Future Work: Sequence alignment algorithm can be implementation in both global alignment (Needleman-Wunsch algorithm) and local alignment (Smith-Waterman algorithm) by using dynamic sequence alignment.

D. *CONVOLUTIONAL NEURAL NETWORK MODEL (CNN)*

Three different Convolutional Neural Network (CNN) models [19] were presented and that take unstructured gene expression data as input to classify tumor and non-tumor samples into their designated cancer types or as normal.

Advantage: Accurate and simultaneous cancer/normal and cancer type's prediction based on gene expression profiles.

Future Work: Explanation of cancer markers for earlier cancer detection

E. *FEED FORWARD NEURAL NETWORK (FFNN)*

A Feed Forward Neural Network approach [20] was developed, that addressed a challenging multipanel biomarker development problem in the early detection of breast cancer. The method combined the three way data split with an optimization model of FFNN.

Advantage: Better solution for multi-biomarker panel discovery.
Future Work: Biological experiments will be done to validate these biomarkers

F. *PROBABILISTIC NEURAL NETWORK (PNN)*

In this method, the effects of dimensionality reduction techniques with probabilistic neural network (PNN) [21] on breast cancer classification are examined. A hybrid method is proposed using the independent component analysis (ICA) and the discrete wavelet transform (DWT) to reduce feature vectors of Wisconsin diagnostic breast cancer (WDBC) data set.
Advantage: This method increases the performance of breast cancer classification as benign and malignant compared to original data set, as well as reducing the computational complexity.

G. *DEEP BELIEF NETWORKS WITH SUPPORT VECTOR MACHINES*

The classification performance of deep belief networks (DBNs) in dependence on its multiple model parameters and in comparison with support vector machines (SVMs) [22] was addressed in this paper. Furthermore, combined classifiers that integrate DBNs with SVMs are also addressed. Such a classifier utilizes a DBN as representation learner forming the input for a SVM. Overall, the results provide guidelines for the complex usage of DBN for classifying gene expression data from complex diseases.

Advantage: Combined classifiers that integrate DBNs with SVM perform well.
Future Work: Other DL classifiers with DBN may lead to different results.

H. *EFFICIENT NEURAL NETWORK*

This paper proposes an automated technique using artificial neural networks [23] as decision making tools in the field of breast cancer. The features extracted from biopsy slide images are converted into matrix and used to train the neural network. Both supervised and unsupervised methods of neural networks are tested to develop the most efficient alternative for breast cancer diagnosis.

Advantage: The proposed work can be used as a decision support system to assist clinicians in cancer detection.
Future Work: instead of extracted features, full dataset can be used.

I. *GENETIC BEE COLONY ALGORITHM*

In this paper, a new hybrid gene selection method is proposed, namely Genetic Bee Colony (GBC) algorithm [24]. The proposed algorithm combines the use of a Genetic Algorithm (GA) along with Artificial Bee Colony (ABC) algorithm. The goal is to integrate the advantages of both algorithms.

Advantage: The proposed work combines the use of a Genetic Algorithm (GA) along with Artificial Bee Colony (ABC) algorithm. It was successfully utilized in high dimensional datasets.
Future Work: Combination of other Evolutionary algorithm may be used.

IV. CONCLUSIONS

A considerable portion of the population is affected by cancer. If cancer is not diagnosed on time, it becomes difficult for the doctor to save the patient's life. In the literature, there are numerous methods available for predicting cancer. However, gene expression data is currently receiving attention. As there is big data, various Machine learning and deep learning methods are used to predict the cancerous and noncancerous genes. Each technique has its pros and cons. When working with large amounts of data, machine learning techniques will overfit the training dataset. According to the findings, gene expression data provides some additional details that assist in the classification and diagnosis of cancer. As a result, it indicates that gene expression data can be used to enhance results. Machine learning algorithms are utilised when the dataset is small. When big data is involved, various deep learning methods can be used to predict the cancerous and noncancerous genes. Therefore, deep learning methods can be employed in the gene expression analysis.

V. REFERENCES

1. https://medlineplus.gov/
2. Doreswamy, H., and U. M. Salma. "A binary bat inspired algorithm for the classification of breast cancer data." *Int J Soft Comput Artifi Intell Appl (IJSCAI)* 5.2/3 (2016): 1-21.
3. Selvaraj, Saravanakumar, and Jeyakumar Natarajan. "Microarray data analysis and mining tools." *Bioinformation* 6.3 (2011): 95.

4. Panigrahi, Priya P., and Tiratha Raj Singh. "Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: revealed association with aging and other diseases." *Journal of theoretical biology* 334 (2013): 109-121.
5. Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." *Advances in bioinformatics* 2015 (2015).
6. Alomari, Osama Ahmad, et al. "MRMR BA: a hybrid gene selection algorithm for cancer classification." *J Theor Appl Inf Technol* 95.12 (2017): 2610-2618.
7. Brown, Patrick O., and David Botstein. "Exploring the new world of the genome with DNA microarrays." *Nature genetics* 21.1 (1999): 33-37.
8. Collins, Francis S. "Microarrays and macroconsequences." *Nature genetics* 21.2 (1999).
9. Lander, Eric S. "Array of hope." *Nature genetics* 21.1 (1999): 3-4.
10. Southern, Edwin, Kalim Mir, and Mikhail Shchepinov. "Molecular interactions on microarrays." *Nature genetics* 21.1 (1999): 5-9.
11. Schena, Mark, et al. "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes." *Proceedings of the National Academy of Sciences* 93.20 (1996): 10614-10619.
12. Lockhart, David J., et al. "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nature biotechnology* 14.13 (1996): 1675-1680.
13. Lipshutz, Robert J., et al. "High density synthetic oligonucleotide arrays." *Nature genetics* 21.1 (1999): 20-24.
14. Su, Andrew I., et al. "Large-scale analysis of the human and mouse transcriptomes." *Proceedings of the National Academy of Sciences* 99.7 (2002): 4465-4470.
15. Cheng, Qing, et al. "Identification and characterization of genes involved in the carcinogenesis of human squamous cell cervical carcinoma." *International journal of cancer* 98.3 (2002): 419-426.
16. Punitha S, Amuthan A, Joseph KS. Enhanced Monarchy Butterfly Optimization Technique for effective breast cancer diagnosis. J Med Syst. 2019 May 29;43(7):206. doi: 10.1007/s10916-019-1348-8.
17. Umamaheswari, T. S., and P. Sumathi. "Enhanced firefly algorithm (EFA) based gene selection and adaptive neuro neutrosophic inference system (ANNIS) prediction model for detection of circulating tumor cells (CTCs) in breast cancer analysis." *Cluster Computing* 22.6 (2019): 14035-14047.
18. RAHEEM, ALI K. ABDUL, and BAN NADEEM DHANNOON. "Predication and classification of cancer using sequence alignment and back propagation algorithms in brca1 and brca2 genes." *International Journal of Pharmaceutical Research* 11.1 (2019).
19. Mostavi, Milad, et al. "Convolutional neural network models for cancer type prediction based on gene expression." *BMC medical genomics* 13.5 (2020): 1-13.
20. Zhang, Fan, et al. "A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer." *BMC proceedings*. Vol. 7. No. 7. BioMed Central, 2013.
21. Mert, Ahmet, Niyazi Kılıç, and Aydın Akan. "An improved hybrid feature reduction for increased breast cancer diagnostic performance." *Biomedical Engineering Letters* 4.3 (2014): 285-291.
22. Smolander, Johannes, Matthias Dehmer, and Frank Emmert-Streib. "Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders." *FEBS Open Bio* 9.7 (2019): 1232-1248.
23. Singh, Seema, et al. "An efficient neural network based system for diagnosis of breast cancer." *Breast cancer* 8.10 (2014): 12.
24. Alshamlan, Hala M., Ghada H. Badr, and Yousef A. Alohali. "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification." *Computational biology and chemistry* 56 (2015): 49-60.

ID: 31

COMPARATIVE STUDY OF COMPRESSION ALGORITHM BASED ON BLOCK TRUNCATION CODING

Dr.T.ILAM PARITHI
Guest Lecturer,
Department of Computer Science,
Government Arts & Science College Sankarankovil,
Tenkasi, Tamil Nadu, India.

Ms.K.ANTONY SUDHA
Assistant Professor,
Department of Computer Science & Engineering,
Einstein College of Engineering,
Tirunelveli, Tamil Nadu, India.

I. INTRODUCTION

Compression is beneficial because it reduces the need for costly resources like hard disc space or transmission bandwidth. One of the key factors that directly affects any communication medium performs is image compression. Using a 64 kbps channel, a 16 MB image requires more than four minutes downloading, However, if the same image is compressed with a rate of 20:1 its size is reduced to 800 KB and takes only around 12 seconds to download. Transform coding and spatial coding are the two fundamental technologies used in image data compression. Therefore, the need to develop efficient image compression algorithms has become quite important. Implementing spatial coding is typically easy in terms of memory requirements and number of operations. It is quite sensitive to changes in data statistics and to the impacts of channel error, which decrease the quality of the image. Multiple coding methods, including Block Truncation Coding (BTC) and Binary Image Compression (BIC), have been examined in the spatial domain. In order to achieve data compression, transform coding must transform the original image to a new space where most of the energy will be focused in a small number of coefficients. Additionally, it is feasible to mix the two methods using a method known as hybrid coding. In order to eliminate any redundant information, spatial coding directly modifies sample image data. Grayscale images can be compressed using the straightforward and simple BTC method.

II. LITERATURE SURVEY

Shih-Lun Chenet al., proposed color and multispectral image compression using Enhance block truncation code [1] These technique are based on the mean and standard deviation. The satellite image is reshaped using this technique. There are several different sub-blocks in the satellite image. After calculating mean values, each sub-block's pixel count is compared to the mean, and each pixel's value is then changed to a binary integer in accordance with the mean. Finally for the Enhance block truncation code for satellite images, MSE, PSNR, and compression ratio are determined.

C.Senthilkumar et al., proposed a hybrid compression technique based on the Walsh Hadamard Transform (WHT) and Block Truncation Coding (BTC) has been proposed for the compression of digital images [2]. By combining BTC and WHT, this hybrid approach seeks to increase compression ratio. To evaluate the coding effectiveness and performance of the hybrid approach and compare it to the BTC and WHT, several grayscale test images are employed. Overall research suggests that the recommended strategy produces superior outcomes.By dividing the input image and changing the layout of nearby reference pixels, the conventional algorithm's processing dependence is eliminated. According to research results, parallel implementation significantly improves visual quality while decreasing processing time by 6–7 times. Jing-Ming Guo et al., proposed A modified Block Truncation Coding using max-min quantizer (MBTC). [3]. The mean and standard deviation of the pixel values in each block are used to determine the quantization in common BTC.

III. COMPARED ALGORITHMS

1. BTC

A simple and basic method for image compression is BTC.The basic premise of this method is to split the original image into several non-overlapping segments, where each block is represented by two different values.Compression in BTC is done block by block.Each block has a size of MxN, then a and b are calculated for each block mean.Every pixel in the block is thresholded using the block mean value to compress the image.If the pixel surpasses the mean, it is given the value 1.The pixel is given the value 0 if not.Bitmap image is the name given to the resulting image.The bit map image and the equivalents "a" and "b" are transmitted together with the text.Consider an 8x8 block size and a standard 8-bit image.In the decoding process, a bitmap picture is used to replace the original image like „1' is changed to the high mean letter 'a' and '0' to the low mean letter 'b'.Block size encourages the compression ratio to rise, but it also results in block distortions and poor image quality.

2. Enhanced BTC

The term "Enhanced BTC" refers to a novel method of image compression that enhances the visual quality of reconstructed images called "Enhanced Adaptive Block Truncation Coding with Edge Quantization Scheme."This technique reduces the amount of bits required for compression and increases compression rate by using a variable-length code to represent the index of each cluster.Through the use of a compressed image's edges, it also improves the visual quality of the reconstructed image. To achieve the optimum balance between compression effectiveness and image quality, an image can be divided into edge blocks and non-edge blocks using an edge detector, and each kind of block can then be compressed using a separate technique.

The following stages are part of the suggested procedure in the PDF file:

1. Dividing the image up into blocks.
2. Recognising between edge and non-edge blocks.
3. Using various quantization methods for edge and non-edge blocks.
4. Reducing the amount of bits required for compression by using a variable-length code.
5. Using the tri-clustering technique on edge blocks to decrease the mean square error
6. Developing quantization ranges to account for the minute differences between two quantization values
7. Increasing the number of clusters to more precisely approximation pixel values
8. Outlining three distinct strategies (Scheme A, Scheme B, and Scheme C) that trade off compression rate versus image quality based on various findings.
9. Producing experimental findings that highlight the benefits and success of the concepts presented offered.

3. Modified Block Truncation Coding (MBTC)

Digital grey level images are compressed using the Modified Block Truncation Coding (MBTC) lossy image compression method, which employs a max-min quantizer.In MBTC, the compressed image is created by splitting the original image into blocks, quantizing each block according to the threshold, and creating a bit map. By conserving the higher mean and lower mean of each block, the compressed image is decoded. quantization is performed based on the average value of the maximum, minimum, and mean of the blocks of pixels rather than the mean and standard deviation.The error between the pixel values of the original and the reconstructed image is reduced as a result of the reduction in the difference between the pixel values in each segment.As a result, the decoding process produces an image of higher quality while using the same bit rate as the traditional BTC.By overcoming the drawbacks of traditional BTC, the MBTC algorithm creates more aesthetically beautiful images with higher resolution while still keeping a good compression ratio.

4. ABTC-EQ ABTC-EQ

(Adaptive BTC – Edge Quantization) is an image compression method that uses edge-based block truncation. In order to quantize the pixel values based on the edge information, the method first uses the Canny edge detector to determine the edge information in an input image.The tri-clustering method matches pixel grayscale values using three values, which can lower mean square error and enhance image quality. This technique improves on BTC by handling more complicated images and achieving higher compression rates without affecting image quality. This method is as follows:

1. Enter a grayscale image with a size of M x N pixels together with the block size k that will be used to separate the image into non-overlapping blocks.
2. Locating the edge map using emap The Canny edge operator is used to derive the edge map of the input image. The Canny edge detector is a superior edge detection technique with multiple stages that can identify a variety of edges in images. The noise reduction phase of the technique involves applying a Gaussian filter on the image in order to smooth it out. Find the intensity gradient in the second stage and use the edge thinning method known as non-maximum suppression. Then use connection analysis and double thresholding to follow the image's edges.
3. The image should be divided into blocks of size k k, where k can have a value of 4, 8, and 16, etc.
4. Block classification : Classify the blocks so that the corresponding image block W is defined as an edge block, otherwise a non-edge block, if any one of the edge values in E is 1, but not all of the edge values are 1. When defining an identifier flag, it is given the values 0 for an edge block and 1 for a non-edge block.
5. Adaptive Encoding.

IV. ADAPTIVE DECODING

The resulting matrix indicates the reconstructed image. Repeat the process for each block. There may be blocks in the edge map that have all values of 0 or 1, in which case there are no edges, creating a visually continuous effect. For these non-edge blocks, quantization is carried out depending on the threshold, which is the average of the maximum, minimum, and means value. The mean values of the upper range and lower range are used as the reconstruction values when recreating such blocks. For the edge block, it is assumed that an edge separates the region into three sections: the background, the foreground, and the section with edge pixels.

V. CONCLUSION

This paper focus on Compression algorithms based on Block truncation coding such as BTC, Enhanced BTC , Modified Block Truncation Coding ,ABTC-EQ. Each Algorithm is an enhanced from BTC and they differ by the approach and they produce better results based on the complexity in their inputs

VI. REFERENCE

1. Shuyuan Zhu, Zhiying He, XiandongMeng, Jiantao Zhou and Bing Zeng, “Compression-dependent Transform Domain Downward Conversion for Block-based Image Coding”, IEEE Transactions on Image Processing, Volume: 27, Issue: 6, June 2018.
2. Sunwoong Kim and Hyuk-Jae Lee, “RGBW Image Compression by Low-Complexity Adaptive Multi-Level Block Truncation Coding”, IEEE Transactions on Consumer Electronics, Vol. 62, No. 4, November 2016.
3. Jing-Ming Guo, Senior Member, IEEE, and Yun-Fu Liu, Member, IEEE, “Improved Block Truncation Coding Using Optimized Dot Diffusion”, IEEE Transactions on Image Processing, Vol. 23, No. 3, March 2014.
4. Jayamol Mathews, Madhu S. Nair, “Modified BTC Algorithm for Gray Scale Images using max-min Quantizer”, 978-1-4673-5090-7/13/$31.00 ©2013 IEEE.
5. Lin, C.; Liu, X. A reversible data hiding scheme for block truncation compressions based on histogrammodification. In Proceedings of the 2012 Sixth International Conference on Genetic and EvolutionaryComputing, Kitakushu, Japan, 25–28 August 2012; pp. 157–160.
6. C. Chang, H. C. Hsia, and T. S. Chen. A progressive image transmission scheme based on block truncation coding. In LNCS Vol 2105, pages 383–397, 2001.

ID: 32

PROFICIENT AUDIT SERVICE OUTSOURCING FOR DATA RELIABILITY IN CLOUDS

Mr.P.DHINESH KUMAR
Assistant Professor,
Department of Computer Science,
Excel College for Commerce and Science,
Komarapalayam, Namakkal, Tamil Nadu, India.

Mrs.B.KAVIPRIYA
Assistant Professor,
Department of Computer Science,
Vidyaa Vikas College of arts and Science,
Tiruchengode, Salem, Tamil Nadu, India.

ABSTRACT

Distributed computing is an impending upset in data innovation (IT) industry in light of its exhibition, openness i.e., distributed storage empowers clients to get to their information anyplace and whenever, pay per use administration. Cloud-based re-appropriated capacity eases the client's weight for capacity the board and upkeep by giving a similarly minimal expense, versatile, area free stage. However, the fact that customers no longer have physical access to data suggests that they face a significant risk of data corruption or missing data.

To stay away from the security gambles, review administrations are basic to ensure the trustworthiness and accessibility of reevaluated information and to accomplish advanced crime scene investigation and believability on distributed computing. Audit services can be created using Provable Data Possession (PDP), a cryptographic method for verifying the integrity of data without retrieving it from an untrusted server. In this, we demonstrated how to construct a well-organized cloud-based audit service for data integrity. We predicted an interactive audit procedure to apply the audit service based on a third party auditor by utilizing the typical interactive verification system. In this review examination, the outsider examiner can concern an occasional verification to check the change of reevaluated information by giving a streamlined plan for the day. We only want to secure the third-party auditor's security and set up a small daemon to carry out the verification protocol in order to implement the audit model. In order to cut down on cloud audit services' computational costs, we present an expert approach for selecting the optimal parameter value. Our outcomes show the adequacy of our methodology.

INTRODUCTION

In current days, the promising distributed computing model is quickly in continue on force as a flighty to conventional data innovation. Through on-demand self-services, cloud computing provides a scalable environment for growing amounts of data and processes that operate on a variety of services and applications. One essential trait of this model moving is that information are being incorporated and rethought into mists. This sort of rethought stockpiling administrations in mists have turned into another benefit development point by giving a similarly minimal expense, versatile, area free stage for dealing with client's information.

The distributed storage administration lets the heap free from stockpiling the board and upkeep. These security gambles with come from the ensuing reasons: the cloud frameworks are considerably more extraordinary and solid than private processing gadgets. However, despite their advantages, they continue to face security risks from both the outside and the inside of the cloud. Additionally, there are a variety of reasons why cloud service providers mistreat cloud users, and disagreements occasionally result from a lack of trust in CSP. Hence, their ways of behaving may not be perceived by the cloud clients, regardless of whether this debate may result from the clients own hostile tasks. Based on hash functions and a signature scheme, traditional cryptographic technologies for data availability and integrity cannot work with outsourced data that does not include a local copy. Due to the complex transaction, it is also not a feasible outcome for downloading data for validation, particularly for large files. Additionally, for cloud users, auditing the accuracy of data in a cloud environment can be challenging and costly. Consequently, it is essential to recognize public audit capability.

The outsourced data cannot be secured using the conventional signature schemes and hash functions-based cryptographic technologies for data availability and integrity. Due to the high cost of communications, especially for large files, downloading them is not a viable option for data validation. In addition, the ability to verify the accuracy of data in a cloud environment can be challenging and costly for cloud users. As a result, CSS must have a public audit capability in order for data owners to use a third party auditor with expertise and capabilities that a typical user lacks to periodically audit the outsourced data. This review administration is fundamentally significant for advanced criminology and believability in mists. To execute public review capacity, the ideas of confirmation of hopelessness and provable information ownership have been proposed by certain analysts. Their methodology depended on a probabilistic confirmation procedure for a capacity supplier to demonstrate that clients' information stay in salvageable shape.

There are some options for outsourced data audit services. *Xie et al.*, for example the proposed an effective technique on happy likeness for re-appropriated data set, yet it was not appropriate for unpredictable information. *Wang et al.* It likewise gave a comparable engineering to public review administrations. To help their engineering, a public review plot was proposed with security safeguarding property. In any case, the absence of thorough execution examination for a developed review framework enormously influences the commonsense utilization of their plan. For example, in this plan a reevaluated document is straightforwardly parted into n blocks, and afterward each block produces a confirmation tag. To keep up with security, the length of block should be equivalent to the size of cryptosystem, that is to say, 160 pieces, which is 20 bytes. This indicates that a file of one million bytes is divided into 50,000 blocks, which generate 50,000 tags and store at least one million bytes of tags. Consequently, developing an audit system based on this plan is inefficient. To resolve such an issue, we acquaint a part method with further develop the framework execution and lessen the additional stockpiling.

One more central issue is the security issue of dynamic information tasks for public review administrations. Dynamic scalability for various applications is one of the fundamental design principles in clouds. This implies that remotely put away information may be gotten to by the clients as well as powerfully refreshed by them, for example, through block tasks like adjustment, cancellation and inclusion. The forgery of the verification metadata (also known as tags) generated by DOs and the leakage of the user's secret key, for instance, may pose security risks in the majority of the schemes currently in use. Consequently, it is urgent to foster a more proficient and secure system for dynamic review administrations, in which an expected enemy's benefit through powerful information tasks ought to be denied.

ACCESSIBLE PROCESS

To safely present a viable outsider examiner, the accompanying two central prerequisites had been met: TPA were had the option to effectively review the cloud information capacity without requesting the nearby duplicate of information, and present no extra on-line weight to the cloud client. The outsider evaluating process had not gotten any new weakness towards client information security. Personal computers lack the power and dependability of cloud infrastructures. However, because of their possession, they continue to be vulnerable to security threats from both the outside and the inside of the cloud. Additionally, there are a variety of reasons why cloud service providers act unfaithfully toward cloud users, and the dispute occasionally results from a lack of trust in CSP. Thusly, their ways of behaving may not be known by the cloud clients, regardless of whether this question might result from the clients' own ill-advised tasks.

The fact that TPA requires the retrieval of user data violates privacy is one of the system's drawbacks. TPA needs to recollect which key had been utilized. These two plans really great for static information not so much for dynamic information.

PROJECTED STRUCTURE

Provable information ownership has been utilized, which is a cryptographic strategy for confirming the trustworthiness of information without recovering it at an un-confided in server; can be utilized to acknowledge review administrations. It is an irregular cover strategy to accomplish a protection saving public examining framework for cloud information capacity security while remembering all above necessities to help productive Treatment of different inspecting undertakings and further investigate the procedure of bilinear total mark to broaden our primary outcome into a multiuser setting, where TPA can play out numerous evaluating errands at the same time. In addition to demonstrating how to expand our main scheme to support batch auditing for TPA upon delegations from multiple users, extensive security and performance analysis demonstrates that the proposed schemes are demonstrated to be highly secure and efficient.

The benefits of the proposed framework are Information checked in the cloud without download the source information. Information protection is ensured in out proposed framework. Key confirmation set is utilized for more significant level information security for record access in cloud. Each client ought to be register in cloud if the client has any desire to get to the information. It is helpful in identifying the global distribution of data.

AUDIT SERVICE SYSTEM

It offers a cryptographic interactive audit scheme for public audit ability that is both effective and safe. The soundness property and zero-knowledge property of proof systems are maintained in an effective and secure cryptographic interactive. In addition to preventing the deception and forgery of cloud storage providers, these two properties also prevent the leakage of outsourced data during verification.

DATA STORAGE SERVICE SYSTEM

This module contains four entities for safe data storage:

Data owner

Do have a lot of data stored in the cloud.

Cloud service provider

CSP gives information capacity administration and has sufficient extra rooms and calculation assets.

Third party auditor

TPA has abilities to oversee or screen - re-appropriated information under the appointment of information proprietor.

Granted applications

GA has the privilege to get to and control put away information. Depending on the specific requirements, these applications can be located inside or outside of clouds.

Audit Outsourcing Service System

In this module the client (data owner) preprocesses a file containing a set of blocks that form a set of public authentication information stored in the TPA using a secret key and sends the file and some authentication tags to the cloud service provider. CSP and its local copy can be deleted. At a later stage the irrevocable proof protocol issues a challenge to audit (or verify) the integrity and availability of external data based on public verification information using the TPA (Client as Audit Agent). It is important to be alerted to unusual events.

Secure and Performance Analysis

It is used to protect the data and provides the following properties:

Audit-without-downloading

Allows the TPA (or another client with the help of the TPA) to verify the correctness of the cloud data on demand without copying all the data or introducing an additional load on internet.

Verification-correctness

Ensure that any rogue CSP cannot pass TPA review without actually storing the user data in its entirety.

Privacy-preserving

Ensure that ATP cannot derive user data from information collected during the audit process.

High-performance

It enables TPAs to perform audits with minimal communication overhead and computational overhead and supports statistical audit sampling and optimization of audit plans over long periods of time.

CONSTRUCTION OF INTERACTIVE AUDIT SCHEME

This section proposes an encrypted interactive auditing scheme to support auditing systems in the cloud. The scheme is based on a standard interactive proof system that guarantees the confidentiality of secret data (zero-knowledge property) and the undecidability of invalid labels (validity property). Full-key encryption consists of five polynomial-time algorithms: data owners set public system parameters via setup and public/master key pairs are generated via keygen. Messages can be encrypted by anyone who specifies the ciphertext class associated with a plaintext message to be encrypted. Data owners can use the master key to generate an aggregate decryption key for a set of ciphertext classes by extraction. The resulting key can eventually be sent securely (via secure email or secure device) to the delegate. If the aggregate key is included in the ciphertext class decrypt any user who has the aggregate key can decrypt the specified ciphertext.

Implementation and Results

We implemented a prototype audit system based on the proposed solution to justify the effectiveness of our approach. The system was developed in an experimental cloud computing system environment built on the IaaS framework to provide powerful virtualization of distributed storage and automated management.

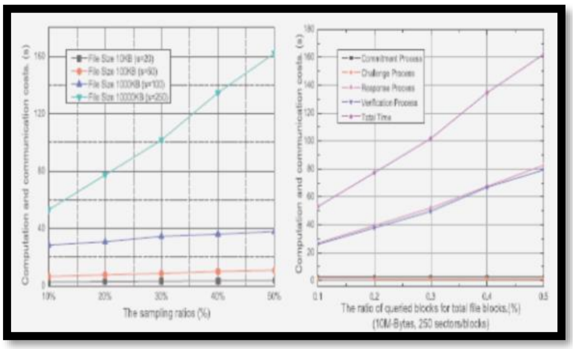


Fig.1: Experiment results under different file size, sampling ratio, and sector number.

CONCLUSION

This article discusses building an effective auditing service for data integrity in the cloud. An interactive audit protocol is proposed to implement standard interactive proof systems and realize audit services based on third-party auditors. In this audit service third-party auditors known as data owner agents can provide a customized schedule to monitor changes in outsourced data and issue periodic verifications. Understanding the audit model requires only the deployment of a lightweight daemon to enforce security and authentication protocols from third-party auditors.

Therefore the technology can be easily adopted in a cloud computing environment. This experiment clearly shows that this approach can reduce computation and communication.

REFERENCES

1. M. Mowbray, “The Fog over the Grimpen Mire: Cloud Computing and the Law,” Technical Report HPL-2009-99, HP Lab., 2009.
2. A.A. Yavuz and P. Ning, “BAF: An Efficient Publicly Verifiable Secure Audit Logging Scheme for Distributed Systems,” Proc. Ann.Computer Security Applications Conf. (ACSAC), pp. 219-228, 2009.
3. H.-C. Hsiao, Y.-H. Lin, A. Studer, C. Studer, K.-H. Wang, H.Kikuchi, A. Perrig, H.-M. Sun, and B.-Y. Yang, “A Study of User-Friendly Hash Comparison Schemes,” Proc. Ann. Computer Security Applications Conf. (ACSAC), pp. 105-114, 2009.
4. A.R. Yumerefendi and J.S. Chase, “Strong Accountability for Network Storage,” Proc. Sixth USENIX Conf. File and Storage Technologies (FAST), pp. 77-92, 2007.
5. Y. Zhu, H. Wang, Z. Hu, G.-J. Ahn, H. Hu, and S.S. Yau, “Efficient Provable Data Possession for Hybrid Clouds,” Proc. 17th ACM Conf. Computer and Comm. Security, pp. 756-758, 2010.
6. Juels and B.S. Kaliski Jr., “PORs: Proofs of Retrievability for Large Files,” Proc. ACM Conf. Computer and Communications Security (CCS ’07), pp. 584-597, 2007.
7. Ateniese, R.C. Burns, R. Curtmola, J. Herring, L. Kissner, Z.N.J. Peterson, and D.X. Song, “Provable Data Possession at Untruste Stores,” Proc. 14th ACM Conf. Computer and Comm.Security, pp. 598-609, 2007.
8. M. Xie, H. Wang, J. Yin, and X. Meng, “Integrity Auditing of Outsourced Data,” Proc. 33rd Int’l Conf. Very Large Databases (VLDB), pp. 782-793, 2007.
9. Wang, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing,” Proc. IEEE INFOCOM, pp. 1-9, 2010.

ID: 33

MOBILE AD HOC NETWORK ROUTING STABLE PATH SELECTION USING ENRDP PROTOCOL

Mr.P.BALAMUTHUKUMAR
Research Scholar,
Department of Computer Science,
Park's College,
Chinnakkarai, Tirupur.

Dr.V.KATHIRESAN
Associate Professor and Principal,
AVP College of Arts and Science,
Tirupur. Tamil Nadu, India.

ABSTRACT:

MANET is a network where no fixed infrastructure exists. Such networks are expected to play vital role in future civilian and military settings, being useful to provide communication support where no fixed infrastructure exists or the deployment of a fixed infrastructure is not economically profitable and movement of communicating parties is possible. The topology of MANETs is dynamic, because the link among the nodes may vary with time due to device mobility, new device arrivals, and the possibility of having mobile devices. Neighbor discovery is the process by which a node in a network determines the total number and identity of other nodes in its neighborhood position. Neighbor discovery is especially important to the proper functioning of wireless networks. In wireless networks, neighbors are usually defined as nodes that lie within radio range of each other. Thus, neighbor discovery can be considered as the exploration of the volume of space or "neighborhood" immediately surrounding a wireless node. The primary objective of this paper is to selects the stable path to reduce the fake position and packet lost using ENRDP protocol. It selects the stable path, so neighbor discovery failure decreases and there by increases the packet delivery ratio and reduces energy consumption, End-to-End delay and packet lost. The proposed work is analyzed and simulates in the software Network Siimulator-2(NS2) and compares its effectiveness with the existing protocol. The simulation will cover different network scenarios with varying network performance parameters such as Packet delivery ratio, End-to-End delay, Packet lost and Energy Consumption.

KEYWORDS: *MANET, Path Discovery, ENDRP, Neighbor, Nodes.*

I. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) have emerged as an advanced networking concept based on collaborative efforts among numerous self-organized wireless devices. MANET is a network where no fixed infrastructure exists. Such networks are expected to play vital role in future civilian and military settings, being useful to provide communication support where no fixed infrastructure exists or the deployment of a fixed infrastructure is not economically profitable and movement of communicating parties is possible. The topology of MANETs is dynamic, because the link among the nodes may vary with time due to device mobility, new device arrivals, and the possibility of having mobile devices.

A MANET is a collection of mobile nodes sharing a wireless channel without any centralized control or established communication backbone. MANET has dynamic topology and each mobile node has limited resources such as battery, processing power and on-board memory. This kind of infrastructure-less network is very useful in situation in which ordinary wired networks is not feasible like battlefields, natural disasters etc. The nodes which are in the transmission range of each other communicate directly otherwise communication is done through intermediate nodes which are willing to forward packet hence these networks are also called as multi-hop networks. MANET.

Fig.1 MANET

The routing protocol design must consider the physical limitations and constraints imposed by the ad hoc environment so that the resulting routing protocol does not degrade system performances. Since in MANET, there is no fixed-infrastructure such as base stations, mobile devices need to operate as routers in order to maintain the information about the network connectivity, as a result the conventional routing protocols cannot be supported easily by ad hoc networks. Several research studies have been launched to study this issue, those defined by the IETF MANET group can be classified into two categories: proactive protocols and reactive protocols. MANET's technology offers both new challenges and opportunities for many applications. The major challenges for ad hoc technology is secure and efficient routing, due essentially to MANET features (e.g., open medium, lack of centralized management, nodes mobility).

Several approaches have been proposed to secure ad hoc routing. Some existing solutions in wireless networks employ mechanisms used to protect routing protocols in wired networks that are based on the presence of a centralized infrastructure. These solutions are not appropriate for a decentralized ad hoc network. In mobile ad hoc networks, neighbor discovery is the process by which a node in a network determines the total number and identity of other nodes in its vicinity.

II. PROBLEM DEFINITION

Neighbor discovery is the process by which a node in a network determines the total number and identity of other nodes in its neighborhood position. It is a fundamental building block of many protocols including localization, routing, leader election, and group management. Time-based communications and many media access control mechanisms rely on accurate neighbor information. Neighbor discovery is especially important to the proper functioning of wireless networks. In wireless networks, neighbors are usually defined as nodes that lie within radio range of each other. Thus, neighbor discovery can be considered as the exploration of the volume of space or "neighborhood" immediately surrounding a wireless node.

Nodes found within the neighborhood are neighbors and, depending on network configuration and topology, may cooperate in the performance of various tasks including communications, sensing and localization. However, wireless communications are susceptible to abuse. Attackers have the freedom to perform malicious activities ranging from simple denial of service to sophisticated deception. NPV solution, instead, allows any node to validate the position of all of its neighbors through a fast, one-time message exchange, which makes it suitable to both static and mobile environments. Discovery and Verification of Neighbor Positions in Mobile Ad Hoc Networks, proposed a fully distributed cooperative solution that is robust against independent and colluding adversaries, and can be impaired only by an overwhelming presence of adversaries. Results show that our protocol can thwart more than 99 percent of the attacks under the best possible conditions for the adversaries, with minimal false positive rates. To the problems identified, there must be a protocol which is fully distributed and light weight to solve the verification of node position in mobile Adhoc networks. It should not depend on trusted nodes and should be secure for various kinds of attacks. A neighbor position verification (NPV) is used to discover and verify the position of the neighbors. Particularly, in a mobile Adhoc network, where a pervasive infrastructure is not present, and the location data must be attained through node-to-node communication. This situation is a chance for the adversarial nodes to misuse the location-based services. By advertise the forged positions, adversaries could bias data gathering processes, attracting network traffic and then discard the data.

III. RELATED WORKS

The primary objective of the research is to selects the stable path to reduce the fake position and packet lost using ENRDP protocol. It selects the stable path, so neighbor discovery failure decreases and there by increases the packet delivery ratio and reduces energy consumption, End-to-End delay and packet lost.

Fiore, M. et.al, 2013, [19] proposed Discovery and Verification of Neighbor Positions in Mobile Ad Hoc Networks, address the problem of a priori trusted node, the discovery and verification of neighbor positions presents challenges that have been scarcely investigated in the literature. This open issue is addressed by proposing a fully distributed cooperative solution that is robust against independent and colluding adversaries, and can be impaired only by an overwhelming presence of adversaries.

Chiang, J., 2009; Capkun, S. et.al, 2008, [5] propose For the verification of Neighbor position, there are methods was studied in the context of ad hoc and sensor networks; however, existing Neighbor Position Verification schemes often rely on fixed or mobile trustworthy nodes, which are assumed to be always available for the verification of the positions announced by third parties. In ad hoc environments, however, the pervasive presence of either infrastructure or neighbor nodes that can be aprioristically trusted is quite unrealistic.

Poturalksi, M. et.al, 2008, [25] proposed discovery of mobile nodes in, they explored the various attacks possible in the physical and communication medium of the mobile Adhoc networks they classified the neighbor discovery as physical and communication neighbor discovery. Protocols aiming at communication ND, which are based on physical ND protocols, often fail to achieve their objective. This is because these two types of discovery are not equivalent. At the same time, protocols for communication ND do not fully address the problem at hand. They are effective only under very specific operational conditions or they do not ensure correctness in all cases.

Priyadarshani, K. et.al, 2013, [26] proposed Dynamic Neighbor Positioning in MANET with Protection against Adversarial Attacks, propose techniques for finding neighbors effectively in a non priori trusted environment are identified. These techniques will eventually provide security from malicious nodes. The protocol is robust to adversarial attacks. This protocol will also update the position of the nodes in an active environment. The performance of the proposed scheme will be effective one.

Seon Yeong Han, et.al 2013, [34] proposed an adaptive Hello messaging scheme to suppress unnecessary Hello messages without reduced detect ability of broken links. An Adaptive Hello Messaging Scheme for Neighbor Discovery in On-Demand MANET Routing Protocols Seon Yeong Han, 2013, proposes an adaptive Hello messaging scheme to suppress unnecessary Hello messages without reduced detect ability of broken links.

RiteshMaheshwari et.al in 2007, [28] proposed a novel algorithm for identifying the wormhole attacks in wireless networks. This novel algorithm utilizes only connectivity information to look forbidden substructures in the connectivity graph. This approach is totally localized and does not use special hardware, making the technique generally applicable. The wireless communication model between the nodes is used in the detection algorithm. Because a communication model can help define what substructures observed in the connectivity graph could be forbidden.

LoukasLazos et.al in 2006, [15] proposed High resolution Range-independent Localization scheme in the wireless sensor networks. This method permits sensors to passively decide the location with high accuracy. This method cannot enlarge the complexity of the hardware of each and every reference point. In the High resolution Range-independent Localization scheme sensors decide their location based on the intersection of the areas covered by the beacons transmitted by multiple reference points.

IV. PROPOSED METHODOLOGY

The proposed ENRDP protocol divides into two tasks. The one is a fully distributed cooperative scheme for NPV, which enables each node, to discover and verify the position of its communication neighbors. Efficient Neighbor Route Discovery Protocol (ENRDP) provides best selection of neighbor based on the stability of the link.

To ensure the reliability and stability of the routing process here Efficient Neighbor Route Discovery Protocol (ENRDP). First it is provide a distributed, lightweight solution to the neighbor position verification problem that need not require infrastructure or a priori trusted neighbors and is robust against several different attacks, including coordinated attacks by independent and colluding adversaries. Next, it provides best selection of neighbor based on the stability of the link. The working procedure of ENRDP protocol is described in the in to two tasks.

1) Distributed cooperative NPV
2) Path stability prediction technique

4.1 Distributed Cooperative Scheme for NPV

A fully distributed cooperative scheme for NPV, which enables a node, called verifier, to discover and verify the position of its communication neighbors. A verifier can initiate the protocol at any time, by triggering the message exchange called POLL, REPLY, REVEAL and REPORT, within its 1-hop neighborhood.

To check with their neighbor position and secure transmission of content to the proper destination.

Step 1: discover nodes in range.
Step 2: send request to nodes
Step 3: wait for connection
Step 4: get location from peers with time.
Step 5: maintain location table
Step 6: broadcast the location to other nodes
Step 7: get response from other
Step 8: verify the destination location and response from other nodes
Step 9: check for location data at every request or operation
Step 10: if the location of peer is invalid mark it as spam (by its mac id)
Step 11: broadcast the spammed peer mac id to all other nodes.

4.2 Path Stability Prediction Technique

The fundamental issue arising in mobile ad hoc networks (MANETs) is the selection of the optimal path between two nodes. Ensuring a path to be valid for adequately longer period of time is a very difficult problem in MANET due to its high mobility nature. A method that has been advocated to improve routing efficiency is to select the most stable path so as to reduce the latency and the overhead due to route reconstruction. As per Distributed cooperative scheme for NPV technique, solves the neighbor verification and this scheme does not concentrate on link failures which is more often in MANET network so neighbor position verification is not get optimized results thus provide solution to link breakages through path quality technique and enhance neighbor position verification technique as per path quality technique which delivers results in efficient manner.

4.3 ENRDP Protocol

Route maintenance and route discovery procedures are similar to the DSR protocol, but with the route selection based on the link breakages. Delivery probabilities are synthesized locally from context information's like value describes the above metrics. A delivery probability of each node is used to select link stability path over dynamic route discovery.

To ensure the reliability and stability of the routing process here Efficient Neighbor Route Discovery Protocol (ENRDP). First it is provide a distributed, lightweight solution to the neighbor position verification problem that need not require infrastructure or a priori trusted neighbors and it is robust against several different attacks, including coordinated attacks by independent and colluding adversaries. Next, it provides best selection of neighbor based on the stability of the link. The working procedure of ENRDP protocol is described in the in to two tasks likely as Distributed co operative scheme for NPV and Path stability prediction techniques. ENRDP protocol is to selects the stable path to reduce the fake position and communication overhead using ENRDP protocol. It selects the stable path, so neighbor discovery failure decreases and there by increases the packet delivery ratio and reduces energy consumption, End-to-End delay and packet lost. Route selection and Route Stability

The selection of stable path among the neighbors which not only describes the selection of correct position neighbors but also best link stability neighbors. Thus overcome the adversary and also link failures. both the availability and the duration probability of a routing path that is subject to link failures caused by node mobility in terms of malicious activities.

V. RESULTS & DISCUSSIONS

The research design is stimulated in NS2. To measure and compare the performance of the proposed scheme the following performance metrics are analyzed.

- Packet delivery ratio
- End-to-End delay
- Packet lost

5.1 Packet Delivery Ratio

The ratio between the numbers of received data packets to the number of total data packets sent by the source.

$$\text{Packet Delivery ratio} = \frac{\text{Number of packet received}}{\text{Numbers of packets send}} \text{ X } 100$$

5.2 End -To-End Delay

The average time elapsed for delivering a data packet within a successful transmission from source to destination.

$$\text{Delay} = \frac{\text{Inter arrival of 1st Packet \& 2nd Packet}}{\text{Total number of packets received by all destination nodes}}$$

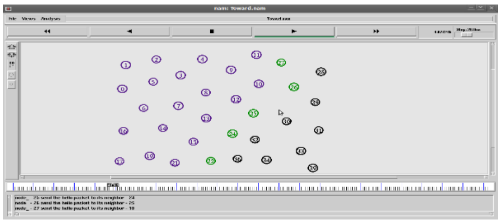
5.3 Packet Lost

Discarding of data packets in a network when a router, is overloaded and cannot accept any incoming data at a given moment.

Packet lost = Total number of packet received – Total no of packet Send

The sample outputs for the proposed design are as follows:

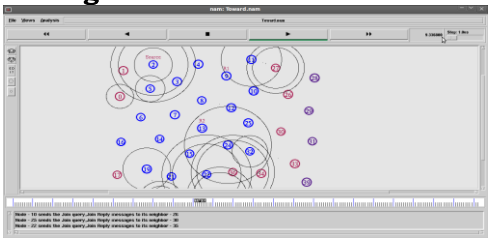
5.3. 1 Nodes Creation



```
node_ - 25 send the hello packet to its neighbor - 24
node - 26 send the hello packet to its neighbor - 25
node_ - 27 send the hello packet to its neighbor - 10
```

Figure 2 Nodes Creation

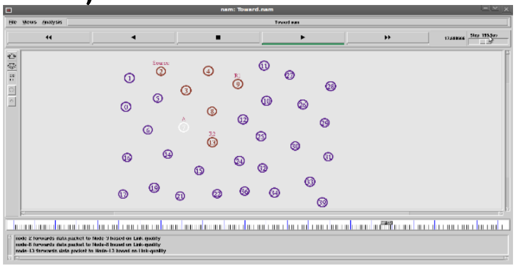
5.3.2 Neighbor Position Verification



```
Node - 10 sends the Join query,Join Reply messages to its neighbor - 26
Node - 25 sends the Join query,Join Reply messages to its neighbor - 30
Node - 22 sends the Join query,Join Reply messages to its neighbor - 35
```

Figure 3 Neighbor Position Verification

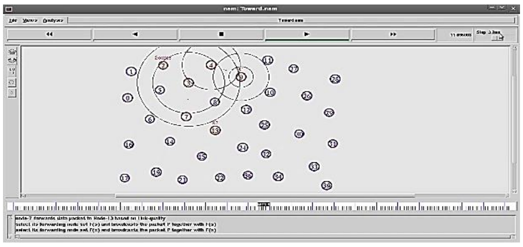
5.3.3 Finding the Malicious Node (Fake Position)



```
node 2 forwards data packet to Node 3 based on Link quality
node-8 forwards data packet to Node-8 based on Link-quality
node-13 forwards data packet to Node-13 based on Link-quality
```

Figure 4 Finding the Malicious Node (Fake Position)

5.3.4 Stable Path Creation



```
node-7 forwards data packet to Node-13 based on Link-quality
select its forwarding node set F(x) and broadcasts the packet P together with F(x)
select its forwarding node set F(x) and broadcasts the packet P together with F(x)
```

Figure 5 Stable Path Creation

VI. CONCLUSIONS

In Mobile ad hoc networks neighbor position verification are important task. Distributed techniques to perform secure neighbor position discovery, suitable for highly mobile ad hoc environments, ENRDP system under discovery of neighbor by avoiding false positions neighbors and also addressed the selection of stable path among the neighbors which not only describes the selection of correct position neighbors but also best link stability neighbors. Thus overcome the adversary and also link failures. both the availability and the duration probability of a routing path that is subject to link failures caused by node mobility in terms of malicious activities. The proposed work is implemented using NS-2. The performance are analyzed and addresses that ENRDP protocol has reduced the packet lost and delay and increases the packet delivery ratio and Energy of the network.

VII. REFERENCES

1. Ahmed Nabet, RidaKhatoun, LyesKhoukhi, Juliette Dromard and Dominique Gaïti “Towards Secure Route Discovery Protocol in MANET,” IEEE 2011.
2. Ankur O. Bang, Prabhakar L. Ramteke “MANET: History, Challenges And Applications”, Volume 2, Issue 9, September 2013.
3. C. LOURDU RAJA “Neighbor Position Verification in Mobile Ad Hoc Network”, ISSN (Online): 2320-9801, Vol.2, Special Issue 1, and March 2014.
4. Chandrasekar P, BeaulahDavid, Shaheen H,” Dissemination of Link State Information for Enhancing Security in Mobile Ad Hoc Networks”, e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VIII ,Feb. 2014, PP 24-31, 2007.
5. Chiang, J., J. Haas and Y. Hu, “Secure and PreciseLocation Verification Using Distance Bounding and Simultaneous Multilateration,” Proc. Second ACM Conf. Wireless Network Security (WiSec), 2009.
6. C-K. Toh, “Associativity-Based Routing for Adhoc Mobile Networks,” IEEE Personal Communications, Vol.4, No. 2, pp.103 – 139, March 1997.
7. E. Ekici, S. Vural, J. McNair, and D. Al-Abri, “Secure Probabilistic Location Verification in Randomly Deployed Wireless Sensor Networks,” Elsevier Ad

Hoc Networks, vol. 6, no. 2, pp. 195-209, 2008.

8. F.D. Rango, F. Guerriero, "Link Stability and Energy Aware Routing Protocol in Distributed Wireless Networks", IEEE Transactions on Parallel and Distributed systems, vol.23, no. 4, April 2012.
9. F.Guerriero "ABiojective Optimization Model for Routing in Mobile Ad-hoc Networks", IEICE Trans. Comm. Pp 4588- 4597, 2007.
10. G.W. Park, S.Lee, "A routing protocol for Extent Network Lifetime through the Residual Battery and Link Stability in MANET", ACC '08, Istanbul, Turkey, May 27-30, 2008.
11. Hwee Xian, Winston Seah, "Limiting Control Overheads Based on Link Stability for Improved performance in Mobile Adhoc Networks", Springer, UNCS3510, pp.258-268. WWIC, 2005.
12. J. Eriksson, S. Krishnamurthy, and M. Faloutsos, "True Link: A Practical Countermeasure to the Wormhole Attack in Wireless Networks," Proc. IEEE 14th Int'l Conf. Network Protocols (ICNP), Nov. 2006.
13. Jubin Sebastian E , Sreeraj V.R, Tauheed Ul Islam, "Location Based Opportunistic Routing Protocol for Mobile Ad Hoc Networks", e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-01, Issue-01, pp-16-21,2005.
14. K. Priyadharshini, V. Kathiravan, S.Karthiga, A.Christopher Paul, "Dynamic Neighbor Positioning In Manet with Protection against Adversarial Attacks". International Journal of Computational Engineering Research, Vol, 03, Issue, 4, 1999.
15. L. Lazos and R. Poovendran, "HiRLoc: High-Resolution Robust Localization for Wireless Sensor Networks," IEEE J. Selected Areas in Comm., vol. 24, no. 2, pp. 233-246, Feb. 2006.
16. Lee, M.-G., & Lee, S. "A link stability model and stable routing for mobile ad-hoc networks", LNCS4096, Seoul, Korea, pp. 904–913, in EUC.2006.
17. M. Poturalski, P. Papadimitratos, and J.-P.Hubaux, "Secure Neighbor Discovery in Wireless Networks: Formal Investigation of Possibility," Proc. ACM Symp. Information, Computer and Comm. Security (ASIACCS), Mar. 2008.
18. M.Abinaya, Mrs.K.Thamaraiselvi, "Effective Neighbor Identification with False Report Verification Using Manets", Vol.2, Special Issue 1, March 2014.
19. Marco Fiore, Claudio EttoreCasetti, Carla-FabianaChiasserini, and PanagiotisPapadimitratos, "Discovery and Verification of Neighbor Positions in Mobile Ad Hoc Networks", IEEE Transactions on Mobile Computing, VOL. 12, No. 2, February 2013.
20. Mohit Kumar, Rashmi Mishra, "An Overview of MANET: History, Challenges and Applications", Vol. 3 No. 1 Feb-Mar 2012.
21. Nitin Goyal, Alka Gaba,"A review over MANET- Issues and Challenges", Vol. 2, Issue 4, ISSN: 2319-7242, April-2013.
22. NityanandaSarma, Sukumar Nandi, "Route Stability Based QoS Routing in Mobile Ad Hoc Networks", Springer, March 2009.
23. T. A. Murshedi and X. Wang, "Mobility adaptive ad-hoc on demand distance vector routing protocol in MANET," *International Journal of Future Generation Communication and Networking*, vol. 8, no. 6, pp. 71–82, 2015.
24. W. A. Jabbar, M. Ismail, and R. Nordin, "Energy and mobility conscious multipath routing scheme for route stability and load balancing in MANETs," *Simulation Modelling Practice and Theory*, vol. 77, pp. 245–271, 2017.
25. W. A. Jabbar, W. K. Saad, and M. Ismail, "MEQSA-OLSRv2: a multicriteria-based hybrid multipath protocol for energy-efficient and QoS-aware data routing in MANET-WSN convergence scenarios of IoT," *IEEE Access*, vol. 6, pp. 76546–76572, 2018.
26. M. Khan, M. F. Majeed, M. F. Majeed, and J. Lloret, "The impact of mobility speed over varying radio propagation models using routing protocol in MANET," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Lecture Notes in Networks and Systems, vol 92*, M. Ezziyyani, Ed., pp. 277–288, Springer, Cham, 2019.

ID: 34

MULTI-AGENT SYSTEMS

Ms.M.ARCHANA
II M.Sc.,
Department of Information Technology,
Rathinam Collage of Arts and Science,
Coimbatore, Tamil Nadu, India.

Ms.C.SRIMATHI
II M.Sc.,
Department of Information Technology,
Rathinam Collage of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT:
Multi-agent systems (MASs) have received tremendous attention from scholars in different disciplines, including computer science and civil engineering, as a means to solve complex problems by subdividing them into smaller tasks. The individual tasks are allocated to autonomous entities, known as agents. Each agent decides on a proper action to solve the task using multiple inputs, e.g., history of actions, interactions with its neighboring agents, and its goal. The MAS has found multiple applications, including modeling complex systems, smart grids, and computer networks. Despite their wide applicability, there are still a number of challenges faced by MAS, including coordination between agents, security, and task allocation. This survey provides a comprehensive discussion of all aspects of MAS, starting from definitions, features, applications, challenges, and communications to evaluation. A classification on MAS applications and challenges is provided along with references for further studies. We expect this paper to serve as an insightful and comprehensive resource on the MAS for researchers and practitioners in the area.

KEYWORDS: *MAS; MAS aplications; Distributed Artificial Intelligence; Distributed Proplem solving*

I. INTRODUCTION

In recent years, Distributed Artificial Intelligence (DAI) has received tremendous attention from academia due to its ability to address complex computing problems. DAI algorithms are classified into three categories, based on the fundamental methods used to solve the tasks, namely: parallel AI, Distributed Problem Solving (DPS), and Multi-Agent Systems (MAS). Parallel AI involves developing parallel algorithms, languages, and architectures for increasing the efficiency of classical AI algorithms by leveraging task parallelism. DPS involves dividing a task into subtasks each of which is allocated to a node among a set of cooperative nodes, known as computing entities.

Multi-Agent Systems (MAS), which are the main focus of this paper, consist of autonomous entities known as agents. Similarly to computing entities in DPS, agents collaboratively solve tasks yet they offer more flexibility due to their inherent ability to learn and make autonomous decisions. Agents use their interactions with neighbouring agents or with the environment to learn new contexts and actions. Subsequently, agents use their knowledge to decide and perform an action on the environment to solve their allocated task. It is this flexibility that makes MAS suited to solve problems in a variety of disciplines including computer science, civil engineering, and electrical engineering. To develop MAS require addressing a diverse range of complex challenges such as coordination among agents, learning and security.

II. AN INTRODUCTION TO AGENTS

In this section, we define agents and their key features. In the literature, there are multiple definitions for agents resulting from diverse application-specific features of agents. Russell. defined an agent as “a flexible autonomous entity capable of perceiving the environment through the sensors connected to it.” This definition has been corroborated by other researchers in the discipline. A different perspective was presented in, where the authors defined an agent as “an encapsulated computational system that is situated in some environment and this is capable of flexible, autonomous action in that environment in order to meet its design objective.” These and most other definitions frame agents in the context of a specific field of study. However, the notion of agents is rather generic and can be broadly applied to many disciplines. Thus, we propose a generalized definition considering the fundamental abilities and features of agents:

Agent: An entity which is placed in an environment and senses different parameters that are used to make a decision based on the goal of the entity. The entity performs the necessary action on the environment based on this decision.

The above definition comprises four keywords which can be further elaborated:

1. *Entity:* Entity refers to the type of the agent. An agent can be a software, e.g. daemon security agents, a hardware component, e.g. thermostat, or a combination of both, e.g. a robot.
2. *Environment:* This refers to the place where the agent is located. The environment can be a network in the case of traffic monitoring agents, a software when the agent is monitoring the actions of software components, etc. An agent uses the information sensed from the environment for decision making.
3. *Parameters:* The different types of data that an agent can sense from the environment are referred to as parameters. For instance, the parameters for a soccer robot agent are the position and speed of the team members and opponents, and the position of the ball.
4. *Action:* Each agent can perform an action that results in some changes in the environment. For example, when a soccer robot kicks a ball the position of the ball changes. An agent can perform a set of discrete or continues actions. In a continues set of actions, the agent can perform unlimited actions, e.g. a soccer game. A discrete set of actions in contrast has a finite set of actions, e.g. an agent controlling a thermostat in a room.

The goal of each agent is to solve its allocated task with some additional constraints, e.g. a deadline. To achieve this aim, the agent first senses parameters from the environment. Empowered with this data, the agent can build up knowledge about the environment (discussed in Section V-B). An agent might also use the knowledge of its neighbors. This knowledge along with the history of the previous actions taken and the goal are fed to an inference engine which decides on the appropriate action to be taken by the agent.

3 Multi-Agent Systems (MAS)

The salient features of MAS, including efficiency, low cost, flexibility, and reliability, make it an effective solution to solve complex tasks. Their efficiency stems from the division of labor inherent in MAS whereby a complex task is divided into multiple smaller tasks, each of which is assigned to a distinct agent. Naturally, the associated overheads, e.g., processing and energy consumption, are amortized across the multiple agents, which often results in a low cost solution as compared to an approach where the entire complex problem is to be solved by one single powerful entity. Each agent can solve the allocated task with any level of pre-defined knowledge which introduces high flexibility [14]. The distributed nature of problem solving adopted in MAS also imparts high reliability. In the event of agent failure, the task can be readily reassigned to other agents.

To study MAS, agents and their relations are modeled using graphs. Graphs have been extensively used in computer science for modeling complex systems, e.g. social media, and analyzing them mathematically. When MAS are modeled as a graph, each vertex represents an agent and an edge between two vertices indicates that the two agents are communicating with each other (see Section VI). The actions taken by an agent may potentially change the relations between agents and thus change the structure of the graph. The final decision made by an agent applies to the corresponding graph that might change the edges or structure of the graph. Details of graphs relevant to MAS are beyond the scope of this paper and readers are referred as complete resources.

In MAS, agents typically only have partial information as an agent mainly communicates with its direct neighbors. On the one hand, this reduces the communication overhead which in turn ensures scalability since the overheads remain relatively low as the number of agents increase. On the other hand, the time and communication overhead associated with finding an agent that provides a particular service increases. To reduce the outlined overheads for finding an agent, particularly for large-scale MAS, the notion of middle agents is introduced. Middle agents maintain a list of services offered by all agents. Any agent that is searching for a particular service first contacts a middle agent which directs it to the appropriate agent offering that service. Depending on the implementation, middle agents can be classified as:

- **Facilitator:** As shown in Figure 3a, a facilitator acts as an intermediary between the agent sending the request (requester) and the agent providing the service (requestee). The facilitator routes the request to the appropriate agent. The response is sent back to the facilitator which relays it to the requester. As is evident, the facilitator becomes a bottleneck and a potential single point of failure. To amortize the effects of a central facilitator, multiple collaborative facilitators are employed to respond to the requests. This method requires the facilitators to communicate to remain synchronized and balance the load, i.e., requests, among themselves.
- **Mediator:** This implementation differs from the above in that the requester and requestee agents can communicate directly with each other as shown in Figure 3b. This reduces the load on the mediator agent as compared to the facilitator implementation.

A. MAS FEATURES

In the following we outline seven important features of MAS and discuss the different categorizations that arise while considering each feature.

1) Leadership

Herein, we consider the existence of a leader, i.e., an agent that defines goals and tasks for the other agents based on one global goal. The presence or absence of such a leader can be used to categorize MAS as leaderless or leader-follow. In leaderless MAS, each agent autonomously decides on its actions based on its own goals. The decision of each agent is affected by the decision of other agents if agents collaborate to reach consensus (see Section V) on a particular feature.

2) Decision Function

Herein, we categorize MAS based on the proportionality of the changes in the decision function output to its input changes. According to this, MAS are categorized as: linear and non-linear. In linear MAS, the decision of an agent is proportional to the sensed parameters from the environment, e.g. a thermostat agent turns off the heater when the temperature reaches a threshold. This feature makes linear agents easy to analyze mathematically.

3) Heterogeneity

Based on the heterogeneity of agents MAS can be divided into two categorize namely: homogeneous and heterogeneous. Homogeneous MAS include agents that all have the same characteristics and functionalities, while heterogeneous MAS include agents with diverse features.

4) Agreement Parameters

In some applications of MAS, agents need to agree on particular parameters known as metrics. Based on the number of metrics, MAS are classified as first, second, or higher order. In first order, agents collaborate to agree on one metric.

5) Delay Consideration

Agents might face multiple sources of delay for performing tasks. For instance, delay in the communication media, e.g. wireless or wired, used by agents to exchange data, or delay in scheduling resources for each agent. Depending on whether the delays are substantial and relevant, MAS can be classified into two groups namely with delay or without delay.

6) Topology

Topology refers to the location and relations of agents. MAS topology can be either static or dynamic (also known as switching in the literature). In a static topology, the position and relations of an agent remains unchanged over the lifetime of the agent. In dynamic topology MAS, the position and relations of an agent changes as the agent moves, leaves or joins the MAS, or establishes new communications, i.e. relations, with other agents.

7) Data Transmission Frequency

Agents sense the environment and share the sensed data with other agents either in a time-triggered or an event-triggered manner. In the former, the agent continuously senses the environment, collects data, and in pre-defined time intervals sends all newly sensed data to other agents. In event-triggered MAS, the agent only senses the environment when a particular event occurs. Then, the agent sends the collected data to other agents.

8) Mobility

Based on their dynamicity, agents can be classified as static or mobile agents. A static agent is always located in the same position in the environment, while mobile agents can move around in the environment. A mobile agent can be hosted by other agents meaning that it uses the resources of other agents, monitors them, or senses the environment from the position of other agents to perform actions. For example, an Intrusion Detection System (IDS) agent moves between multiple servers (agents) in the network to analyze the server processes and communications and thus detect attacks.

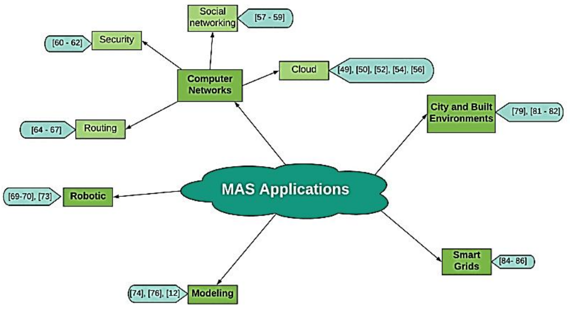
B. DIFFERENTIATING MAS WITH SIMILAR SYSTEMS

Wooldridge and Sadeghi compared MAS with expert systems and object-oriented programming language, two concepts that also involve decision making and sharing of knowledge. An expert system senses the environment and learns knowledge, then makes a decision to solve a task. Unlike MAS where each agent can communicate with any other agent, an expert system can communicate and exchange data with pre-defined entities. Although both MAS and expert systems use a decision making function, the input differs which affects the final decision. In an expert system the decision is based on the sensed data from the environment and the knowledge of the expert system, while an agent also uses its goal. According to the decision, an expert system advises a controller to perform an action. The controller can reject the decision of the expert system as it is a separate system that uses other inputs as well. However, an agent directly acts on the environment after making a decision.

In object-oriented programming, an object (e.g. object A) can communicate (or share knowledge or resources) with a pre-defined limited set of objects by creating a public function. Other objects invoke the function to communicate with object A. When object A permits other objects to access one of its functions (by making the function publicly available), it cannot control the frequency with which the function is accessed. However, in MAS an agent can communicate with any node in the network and can control the frequency with which other agents request its resources. Objects have limited pre-defined inputs that can be passed from the main object, while agents use multiple inputs.

IV. MAS Applications

In this section, we present a taxonomy of MAS applications based on broad discipline: i) Computer networks, ii) Robotics, iii) Modeling, iv) City and built environments, and v) Smart grids. A summary of these applications is outlined in,



A. Computer Networks

The complexity in computer networks significantly increases due to the emergence of new technologies and proliferation of Internet-connected devices. Agents are widely used to overcome this complexity. Due to the broad range of applications of MAS in networks, we further classify them into four sub-categorize:

1) Cloud Computing

Cloud computing enables ubiquitous access to the configurable system resources (e.g., CPU, GPU, and memory) and computing services (e.g., servers, databases, networking, and software) often over the Internet. Cloud computing uses *virtualization* as the underlying technology to provide service to the users. Using virtualization, a physical machine is shared among multiple customers as multiple Virtual Machines (VMs), each of which emulates a distinct machine. Cloud computing has multiple advantages compared to the traditional approach wherein each user maintains dedicated resources for himself:

- **Reduction of the monetary cost:** The cloud users rent the resources they require which is managed by the cloud provider. This eliminates the need for buying and maintaining resources and thus reduces the monetary cost for the user.
- **Reliability:** The existence of multiple replications for each resource makes cloud computing resilient against resource failure and thus increases its reliability. In event of a resource, e.g., a CPU, failure its tasks can be offloaded to other similar resources.

2) Social Networking

The popularity of social networks has increased exponentially with the growth in Internet users. A social network is comprised of actors, e.g. users, groups, and services. The complexity of the social networks is derived from its dynamicity, i.e., large number of participants joining or leaving the network or establishing new connections with other participants, and its broad range of applications and services. MAS can be a potential solution to overcome the complexity of social networks.

Gatti proposed an agent-based method to predict user behavior, e.g. likes, posts, and follows, in social networks, e.g. Twitter. The authors proposed the use of multiple agents, i.e., actors, which are distributed in the social network to collect a dataset of the behavior of the users. The agents then perform topic and sentiment classification on the data of each particular user which is then used to build the user profile. Finally, the user profile is fed into a prediction system that predicts the future behavior of the user including likes, topics, replies, posts, and shares. A social network needs not to be necessarily web-based. Any location that humans with similar interests gather together to interact and share information for a particular reason can be considered as a social network.

3) Security

Security applications of MAS in networks have been studied since 2002. MAS are an effective solution to network security as they can proactively learn about and thus detect new security threats.

An autonomous agent-based Intrusion Detection System (IDS) is proposed. The proposed IDS consists of five agents namely collective, detection, decision, response, and collaborative agents. The collective agent collects Simple Network Management Protocol (SNMP) and routing tables content from the network, and sends this data to the detection agent. The latter uses a misuse and anomaly detection engine to detect unusual packets or communications. The results of the detection engine are fed into a decision agent that decides if any malicious activity has occurred in the network and if so, it decides on a proper way to mitigate its affects. The decided action is then passed to a collaborative agent which is a mobile agent that delivers the decision to a response agent. The response agent enacts the appropriate action in the network.

4) Routing

Routing refers to finding a path for packets from a source node towards a destination based on specific metrics, e.g. the number of hops between the source and the destination. Using agents for routing is among the first applications of MAS studied since 1998. MAS have been evolved since then as new challenges emerge for routing protocols.

B. Agents in Robotics

Using agents for robotics has been studied for over two decades with the first article published in 1996 outlining the the pros and cons of agents in robotics. Cena argued that there exists two main challenges in robotics namely: (1) cooperation and coordination between robots, and (2) planning their movement trajectory. The authors then proposed a method that uses hardware and software agents to overcome the outlined challenges. Hardware agents refer to the physical hardware that makes up the robot, while software agents are decision making, path planning, task management, and communication agents. A hardware agent uses its sensors, e.g. cameras, to capture images of the environment. Then, a communication agent sends the images to an image processing agent. The latter processes images to find the location of the robot and obstacles in the environment. This information is then sent to the decision maker agent that finds a path with minimum obstacles toward the destination. The implementation results demonstrated that the proposed method could detect obstacles and find an optimized path (without any obstacle) to reach the destination.

Robots may be deployed in non-deterministic dynamic environments which increases the complexity of their decision-making. To study such complexity, an agent-based soccer robot is proposed by Duan. Agents (i.e., players) are grouped into teams. Agents in a team learn knowledge regarding the opponent team and possible actions by interacting with the environment, then they share learnt policies with other agents in their team. Reinforcement learning is used along with Probabilistic Neural Networks (PNN) to increase the accuracy of the final decision made by the agent. Implementation results show that agents predict the correct actions leading to an increased ball possession percentage (an important performance measure in soccer) in the agent team compared with the non-agent team.

C. Agents for Modeling Complex Systems

Modeling complex dynamic systems is costly and incurs significant processing overhead due to the demand for powerful modeling platforms and high complexity. The flexibility, autonomy and scalability afforded by agents makes Agent Based Modeling (ABM) a low-cost and low-resource solution for modeling complex systems. ABM uses a rule-based methodology for modeling the environment in contrast with other modeling methods that use equations.

The most important advantages of ABM are:

i) Ability to be aggregated and combined with other modeling methods,
ii) Flexibility in assumptions for modeling a MAS,
iii) Flexibility in pre-defined knowledge as agents can gain knowledge by learning from the environment,
iv) Possibility of parallel execution which can speed up the modeling process, and
v) Ability to explore emergent behaviors due to agent proactivity.

V. MAS CHALLENGES

Although the salient features of MAS increase its applicability in multiple disciplines, significant research challenges need to be addressed which include: coordination between agents, learning, fault detection, task allocation, localization, organization, and security. MAS challenges are typically application-specific. In this section, we outline the key challenges that apply to the vast majority of applications as summarized.

A. Coordination Control

The action performed by each agent affects the environment and thus the decision made by other agents. Coordination control refers to managing agents to collaboratively reach their goals. Multiple challenges arise from coordination including consensus, controllability, synchronization, connectivity, and formation.

B. Learning

In MAS, each agent autonomously decides on the appropriate action to reach to its goal based on multiple metrics. Agents can leverage machine learning algorithms to discover and forecast the changes in the environment and adapt to unforeseen situations and thus form Multi Agent Learning (MAL) systems. The following challenges increase the complexity of adopting learning systems for MAS: i) Processing and communication overhead of learning methods which consumes the resources of agents, ii) MAS environment may be dynamic. Thus, the agents must frequently sense updated information to be used by the learning machine, which in turn consumes a significant amount of agent resources, iii) The topology of MAS may change which requires reconnecting with neighbor agents, iv) Protecting agents against malicious agents which inject false information, and v) Scalability of the learning method for large scale MAS.

C. Fault Detection

Detecting and isolating faulty agents is a fundamental task as a faulty agent may infect other agents that it collaborates with. Current methods for Fault Detection and Isolation (FDI) are mainly centralized, where a central agent aims to detect and then isolate faulty agents. Centralized methods are suboptimal for large-scale and distributed systems such as MAS. This is attributed to the typical issues that plague centralized approaches including single point of failure and the potential for overwhelming a single agent (if many agents send requests to this agent). Consequently, FDI demands distributed solutions where all agents exchange data to detect and isolate faulty nodes.

VI. AGENT COMMUNICATION

Communication between agents has been studied for over 50 years. Three widely used approaches for communication include: Speech act: In John Austin, the pioneer researcher in speech act communication, identified that some utterance verbs or sentences, referred to as speech acts, change the physical environment, e.g. in the proper circumstances if a proper person says “I now make you man and wife,” then this sentence affects the physical environment by defining new roles and conditions. An agent can act as a speaker (S) that produces utterance to change the beliefs of the hearer (H) . Agents perform utterance by performing primary actions that are perceived as an utterance according to some language grammar.

Message passing: In this method agents directly message each other as shown. The agents use point-to-point or broadcast communication to talk to other agents. In the former, agent A can directly talk to agent B if it knows B’s address. In broadcast communication model, agent A sends a message to all its neighbors. To ensure message interpretability, the agents in a communication must use an agreed structure which is further discussed below.

Blackboard: In this communication method, agents can collaboratively share data with each other using a central repository called Blackboard as shown. Each agent stores its data on the blackboard that is readable by other agents. To control the access of agents, the blackboard uses a control knowledge. Each agent can access multiple data defined in the control knowledge.

VII. MODELING AND SIMULATION ENVIRONMENTS FOR AGENT-BASED SYSTEMS

This section outlines multiple modeling and evaluation methods used to analyze performance metrics, which vary depending on the MAS application and goal, of the designed agent-based system compared to the state-of-the-art. In the following we outline three fundamental evaluation methods:

Java Agent Development framework (JADE): JADE is among the most widely used simulators in MAS. The popularity of JADE stems from the following features: i) it is Java based and benefits from third-party libraries, ii) it is written based on FIPA standard (see Sec VI), iii) it supports simulating distributed systems, iv) it has a graphical interface for designing MAS, v) it hides MAS complexity from the designer, vi) it is open-source, and vii) it can be linked to Matlab. The authors in provide a complete instruction to MAS implementation using JADE.

GAMA: GAMA is a modeling and simulation platform for building agent-based systems [160]. GAMA has a number of advantages including: i) it can be used to model/simulate MAS in any application, ii) it supports GAML, a high-level and intuitive agent-based language, that can be readily used to simulate MAS, and iii) it supports large scale MAS which are comprised of millions of agents.

Matlab: Matlab is used to study the performance of MAS especially with respect to mathematical complex evaluations. Additionally, Matlab is linkable to JADE for further studies on MAS performance. Mathematical analysis: MAS are representable using graphs (see Section III). Consequently, using mathematical analysis are employed to evaluate MAS performance.

VIII. CONCLUSION

In this survey, we proposed a high-level comprehensive discussion regarding diverse aspects of MAS which helps newcomers to grasp basic concepts of MAS, study existing applications in multiple disciplines, the challenges in developing MAS, and the methods to study MAS performance. We first provided a definition of agents and MAS and outlined their key features. We then discussed the main applications and challenges of MAS while introducing references for further studies. Next, we discussed communications between agents and concluded the paper by a discussion on evaluation methods to analyze the effectiveness of an agent-based system. We expect this article to serve as an insightful and comprehensive resource on MAS for researchers and practitioners in the area.

IX. REFERENCES

1. M. Wooldridge, An Introduction to Multiagent Systems, New York, NY, USA:Wiley, 2009.
2. A. H. Bond and L. Gasser, Readings in Distributed Artificial Intelligence, San Mateo, CA, USA:Morgan Kaufmann, 2014.
3. S. Shamshirband, N. B. Anuar, M. L. M. Kiah and A. Patel, "An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique", *Eng. Appl. Artif. Intell*, vol. 26, pp. 2105-2127, 2013.

4. A.-M. Zou, K. D. Kumar and Z.-G. Hou, "Distributed consensus control for multi-agent systems using terminal sliding mode and Chebyshev neural networks", *Int. J. Robust Nonlinear Control*, vol. 23, no. 3, pp. 334-357, Feb. 2013.
5. M. H. Bowling, "Convergence and no-regret in multiagent learning", *Proc. NIPS*, pp. 209-216, 2004.
7. A. Zidan et al., "Fault detection isolation and service restoration in distribution systems: State-of-the-art and future trends", *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2170-2185, Sep. 2016.
8. P. Balaji and D. Srinivasan, "An introduction to multi-agent systems" in Innovations in Multi-Agent Systems and Applications, Berlin, Germany:Springer, pp. 1-27, 2010.
9. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Egnlewood Cliffs, NJ, USA:Prentice-Hall, vol. 25, pp. 27, 1995.
10. L. C. Jain and D. Srinivasan, Innovations in Multi-Agent Systems and Application, Springer, 2010.
11. D. Ye, M. Zhang and A. V. Vasilakos, "A survey of self-organization mechanisms in multiagent systems", *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 47, no. 3, pp. 441-461, Mar. 2017.
12. A. P. Garcia, J. Oliver and D. Gosch, "An intelligent agent-based distributed architecture for smart-grid integrated network management", *Proc. IEEE 35th Conf. Local Comput. Netw. (LCN)*, pp. 1013-1018, Oct. 2010.
13. S. D. J. McArthur et al., "Multi-agent systems for power engineering applications—Part I: Concepts approaches and technical challenges", *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1743-1752, Nov. 2007.
14. H. Rezaee and F. Abdollahi, "Average consensus over high-order multiagent systems", *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 3047-3052, Nov. 2015.
15. L. Ma, H. Min, S. Wang, Y. Liu and S. Liao, "An overview of research in distributed attitude coordination control", *IEEE/CAA J. Autom. Sinica*, vol. 2, no. 2, pp. 121-133, Apr. 2015.
16. J. Qi, R. Vazquez and M. Krstic, "Multi-agent deployment in 3-D via PDE control", *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 891-906, Apr. 2015.
17. R. Merris, "Laplacian matrices of graphs: A survey", *Linear Algebra Appl.*, vol. 197, pp. 143-176, Jan./Feb. 1994.
18. C. Godsil and G. F. Royle, Algebraic Graph Theory, Springer, vol. 207, 2013.
19. H. F. Ahmad, "Multi-agent systems: Overview of a new paradigm for distributed systems", *Proc. 7th IEEE Int. Symp. High Assurance Syst. Eng.*, pp. 101-107, Oct. 2002.
20. Q. Liu, L. Gao and P. Lou, "Resource management based on multi-agent technology for cloud manufacturing", *Proc. Int. Conf. Electron. Commun. Control (ICECC)*, pp. 2821-2824, Sep. 2011.
21. F. M. Al-Shrouf, "Facilitator agent design pattern of procurement business systems", *Proc. 32nd Annu. IEEE Int. Comput. Softw. Appl. (COMPSAC)*, pp. 505-510, Aug. 2008.

ID: 35

ARTIFICIAL NEURAL NETWORK TECHNIQUES IN EDUCATION

Ms.N.VAISHNAVI
Ph.D Scholar,
Department of Computer Science,
Dr. SNS Rajalakshmi College of Arts and Science,
Coimbatore, Tamil Nadu, India.

Dr.A.DEVI
Associate Professor,
Department of Computer Applications,
Dr. SNS Rajalakshmi College of Arts and Science
Coimbatore, Tamil Nadu, India.

ABSTRACT:

The education is the first and foremost asset for every human beings. There are many type of learners and every student follow unique learning style. This work describes how Artificial Neural Networks can be used and is being used in Educational sector. The AIED (Artificial Intelligence in Education) is one of the currently emerging fields in Educational technologies. The use of AI can impact on teaching and learning in higher education. The techniques are used in AIED with the help of neural networks are presented here.

***KEYWORDS**: Artificial Intelligence, Neurocomputing, Education, Virtual learning environment, Multiple linear regression.*

I. INTRODUCTION

Education 4.0 enables anytime, anywhere learning tools and the smart applications which provides opportunities for remote learning systems. In recent years, the digitalization of society transformed the nature of every job sectors in a different way. The education sector also evolved with education 4.0. Anytime, anywhere learning system is also applicable for the students. It is also known as E-Learning. This Covid-19 pandemic situation opens the platform for the new learning methodologies. Now-a-days all the school and college students even a primary school kids are forced to practice with the virtual learning system. The Instructor also had a tough time to transform the knowledge to all the students.

II. REVIEW OF LITERATURE

Alberto Rivas [1], the survey conducted to identify the important parameters to predict the performance of the students in the virtual learning environments. The prediction of the students are mainly based on the 5 parameters namely the course attendance, the attendance in the live session, the discussion of the students in a session, submission of the answers during the session, the participation module. The pass rate is increased by the identification of the outcomes from these parameters.

Kanu Ratan Butani [2], the neural network preceptors plays a vital role in the predictive analysis mechanism that became the trends of AI revolutionizing modern industries. The predictive analysis, suggested as the best and foreword techniques to find the root cause of the future problem. It helps to maintain the system of all the sectors before the problem arises.

Ahmed Dakkak [3], the student's performance is predicted by using multiple linear regression method. The supervised machine learning algorithm is used for the prediction. The MLR models are constructed and the performance are compared for each models and pick the best model with good accuracy. It is used to find the relationship between the independent and dependent variables by the method of exploratory analysis. The MSE and MAE are above 1%, but in the acceptance range, in future the model can be altered for the best. The R-squared range from 0.1 to 0.8, the maximum value denotes the strong effect size. Multivariate adaptive regression splines (MARS) Method is used for the prediction.

III. NEUROCOMPUTING IN PREDICTION

Neurocomputing may be considered part of computing (AI). In 1959, Arthur Samuel suggested the brilliant concept that human should always not need to teach computers, but rather, we could allow them to learn on their own. He coined the term “Neurocomputing” to explain his theory, which is now a regular definition for the flexibility of computers to be told autonomously [4]. The data are learned by the system, it automatically identify the patterns and then provide the result regarding to the learned dataset by the machine with minimum human intervention. The education, medical sector, bioinformatics, recognitions of patterns, financial, military etc. are just a few of fields where machine learning is applied. The computers can be trained to make prediction like humans is called neurocomputing prediction.

Neurocomputing Methodologies in Education

In the field of education, the dataset is classified in the terms of patterns for the implementation of neural network to predict the grade. The techniques used to classify the data are statistical and probability methods [5]. The below mentioned are the popular methods used to implement the artificial neural network

A. KNN Neural Network Prediction

K-nearest neighbor, used for the classification of the data. The K-nearest neighbor algorithm takes a less time to train a data when compared to the ANN. The KNN classifier uses the Euclidean distance metric is very easy to implement in the multidimensional input and yield best results. The study in used the methodology to compare the ANN and KNN, the result shows the artificial neural networks shows more accuracy than the KNN. K-nearest neighbor, used for the classification of the data. The K-nearest neighbor algorithm takes a less time to train a data when compared to the ANN. The KNN classifier uses the Euclidean distance metric is very easy to implement in the multidimensional input and yield best results. This algorithm will provides the similarities between the actual and the predicted data. The predicted data will automatically assume to the similar category of the dataset.

The general form of KNN-Nearest Neighbour

$$\mathbf{KNN} = \sum_{i=1}^{k}(\mathbf{Xi} - \mathbf{yi})$$

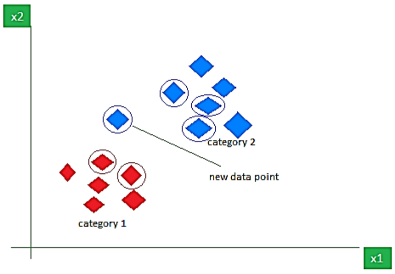


Fig 1: Representation of KNN

B. Random Forest Prediction

The random forest algorithm based on the multiple decision tress to predict the student interaction the virtual learning environments. It is use to fetch the data of the previous grade and will predict the result according to the classified data.

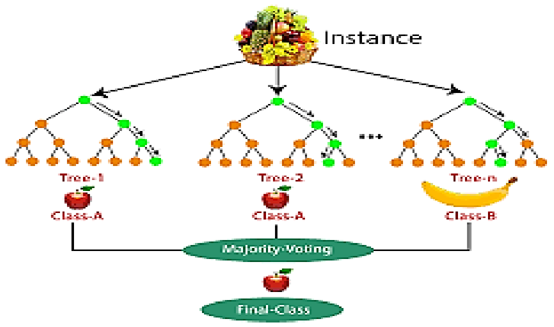


Fig 2: Representation of RFP

C. ANNs Regression Prediction

The regression network used with the perceptron. The perceptron is a basic building block of ANN. The binary classifier named as the perceptron. It provides a linear equation for the input data corresponding to the output. The linear equation is the straight line it is used for the separable of data. However, the classes can’t able to separable. The neuron are used to classify the data with weights and forward to the hidden layer that provides an activation function.

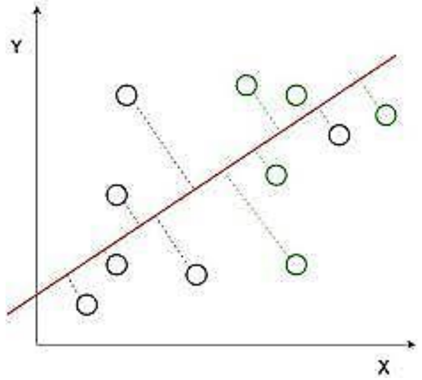


Fig 3: Representation of ANN

D. Gaussian Naive Bayes

The Bayes theorem is applied for the regression technique. It will give a result that all the predictors are independent to each other in a numeric values. The independent variables are takes as the input data features and will predict the corresponding output values which is dependent or outcome variables. It will provide the association between the input and corresponding output data The Gaussian Naive Bayes is the simplest way for the prediction of student performance. The Baye's theorem is applied for the regression technique. It will give a result that all the predictors are independent to each other in a numeric values. The independent variables are takes as the input data features and will predict the corresponding output values which is dependent or outcome variables. I will provide the association between the input and corresponding output data.

The general form of Gaussian Naïve Bayes

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

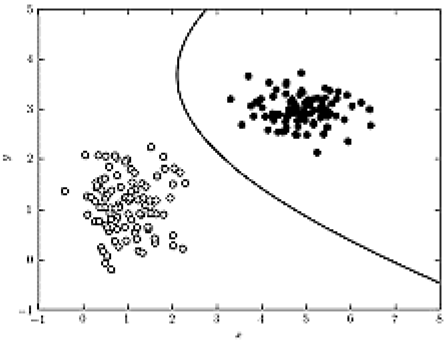
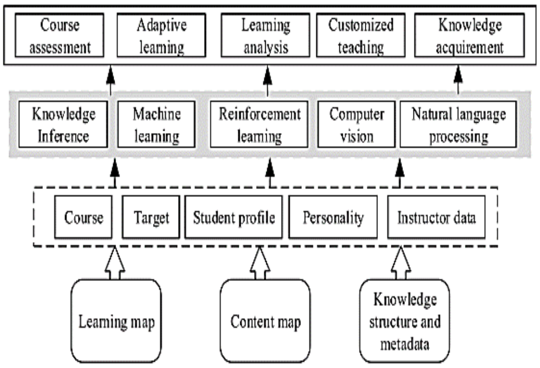


Fig 4: Representation of GNB

Scientific Structure of AI in Education



IV. REFERENCES:

1. S. Kotsiantis, C. Pierrakeas (2004), "Predicting Student performance In Distance Learning Using Machine Learning Techniques", Taylor & Francis Inc, (Vol. 18) (pp.411-426).
2. Abdullah Alshanqiti, Abdallah Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification", IEEE Access, (Vol 8, 2020) (pp.203827-203844). November, 2020.
3. Alberto Rivas, Alfonso González-Briones, Pablo Chamoso"Artificial neural network analysis of the academic performance of students in virtual learning environments", Elsevier, (pp. 0925-2312). February, 2020.
4. Kanu Ratan Butani, "Electronics for you", Electronics Magazine, an EFY GROUP Publication, (Vol. 52 No.5. ISSN 0013-516X) (pp.38-40) . May, 2020.
5. Ahmed Dakkak, "A Multiple Linear Regression-Based Approach to Predict Student Performance", Springer, (AI2SD'2019) (pp.9-23). January, 2020.

ID: 36

MITIGATING CONGESTION BASED DOS ATTACKS WITH AN ENHANCED AQM TECHNIQUE

Dr.S.THILAGAVATHI
Assistant Professor,
Department of Computer Science,
NIFT-Tea College of Knitwear Fashion,
Tirupur, Tamilnadu, India.

Dr.S.NIVAS
Assistant Professor & Head,
Department of Computer Science,
NIFT-Tea College of Knitwear Fashion,
Tirupur, Tamilnadu, India.

ABSTRACT

As widely evidenced, Denial of Service is one of the most prominent attack mechanisms on the Internet. In a recent study conducted by Radware and Ponemon Institute that consisted of surveying 705 IT practitioners, it was observed that 65% of the represented organizations suffered from three DoS attacks on average in 2012. Their average downtime lasted 54 min, resulting in an estimated cost of $22 K per minute, including the loss in revenue, traffic and end user productivity. As computer hardware becomes cheaper and social networking becomes more accessible through the cyberspace, organization and execution of such distributed attacks become significantly easier.

KEYWORDS: *DOS Attacks***,** *AQM Technique, TCP, UDP Attacks, Weighted Fair Share, Per-Flow* State.

I. INTRODUCTION

A novel congestion identification and mitigation technique is proposed that works at the server level. It aims to (a) ensure that each network flow gets its fair share of bandwidth, and (b) identify attack flows so that corrective actions (e.g. drop flooding traffic) can be explicitly taken against them. In particular, we develop an Adaptive Queuing Mechanism (AQM) is developed based on Robust Equal Sharing (RES) that maintains per-flow state such that DoS attack traffic can be precisely identified and effectively mitigated while ensuring fairness. Internet traffic can be broadly categorized into two major categories.

II. DOS ATTACKS

DoS attack flows can be defined as:

(i) The set of unresponsive flows that use more than their fair share of the bandwidth;

(ii) The set of responsive flows that do respond to congestion notifications, but not in a fair manner. Examples include flows using a hacked version of the TCP protocol with the intent to make the flows behave selfishly (or unresponsively); and

(iii) The set of responsive flows that originate from a single client where their cumulative share is more than the fair share per client. In this case, the attacker generates a number of parallel fair TCP connections to the target server with the intent to use a major portion of the bandwidth as a whole. Examples include download accelerator tools that create multiple parallel connections to a target server requesting for the same file (different parts) in order to increase the client download speed[1][2].

In this proposed approach it is aimed to build an AQM scheme that tries to meet the above ideal requirements while minimizing overhead. In case of no attack, the incoming traffic consists of only legitimate flows, that is, responsive (TCP) flows and unresponsive (UDP) flows within a fair rate. TCP uses a feedback control algorithm known as additive increase / multiplicative-decrease (AIMD) that makes these TCP flows highly sensitive to packet drops. Hence to ensure high degree of fairness among such competing flows, the packet drop (or ECN mark) decisions should be evaluated for each incoming packet, for each incoming flow. Dividing the available buffer space fairly among all incoming flows does not always guarantee an optimal solution. This is because if all flows are to use their fair buffer share, the buffer would always remain full. This would lead to higher latency and make the buffer unable to handle surges/fluctuations in incoming traffic. Moreover, it could lead to situations where packets from many flows will not be accepted even if the buffer has available empty space.

Denial of Service (DoS) attacks are currently one of the biggest risks of any organization connected to the Internet can face. Hence, the congestion handling techniques at the edge router(s), such as Active Queue Management (AQM) schemes must take into account such attacks. Ideally, an AQM scheme should (a) ensure that each network flow gets its fair share of bandwidth, and (b) identify attack flows so that corrective actions (e.g. drop flooding traffic) can be explicitly taken against them to further mitigate the DoS attacks. This paper presents a proof-of-concept work on devising such an c do not achieve the above goals or have significant room for improvement. DFS uses the concept of weighted fair share (*wfs*) that allows it to dynamically self-adjust the router buffer usage based on the current level of congestion, while aiding in identifying malicious flows. By using multiple data structures (a comprehensive repository and a cache) for keeping state of legitimate and malicious flows, DFS is able to optimize its runtime performance (e.g. higher bandwidth flows being handled by the cache). The performance advantage of DFS is demonstrated via extensive simulation while comparing against other existing AQM techniques [3].

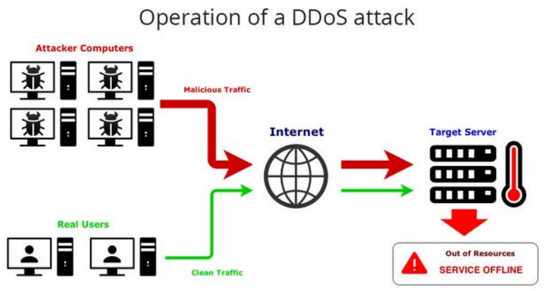


Figure 1. Operation of DDoS Attack

In Figure 1, the Internet traffic generates stream of data packets in the network with different traffic rate and leads to congestion. During congestion, the network throughput drops whereas end to end delay increases. Congestion is an important issue in this research which focus on in the TCP network environment. Most used protocols on internet are UDP and TCP. Queue management algorithm by the routers is one of the important issues in the congestion control study. These routers are augmented to monitor traffic and grant requests for rate-limiting of the streams they deliver to their peers.

Thus an AQM technique is required to share the buffer fairly among all competing flows, while ensuring that the queue length always remains stable. Moreover, the buffer should not be underutilized or overflown. To obtain these properties, the concept of weighted fair share (WFS). WFS determines the fair buffer share for each competing flow by using a scaling factor that is based on the available empty buffer space. It is a dynamic value that is updated for each incoming packet and is used to decide whether this packet should be allowed or dropped [4].

Legitimate TCP and UDP users are sending data respectively on their specified rate and attacker is sending data on varying rate. Result shows that TCP attack traffic does not have any effect on legitimate UDP user. It affects only legitimate TCP user .Simulation results shows that UDP attack traffic has greater effect on both users as compared to TCP type attack traffic. TCP user is being affected much as compared to UDP user. Therefore it is analyzed that there is more packet loss, delay and lesser throughput is achieved in TCP type of attack Traffic. TCP has no explicit congestion signal defined [5].

According to the consideration node is representing a system in the internet; node 0, node 1, node 2, node 3 and node 4 represent the legitimate UDP user, legitimate TCP user, attacker, router and receiver respectively. Link bandwidth for node 0, node 1, node 2, node 3 and node 4 is 1Mbps with 100ms of propagation delay. Drop Tail is used as queuing algorithm. Most used protocols on internet are UDP and TCP. First of all we have perform UDP flood attack and TCP attack to find out which one is more powerful attack in terms of affecting the legitimate users and consuming the more bandwidth as much as possible. We consider node 0 sends50% data that means it will occupies 0.5Mbps bandwidth.

Therefore, concurrently if node 0 sends the 30% data to node 4, and node 2 sends 30% data to node 4 and node 2 sends 60%data to node 4, the total coming traffic at node 3 is 140 %means coming traffic will use 1.3Mbps bandwidth but here we have 1Mbps link between node 3 and node 4. So data capable of 1Mbps can be transferred by node 3 therefore 40% data will be dropped and also called 40% attack intensity. These data may belong to any of users, may be of TCP user, UDP user or attacker. The above node1, 2, 3, and 4 are used in simulation environment to monitor and detedt DDOS attacks. So finally attacker gets success in consuming the bandwidth.

During this operation, the field of each incoming packet is extracted and the B-tree and Cache are search in sequence for any potential matches. Since these two data structures are mutually exclusive in terms of flows stored, if a match is found in Cache, then the B-tree is not searched. If a match is found, then the corresponding flow is updated and it is determined whether the current packet should be allowed through the buffer or not. This action is performed by the Process Packet and Decide function. If the packet is to be allowed, then the value of the matching flow i's size in buffer (sz_i) is incremented by the packet's size ($size_p$). Similarly, when the packet is dequeued from the buffer, the Deque function decrements sz_i by $size_p$.

Flows are shifted between the B-tree and Cache based on their behavior – which is defined by the value of their pm. If a flow is found in B-tree and has pm ¼ maxth, it is moved to Cache since it is a high bandwidth flow. If a flow is found in Cache and has pm ¼ minth, it is moved to B-tree since this particular flow has changed from being malicious to a legitimate one. If the fidp of the incoming packet does not match with any flow in B-tree or Cache, a new flow entry with matching fidp is made in the B-tree. For deletion of inactive flows from the Cache and B-tree, the Purge operation is performed periodically as illustrated in Algorithm 2. Its execution frequency can be adjusted by the variable prune interval. During each operation, Purge deletes all flows that were not updated since it was last run. Since a flow is only updated when it transmits a packet, this operation ensures that inactive flows are effectively deleted.

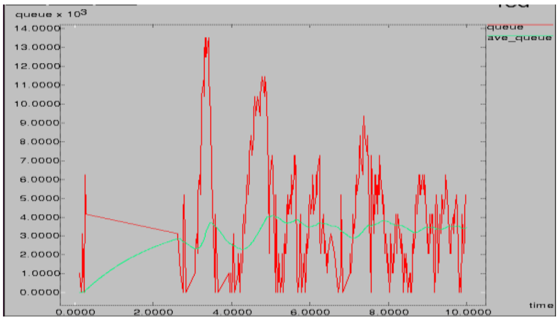


Figure 2 Average Queue Length

Average Queue Length

Provide congestion avoidance by controlling the average queue size at a network node. Avoid global synchronization and a bias against burst traffic sources. Maintain the network in a region of low delay and high throughput by keeping the average queue size low while fluctuations in the instantaneous queue size are allowed to accommodate burst traffic and transient congestion.

Figure 2, shows the Graphical Visualization of raw or processed data which is collected in a simulation and it can be graphed using tools X graph. Red color represents scenario with 'n' number of users in the queue. Green color represents the scenario is in average queue length.

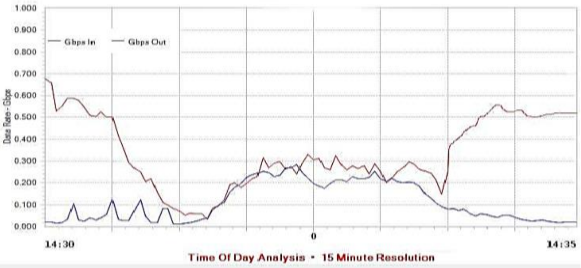


Figure 3. Evolution of the queue length at varying loads for PI

Figure 3 shows that the queue length that may vary the different loads in DDOS attacks. A Internet like topology of comprising of the attackers, legitimate users, router and the destination node is put through the flooding based DDOS attack and different AQM techniques are implemented on the router one by one to study their impact on the different parameters like throughput, Delay and Packet loss. An attempt has been made to Mitigate the effect of DDoS attack by applying different time rates of each sender node, and setting threshold value and evaluate performance.

CONCLUSION

In this work, present a proof-of-concept towards devising AQM technique that can successfully identify and mitigate congestion oriented DoS attack traffic. The proposed novel concept of weighted fair share that dynamically determines the fair buffer share for each competing flow to ensure optimal fairness. This concept is realized in DFS which uses a set of data structures in combination to provide low operational overhead while maintaining limited per-flow state and offer high DoS attack identification capability. The results demonstrate that DFS is able to provide a higher degree of fairness and throughput to legitimate flows while stabilizing the router queue length and allowing the least bandwidth to the attack traffic.

REFERENCES

1. S. T. Zargar, and J. B. D. Joshi, A Collaborative Approach to Facilitate Intrusion Detection and Response against DDoS Attacks, the 6th Int'l Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010), Chicago, IL, October 9-12, 2010.
2. M. R. Sharma, and J. W. Byers, Scalable Coordination Techniques for Distributed Network Monitoring, in Proc. PAM, pp. 349-352, 2005.
3. B. Claise, Cisco Systems NetFlow Services Export Version 9, RFC 3954, 2004.
4. R. Mahajan, S. M. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, and S. Shenker, Controlling high bandwidth aggregates in the network, presented at Computer Communication Review, pp.62-73, 2002.
5. D. Yau, J. C. S. Lui, and F. Liang, Defending against distributed denial of service attacks using max-min fair server centric router throttles, IEEE international conference on Quality of Service. 2002.
6. R. Chen, and J. M. Park, Attack Diagnosis: Throttling distributed denial-of-service attacks close to the attack sources, IEEE Int'l Conference on Computer Communications and Networks (ICCCN'05), Oct. 2005.
7. R. Chen, J. M. Park, and R. Marchany, TRACK: A novel approach for defending against distributed denial-of-service attacks, Technical Report TR-ECE-06-02, Dept. of Electrical and Computer Engineering, Virginia Tech, Feb. 2006.
8. J. Mirkovic, P. Reiher, and M. Robinson, Forming Alliance for DDoS Defense, in Proc. New Security Paradigms Workshop, Centro Stefano Francini, Ascona, Switzerland, 2003.
9. B. Parno et al., Portcullis: protecting connection setup from denial-of-capability attacks, SIGCOMM Comput. Commun. Rev., vol. 37, no. 4, pp. 289-300, 2007.
10. X. Yang, D. Wetherall, and T. Anderson, TVA: a DoS-limiting network architecture, IEEE/ACM Trans. Netw., vol. 16, no. 6, pp. 1267-1280, 2008.
11. X. Yang, D. Wetherall, and T. Anderson, A DoS-limiting Architecture, in ACM SIGCOMM, Philadelphia, PA, USA, August 2005.
12. A. Yaar, A. Perrig, and D. Song, SIFF: a Stateless Internet Flow Filter to Mitigate DDoS Flooding Attacks, in Proc. 2004 IEEE Symposium on Security and Privacy, pp. 130-143, May 2004.
13. X. Liu, A. Li, X. Yang, and D. Wetherall, Passport: secure and adoptable source authentication, in Proc. 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), San Francisco, CA, USA, pp. 365-378, 2008. G.F.Ali Ahammed, ReshmaBanu, "Analyzing the performance of Active Queue Management", Journal of Information & Computational Science Vol.2, No.2, March 2010.
14. XiapuLuo, Rocky K. C. Chang, and Edmond W. W. Chan, "Performance Analysis of TCP/AQM Under Denial-of-Service Attacks", 13th International Conference on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2005. [5] Changwang Zhang, Zhiping Cai, Weifeng Chen, Xiapu Luo, Jianping Yin, "Flow Level detection and filtering of l
15. Harkeerat Bedi, Sankardas Roy, Sajjan Shiva, "Mitigation congestion based DOS attacks with an enhanced AQM technique", SEPTEMBER, ELSVIER 2014.
16. Siddharth Ghanesla, "Network Security: Attacks, Tools and Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

ID: 37

SURVEY ON MACHINE LEARNING AND DEEP LEARNIG APPLICATIONS

Mr.A.MOHAMMED ARSHATH
Student,
Department of Information Technology,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

Ms.R.DEEPA
Student,
Department of Information Technology,
Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamil Nadu, India.

ABSTRACT:

The application of artificial intelligence is machine learning which is one of the current topics in the computer field as well as for the new COVID-19 pandemic. Researchers have given a lot of input to enhance the precision of machine learning algorithms and lot of work is carried out rapidly to enhance the intelligence of machines. Learning, a natural process in human behaviour that also becomes a vital part of machines as well. Besides this, another concept of deep learning comes to play its major role. Deep neural network (deep learning) is a subgroup of machine learning. Deep learning had been analysed and implemented in various applications and had shown remarkable results thus this field needs wider exploration which can be helpful for further real-world applications. The main objective of this paper is to provide insight survey for machine learning along with deep learning applications in various domains.

Also, some applications with new normal COVID-19 blues. A review on already present applications and currently going on applications in several domains, for machine learning along with deep neural learning are exemplified.

Keywords: *Deep neural learning (DL), Machine Learning (ML), Machine intelligence (artificial intelligence-AL)*

I. INTRODUCTION:

Machine learning, a sub-class of artificial intelligence as shown in Fig. 1. It is self-learning based on algorithms that mean the system learns from its experience. For instance, the type of data given input to the system learns the pattern and responds from its learning at the output.

In this case, the system becomes smart, smarter, and smartest with time without human involvement. It uses a statistical learning algorithm that automatically learns and improves without human help. On the other side in a deep learning system, it learns from its experience but a large database or large information provided at input**.** Deep is the term that refers to several layers in between the input and output of a neural network whereas in shallow neural networks maximum of two layers are present in between the input and output neural network.

Artificial intelligence is a wide discipline of generating intelligent machines. Mostly artificial intelligence work includes machine learning as intelligent behaviour needs extensive information or knowledge. Technologies never stop to imitate human intelligence, that's why AI has gained tremendous attention now. Most researchers in the computer filed have done work since the 1950s on machine learning. So, the expectation from the machines gets higher whereas deep learning attempt is continued in this direction and also gained a lot of attention with the effect of COVID-19 pandemic.

Various applications like face recognition have gained attention in the contact-free world because it provides great authenticity to humans as every person's face is unique. Moreover, the health sector, face recognition attendance system, banks with face ageing and face recognition, retailer, travel, airports, amazon Alexa, google home for voice help, YouTube, Netflix for recommendation system, search engines for a translator, and much more business are switched to use machine learning technology to enhance their system during and post-COVID-19 pandemic [2], [3]. Plenty of work has been in various areas using machine learning. Doing the research with new, upcoming areas and existing areas for improvement is always a continuous process in the research community.

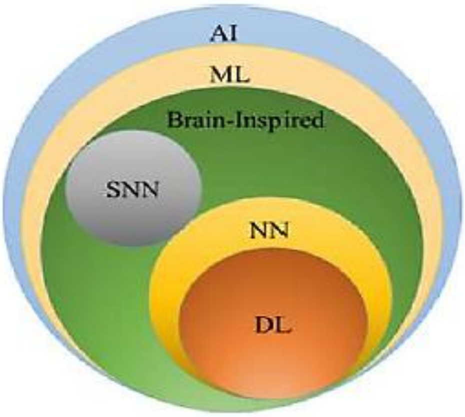


Fig.1. Shows the correlation between artificial intelligence (AL), machine learning (ML), and deep learning (DL).

II. APPROACHES IN MACHINE LEARNING:

ML algorithms wide classification is done in three categories which are supervised, unsupervised, and reinforcement learning as shown Fig. 2.

A. Supervised Learning uses an algorithm that requires external help. The provided input database is separated into training and testing datasets. The output variable is predicted or classified from the training database. Algorithms try to learn some shapes during training the database and implement these learned patterns to the testing database which provides results in estimation [4].

B. Unsupervised learning is a machine learning algorithm that learns some characteristics of input information. After providing anew database, it utilizes formerly learned characteristics for the identification of the class of data. It is mostly preferred for feature reduction also for clustering.

C. Reinforcement learning is action based on decision concept learning. In this learning, actions are based as per the decision taken so that the results become more valuable at the output or desired favourable condition.

However, the learner doesn't have prior information of data. After providing the situation, it learns to decide which action to be taken according to the given situation. The present and future situation is affected by the learner's decision i.e., action taken. Reinforcement learning exclusively relies on two conditions: delayed outcome and trial and error search [5].

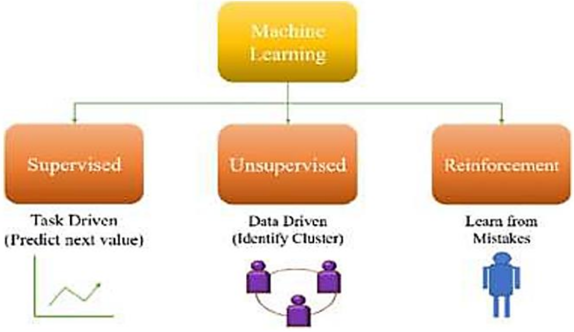


Fig. 2. Machine learning approaches.

MACHINE LEARNING OVERVIEW:

The evolution of machine was done by Arthur Samuel in 1959 who introduced the term machine learning, a pioneer in the area of AI, computer gaming. Before that, foremost chess game based on the computer program with paper and pencil was created in 1948 by Turing and Champernowne. Then, in 1951 Dietrich Prinz's new chess game machine was introduced. In 1952, Christopher Strachey created the foremost draughts algorithm. The entire draught session was played at a good speed. In the 1960s, Nilsson published a book on machine learning with pattern classification. In 1970, Duda and Hart explained that interest in pattern classification continued. Further, 1981 from the computer terminal neural network begin to learn 40 characters. In 1985 and 1986 neural network scientists Hinton, Nielsen, Rumelhart, Williams-Hetch, consecutively shown the idea which is the multilayer perceptron (MLP) with practical backpropagation (BP) training. As coming towards today's time, a new era of neural network that is termed as deep learning had grown. The third next rise of the neural network had begun in the year 2005 with researchers Andrew Ng, Hinton, Bengio, LeCun, and various other researchers.

III. MACHINE LEARNING APPLICATIONS:

The literature has shown several application fields, sub-fields with machine learning. The realworld applications are listed below and shown in Fig. 3

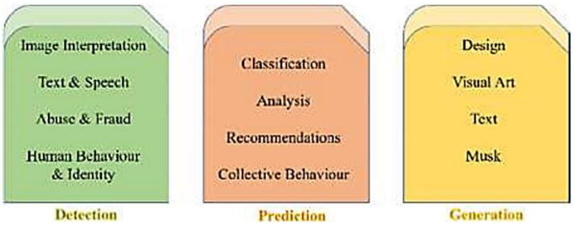


Fig. 3. Applications of machine learning and deep learning.

Computer vision is a versatile domain of machine learning which trains the machines for processing, analysing, and recognizing visual data. The various key algorithm in computer vision is KNN, SVM, Naïve Bayes. The sub-domains of this field are object detection, object processing, recognition.

Today because of the COVID-19 pandemic, new era technologies such as facial recognition and iris scans are at topmost demand as fingerprints authentication is not in line along with distance norms. For Aadhar cards in India and Banks, this face recognition with ML will be beneficial.

Machine learning-based face recognition technology is used to recognize extremists in the crowded places from the visitors at congress centres, airports, and various other important events. Now at the pandemic situation of COVID-19, this technology is proving very helpful in contactless communication and security. Thus, currently used in many businesses. Also, computer vision is used in face recognition for security purposes. An algorithm recognizes the faces of the person then allow permission for further accessibility. Moreover, used for automatic attendance system checking in professional institutes. That provides ease over the conventional methods such as keys, identity cards which can be easily stolen. Pardo, F. 2020 [6] introduces a deep reinforcement learning library called Tonic for further fast implementations in research. Various other applications such as FacePRO, Waymo are used for facial recognition and car safely driving apps respectively. Handwritten recognition application makes the work easy for organizations where handwritten documents are large. For example, universities, exam centres, police, etc. It is a process of scanning and digitizing documents in a few minutes.

Speech recognition is a process of translating spoken words to text. It is providing benefits to healthcare, military, in cars systems or to create voice interfaces and voice assistants in everyday life as it helps to improve accessibility. Speech recognition is also known as speech to text and automatic speech recognition. Various algorithms used are artificial neural networks, vector quantization, dynamic time wrapping. Wibowo, H. et al 2020 [7] provides the forward-translation of Indonesian informal to formal language with semi-supervised learning and shown the improvement in results. With research, special software can detect any variance in humans accurately in the healthcare department. It can detect various parameters at the same time and process them for medical records in real-time applications. Also, statistical analysis of medical documentation is proofing itself a great benchmark.

Predictions based on historical data can be done using machine learning. Various applications like stock pricing predictions, scientific research, marketing campaigns, and many more cases.

Generally, artificial neural networks and random forest algorithms are used for predictions. Its various sub-domains are text classification, image classification, medical diagnosis, etc. One of the beneficial domains in machine learning is the bank sector and financial sector where the chances are high for fraud detection in case money transactions are going digital. Fraud detection and prevention are done based on the identification of patterns in client transactions, identifying strange behaviour, credit scores. Mostly, classification and regression techniques of machine learning as well as neural networks are used in scam detection. Auto encoders technique with Tensor flow and Keras are being established for the detection of credit card frauds that saves a huge amount of money for cost recoveries and insurance for financial institutions.

3.2. FOR COVID-19 APPLICATIONS OF MACHINE LEARNING:

Diagnosing patients, identifying who is at most risk, better understand viruses, predict the spread of the disease, Map from where the viruses come, discovering existing drugs that can help, developing drugs at the fastest speed, Predicting the next pandemic is the crucial task done by machine learning shown in Fig. 4.

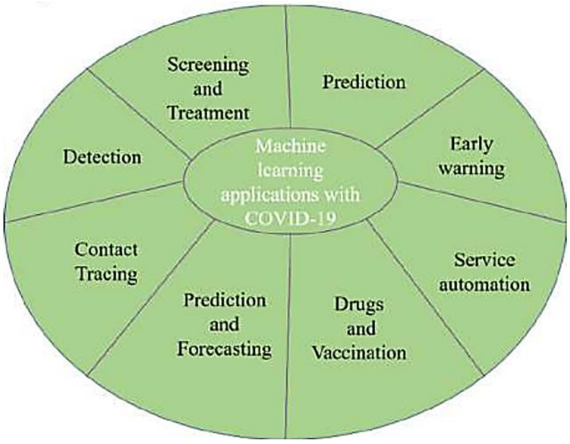


Fig. 4. Machine Learning Applications with COVID-19.

Machine learning is used for infection prediction in COVID-19. This virus spread worldwide very rapidly which needs international actions. Therefore, ML can be used to predict the behaviours of new cases to stop the disease from spreading as machine learning train the machines with mathematical models for learning and analysing. After training the machine, an interesting pattern can be detected. Researcher's Li et al. [8] developed a prediction model with machine learning to detect the reported cases in China and the world and Kumar et al. [9] applied the ARIMA (autoregressive integrated moving average) model to predict the coronavirus spread in 15 most infected countries. Huang et. al [10] used CNN and proved the CNN model efficient when compared with the model MLP, LSTM, and gated recurrent units. Pandey et al. applied two statistical algorithms— regression models and the susceptible-exposed-infectious-recovered (SEIR) for evaluating and forecasting the distribution of COVID-19 in India.

Machine learning is an unbelievable sub-domain of AI, and its involvement in the development of technologies is precious and continuous. It has created many beneficial applications in various fields and made many developments simpler where it is included.

IV. REVIEW ON DEEP LEARNING:

Deep neural networks, a subcategory of machine learning. It is a network model with neurons having several parameters and layers in between input and output. DL follows the neural network architectures approach. Thus, called deep neural networks. DL provides automatic learning of features and their representation in a hierarchical manner at various levels. This powerful process of deep learning makes it robust in contradiction of traditional machine learning methods, in short, deep learning complete architecture is used for feature extraction and alteration process. The initial layers perform simple processing of input data or learn the easy features and that output goes to the upper layers which performs complex features learning. Therefore, deep learning is suitable for dealing with larger data and complexity.

4.1. HISTORY:

McCulloch & Pitts (1943) demonstrated the Turing machine constructed from joining the neurons. In 1958, Rosenblatt showed that the perceptron's would intersect if what they were seeking to learn was able to be represented. Minsky & Papert in 1969 had presented the drawback of perceptron's, cease the investigate in neurons working for at least a decade. After that Geoffrey Hinton et al. (1985) introduced the back-propagation algorithm. Then a hierarchical neural network in 1988 called Neocognitron was skilled for visual pattern recognition. Further, Yan LeCun in 1998 [18] analysed back propagation with CNN for document analysis. Then, the Hinton laboratory in 2006 solved the training issue for DNNs from 2012 till now 2020 a deep learning algorithm is increasingly developing in various domains.

The need to use DL

Presently, DL is applied in almost all fields. Thus, this method is frequently termed as a universal learning method. DL is being used in numerous situations where machine intelligence can be beneficial such as navigation on Mars where there is the absence of a human expert, vision, speech recognition, and language understanding and biometrics, personalization for solutions in particular cases.

A. Universal learning method: The DL method is now occasionally termed as universal learning as it is proving itself a useful technique to almost every application field.
B. Robust deep learning methods do not need a precise designing feature. Instead, its process of automatically learning and representing the best features for any task make it robust.
C. The deep learning approach is generalized that means the same DL method can be applied with different

types of datasets or in different applications. That method is also termed as transfer learning. Moreover, this method is supportive where the problem has insufficient data.

D. The DL method is extremely scalable in terms of data and computation. Microsoft created a deep network called ResNet and was implemented at a supercomputing scale.

Along with that, there are plenty of challenges in DL such as Big data analytics with DL. A survey in this context was conducted by author of the paper explained the various principles like velocity, volume, and veracity of big data and also describe the merits of DL with big data. Scalability in DL methods, the capability to produce data that is significant where adequate information is not accessible for learning the system e.g. in inverse graphics, a computer vision task, mobile intelligence which are energy-efficient techniques, FPGAs, and many more, deals with causal learning.

All the above discussed challenges are faced by DL researchers until now. The performance of traditional ML methods had presented better performance with a minimum intake of data. After crossing the threshold point, traditional machine learning methods performance becomes stable, on the contrary, DL methods performance increases with increasing the amount of data. The key motives in the acceptance of deep learning nowadays are because of enlarged chip processing capabilities such as GPU units, cost-effective computer hardware, and create current advancements in machine learning projects GPU speed up the research as the highspeed evaluation is required in deep learning because of the large volume of data.

With a larger dataset, speed increment is also required which shifts the GPU to TPU (Tensor processing unit) and then HPC which high - performance computing / supercomputing, where HPC takes care of both computational improvements in terms of hardware and software that will scale up deep learning. In GPU each set of weights can be stored as a matrix (m, n). GPUs are beneficial to perform common problems parallelly problems faster. Also, all alike calculations can be at the same time. This enormously boosts the performance in parallel computations.

Various companies due to the COVID-19 blues adopting new era technologies investing hugely in ML, DL techniques to make a better prediction, analysis, and communication with customers. For example, American express sets an AI, DL lab in IIT Madras, Prato a medical app that uses artificial intelligence, financial sector, new-age security systems, and leveraging many businesses and creating with new normal COVID-19 time. Because many enterprises viewing artificial intelligence as a planned opportunity.

4.2. DEEP LEARNING APPLICATIONS:

In Big data, MAVIS-Microsoft speech recognition is done with the help of deep learning. In this learning human voices and speeches help the search of audios and video files. Also, Google applied the deep learning method on the Bigdata environment for image searching which helps to create an understanding of images so that image tagging, indexing, and annotation can be made easy. A deep dream is Google's software that can classify the images and can generate artificial and strange paintings from its knowledge. Further, deep text announced by Facebook is also an artificial intelligence arrangement. In this machine understand the text with a deep learning-based technique that can classify large data, also various corresponding services such as cleaning up spam messages and identifying handlers chatting messages.

Nowadays Google maps are using DeepMind artificial intelligence to predict arrival time an this technology is adopted by PSUs for cybersecurity with the outbreak of new normal COVID - 19 blues. Beattie, C. et al. 2020 presented DeepMind Lab2D, a simulator which provides a scalable environment for artificial intelligence research.

Google in collaboration with Improbable company recreates the real-world with augmented reality (AR). For example, an improved navigation system uses AR to superimpose the route over the live view of the roadway, AR glasses are used by Singapore airport ground staff to see more information regarding cargo containers and speed up load times.

The biggest challenge for the future nation and parents to detect and handle developmental delay in children. Researchers at MIT had established a computer system that is capable of identifying language as well as speech disorders even before kindergarten. Autism, developmental disorders, and speech disorders are a barrier to a children quality full life. Initial stage analysis and its cure can generate a brilliant outcome such as physical, emotional, and mental health.

Photo Descriptions has used deep learning methods amazingly. Andrej Karpathy and Li Fei- Fei worked on deep neural network training to recognize lots of exciting zones in the image than writing a sentence that elaborates each portion of the image. Computers tend to automatically classify photographs. For example, Facebook and Google Photos. Wan, Z. et al. 2020 have shown critical degraded old photos restoration via deep learning approach.

Pixel Restoration or Pixel Recursive Super-Resolution with deep learning made possible the zooming concept in videos. Researcher Google Brain in 2017, a deep learning network was trained with extremely low-resolution face images then predict the individual's face through it.

The recurrent neural network with CNN and LSTM generating sounds to silent movies or videos. The video frames are synchronized with a dataset of pre-recorded sounds to do a choice for the appropriate sounds for a particular scene. Thus, it tells whether the sound is real or fake also for good results a Turing-test setup created. Zou, Z. 2020 present a sky replacement dynamically and videos harmonization with deep learning.

Complexities of language are difficult to understand whether it is syntax, tonal nuances, semantics, expressions, which are the toughest tasks for a human being to learn. With the help of deep learning, Natural language processing (NLP) is trying to achieve the optimal level of success. SVM, logistic regression was time-consuming but now CNN, RNN, reinforcement learning has achieved a significant level of results. Ramamurthy, R. et al.2020 provides a new toolkit for the evaluation of reinforcement learning on NLP tasks.

Self-Driving Cars are new with artificial intelligence. The AI lab of Uber is working on creating driverless cars with several smart features. As per Forbes, MIT is trying to develop a system that will permit navigation of independent cars without a map. The restriction is still present for 3-D mapping at some prime zones in the world as it is not effective to avoid accidents. Zhou, M. et al. 2020 develop open-source SMARTS which is “scalable multi-agent reinforcement learning training school for autonomous driving” that helps in training, how to utilize diverse behaviour model of road users, accumulation.

4.3. APPLICATIONS WITH COVID-19 FOR DEEP LEARNING:

Deep learning is used for X-ray diagnosis to study the affected area of the body as cancers, bone problems, injuries, and lung diseases. CNN based models ResNet-50, Inception-v3, and Inception-ResNet-v2 used to predict COVID-19 patients with chest X-ray images by Narin et al. and reported that ResNet-50 had the best detection accuracy (98%). Further, a deep learning algorithm and support vector machine (SVM) is used to categorize the image as healthy or infected by the feature extraction process with chest X-ray images. Various deep learning models like Inception-v3, AlexNet, VGG16, Inception-ResNet-v2, VGG19, ResNet- 18, ResNet-50, Google Net, ResNet-101, DenseNet201, and XceptionNet were used and achieved a 95.38% accuracy with ResNet50 and SVM. Godfrey Hounsfield and Allan Cormack in 1972 developed a CT Scan. The CT scan diagnosis with deep learning uses X-ray technology to carefully diagnose delicate internal organs .Waheed, A. et al. 2020 presents a Covid GANfor improved covid-19 detection.

V. DISCUSSION:

Various applications such as computer vision, natural language processing, semantic analysis, prediction fields with machine learning, and deep learning methods. ECRM (electronic customer relationship management) the newest filed as an application of deep learning. The main aim of using deep learning is GPU (Graphics Processing Unit) hardware, data dependencies, and feature engineering. Data dependencies means which works with a large amount of data. The distinguishing point of DL in comparison to ML is its capability to acquire high-level characteristics from given data called as feature engineering. Thus, deep learning is emerging with new applications in several areas in the future. It is said by Andrew Ng in his Quora post “There're many projects in precision agriculture, consumer finance, medicine, where deep learning has a great impact and grows rapidly.

VI. CONCLUSION:

Machine learning practices with a set of algorithms to analyse and interpret data, learn from it, and based on that learnings, makes the best possible decisions. In the case of deep learning, the system depends upon layers of artificial neural networks. The in-depth study of deep learning and machine learning with its applications are discussed. Nowadays every individual is doing utilization of machine learning directly or indirectly. From receiving recommendations on a product in electronic shopping to update the pictures on social network sites. The development of both machine learning and deep learning, their key features, common features, and difference are also elaborated. That tells us that there is a new scope of deep learning with various applications that can produce remarkable results in the future. As research is a continuous process, a new architecture may also evolve.

CONFLICT OF INTEREST:

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

VII. REFERENCES:

1. M. Arun, E. Baraneetharan, A. Kanchana, S. Prabu Detection and monitoring of the asymptotic COVID-19 patients using IoT devices and sensors
2. S. Prabu, B. Velan, F.V. Jayasudha, P. Visu, K. Janarthanan Mobile technologies for contact tracing and prevention of COVID-19 positive cases: a cross-sectional study
3. M. Visweswaraiah, K. Somashekar, N.V. Babu Test mode power computation and IR drop analysis of application specific integrated circuits implementing face detection algorithms
4. S.B. Kotsiantis Supervised machine learning: a review of classification techniques
5. R.S. Sutton Introduction: the challenge of reinforcement learning
6. F. Pardo Tonic: A deep Reinforcement Learning Library For Fast Prototyping and Benchmarking
7. Wibowo H., Prawiro T.A., Ihsan M., Aji A.F., Prasojo R.E., Mahendra R., Semisupervised Low-Resource Style Transfer of Indonesian informal to Formal Language With Iterative Forward-Translation.
8. M. Li, Z. Zhang, S. Jiang, Q. Liu, C. Chen, Y. Zhang, *et al.* Predicting the epidemic trend of COVID-19 in China and across the world using the machine learning approach
9. P. Kumar, H. Kalita, S. Patairiya, Y.D. Sharma, C. Nanda, M. Rani, *et al.* Forecasting the dynamics of COVID-19 pandemic in top 15 countries in April 2020 through ARIMA model with machine learning approach
10. C.-J. Huang, Y.-H. Chen, Y. Ma, P.-H. Kuo Multiple-input deep convolutional neural network model for COVID-19 forecasting in China

ID: 38

SURVEY ON DATA MINING APPROACH FOR ANALYSIS AND PREDICTION OF STUDENT PERFORMANCE

Dr.P.LOGESWARI
Assistant Professor,
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab, India.

Ms.J.GOKULAPRIYA
Research Scholar,
Department of Computer Science,
Sri Krishna Arts & Science College,
Coimbatore, Tamil Nādu, India.

ABSTRACT

The quality of education is measured by the academic performance of scholars and the results they produce. Since the pupil academic performance is the made up of the environmental, cerebral, socio-profitable and other factors, it's challenging to measure the academic performance of scholars. Similar difficulties can be reduced by disquisition of colorful factors that impact the pupil performance. Numerous experimenters have been used different approaches to relating the variables that help to prognosticate scholars' performance. This check paper examines colourful datamining methodologies that have been used to dissect and prognosticate scholars' performance.

KEYWORDS: *Student Performance, Data Mining, Prediction, Classification Algorithm*

I. INTRODUCTION

Student Performance is defined as how well the student has per- formed in class, and how well the student has grasped the offered material. The performance achieved by the students measures the quality of education delivered by educational institutes. A quality education gives students the skills they have to partake as productive, innovative and responsible individuals from society. Besides various factors may influence the quality of education as well as the student motivation and potential. Deeply analysing and identifying this factor enables to build a new strategy and better decisions for the future development of quality education specifically and for nation's future social and economic prosperity in general. To achieve this objective a careful information assessment or data about the institute, teachers, the profile of students and data mining technique are crucial for better decision.

Data mining is the process of identifying relevant patterns from huge data's along with the various machine learning and statistical methods to support future decision-making. Today, due to availability of vast amount of data and the need for changing such data into helpful information and knowledge, data mining has pulled a lot of consideration in research industry and society as a whole.

Knowledge Discovery in Databases (KDD) is an-other name of datamining, which is the area of discovering new and potentially valuable information from enormous databases Arockm, L., et al. [2], When it comes to the education sector, it can provide the tasks that are used to study the student performance, i.e., prediction and analysis with the data available about all factors that in- fluence the quality of education. It means the academic performance of the student isn't an effect of just a single main factor besides it heavily depends on various factors like personal, socio- economic, psychological and other environ-mental variables. Examining and choosing the most relevant and influential variables from this factor is better to know what it will come next.

Despite this, prediction and analysis are an imperative point of reference in an educational environment for improvement of student performance. Student's academic performance is an essential factor in building their future Baker, et al., Tang et al. In this work, authors have summarized various classification ap- preach used to evaluate students' performance using important variables or predictors. The ultimate aim of classification method is to make a possible prediction of main class with a higher accuracy in given dataset.

II. LITERATURE REVIEW

There are a number of datamining techniques used in analysis and prediction based on the dataset. Association rule mining is used to analyse the relationship between dependent variables and independent variables. Classification and Regression models are the most commonly used in predicting the target class from the given data. This literature review is used to explore the important factors and different data mining algorithms that have been used in predicting students' performance.

In Angeline, et al [4], student's assignment marks, class tests, at-tendance, lab-work, previous semester grade and their participation in extra-curricular tasks are the basis for internal assessment. In addition, an external assessment of a student is on the bases of marks scored on the final exam. The proposed model used in this Paper helps to predict the students about poor, normal and great in light of class performance additionally class attendance from the generated rules. Result: From the extracted pattern Apriorism algorithm is found to be effective in predicting the student under three categories: good, average and poor.

Quadri, et al. [5] presents decision tree method to measure the performance of academic based on their cumulative grade point aver- age (CGPA). In this paper the factors that can influence students' dropouts selected by the decision tree and the effect of each risk factor is quantified by logistic regression.

Thiele, Tamara, et al. [6] utilizes the relationship between type of school, school performance, socio-economic deprivation, school grades, neighbourhood involvement, gender and academic success. Multivariable logistic regression is presented to identify the factors, which were independently associated with academic performance. According to this paper, more affluent students per- formed better than students from lowest income.

In Ahmed, et al. [7] presents prediction the final grade of students, which based on the decision tree (ID3) classification. Some variables were collected from the database of students to predict the student's final grade. This study will benefit the students to improve the student's performance, to identify those students who needed additional attention to reducing failing ration and taking appropriate action at the right time.

Bunker, Kamal et al. [8] generates the classification rule, which is based on decision tree to classify the student performance and identify the probability of failing students. It uses generated rules to which allows predicting the final result in un-der studying course. The classification algorithm successfully identifies the students who are expected to fail.

In Bhardwaj, et al. [9] authors are used in student database to predict the students' division based on last year database. This study will benefit the students and the teachers to improve the division of the student. The study investigates that not only the effort of the student's influential factor but other factors have significant influence over students' performance. This proposal will improve the insights of previous methods.

Yadav, et al [10] paper used C4.5, ID3, and CART decision tree algorithms predict performance of engineering students in the final exam. The result of the decision tree method shows the number of students who are expected to fail or pass to next year. This study helps to improve the performance of the students who were predicted to fail or pass. The results of the comparative analysis show that low achiever students can be benefited from the prediction to improve their potential.

In Satyanarayana, et al. [11] Decision Trees - J48, Naïve Bayes and Random Forest are used to increase the quality of student data by eliminating the noisy classes, and hence improving pre-diction accuracy. Personal, socio-economic, psychological and other environmental attributes are used to measure student academic performance. Apriorism, Filtered Associator and Tertius are used to identify factors that can affect student result. The paper empirically com- pares the selected technique with single model-based techniques and show that using hybrid models, it gives better pre-

Diction accuracies and also provides better rules for identifying the factors that influence student results.

Table 1: Summary of Previous Works

Quadri, M. M., & Kal- yankar,N.V. [5]	Decision tree tech- niques logistic regres- sion	A decision tree algorithm picks the factors that can influence dro- pouts. The dropouts and the consequence of each risk factor measured by logistic regression.
Thiele, T., et al. [6]	Multivariable logistic regression	School grades are representative of 'true academic' potential by comparing group differences in attainment at school compared to university.
Bunkar, Kamal, et al. [8]	Decision tree	It uses generated rules to predict the final grade in a understudy- ing course. Identifying the students who are probable to fail.
Bhardwaj, Brijesh Kumar, and Saurabh [9]	Bayesian classification	The result achieves that the factors like students' grade in the se- nior secondary exam, living location, the medium of teaching, family annual income, qualification of mother's, other habit of students, and family status of the students were highly correlated with the academic performance of students.
Ahmed, Abeer Badr El Din [7]	Decision tree(ID3)	It predicts student final grade data from student database. This study will help the students to improve the student's perfor- mance. Identify students who needed additional help not to fail in exam.
Yadav, Surjeet Ku- mar, and Saurabh Pal [10]	C4.5, ID3, CART decision tree	The result shows that, the decision tree can identify the number of students who are expected to fail or pass to next year.
Ashwin Satyana- rayana, Mariusz Nuck- owski [11]	Apriori, Filtered Asso- ciator and Tertius Decision Trees-J48 Na- ïve Bayes, and Random Forest	Various approaches are used for removing noisy data and result prediction. Apriori, Filter Associator, and Tertius are used to see- lect the best predictor variables. This provides better rules for understanding the factors those influ- ence better student outcomes.
Mythili, M. S., and AR Mohamed Shana- vas. [12]	C4.5 J48 Random Forest, Multi- layer Perception IB1 Decision Table	Various methods are compared and investigated that Random For- est performance is best than that of other algorithms employed in the study based on accuracy achieved, confusion matrix result and time is taken to execute.
Osmanbegović, E., Suljić, M., & Agić, H. (2015). [13]	Random Forest J48	The methodology is used in the reduction of data dimensionality and prediction of student's performances. Their personal demographic and societal features use data. Achieves results higher than 71%.
Al-Radaideh, et al. [14]	ID3, C4.5, NaïveBayes	It compared decision tree and naive Bayes data mining techniques and summarized that decision tree achieves best.
Shiv Pratap Singh Kush- wah, et al. [15]	K-nearest neighbor (KNN)	Compare several techniques and states that K-nearest neighbor (kNN) classification algorithm is easy to understand and to im- plement and also more powerful than others

III. DATA MINING METHODS

3.1 Association rule mining

Association rule mining is a method in which frequent patterns, correlations, associations, or causal structures are extracted from data sets.

Apriorism Algorithm- is the most popular tool used by different researchers in frequent pattern and association rule mining. It is used to discover a correlation between variables in alarge dataset.

FP-Growth Algorithm - is fast algorithm for determining frequent item sets in the record. It is also memory efficient but expensive to build FP-Tree.

3.2 Classification algorithms

There are a number of classifier algorithms used in predicting student academic performance.

Decision tree - A decision tree is the most effective and well- known method for classification and prediction in educational data mining. A Decision tree normally begins with asingle node, which branches into the conceivable outcome.

Naive Bayes - The Naive Bayes Classifier is based on Bayeda theorem and is specifically for high dimensional inputs. In addition to its simplicity, Naive Bayes can often perform better classification methods. Naive Bayes classifier is widely used in prediction students' academic performance.

Support Vector Machines - Support Vector Machines (SVM) are based on the notion of decision planes that states decision boundaries. A decision plane is used to separate the objects from contained under different classes. The standard.

SVM takes a sorted dataset and predicts, for each given in- for motion, which of two conceivable classes includes the in- for motion, making the SVM a non-probabilistic binary linear classifier.

Random Forest-Random Forest algorithm is the most popular algorithm, which can use for both classification and the regression kind of problems. Random forest algorithm is one of a supervised classification algorithm, which creates a forest using a number of trees. The forest gives the high ac- cu-racy results since it is constructed from the higher the number of trees.

IV. Data preparation

In this study, the dataset obtained from the Kaggle educational dataset, which is collected from learning management system, will be used. The dataset contains 480 student data rows and 16 variables. The dataset consists of three main groupings of features:

- Demographic: - gender, nationality
- Academic background: - educational stage, grade Level, and section

Behavioural features: - raised hand in class, opening re- sources, answering the survey by parents, and school satisfaction.

Since real-world data is often inconsistent, incomplete, and lacking in certain behaviours and is probably to contain several errors, Data preprocessing is a known method to solve such problems. Data preprocessing prepares unprocessed data for further processing. In this work, the dataset obtained will be preprocessed, such as data cleaning, transformation, reduction and other tasks will be applied. Many Data mining techniques used by many researchers. However, the proposed method selects the best classifier based on comparative analysis.

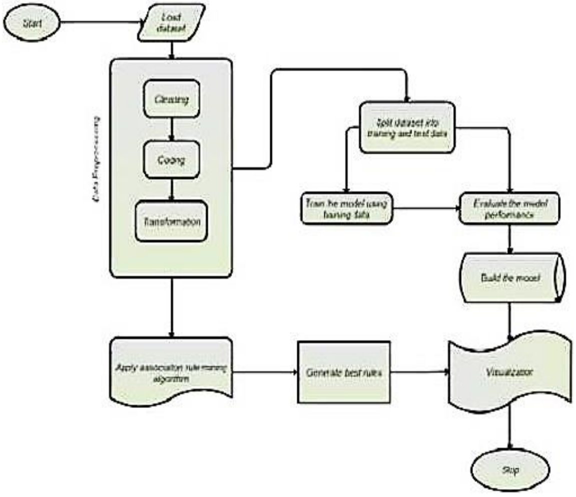

Start
Load dataset
Data Preprocessing
Cleaning
Coding
Transformation
Split dataset into training and test data
Train the model using training data
Evaluate the model performance
Build the model
Apply association rule mining algorithm
Generate best rules
Visualization
Stop

Fig. 1: Flow Diagram of Proposed System.

1. Conclusion and future work

This paper aims to identify related factors and several data mining algorithms used in predicting the performance of students. Predicting the student performance is the most important tool to help the students and schools in improving their learning and teaching process. Various studies have been reviewed that works on predicting students' performance with various methods. A number of factors that can affect the student performance also have been identified. From this review, most of the studies have been achieved a better result by using Apriorism algorithm in identifying the relationship be- tween different factors that can affect the students' performance. And also, a decision tree is the most used algorithm in predicting the student performance. The future work of this paper will be com- paring and using advanced machine learning algorithm to get better prediction result.

REFERENCES

1. Tang, Tiffany Ya, and Gordon McCalla. "Smart recommendation for an evolving e-learning ystem: Architecture and experiment." In- ter- national Journal on elearning 4.1 (2005): 105.
2. Arockiam, L., et al. "Deriving Association between Urban and Ru- ral Students Programming Skills." International Journal on Com- puter Science and Engineering 2.3 (2010).
3. Baker, Ryan Shaun, Albert T. Corbett, and Kenneth R. Koedinger. "Detecting student misuse of intelligent tutoring systems." Interna- tional conference on intelligent tutoring systems. Springer, Berlin, Heidelberg, 2004.
4. [Angeline, D. Magdalene Delighta. "Association rule generation for student performance analysis using apriori algorithm." The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1.1 (2013): 12-16.
5. Quadri, Mr MN, and N. V. Kalyankar. "Drop out feature of student data for academic performance using decision tree techniques." Global Journal of Computer Science and Technology (2010).
6. Thiele, Tamara, et al. "Predicting students' academic performance based on school and socio-demographic characteristics." Studies in Higher Education 41.8 (2016): 1424-1446.
7. Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Min- ing: A prediction for Student's Performance Using Classifica- tion Method." World Journal of Computer Application and Tech- nology 2.2 (2014): 43-47.
8. Bunkar, Kamal, et al. "Data mining: Prediction for performance im- provement of graduate students using classification." Wireless and Optical Communications Networks (WOCN), 2012 Ninth Inter- na- tional Conference on. IEEE, 2012.
9. Bhardwaj, Brijesh Kumar, and Saurabh Pal. "Data Mining: A pre- diction for performance improvement using classification." arXiv preprint arXiv: 1201.3418 (2012).
10. Yadav, Surjeet Kumar, and Saurabh Pal. "Data mining: A predic- tion for performance improvement of engineering students using classification." arXiv preprint arXiv: 1203.3832 (2012).
11. Satyanarayana, Ashwin, and Mariusz Nuckowski. "Data Mining us- ing Ensemble Classifiers for Improved Prediction of Student Aca- demic Performance." (2016).
12. Mythili, M. S., and AR Mohamed Shanavas. "An analysis of stu- dents' performance using classification algorithms." IOSR Journal of Computer Engineering 16.1 (2014): 63-9.
13. Osmanbegović, Edin, Mirza Suljić, and Hariz Agić. "Determining dominant factor for student's performance prediction by using data mining classification algorithms." Tranzicija 16.34 (2015): 147-158.
14. Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al- Najjar. "Mining student data using decision trees." International Arab Conference on Information Technology (ACIT'2006), Yar- mouk University, Jordan. 2006.
15. Kushwah, Shiv Pratap Singh, Keshav Rawat, and Pradeep Gupta. "Analysis and comparison of efficient techniques of clustering algo- rithms in data mining." International Journal of Innovative Tech- nol- ogy and Exploring Engineering (IJITEE) 1.1 (2012): 2278-3075.
16. Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint ar- Xiv: 1201.3417 (2012).

ID: 39

MACHINE LEARNING AND DEEP LEARNING

Mr.S.SACHINDEVARAJAN
Department of Information Technology,
Rathinam Collage of Arts and Science,
Coimbatore, Tamil Nadu, India.

Ms.T.DEVAKI
Department of Information Technology,
Rathinam Collage of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT:

With the development of the Internet, cyber-attacks are changing rapidly and the cyber security situation is not optimistic. This survey report describes key literature surveys on machine learning (ML) and deep learning (DL) methods for network analysis of intrusion detection and provides a brief tutorial description of each ML/DL method. Papers representing each method were indexed, read, and summarized based on their temporal or thermal correlations. Because data are so important in ML/DL methods, we describe some of the commonly used network datasets used in ML/DL, discuss the challenges of using ML/DL for cybersecurity and provide suggestions for research directions.

KEYWORDS: *ML; DL; Dataset; Cyber Security*

I. INTRODUCTION

With the increasingly in-depth integration of the Internet and social life, the Internet is changing how people learn and work, but it also exposes us to increasingly serious security threats. How to identify various network attacks, particularly not previously seen attacks, is a key issue to be solved urgently. Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction. A network security system consists of a network security system and a computer security system. Each of these systems includes firewalls, antivirus software, and intrusion detection systems (IDS). IDSs help discover, determine and identify unauthorized system behavior such as use, copying, modification and destruction.

Security breaches include external intrusions and internal intrusions. There are three main types of network analysis for IDSs: misuse-based, also known as signature-based, anomaly-based, and hybrid. Misuse-based detection techniques aim to detect known attacks by using the signatures of these attacks. They are used for known types of attacks without generating a large number of false alarms. However, administrators often must manually update the database rules and signatures. New (zero-day) attacks cannot be detected based on misused technologies. Anomaly-based techniques study the normal network and system behavior and identify anomalies as deviations from normal behavior. They are appealing because of their capacity to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, therefore making it difficult for attackers to know which activities they can perform undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviors can be categorized as anomalies.

Hybrid detection combines misuse and anomaly detection. It is used to increase the detection rate of known intrusions and to reduce the false positive rate of unknown attacks. Most ML/DL methods are hybrids. This paper presents a literature review of machine learning (ML) and deep learning (DL) methods for cybersecurity applications. ML/DL methods and some applications of each method in network intrusion detection are described. It focuses on ML and DL technologies for network security, ML/DL methods and their descriptions. Our research aims on standards-compliant publications that use “machine learning”, “deep learning” and cyber as keywords to search on Google Scholar. In particular, the new hot papers are used because they describe the popular techniques.

II. SIMILARITIES AND DIFFERENCES IN ML AND DL

There are many puzzles about the relationship among ML, DL, and artificial intelligence (AI). AI is a new technological science that studies and develops theories, methods, techniques, and applications that simulate, expand and extend human intelligence. It is a branch of computer science that seeks to understand the essence of intelligence and to produce a new type of intelligent machine that responds in a manner similar to human intelligence. Research in this area includes robotics, computer vision, nature language processing and expert systems. AI can simulate the information process of human consciousness, thinking. AI is not human intelligence, but thinking like a human might also exceed human intelligence.

ML is a branch of AI and is closely related to (and often overlaps with) computational statistics, which also focuses on prediction making using computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. ML is occasionally conflated with data mining, but the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. ML can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies. The pioneer of ML, Arthur Samuel, defined ML as a “field of study that gives computers the ability to learn without being explicitly programmed.” ML primarily focuses on classification and regression based on known features previously learned from the training data.

DL is a new field in machine-learning research. Its motivation lies in the establishment of a neural network that simulates the human brain for analytical learning. It mimics the human brain mechanism to interpret data such as images, sounds and texts. The concept of DL was proposed by Hinton based on the deep belief network (DBN), in which an unsupervised greedy layer-by-layer training algorithm is proposed that provides hope for solving the optimization problem of deep structure. Then the deep structure of a multi-layer automatic encoder is proposed. In addition, the convolution neural network proposed by LeCun et al. is the first real multi-layer structure learning algorithm that uses a space relative relationship to reduce the number of parameters to improve the training performance. DL is a machine-learning method based on characterization of data learning. An observation, such as an image, can be expressed in a variety of ways, such as a vector of each pixel intensity value, or more abstractly as a series of edges, a region of a particular shape, or the like. Using specific representations makes it easier to learn tasks from instances. Similarly to ML methods, DL methods also have supervised learning and unsupervised learning. Learning models built under different learning frameworks are quite different. The benefit of DL is the use of unsupervised or semi-supervised feature learning and hierarchical feature extraction to efficiently replace features manually.

The differences between ML and DL include the following:

- Data dependencies. The main difference between deep learning and traditional machine learning is its performance as the amount of data increases. Deep learning algorithms do not perform as well when the data volumes are small, because deep learning algorithms require a large amount of data to understand the data perfectly. Conversely, in this case, when the traditional machine-learning algorithm uses the established rules, the performance will be better.
- Hardware dependencies. The DL algorithm requires many matrix operations. The GPU is largely used to optimize matrix operations efficiently. Therefore, the GPU is the hardware necessary for the DL to work properly. DL relies more on high-performance machines with GPUs than do traditional machine-learning algorithms.
- Feature processing. Feature processing is the process of putting domain knowledge into a feature extractor to reduce the complexity of the data and generate patterns that make learning algorithms work better. Feature processing is time-consuming and requires specialized knowledge. In ML, most of the characteristics of an

application must be determined by an expert and then encoded as a data type. Features can be pixel values, shapes, textures, locations, and orientations. The performance of most ML algorithms depends upon the accuracy of the features extracted. Trying to obtain high-level features directly from data is a major difference between DL and traditional machine-learning algorithms. Thus, DL reduces the effort of designing a feature extractor for each problem.

- Problem-solving method. When applying traditional machine-learning algorithms to solve problems, traditional machine learning usually breaks down the problem into multiple sub-problems and solves the sub-problems, ultimately obtaining the final result. In contrast, deep learning advocates direct end-to-end problem solving.
- Execution time. In general, it takes a long time to train a DL algorithm because there are many parameters in the DL algorithm; therefore, the training step takes longer. The most advanced DL algorithm, such as ResNet, takes exactly two weeks to complete a training session, whereas ML training takes relatively little time, only seconds to hours. However, the test time is exactly the opposite. Deep learning algorithms require very little time to run during testing. Compared with some ML algorithms, the test time increases as the amount of data increases. However, this point does not apply to all ML algorithms, because some ML algorithms have short test times.
- Interpretability. Crucially, interpretability is an important factor in comparing ML with DL. DL recognition of handwritten numbers can approach the standards of people, a quite amazing performance. However, a DL algorithm will not tell you why it provides this result. Of course, from a mathematical point of view, a node of a deep neural network is activated. However, how should neurons be modeled and how do these layers of neurons work together? Thus, it is difficult to explain how the result was generated. Conversely, the machine-learning algorithm provides explicit rules for why the algorithm chooses so; therefore, it is easy to explain the reasoning behind the decision.

III. NETWORK SECURITY DATA SET

Data constitute the basis of computer network security research. The correct choice and reasonable use of data are the prerequisites for conducting relevant security research. The size of the dataset also affects the training effects of the ML and DL models. Computer network security data can usually be obtained in two ways: 1) directly and 2) using an existing public dataset. Direct access is the use of various means of direct collection of the required cyber data, such as through Win Dump or Wireshark software tools to capture network packets. This approach is highly targeted and suitable for collecting short-term or small amounts of data, but for long-term or large amounts of data, acquisition time and storage costs will escalate. The use of existing network security datasets can save data collection time and increase the efficiency of research by quickly obtaining the various data required for research. This section will introduce some of the Security datasets that are accessible on the Internet and facilitate section IV of the research results based on a more comprehensive understanding.

A. DARPA Intrusion Detection Data Sets

DARPA Intrusion Detection Data Sets, which are under the direction of DARPA and AFRL/SNHS, are collected and published by The Cyber Systems and Technology Group (formerly the DARPA Intrusion Detection Evaluation Group) of MIT Lincoln Laboratory for evaluating computer network intrusion detection systems. The first standard dataset provides a large amount of background traffic data and attack data. It can be downloaded directly from the website.

B. KDD Cup 99 Dataset

The KDD Cup 99 dataset is one of the most widely used training sets; it is based on the DARPA 1998 dataset. This dataset contains 4 900 000 replicated attacks on record. There is one type of the normal type with the identity of normal and 22 attack types, which are divided into five major categories: DoS (Denial of Service attacks), R2L (Root to Local attacks), U2R (User to Root attack), Probe (Probing attacks) and Normal. For each record, the KDD Cup 99 training dataset contains 41 fixed feature attributes and a class identifier. Of the 41 fixed feature attributes, seven characteristic properties are the symbolic type.

C. NSL-KDD Dataset

The NSL-KDD dataset [7] is a new version of the KDD Cup 99 dataset. The NSL-KDD dataset improves some of the limitations of the KDD Cup 99 dataset. The KDD 1999 Cup Dataset Intrusion Detection Dataset was applied to the 3rd International Knowledge Discovery and Data Mining Tools Contest. This model identifies features between intrusive and normal connections for building network intrusion detectors. In the NSL-KDD dataset, each instance has the characteristics of a type of network data.

D. ADFA Dataset

The ADFA data set is a set of data sets of host level intrusion detection system issued by the Australian defence academy (ADFA), which is widely used in the testing of intrusion detection products. In the dataset, various system calls have been characterized and marked for the type of attack. The data set includes two OS platforms, Linux (ADFA-LD) and Windows (ADFA-WD), which record the order of system calls. In the case of ADFA-LD, it records the invocation of operating system for a period of time. Kernel provides the user space program and the kernel space interact with a set of standard interface, the interface to the user program can be restricted access hardware devices, such as the application of system resources, operating equipment, speaking, reading and writing, to create a new process, etc. User space requests, kernel space is responsible for execution, and these interfaces are the bridge between user space and kernel space. ADFA-LD is marked for the attack type, as shown in the figure. Linux system, user space by making system calls to kernel space to produce soft interrupts, so that the program into the kernel state, perform corresponding operations. There is a corresponding system call number for each system call.

IV. ML AND DL ALGORITHM FOR CYBER SECURITY

This section is divided into two parts. The first part introduces the application of traditional machine-learning algorithms in network security. The second part introduces the application of deep learning in the field of cybersecurity. It not only describes the research results but also compares similar studies.

A. Support Vector Machine

Support Vector Machine (SVM) is one of the most robust and accurate methods in all machine-learning algorithms. It primarily includes Support Vector Classification (SVC) and Support Vector Regression (SVR). The SVC is based on the concept of decision boundaries. A decision boundary separates a set of instances having different class values between two groups. The SVC supports both binary and multi-class classifications. The support vector is the closest point to the separation hyperplane, which determines the optimal separation hyperplane. In the classification process, the mapping input vectors located on the separation hyperplane side of the feature space fall into one class, and the positions fall into the other class on the other side of the plane. In the case of data points that are not linearly separable, the SVM uses appropriate kernel functions to map them into higher dimensional spaces so that they become separable in those spaces.

B. K-Nearestneighbor

The kNN classifier is based on a distance function that measures the difference or similarity between two instances. The standard Euclidean distance d(x, y) between two instances x and y is defined as:

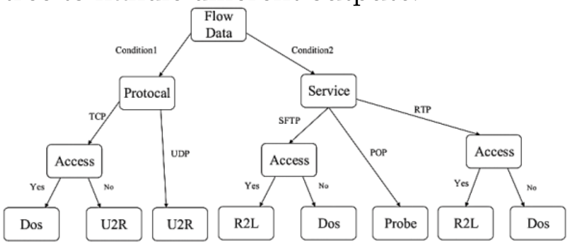
$$d(x,y)=\sqrt{\ }\ \sum_{k=1}^{n}(xk-yk)^{2}$$

Where, $\mathbf{x_k}$ is the kth featured element of instance **x**, $\mathbf{y_k}$ is the kth featured element of the instance **y** and **n** is the total number of features in the dataset.

Assume that the design set for kNN classifier is U. The total number of samples in the design set is S. Let C = {C1,C2,…CL } are the L distinct class labels that are available in S. Let **x** be an input vector for which the class label must be predicted. Let $\mathbf{y_k}$ denote the kth vector in the design set S. The kNN algorithm is to find the k closest vectors in design set S to input vector **x**. Then the input vector **x** is classified to class Cj if the majority of the k closest vectors have their class as Cj.

C. Decision Tree

A decision tree is a tree structure in which each internal node represents a test on one property and each branch represents a test output, with each leaf node representing a category. In machine learning, the decision tree is a predictive model; it represents a mapping between object attributes and object values. Each node in the tree represents an object, each divergence path represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree only has a single output; if you want complex output, you can establish an independent decision tree to handle different outputs.



D. Deep Belief Network

Deep Belief Network (DBN) is a probabilistic generative model consisting of multiple layers of stochastic and hidden variables. The Restricted Boltzmann Machine (RBM) and DBN are interrelated because composing and stacking a number of RBMs enables many hidden layers to train data efficiently through activations of one RBM for further training stages. RBM is a special topological structure of a Boltzmann machine (BM). The principle of BM originated from statistical physics as a modeling method based on an energy function that can describe the high-order interactions between variables. BM is a symmetric coupled random feedback binary unit neural network composed of a visible layer and a plurality of hidden layers. The network node is divided into a visible unit and a hidden unit, and the visible unit and the hidden unit are used to express a random network and a random environment. The learning model expresses the correlation between units by weighting.

V. DISCUSSION AND FUTURE DIRECTION

Our work examines a large number of academic intrusion detection studies based on machine learning and deep learning as shown in Table 5. In these studies, many imbalances appear and expose some of the problems in this area of research, largely in the following areas: (i) the benchmark datasets are few, although the same dataset is used, and the methods of sample extraction used by each institute vary. (ii) The evaluation metrics are not uniform, many studies only assess the accuracy of the test, and the result is one-sided. However, studies using multi-criteria evaluation often adopt different metric combinations such that the research results cannot be compared with one another. (iii) Less consideration is given to deployment efficiency, and most of the research stays in the lab irrespective of the time complexity of the algorithm and the efficiency of detection in the actual network.

A. Data Sets

Existing datasets have the defects of old data, redundant information and unbalanced numbers of categories. Although the data can be improved after processing, there is a problem of insufficient data volume. Therefore, establishing network intrusion detection datasets with large amounts of data, wide-type coverage and balanced sample numbers of attack categories becomes a top priority in the field of intrusion detection.

B. Hybrid Method

Hybrid detection methods mostly combine machine-learning methods such as those described by, whereas intrusion detection with a combination of deep learning and machine-learning methods is less studied. AlphaGo has validated the validity of this idea, which is an exciting research direction.

C. Detection Speed

By reducing the detection time and improving the detection speed from the algorithm and hardware aspects, the algorithm can be used less time given the complexity of the machine-learning algorithm and deep learning algorithm. Hardware can use multiple computers for parallel computing. Combining the two approaches is also an interesting topic.

D. Online Learning

The means of network intrusion is increasing day by day. How to fit the new data better with the trained model is also an exciting research direction. At present, transfer learning is a viable means to fine-tune the model with a small amount of labeled data, which should be able to achieve better results in actual network detection.

VI. CONCLUSION

This paper presents a literature review of ML and DL methods for network security. The paper, which has mostly focused on the last three years, introduces the latest applications of ML and DL in the field of intrusion detection. Unfortunately, the most effective method of intrusion detection has not yet been established. Each approach to implementing an intrusion detection system has its own advantages and disadvantages, a point apparent from the discussion of comparisons among the various methods. Thus, it is difficult to choose a particular method to implement an intrusion detection system over the others. Datasets for network intrusion detection are very important for training and testing systems. The ML and DL methods do not work without representative data, and obtaining such a dataset is difficult and time-consuming. However, there are many problems with the existing public dataset, such as uneven data, outdated content and the like. These problems have largely limited the development of research in this area.

Network information update very fast, which brings to the DL and ML model training and use with difficulty, model needs to be retrained long-term and quickly. So incremental learning and lifelong learning will be the focus in the study of this field in the future.

VII. REFERENCES

1. S. Aftergood, "Cybersecurity: The cold war online", *Nature,* vol. 547, pp. 30-31, Jul. 2017.
2. Milenkoski, M. Vieira, S. Kounev, A. Avritzer and B. D. Payne, "Evaluating computer intrusion detection systems: A survey of common practices*", ACM Comput. Surv.*, vol. 48, no. 1, pp. 1-41, 2015.
3. N. Modi and K. Acha, "Virtualization layer security challenges and intrusion detection/prevention systems in cloud computing: A comprehensive review", *J. Supercomput.*, vol. 73, no. 3, pp. 1192-1234, 2017.
4. Viegas, A. O. Santin, A. França, R. Jasinski, V. A. Pedroni and L. S. Oliveira, "Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems*", IEEE Trans. Comput.*, vol. 66, no. 1, pp. 163-177, Jan. 2017.
5. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends", *Comput. Netw.*, vol. 51, no. 12, pp. 3448-3470, Aug. 2007.
6. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel and M. Rajarajan, "A survey of intrusion detection techniques in Cloud", *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 42-57, 2013.
7. S. Revathi and A. Malathi, "A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection", *Proc. Int. J. Eng. Res. Technol.,* pp. 1848-1853, 2013.
8. Sahoo, C. Liu and S. C. H. Hoi, Malicious URL detection using machine learning: A survey, 2017, [online] Available: https://arxiv.org/abs/1701.07179.
9. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection", *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153-1176, 2nd Quart. 2016.
10. M. Soni, M. Ahirwa and S. Agrawal, "A survey on intrusion detection techniques in MANET", *Proc. Int. Conf. Comput. Intell. Commun. Netw.*, pp. 1027-1032, 2016.
11. R. G. Smith and J. Eckroth, "Building AI applications: Yesterday today and tomorrow", *AI Mag.,* vol. 38, no. 1, pp. 6-22, 2017.
12. P. Louridas and C. Ebert, "Machine learning", *IEEE Softw.,* vol. 33, no. 5, pp. 110-115, Sep./Oct. 2016.
13. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends perspectives and prospects", *Science,* vol. 349, no. 6245, pp. 255-260, 2015.
14. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", *Nature,* vol. 521, pp. 436-444, May 2015.
15. G. E. Hinton, "Deep belief networks", *Scholarpedia,* vol. 4, no. 5, pp. 5947, 2009.
16. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
17. L. Deng and D. Yu, "Deep learning: Methods and applications", *Found. Trends Signal Process*, vol. 7, no. 3, pp. 197-387, Jun. 2014.
18. M. Coelho, V. N. Coelho, E. J. da S. Luz, L. S. Ochi, F. G. Guimarães and E. Rios, "A GPU deep learning metaheuristic based model for time series forecasting", *Appl. Energy,* vol. 201, no. 1, pp. 412-418, 2017.
19. Žliobaitė, A. Bifet, J. Read, B. Pfahringer and G. Holmes, "Evaluation methods and decision theory for classification of streaming data with temporal dependence", *Mach. Learn.,* vol. 98, no. 3, pp. 455-482, 2015.

ID: 40

USING THE RADON TRANSFORM, IDENTIFY THE TAMIL CHARACTERS IN INSCRIPTION

Mr.C.BALAJI
Assistant Professor,
Department of Computer Science,
Dr.NGP Arts and Science College,
Coimbatore, Tamil Nadu, India.

Dr.P.LALITHA
Professor,
Department of BCA,
Hindusthan College of Arts and Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT:

Numerous palm leaf manuscripts have been used to preserve the ancient knowledge and cultural history of Southeast Asian nations. They are easily prone to degradation because of different factors like aging, bug bites, stains, and other issues. Therefore, it is crucial to preserve and digitize such fragile documents. Such materials suffer from numerous noise artifacts when scanned or photographed traditionally. For such texts, a depth sensing method is suggested to reduce background noise. Utilizing statistical methods, the segmented characters derived from Tamil palm scripts are further identified. Utilizing the 3D feature (depth), it is believed that recognition accuracy has improved.

KEYWORDS: *Palm Leaf Character Recognition (PLCR), Radon Transform, 3D Features, Tamil Characters, Hand-Written Characters (HWC), character recognition (CR).*

I. INTRODUCTION:

One of the first uses of automatic pattern recognition is character recognition (CR). Character recognition has several uses, including automating libraries, banks, the military, reading aids for the blind, post offices, language processing, and multi-media design. Consequently, character recognition research is quite popular. For people, it is simple to recognize Hand-Written Characters (HWC), but for computers, it can be very challenging. This is mostly caused by the striking variations or imprecision linked to the handwritten designs created by various people [1].

Machine recognition, a field of continuing research, refers to a computer's capacity to accept input from sources like paper and other documents, photos, touch displays, and other devices. The two types of handwritten character recognition (HWCR) are "offline" and "online," where the image is sensed "off-line" from a portion of a document, and "online" handwritten character recognition, which records the movements of the pen/tip "on-line" as used in pen-based computer screen systems. The dynamic properties of the writing are provided via online recognition, which also sends temporal or dynamic information such as the quantity and sequence of pen-on and pen-off motions, the direction and speed of writing, and in certain circumstances, the pressure used to make a character, among other things. Prior to feature extraction and recognition stages, off-line recognition frequently necessitates faulty pre-processing approaches.

Numerous [1] applications for automated handwriting recognition include reading antique manuscripts HWF, digitizing handwritten text, reading forms, copying engravings, and writer identification. Religious scriptures and treatises on a variety of topics, including art, medicine, astronomy, astrology, mathematics, law, and music, can be found in palm leaf manuscripts [3]. These texts are very difficult to preserve for future years. Weather conditions (such as temperature and relative humidity), light, insects, and degradation are the main reasons [3].

A. Difficulties encountered in Palmleaf Handwriting Recognition

The enormous variety of HWCR affecting elements are directly related to the various experimental settings. The type of handwriting database used for experiments is one of the key variations. In some instances, researchers severely restricted their experiments by only using one person's handwriting, and in other instances, researchers did not conduct their tests using benchmark databases. The user's flexibility of expression when writing is the biggest obstacle to handwriting recognition.

The difficulties mentioned above are made more difficult by irregular handwriting, which also makes it more difficult to group symbols and identify relationships among them. It causes "Layout" issues that make it difficult to recognize the entire statement. This can be attributed to novice users, who typically give handwritten symbols too much room for placement and alignment. When symbols are corrected, deleted, or added, other types of irregular writings develop [6]. A situation like this could generate a very complex expression, which cannot be recognized even by humans [6] as shown in the figure 1.

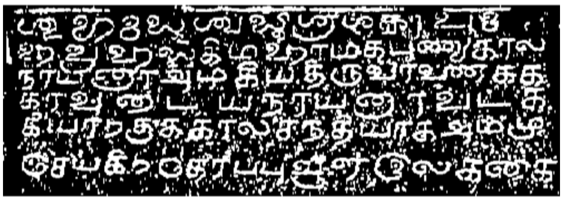


Figure 1

Preprocessing challenges include deciding whether to binarize raw handwriting or leave it in grey-scale form in order to transform it into a more efficient format. Another question is whether to thin out the handwriting or leave it alone in order to keep the features. The choice of the appropriate features to extract and the appropriate method to complete the task present additional challenges for feature extraction. For instance, researchers can decide between extracting a single feature, like the full contour of a character, or multiple features, including endpoints, loops, holes, and so on. The search for an appropriate classification method (for both individual characters and entire words) has now been thoroughly pursued. Again, though, the variety of handwriting and the dearth of effective feature extraction and preprocessing methods have limited the effectiveness of many unrestricted approaches. Pre-processing techniques like noise removal, skew correction, and segmentation add additional complexity to the majority of the aforementioned issues, including feature extraction and classification.

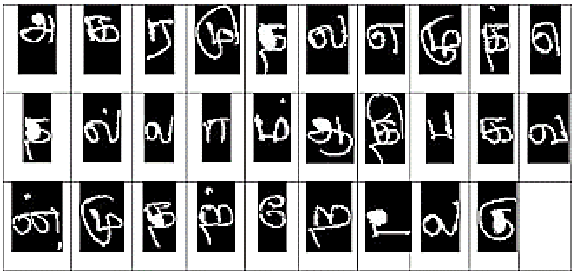
II. CHALLENGES WITH INDIAN LANGUAGES (LIKE TAMIL)

Over the past few decades, there has been a lot of research done on recognizing off-line handwritten characters in various scripts. These works are accessible for the English [7], Chinese [8], Arabic [9] and Kanji [10] scripts, among others. As compared to English, which is at a mature stage of development, character recognition in Asian and, in particular, Indian scripts is still in a very incipient stage [11, 12, 13]. This is because of the following factors:

- Compared to English, Indian languages have composite characters.
- Because of the sheer, number of English speakers, OCR in English is highly developed. With smaller number of speakers, languages like Tamil have not attracted equivalent efforts.
- In Tamil, consonants take modified shapes when attached with the vowels. Additionally, vertical extent of the character varies depending on the modifying vowel or consonant. Such characters are even more difficult for a machine to recognize.
- Non-uniformity in the spacing of the characters within a word due to the presence of consonant conjuncts (vowel + consonant) makes HWCR more difficult. Also, the presence of consonant conjuncts results in improper line segmentation. Recognition programs need to perform further processing to segment the lines.
- In scripts like Devnagari, all the characters in a word are connected by a unique line called shirorekha (also called head line). Word separation and line separation is easy in these cases.

Tamil is one of the prominent scripts in India with more than 82 million worldwide speakers [11]. There are 12 vowels, 18 Consonants, and 216 consonantal vowels in this language. Some of the Tamil characters (vowels) are shown below in table 1.

Table 1



III. RADON TRANSFORM

Let ¦([) = ¦([, \) be a continuous function vanishing outside some large disc in the Euclidean plane R^2. The Radon transform [14], 5¦, is a function defined on the space of straight lines ⁄ in R^2 by the line integral along each suchline:

| | 5I (⁄) □ □ I ([) G[

Concretely, any straight line / can be parameterized by

([(W), \(W) □ ((W sin□ □ V cos□), (□W cos□ □ V sin□)) (2)

Where V is the distance of ⁄ from the origin and is the angle the normal vector to ⁄ makes with the [axis. It follows that the quantities (Į,V) can be considered as coordinates on the space of all lines in R^2, and the Radon transform can be expressed in these coordinates by

$$\infty \quad 5I(\alpha, V) = \int I([(W), \backslash(W))GW$$

$$-\infty\ \infty = \int I((W\sin\alpha + V\cos\alpha), (-W\cos\alpha + V\sin\alpha))GW$$

$$-\infty$$

More generally, in the Q-dimensional Euclidean space "R^Q,the Radon transform of a compactly supported continuous function ¦ is a function 5¦ on the space Ì$_Q$ of all hyper planes in R^Q. It is defined by

$$5I(\xi) = \int I([)G\sigma([)$$

for $\xi \in \sum_Q$, where the integral is taken with respect to the natural hyper surface measure, G1 (generalizing the |G[| term from the 2-dimensional case). If any element of Ì$_Q$ is characterized as the solution locus of an equation

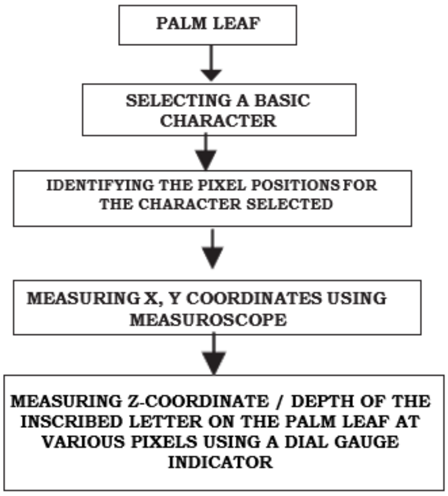
$$[.\alpha = V$$

Where $\alpha \in V^{Q-1}$ is a unit vector and $V \in 5$. Thus the Q- dimensional Radon transform may be rewritten as a function on $6^{Q\hat{1}1}$×R via

$$5I(\alpha, V) = \int_{[.\alpha = V} I([)G\sigma([)$$

It is also possible to generalize the Radon transform still further by integrating instead over N-dimensional affine subspaces of R^Q. The X-ray transform is the most widely used special case of this construction, and is obtained by integrating over straight lines.

IV. DATA ACQUISITION AND DIGITIZATION


PALM LEAF
SELECTING A BASIC CHARACTER
IDENTIFYING THE PIXEL POSITIONS FOR THE CHARACTER SELECTED
MEASURING X, Y COORDINATES USING MEASUROSCOPE
MEASURING Z-COORDINATE / DEPTH OF THE INSCRIBED LETTER ON THE PALM LEAF AT VARIOUS PIXELS USING A DIAL GAUGE INDICATOR

The first step involves measuring and recording the distance to the bottom of the pixel point. This separation is known as D1. Then, a distance known as D2 is determined by placing this needle on the palm leaf's surface closer to the pixel point. Subtracting D2 from D1 yields the indentation depth D for the chosen pixel position. As a result, D = D1 - D2 is the depth of indentation at any given pixel coordinate. At each pixel location where an X and Y measurement was acquired for the Palm Leaf Character, this operation is repeated. Therefore, there are three dimensions—X, Y, and Z—for every Palm leaf character at every given pixel location.

TABLE. 2: XYZ coordinates of Aa

3i[H0 SRiQWV	; (PP)	< (PP)	= (□P)
1	1.091	0.16	25
2	1.456	0.49	24
3	0.925	0.999	26
4	0.338	0.725	29
5	0	0	29
6	0.338	-0.547	28
7	1.832	-0.825	27
8	2.797	-0.547	28
9	3.002	0.396	29
10	2.51	0.756	33
11	2.281	0.306	28
12	3.042	-0.087	34
13	2.098	-0.047	34
14	0.741	-0.06	36
15	0.741	-0.08	38

TABLE 3: XYZ coordinates of Tha

3i[H0 SRiQWV	; (PP)	< (PP)	= (□P)
1	0.308	-0.400	26
2	0.687	0.066	26
3	0.300	0.327	34
4	0.000	0.000	31
5	0.188	-0.426	20
6	0.418	-0.789	25
7	0.833	-0.658	25
8	1.428	-0.618	39
9	1.620	-0.423	38
10	1.670	-0.144	35
11	1.180	0.412	38
12	1.310	-0.127	39
13	1.400	0.390	28
14	0.842	0.798	34
15	0.284	0.876	96
16	0.842	0.812	94
17	1.345	1.342	48
18	1.949	1.554	59

The various steps in the preprocessing stage of the image obtained from *0iFURVRIW ([FH0* chart wizard are as follows:

1. The image is first copied into a file of *$GREH 3KRWRVKRS* and an image name is assigned.
2. Using the concept of minimum boundary rectangle and the rectangular cropping tool of the *$GREH 3KRWRVKRS* the image is cropped. This image is first changed to gray scale mode and then brightness and contrast is adjusted to maximum value. The mode of this image is then changed to bitmap with a final resolution of 72 pixels / inch obtained by diffusion dither method.
3. In the next step the size of the image is converted to a 50 X 50 pixel size using Bicubic method, each image contained in a height and width of 50 pixels each. This image is further stored at the appropriate folder in the system.
4. The process described in the above steps is repeated for all the images of all the characters using all the three types of images (XY, YZ and ZX).

IV. Implementation and Experimental Results of Radon Transform

The first step involves measuring and recording the distance to the bottom of the pixel point. This separation is known as D1. Then, a distance known as D2 is determined by placing this needle on the palm leaf's surface closer to the pixel point. Subtracting D2 from D1 yields the indentation depth D for the chosen pixel position. As a result, D = D1 - D2 is the depth of indentation at any given pixel coordinate. At each pixel location where an X and Y measurement was acquired for the Palm Leaf Character, this operation is repeated. Therefore, there are three dimensions—X, Y, and Z—for every Palm leaf character at every given pixel location. The measureoscope and the previously stated equipment were used to methodically compile the pixel co-ordinates of the Tamil letters written on palm leaves. Different patterns of the Tamil palm leaf characters are formed in the XY, YZ, and XZ planes using the pixel coordinate values (X, Y, Z). Tamil characters are divided into twenty-nine different classes, with four samples of each class being used to train the database for the suggested approach.

TABLE.4 Percentage accuracy of recognition for two sets of palm leaves

	Palm Leaf Set -1 (% Accuracy ofrecognition)	Palm Leaf Set -2 (% Accuracy ofrecognition)
XY Plane	76	71
YZ Plane	89	82
XZ Plane	80	83

REFERENCES

1. Senior and Robinson, "An Off-Line Cursive Handwriting Recognition System", IEEE Transactions on Pattern analysis and Machine Intelligence, Vol.20, No.3, 1998, pp. 309-321.
2. Wakahara et al, "On-line handwriting recognition", Special Issue of Proc. Of the IECC, Vol.80, No. 7, 1992, pp.1181-1194.
3. Shi Zhixin, Setlur Srirangaraj and Govindaraju Venu. 2005. Digital Image Enhancement Using Normalization Techniques and their Application to Palm Leaf Manuscripts. CEDAR. Center For Excellence for Document Analysis and Recognition. New York. U.S.A.
4. Ujjwal Bhattacharya and B.B.Chaudhuri, Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals, IEEE transcations on pattern analysis and machine intelligence, Vol.31 No.3, pp.444-457, March 2009.
5. V.N.Manjunath Aradhya, G.Hemantha Kumar, S.Noushath, "Multilingual OCR system for South Indian Scripts and English documents: An approach based on Fourier transform and PCA", Elsevier, Engineering applications of artificial intelligence, 2008, pp. 658-668.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. S.N. Srihari, E.Cohen, J.J.Hull and L..Kaun, "A System to locate and Recognize ZIP Codes in Handwritten Addresses," Int'l J. Research and Eng.-Postal Applications, Vol. 1, 1989, pp. 37-45.
8. J.Tsukumo and H.Tanaka, "Classification of Hand printed Chinese Characters Using Nonlinear Normalization Methods," Proc. Ninth Int'l Conf. Pattern Recognition, 1988, pp. 168-171.

A. Amin and H.B. Al-Sadoun, "Hand Printed Arabic Character Recognition System," Proc. 12th Int'l Conf. Pattern Recognition, 1994, pp. 536-539.

9. H.Yamada, K.Yamamoto and T.Saito, "A Non-Linear Normalization Method for Hand printed Kanji Character Recognition—Line Density Equalization," Pattern Recognition, Vol.23, 1990, pp.1023-1029.
10. Panyam Narahari Sastry, Ramakrishnan Krishnan, Bhagavatula Venkata Sanker Ram, Tamil Character Recognition on Palm Leaves- A three dimensional Approach Technology Spectrum (JNTU Hyderabad), Vol. 2, No. 3, pp.19-26, November 2008.
11. Panyam Narahari Sastry, Ramakrishnan Krishnan and Bhagavatula Venkata Sanker Ram, Classification and Identification of Tamil hand written characters extracted from palm leaves using decision tree approach, ARPN Journal of Engineering and Applied Sciences, Vol. 5, No. 3, March 2010.
12. Panyam Narahari Sastry, Ramakrishnan Krishnan and T.V.Rajinikanth "Palm leaf Tamil Character Recognition using Hough Transform" Proceedings of International Conference on Advanced Computing Methodologies (ICACM-2011), Elsevier Publication, December 2011, pp 21-28.
13. V.N.Manjunath Aradhya, G.Hemantha Kumar, S.Noushath, "Robust Unconstrained Handwritten Digit Recognition Using Radon Transform", IEEE-ICSCN, 2007, pp. 626-629

ID: 41

LIGHTWEIGHT CRYPTOGRAPHY FOR RESOURCE-CONSTRAINED DEVICES: A BALANCING ACT BETWEEN SECURITY AND EFFICIENCY

Ms.B.USHA
Research Scholar,
Kalasalingam Academy of Research and Education,
Assistant Professor, Department of MCA,
Fatima College (Autonomous),
Madurai, Tamil Nadu, India.

ABSTRACT:
With the rapid proliferation of Internet of Things (IoT) devices, there is an increasing demand for secure communication and data protection in resource-constrained environments. However, traditional cryptographic algorithms often impose excessive computational and memory overhead on IoT devices with limited resources. This paper presents a comprehensive analysis of lightweight cryptographic algorithms specifically tailored for resource-constrained IoT devices with a focus on their efficiency, security, and suitability for IoT applications.

KEYWORDS: *Block cipher, Stream cipher, Feistel structure, SPN, Key size*

I. Introduction:

In today's hyper-connected world, where Internet of Things (IoT) devices and wearables have become an integral part of our lives, the demand for secure communication and data protection has surged exponentially. However, these resource-constrained devices, such as smart sensors, RFID tags, and low-power microcontrollers, often face a significant challenge in implementing strong cryptographic algorithms due to their limited computational capabilities and memory constraints. To address this issue, researchers and cryptographic experts have been developing "Lightweight Cryptography" – a specialized area of study that focuses on striking a delicate balance between security and efficiency for devices with scarce resources.

II. UNDERSTANDING LIGHTWEIGHT CRYPTOGRAPHY:

Lightweight cryptography aims to provide robust security while minimizing the computational burden and memory requirements of cryptographic operations. Traditional cryptographic algorithms, like AES (Advanced Encryption Standard) and RSA (Rivest-Shamir-Adleman), are designed to provide a high level of security but demand considerable computational resources, which is unsuitable for devices with limited processing power.

In contrast, lightweight cryptographic solutions are tailored to meet the specific needs of resource-constrained devices. These solutions encompass a range of algorithms and techniques, including lightweight block ciphers, hash functions, authentication protocols, and key exchange mechanisms.Lightweight cryptography,is largely divided into two categories, symmetric andasymmetric algorithms.The symmetric lightweight algorithmsare further divided into Lightweight Block Ciphers (LWBC) andLightweight Stream Ciphers (LWSC). Elliptic curve cryptography (ECC) falls under asymmetric cryptography. The factors of thelightweight cryptographic primitives are evaluated by the keysize, block size, number of rounds, and structures.

1.1 Lightweight block cipher

Block cipher is a symmetric encryption that processes a whole block of data all at once. Substitution-permutation networks (SPN) and Feistel networks (FN) are two types of networks that use lightweight block ciphers. The same circuit is used for both encryption and decryption in a Feistel structure, which reduces expenses. Feistel structures guarantee low memory requirements by employing the same computer code for both encryption and decryption operations. Without a key schedule, the SPN is quicker, but this leaves the system open to intrusions. Due to its lower power consumption and less execution needs, the SPN structure is more suited for security. The main factors to consider while assessing a lightweight block cipher are key size, block size, structure type, and the encryption/decryption rounds.

1.2 Lightweight stream cipher

Data is bit by bit encrypted and decrypted using a lightweight stream cipher. When compared to other cipher techniques, it is incredibly rapid and easy. These ciphers are widely used in mobile devices and wireless sensor networks. They are favored because they require less calculation. They create a pseudo-random keystream, which is XORed with the plaintext to create ciphertext. XOR is suited for devices with low processing and memory capacity due to its simplicity and efficient architecture. A secret key and careful design are required for security to ensure a lengthy, unpredictable keystream time. For lightweight applications, these ciphers promote security and efficiency.

1.3 Lightweight elliptic curve cipher

ECC and other asymmetric ciphers are also used to protect IoT networks. ECC guarantees confidentiality and authenticity. Asymmetric cryptography tools like RSA and ECC can be used to secure IoT networks. Compared to RSA, ECC achieves a same level of security using a smaller key size. Compared to more conventional techniques like RSA or DSA, ECC offers strong security with reduced key lengths by utilizing the mathematical features of elliptic curves. When it comes to assuring security in devices with limited resources, ECC is currently emerging as the method of choice.

II. KEY FEATURES OF LIGHTWEIGHT CRYPTOGRAPHY:

a. Minimal Resource Utilization: The foremost characteristic of lightweight cryptography is its minimal resource usage. These algorithms are specifically designed to require less memory and processing power, ensuring efficient execution on devices with restricted hardware capabilities.

b. Fast Execution: Lightweight cryptographic primitives are optimized for speed, ensuring that cryptographic operations can be performed swiftly even on low-powered devices. The reduced computation time contributes to energy conservation, extending the battery life of the resource-constrained devices.

c. Hardware-Friendly: Lightweight cryptographic algorithms are crafted with hardware implementations in mind, allowing for efficient execution in hardware modules, such as integrated circuits and Field-Programmable Gate Arrays (FPGAs).

d. Small Code Footprint: Another critical aspect of lightweight cryptography is the small code footprint. This ensures that the cryptographic libraries and protocols occupy minimal memory space on the device, freeing up valuable memory for other essential functionalities.

III. Challenges and Trade-Offs:

Developing lightweight cryptography involves a delicate trade-off between security and efficiency. While these algorithms provide a practical solution for resource-constrained devices, they inherently sacrifice some level of security to achieve their lightweight nature. Striking the right balance becomes paramount to safeguarding sensitive data without compromising the device's performance.

Some of the challenges faced in lightweight cryptography include:

a. Reduced Key Sizes: Lightweight ciphers often employ shorter key lengths, making them susceptible to brute-force attacks. Therefore, careful design and analysis are crucial to maintain an acceptable security level.

b. Vulnerability to Side-Channel Attacks: Resource-constrained devices are more susceptible to side-channel attacks, where attackers exploit power consumption or timing information to extract cryptographic keys. Implementing countermeasures against these attacks becomes a priority.

c. Standardization and Adoption: Establishing widely accepted standards for lightweight cryptography is essential to ensure interoperability and widespread adoption among different devices and platforms.

IV. EXAMPLES OF LIGHTWEIGHT CRYPTOGRAPHIC ALGORITHMS FOR IOT:

a. Tiny Encryption Algorithm (TEA): TEA is a simple and compact block cipher designed for efficient encryption and decryption on low-resource devices.

b. SIMON and SPECK: These are families of lightweight block ciphers developed by the NSA, offering a range of key sizes and block sizes suitable for different IoT applications.

c. QUAD: QUAD is a lightweight hash function designed for IoT devices with a small code footprint.

d. HC-128 and HC-256: These are lightweight stream ciphers with high security levels, suitable for resource-constrained devices.

e. PRESENT: It is a light weight block cipher that uses bit-oriented permutations. It is mostly used in hardware where in simple wiring is used to implement bit permutations.

f. SLIM: SLIM is a light weight stream cipher particularly targeted on data integrity.It uses short keys for more efficiency.

g. TWINE: TWINE is a block cipher that supports varying key lengths. It supports Feistel network structure that includes both confusion and diffusion layers to enhance security

h. RECTANGLE: RECTANGLE is a block cipher employs a Substitution-Permutation Network structure, RECTANGLE has variants with different block sizes and key lengths, allowing users to choose the appropriate configuration for their specific use cases.

i. Klein: KLEIN (Keyed-permutation Lightweight block cipher for Embedded Networks) is a lightweight block cipher designed primarily as a lightweight alternative to more established block ciphers like AES (Advanced Encryption Standard).KLEIN is based on a Feistel network structure, that uses a round function that combines the input data and the subkey to produce the ciphertext.

j. BORON: BORON is a block cipher that uses a Feistel Network structure which involves multiple rounds of operations on the input block using round keys. The design aims to provide both confusion and diffusion properties for security.

k. CLEFIA: CLEFIA uses a substitution-permutation network (SPN) structure, which employs various rounds of substitutions and permutations to achieve confusion and diffusion properties.CLEFIA supports both big-endian and little-endian platforms, making it suitable for a wide range of computing environments.

l. LICI: LICI is a Feistel based block cipher that performs without flaws on both hardware and software platforms. This algorithm has a good resistance to the linear differential attacks.

VI. CONCLUSION:

Lightweight cryptography represents a critical field of research and development, catering to the security needs of resource-constrained devices in an increasingly connected world. Striking a balance between efficiency and security is the primary challenge faced by cryptographers and researchers in this domain. As the demand for secure communication and data protection in IoT and wearable devices continues to grow, ongoing efforts to refine and standardize lightweight cryptographic algorithms will be paramount in safeguarding our data and privacy in a resource-limited environment.

VII. REFERENCES:

1. Regla, A. I., &Festijo, E. D. (2022, April). Performance analysis of light-weight cryptographic algorithms for internet of things (IOT) applications: A systematic review. In 2022 IEEE 7th International conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.
2. Naser, N. M., & Naif, J. R. (2022). A systematic review of ultra-lightweight encryption algorithms. International Journal of Nonlinear Analysis and Applications, 13(1), 3825-3851.
3. Hasan, H., Ali, G., Elmedany, W., & Balakrishna, C. (2022, November). Lightweight Encryption Algorithms for Internet of Things: A Review on Security and Performance Aspects. In 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 239-244). IEEE.
4. Jammula, M., Vakamulla, V. M., &Kondoju, S. K. (2022). Performance evaluation of lightweight cryptographic algorithms for heterogeneous IoT environment. Journal of Interconnection Networks, 22(Supp01), 2141031.
5. Rana, M., Mamun, Q., & Islam, R. (2022). Lightweight cryptography in IoT networks: A survey. Future Generation Computer Systems, 129, 77-89.

ID: 42

UNDERSTANDING NATURAL LANGUAGE PROCESSING FOR TEXT AND SPEECH

Dr.A.INDUMATHI
Associate Professor and Head,
Department of Computer Technology,
Kongunadu Arts and Science College,
Coimbatore, Tamil Nadu, India.

Dr.R.TAMILSELVI
Associate Professor and Head,
Department of Artificial intelligence and Data Science,
VET Institute of Arts and Science College,
Erode, Tamil Nadu, India.

ABSTRACT

Natural language processing (NLP) is the set of methods for making human language accessible to computers. In the past decade, natural language processing has become embedded in our daily lives: automatic machine translation is ubiquitous on the web and in social media; text classification keeps our email in boxes from collapsing under a deluge of spam; search engines have moved beyond string matching and network analysis to a high degree of linguistic sophistication; dialog systems provide an increasingly common and effective way to get and share information. These diverse applications are based on a common set of ideas, drawing on algorithms, linguistics, logic, statistics, and more. NLP is a type of Artificial Intelligence.

KEYWORDS: *Natural language processing (NLP), Artificial Intelligence, Text Classification*

I. INTRODUCTION

Natural Language Processing (NLP) – the application of software systems to examining, interpreting and accurately responding to speech is viewed as the next big leap in user interface technology. However, human speech is far more complex than most people realize. There are rules, such as spelling and grammar. In human language, the words can say one thing, but the context and the tone make those words mean something else. It takes humans a half a lifetime to learn the subtle nuances of language. Even then, there will be words and phrases that some of us don't understand. Then, there are further complications in understanding language, such as dialects and colloquialisms. So, how can a computer that "thinks" in binary be programmed, line by line, to become fluent in any language? But with the help of Artificial Intelligence (AI), a computer can now learn how to understand a language.

What is Natural Language Processing?

Natural language processing, shortened to NLP, is the science of building machines that can read and write similar to how humans do. The goal of NLP is to organize massive amounts of text data (ie) millions of documents into information so humans can use it to solve problems. Since machines can perform tasks much faster than humans, we can work with machines to read massive amounts of text in a fraction of the time. While NLP is just *one* type of artificial intelligence that is becoming important for analysts to understand, it's a critical piece to solving lots of problems.

II. RELATIONSHIP BETWEEN AI, ML, DL AND NLP

Today, Technology has been revolutionized in our lives and its constantly changing and progressing. The most flourishing technologies include Artificial Intelligence, Machine Learning, Natural Language Processing, and Deep Learning.These terms are generally used together in some contexts but do not mean the same and are related to each other in some or the other way. ML is one of the leading areas of AI which allows computers to learn by themselves and NLP is a branch of AI. The below Figure illustrates how AI, ML, DL and NLP overlap with each other.

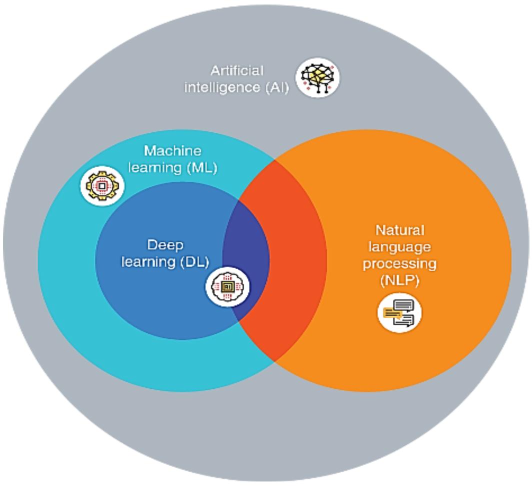

Artificial intelligence (AI)
Machine learning (ML)
Deep learning (DL)
Natural language processing (NLP)

Fig.1 Relationship between AI, ML, Dl and NLP

Artificial Intelligence (AI) is a broad umbrella field that is concerned with achieving human like intelligence using machines. It encompasses a wide range of subfields, including Machine Learning, Natural Language Processing, Computer Vision and Speech Recognition. The field also includes subfields such as reasoning, planning and search, which do not fall under either Machine Learning or Natural Language Processing.

Machine Learning (ML) is usually considered a subfield of Artificial Intelligence that is about improving computer algorithms through experience and data.This includes learning a general function that maps inputs to outputs based on past experience (Supervised Learning), drawing hidden patterns and structures from data (Unsupervised learning),and learning how to act in a dynamic environment based on indirect rewards (Reinforcement Learning). Supervised Machine Learning is the main paradigm for training NLP Models.

Deep Learning (DL) is a subfield of machine learning that usually uses deep Neural Networks.These Neural Network models are called "deep" because they consist of a number of layers. A Layer is a substructure of Neural Networks. With Many stacked Layers, deep Neural Networks can learn complex representations of data and can capture highly complicated relationships between the input and the output.

When the amount of available data and computational resources increases, NLP makes a heavier and heavier use of Machine Learning and Deep Learning. NLP applications and tasks are usually built on top of Machine Learning Pipelines and trained from data. But NLP does not overlap with Machine Learning.

Computational Linguistics (CL) and Text Mining are other fields that are related to NLP. Computational Linguistics is a subfield of linguistics that uses computational approaches to study human language.CL is concerned with engineering approaches for making computers perform something useful related to language. Text Mining is a type of data mining targeted at textual data. Its focus is on drawing useful insights from unstructured textual data, which is a type of text data that is not formatted in a form that is easily interpretable by computers. Such data is usually collected from various sources like Web and Social media.

Why is NLP important?

In the past, computers could only work with structured languages. The language had to be precise and unambiguous. To program a computer to perform any task, you had to give it clear instructions. You could only use the limited number of commands that the computer understood. The syntax had to be perfect as well.

Even an end-user of a computer program needs to give the computer precise commands. Those who are old enough will remember that to use a PC you once had to know the common MS-DOS commands. That barrier was overcome, to a degree, with graphical user interfaces, such as Windows. Now, we can point to a file with a mouse, instead of having to know the name of the file. NLP promises to remove the need for being so precise. Instead of having to learn the computer's language, the computer will learn how to understand ours. A very basic application of NLP will be how we interface with computers.

How does Natural Language Processing work ?

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors such as ears to hear and eyes to see computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.There are two main phases to natural language processing: data preprocessing and algorithm development.

Data preprocessing involves preparing and "cleaning" text data for machines to be able to analyze it. Preprocessing puts data in workable form and highlights features in the text that an algorithm can work with. There are several ways this can be done, including:

- **Tokenization.** This is when text is broken down into smaller units to work with.
- **Stop word removal.** This is when common words are removed from text so unique words that offer the most information about the text remain.
- **Lemmatization and stemming.** This is when words are reduced to their root forms to process.
- **Part-of-speech tagging.** This is when words are marked based on the part-of speech they are such as nouns, verbs and adjectives.

Once the data has been pre-processed, an algorithm is developed to process it. There are many different natural language processing algorithms, but two main types are commonly used:

- **Rules-based system.** This system uses carefully designed linguistic rules. This approach was used early on in the development of natural language processing, and is still used.
- **Machine learning-based system.** Machine learning algorithms use statistical methods. They learn to perform tasks based on training data they are fed, and adjust their methods as more data is processed. Using a combination of machine learning, deep learning and neural networks, natural language processing algorithms hone their own rules through repeated processing and learning.

Techniques and Methods of Natural Language Processing

Syntax and semantic analysis are two main techniques used with natural language processing.

Syntax is the arrangement of words in a sentence to make grammatical sense. NLP uses syntax to assess meaning from a language based on grammatical rules. Syntax techniques include:

- **Parsing.** This is the grammatical analysis of a sentence. *Example*: A natural language processing algorithm is fed the sentence, "The dog barked." Parsing involves breaking this sentence into parts of speech i.e., dog = noun, barked = verb. This is useful for more complex downstream processing tasks.
- **Word segmentation.** This is the act of taking a string of text and deriving word forms from it. *Example:* A person scans a handwritten document into a computer. The algorithm would be able to analyze the page and recognize that the words are divided by white spaces.
- **Sentence breaking.** This places sentence boundaries in large texts. *Example:* A natural language processing algorithm is fed the text, "The dog barked. I woke up." The algorithm can recognize the period that splits up the sentences using sentence breaking.
- **Morphological segmentation.** This divides words into smaller parts called morphemes. *Example:* The word untestably would be broken into [[un[[test]able]]ly], where the algorithm recognizes "un," "test," "able" and "ly" as morphemes. This is especially useful in machine translation and speech recognition.
- **Stemming.** This divides words with inflection in them to root forms. *Example:* In the sentence, "The dog barked," the algorithm would be able to recognize the root of the word "barked" is "bark." This would be useful if a user was analyzing a text for all instances of the word bark, as well as all of its conjugations. The algorithm can see that they are essentially the same word even though the letters are different.

Semantics involves the use of and meaning behind words. Natural language processing applies algorithms to understand the meaning and structure of sentences. Semantics techniques include:

- **Word sense disambiguation.** This derives the meaning of a word based on context. *Example:* Consider the sentence, "The pig is in the pen." The word pen has different meanings. An algorithm using this method can understand that the use of the word *pen* here refers to a fenced-in area, not a writing implement.

- **Named entity recognition.** This determines words that can be categorized into groups. *Example:* An algorithm using this method could analyze a news article and identify all mentions of a certain company or product. Using the semantics of the text, it would be able to differentiate between entities that are visually the same. For instance, in the sentence, "Daniel McDonald's son went to McDonald's and ordered a Happy Meal," the algorithm could recognize the two instances of "McDonald's" as two separate entities one a restaurant and one a person.
- **Natural language generation.** This uses a database to determine semantics behind words and generate new text. *Example:* An algorithm could automatically write a summary of findings from a business intelligence platform, mapping certain words and phrases to features of the data in the BI platform. Another example would be automatically generating news articles or tweets based on a certain body of text used for training.

Current approaches to natural language processing are based on deep learning, a type of AI that examines and uses patterns in data to improve a program's understanding. Deep learning models require massive amounts of labeled data for the natural language processing algorithm to train on and identify relevant correlations, and assembling this kind of big data set is one of the main hurdles to natural language processing.

Earlier approaches to natural language processing involved a more rules-based approach, where simpler machine learning algorithms were told what words and phrases to look for in text and given specific responses when those phrases appeared. But deep learning is a more flexible, intuitive approach in which algorithms learn to identify speakers' intent from many examples almost like how a child would learn human language.

Challenges of Natural Languages Processing

There are a number of challenges of natural language processing and most of them boil down to the fact that natural language is ever evolving and always somewhat ambiguous. They include:

- **Precision.** Computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise; it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.
- **Tone of voice and inflection.** Natural language processing has not yet been perfected. For example, semantic analysis can still be a challenge. Other difficulties include the fact that the abstract use of language is typically tricky for programs to understand. For instance, natural language processing does not pick up sarcasm easily. These topics usually require understanding the words being used and their context in a conversation. As another example, a sentence can change meaning depending on which word or syllable the speaker puts stress on. NLP algorithms may miss the subtle, but important, tone changes in a person's voice when performing speech recognition. The tone and inflection of speech may also vary between different accents, which can be challenging for an algorithm to parse.
- **Evolving use of language.** Natural language processing is also challenged by the fact that language and the way people use it is continually changing. Although there are rules to language, none are written in stone, and they are subject to change over time. Hard computational rules that work now may become obsolete as the characteristics of real-world language change over time.

CONCLUSION

Natural language processing is a major leap forward in AI technology. It removes the communication barrier that has always existed between machines and humans. The potential for the application of NLP in business is immense. A computer could now answer customer queries and take orders. Even if the customer uses obscure language. NLP is likely to remove the need for input devices, such as the keyboard and mouse as well. NLP matters, because it is about to revolutionize the way that we communicate with machines, and how they communicate with us.

REFERENCE

1. Liddy, E. D. (2001). Natural language processing.
2. Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. ONLINE-WESTON THEN WILTON-, 23, 62-73.
3. "Natural Language Processing." Natural Language Processing RSS. N.p., n.d. Web. 25 Mar. 2017
4. Hutchins, W. J. (1986). Machine translation: past, present, future (p. 66). Chichester: Ellis Horwood.
5. Hutchins, W. J. (Ed.). (2000). Early years in machine translation: memoirs and biographies of pioneers (Vol. 97). John Benjamins Publishing.
6. Green Jr, B. F., Wolf, A. K., Chomsky, C., & Laughery, K. (1961, May). Baseball: an automatic question-answerer. In Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference (pp. 219-224). ACM.

ID: 43

MULTIPLE CHAOTIC 3D KEYS TO PROTECT YOUR DIGITAL IMAGES

Mr.S.RAJA
Research Scholar,
Department of Computer Science,
Annamalai University,
Annamalainagar, Tamil Nadu, India.

Mr.C.ASHOK KUMAR
Assistant Professor,
Department of Computer and Information Science,
Annamalai University,
Annamalainagar, Tamil Nadu, India.

ABSTRACT

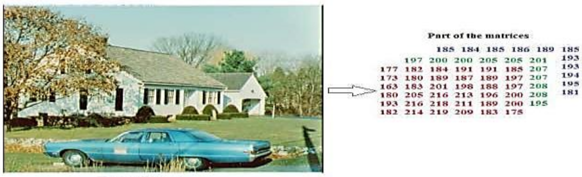
This research report presents a simple, versatile, efficient, sensitive, and high-quality image encryption method. Using this method, you can easily encrypt and decrypt any kind of image (gray and color images) of any size. The proposed method protects images by providing a high level of image protection based on the use of complex secret keys containing eight values each with dual data types. A private key provides the key space needed to withstand hacker attacks. The private key component is used to calculate the size of the generated chaotic 3D logistic key. The private key is also used to run two chaos logistic map models to obtain two different chaos keys based on the chaos parameter values contained in the private key. Two rounds of her XOR operation are used to produce the encrypted/decrypted image. The output produced by the proposed method is affected by the chosen private key value. The proposed method is very flexible because changing the source image or the secret key does not require changing the algorithm of the method. The proposed method is tested and implemented using different images. The results obtained are discussed to demonstrate the properties such as quality, sensitivity, safety and efficiency provided by this method.

KEYWORDS: *Cryptography, quality, sensitivity, performance, CLK, CLMM, MSE, PSNR, NPCR.*

I. INTRODUCTION:

Digital images, regardless of their nature (grey or color), are one of the most important types of data propagated through various means of communication, as they are used in many important applications. It is believed that there are many reasons for the popularity of digital images, the most important of which are: - Digital images can be used in a myriad of computer applications including, but not limited to, medical, educational, security, banking, and other applications. - Ease of retrieval and storage with many different sources available for free.

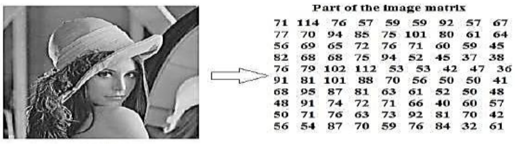
Availability of different devices with which any digital image can be easily created. For ease of digital image processing, gray images are represented by a 2D matrix and color images are represented by a 3D matrix, thus transforming the image processing into a matrix processing process. Increase the size of the digital image. It becomes a large data store that can be used for various purposes.



Part of the matrices

```
                185 184 185 186 189 185
        197 200 200 205 205 201     193
177 182 184 191 191 185 207         193
173 180 189 187 189 197 207         194
163 183 201 198 188 197 208         195
180 205 216 213 196 200 208         181
193 216 218 211 189 200 195
182 214 219 209 183 175
```

House.tiff image

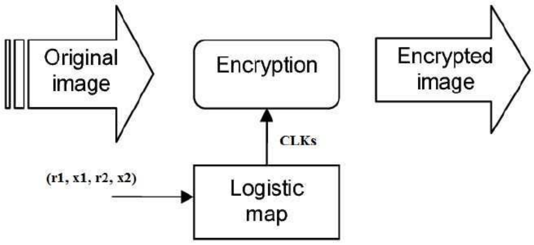


Part of the image matrix

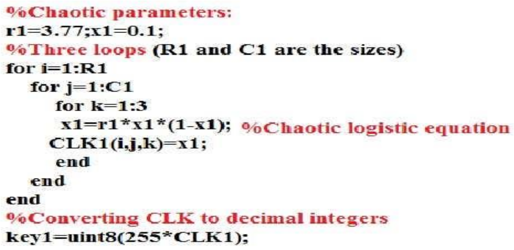
71	114	76	57	59	59	92	57	67
77	70	94	85	75	101	80	61	64
56	69	65	72	76	71	60	59	45
82	68	68	75	94	52	45	37	38
76	79	102	112	53	53	42	47	36
91	81	101	88	70	56	50	50	41
68	95	87	81	63	61	52	50	48
48	91	74	72	71	66	40	60	57
50	71	76	63	73	92	81	70	42
56	54	87	70	59	76	84	32	61

Lena image

Digital images must be protected because may be confidential. - may be of a special nature. May contain confidential information and is used as a stego image. For these reasons, and the importance of digital images, the number of people hiding in them is increasing, and it is necessary to provide the necessary protection, especially when the communication environment is insecure. Protection should prevent images from being hacked. One of the most popular methods of image protection is image encryption. It encrypts the image before sending it and decrypts it after receiving it. The image encryption system includes an original image, an encrypted (encrypted) image, an encryption function, and one or more private keys. The decryption system contains the encrypted image, the decrypted image, the decryption function, and the same PK used by the encryption system.



The encryption function manipulates the original image using the PK to produce a corrupted, scrambled, scrambled image. The decryption function, on the other hand, operates on the scrambled image using the same PK to produce a decrypted image that is identical to the original image. The chaotic logistic map makes it easy to generate the private keys required for encryption and decryption functions. The chaos parameters r and x are used to run a chaotic logistic map model (CLMM) to generate a chaotic logistic key (CLK). This key can be easily converted to a decimal integer to match the value of the digital image. You can generate a 1D matrix CLK with chaos parameters. These parameters can also be used to generate a 3D CLK by setting the size to a value of your choice. You can generate a 3D CLK using the following sequence of Mat lab operations.



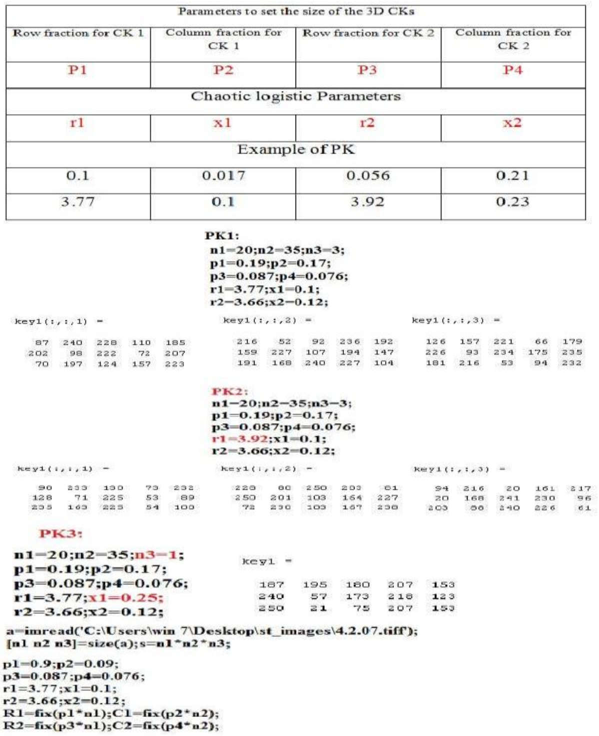
```
%Chaotic parameters:
r1=3.77;x1=0.1;
%Three loops (R1 and C1 are the sizes)
for i=1:R1
   for j=1:C1
      for k=1:3
       x1=r1*x1*(1-x1); %Chaotic logistic equation
      CLK1(i,j,k)=x1;
      end
   end
end
%Converting CLK to decimal integers
key1=uint8(255*CLK1);
```

The CLK technology system require a time and this time will develop up while the important thing length increases, so it's far encouraged to generate a small in length chaotic key, then the acquired key may be without difficulty resized to fit the dimensions of the photograph .Many authors used the concept of chaotic logistic idea to construct diverse schemas of photograph encryption-decryption; those strategies percentage a few precise capabilities which include pleasant of cryptography, protection with the aid of using supplying suitable key area. The delivered strategies furnished throughputs, which a few instances have been low and that they have been numerous from 19.7470 K bytes in line with second to 231.0469 K bytes in line with second .The purpose of this paper studies is to introduce a technique of photograph cryptography to obtain the following:- Multiple functions for use to encrypt-decrypt any photograph with any length and type.- Simplicity and flexibility, converting the photograph or/and converting the PK need to know no longer have an effect on the approach algorithm; additionally simplify the encryption and decryption features with the aid of using decreasing the range of used operations to use encryption and decryption processes. High performance, lowering the encryption-decryption time and growing the throughput of virtual photograph cryptography .High pleasant, the encrypted photograph need to be completely broken and scrambled, whilst the decrypted photograph need to be same to the supply photograph . High protection, growing the important thing area and making the produced outputs very touchy to the selected values of the PK.

Proposed Method The proposed method uses a complex PK containing eight double values,. Parameters P1, P2, P3, and P4 are split fractions used to select the row and column sizes of the two chaotic logistic keys (CLK) to be generated, and r and x drive the chaotic equations is the chaotic logistic parameter required for Check out the Chaos Logistic Map Model (KLMM).

The generated CLK must be converted to a decimal integer to form the secret key needed to apply the XOR operation. The size of the generated secret should be adjusted to the image size. The generated private key is very sensitive to the chosen values of the PK components.

If there is a small change in one or more values, a new key is generated as shown in the



Parameters to set the size of the 3D CKs			
Row fraction for CK 1	Column fraction for CK 1	Row fraction for CK 2	Column fraction for CK 2
P1	P2	P3	P4
Chaotic logistic Parameters			
r1	x1	r2	x2
Example of PK			
0.1	0.017	0.056	0.21
3.77	0.1	3.92	0.23

```
PK1:
n1=20;n2=35;n3=3;
p1=0.19;p2=0.17;
p3=0.087;p4=0.076;
r1=3.77;x1=0.1;
r2=3.66;x2=0.12;

key1(:,:,1) =                     key1(:,:,2) =                     key1(:,:,3) =

   87   240   228   110   185       216    52    92   236   192       126   157   221    66   179
  202    98   222    72   207       159   227   107   194   147       226    93   234   175   235
   70   197   124   157   223       191   168   240   227   104       181   216    53    94   232

PK2:
n1=20;n2=35;n3=3;
p1=0.19;p2=0.17;
p3=0.087;p4=0.076;
r1=3.92;x1=0.1;
r2=3.66;x2=0.12;

key1(:,:,1) =                     key1(:,:,2) =                     key1(:,:,3) =

   90   233   130    73   232       220    80   250   203    81        94   216    20   161   217
  128    71   225    53    89       250   201   103   164   227        20   168   241   230    96
  233   163   225    54   100        72   230   103   167   238       203    88   240   226    61

PK3:
n1=20;n2=35;n3=1;                 key1 =
p1=0.19;p2=0.17;
p3=0.087;p4=0.076;                  187   195   180   207   153
r1=3.77;x1=0.25;                    240    57   173   218   123
r2=3.66;x2=0.12;                    250    21    75   207   153

a=imread('C:\Users\win 7\Desktop\st_images\4.2.07.tiff');
[n1 n2 n3]=size(a);s=n1*n2*n3;

p1=0.9;p2=0.09;
p3=0.087;p4=0.076;
r1=3.77;x1=0.1;
r2=3.66;x2=0.12;
R1=fix(p1*n1);C1=fix(p2*n2);
R2=fix(p3*n1);C2=fix(p4*n2);
```

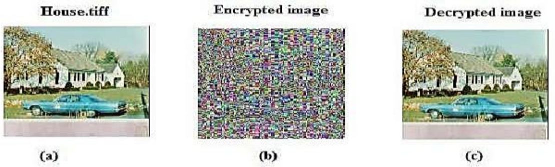
Step 4: Encrypt: XOR the image with the first key, XOR the result with the second key to get the encrypted image.

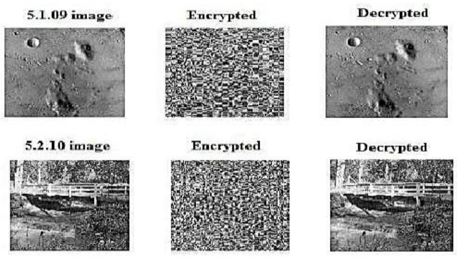
```
for i=1:R1
  for j=1:C1
    for k=1:n3
     x1=r1*x1*(1-x1);
    CLK1(i,j,k)=x1;
    end
  end
end
key1=uint8(255*CLK1);
key11=imresize(key1,[n1 n2]);
for i=1:R2
  for j=1:C2
    for k=1:n3
     x2=r2*x2*(1-x2);
    CLK2(i,j,k)=x2;
    end
  end
end
key2=uint8(255*CLK2);
key22=imresize(key2,[n1 n2]);

e1=bitxor(a,key11);
e=bitxor(e1,key22);
```

This step can be implemented using the following Mat-Lab sequence of operations: The decryption phase can be implemented using the same steps as for encryption, but with the encrypted image as the input image.

II. IMPLEMENTATION AND EXPERIMENTAL RESULTS

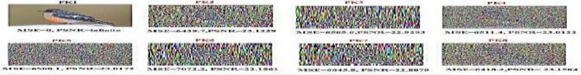
To demonstrate the proposed method, the method was applied using several image files from http://sipi.usc.edu/database/, also known as the USC-SIPI Image Database, and proposed the usefulness and feasibility of the zu analysis were demonstrated. Encryption scheme for gray and RGB images. All relevant experiments and simulations were performed in the MATLAB environment. For image encryption I chose a color image house.tiff of size 512x512. The original house image was encrypted using the proposed method the encrypted image, and decrypted image. The encrypted image is fully encrypted and corrupted do not leave a note revealing the original image information. The decoding results show that the proposed method works effectively and can perfectly restore the original image without losing any information from it.





A gray image was also implemented using the proposed method. The encrypted image obtained was completely corrupted, but the decrypted image was always identical to the source image. An example of printing the result of implementing the proposed method.

The quality of image cryptography can be measured using the computed mean squared error (MSE) between two images and the peak signal-to-noise ratio (PSNR) between the two images. In the encryption phase, MSE values should be high and PSNR values should be low. In the decoding phase, the value of MSE must be zero, but the value of PSNR must be infinite to prove that the proposed method is of sufficient quality As a requirement, some images were taken and MSE and PSNR values were calculated. Table 1 shows the results obtained during the encryption phase. In the decoding phase, MSE was always zero and PSNR was always infinite.



$Z(a,b) = 1$ if $C1 \neq C2$ and $= 0$ if $C1 = C2$ and H is the row size, W is the column size

$$NPCR = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} Z(a,b)}{H*W} \quad (4)$$

The resulting images demonstrate that the proposed method is highly sensitive to the chosen value of PK. Changing one or more values will produce a corrupted decoded image. The algorithm of the proposed method should be able to withstand differential attacks. This is he one of the most efficient and widely used methods for hackers to find important information between plain source images and encrypted images. Metrics such as the number of pixel changes (NPCR) are used to quantify the power of cryptographic algorithms and the effectiveness of their sensitivity when plain images are altered or modified at any stage during transmission or by various attacks. to test. NPCR determines the rate of change in the number of pixels between the two images C1 and C2.

III. CONCLUSION

An efficient method of image encryption was introduced, tested and implemented. The proposed method showed good performance by shortening the encryption/decryption time and improving the throughput of image encryption. Compared to other image encryption methods, this procedure provided significant speed improvements. The proposed method used a secret key with a complicated structure. This key was used to set the length of the 3D chaotic logistic keys and generate them. To minimize and save key generation time, the key size was kept small and adapted to the size of the image being encrypted/decrypted. The proposed technique provided a better level of safety than other existing techniques. Key space and sensitivity analysis showed that the proposed method can resist hacking attacks. The proposed method uses simplified encryption/decryption functions. These functions used two rounds of XOR operations to encrypt and decrypt the image. The proposed method was flexible, using grayscale and color images, and changing the image or private key did not require changing the encryption and decryption functions. The proposed method was tested and implemented using different images. The results obtained demonstrated the quality, sensitivity and efficiency of the method.

IV. REFERENCES

1. Ahmad, M., Alsharari, H. D., & Nizam, M. (2014). Security improvement of an image encryption based onmPixel-chaotic-shuffle and pixel-chaotic-diffusion. arXiv:1403.6626.
2. Thiyagarajan J, Murugan B, Gounden NGA. A chaotic image encryption scheme with complex diffusion matrixfor plain image sensitivity.*Serbian Journal of Electrical Engineering.* 2019;16(2):247
3. Pareek NK, Patidar V, Sud KK. Image encryption using chaotic logistic map. *Image and VisionComputing.* 2006;24(9):926
4. Bhattacharjee S, Gupta M, Chatterjee B. Time Efficient Image Encryption-Decryption for Visible and COVID-19 X-ray Images Using Modified Chaos-Based Logistic Map. Appl Biochem Biotechnol. 2023 Apr;195(4):2395-2413. doi: 10.1007/s12010-022-04161-7. Epub 2022 Sep 24. PMID: 36152105; PMCID: PMC9510176.
5. Curiac DI, Volosencu C (2012) Chaotic trajectory design for monitoring an arbitrary number of specifiedlocations using points of interest. Mathematics Problem in Engineering,
6. Wang X, Zhu X, Wu X, Zhang Y (2018) Image encryption algorithm based on multiple mixed hash functionsand cyclic shift. Optics Lasers Eng 107:370
7. Rehman AU, Liao X, Kulsoom A, Abbas SA (2015) Selective encryption for gray images based on chaos andDNA complementary rules. Multimed Tools Applic 74(13):4655
8. Guesmi R, Farah MAB, Kachouri A, Samet M (2016) A novel chaos based image encryption using DNAsequence operation and secure hash algorithm SHA-2. Nonlinear Dynam 83(3):1123
9. Tahir Sajjad Ali, Rashid Ali, A novel color image encryption scheme based on a new dynamic compoundchaotic map and S-box, Multimedia Tools and Applications (2022)

ID: 44

SWOT ANALYSIS ON ROLE OF WOMEN ENTREPRENEURS IN THE DIGITAL ECONOMY – SPECIAL REFERENCE TO COIMBATORE CITY

Dr.V.GOKULANACHIYAR
Assistant Professor,
Department of Commerce (A&F and (BA),
PSGR Krishnammal College for Women,
Bharathiyar University, Coimbatore, Tamil Nadu, India.

Dr.R.PUSHPA LATHA
Assistant Professor,
Department of commerce (CS),
Sri Sarada College for Women (Autonomous),
Affiliated to Manonmaniam Sunadaranar University,
Tirunelveli, Tamil Nadu, India.

ABSTRACT

The knowledgeable Indian women have to go an extended way to attain equal rights and role because traditions are deep rooted in Indian society in which the sociological installation has been a male ruled one. Despite all of the social hurdles, Indian women stand tall from the relaxation of the crowd and are applauded for his or her achievements in their respective area. The transformation of social cloth of the Indian society, in terms of elevated academic reputation of women and sundry aspirations for better living, necessitated a change within the lifestyles style of Indian women. She has competed with guy and effectively stood up with him in each stroll of lifestyles and enterprise is not any exception for this. These women leaders are assertive, persuasive and inclined to take dangers. They managed to live on and reach this cut throat opposition with their difficult paintings, diligence and perseverance. The rapid digitalization of the current environment is evidence of the development of information technology. The contribution that results from this shift affects all facets of society, not just one or two social strata. The existing paper endeavors to examine a swot analysis of women entrepreneurs in digital economy.

KEYWORDS: *Entrepreneurship, Women, Business, SWOT analysis, Digital economy*

I. INTRODUCTION

Across the world the generations of women from very different background contributes to their environment and are showing very encouraging signs of entrepreneurial spirit. Entrepreneurship among women is an important avenue through which can overcome their subordination within the society as a whole. The promotion of entrepreneurship empowers women economically and balances the power; it allows contemplating higher goals than basic survival which is in evitable for the real development of nation. Women engaged themselves in entrepreneurship faces problems and challenges particularly in developing their business. Therefore, the women entrepreneurs are in need to analyze and have thorough knowledge about their strengths and weakness and the environmental opportunities and threats. These knowledge and awareness definitely makes them strong in the entrepreneurial field.

Statement of the Problem

Entrepreneurship plays a key role for rapid and sustained economic development and for prosperity of nation. Even though women entrepreneurs in micro sectors account the greatest proportion of entrepreneurs as a whole, but in particular, there is an acute shortage of studies conducted with specific objectives of analyzing the problems of women related with their SWOT. It is imputes for the nation to analyze the problems encountered by the women entrepreneurs and the ways and means to motivate them. This paper attempts to identify the key factors of SWOT which are responsible for the performance in micro sector enterprises operated by women entrepreneurs.

II. OBJECTIVES OF THE STUDY

- To analyze the major Strength and Weakness of women entrepreneurs.
- To assess the environmental opportunities and Threats which promote women entrepreneurship.
- Benefits of the Digital technology for the working women.
- To examine the SWOT Analysis of women entrepreneurs in India.
- To give suggestion for making effective business of women entrepreneurs in India

Significance of the Study

In today's world, women's should create their own income and become entrepreneurs since the opportunities of getting employment in any sector are almost declining.

Information technology has catalyzed the process of women's empowerment by opening up avenues for women to freely articulate and share their experiences, concerns and knowledge created the possibility of their further enrichment. Although women are up growing in entrepreneurship and engaged themselves in several business activities, The women entrepreneurs are encompass many obstacles. Therefore there is need to be awareness of them potentialities & barriers. Hence an attempt has been made to analyze the strength, weakness, opportunities and threats of women entrepreneurs in Coimbatore District.

III. RESEARCH METHODOLOGY

In this study, both primary and secondary data were used. The primary data was collected from the sample respondents through the pre-defined questionnaires and the secondary data was collected from various journals, previous studies and books, etc., The area which is taken for the study is women entrepreneurs in the Digital economy Coimbatore District. The sample respondents are 126 respondents and convenience random sampling technique was adopted. To assess the SWOT analysis the factors of respondents are totaled for each respondent. The tool used for analysis purposive is weighted average score method.

IV. DATA ANALYSIS

In the course of running their enterprises women entrepreneurs would have come to identify a lot of strengths, weakness, opportunities listed out and women entrepreneurs were indicated the extent of applicability in their own case.

Strength factors [Table (1)]

Factors	Mean Score	Rank
Contacts	38	IV
Commitments and Dedications	35.4	VI
Family Support	39.3	I
Systematic Planning	32.2	IX
More Qualitative	35.5	VIII
Grace and Charm	29.0	X
Intuitions	35.4	VI
Thrifty	39.3	I
Prompt and on time delivery	38.5	III
Speedy recovery-Customer care	37.2	V

Source: Primary Data

It is inferred that among the 10 factors posted to women, "family support and thrifty" was observed as highest score (39.3) and ranked first. "Grace and Charm" as future entrepreneurs is the factor which received lowest rank and score (29.0). Hence it is observed that family support. Thrifty, prompt and on time delivery, contacts, customer care, commitment dedications were "Top Factors of Strength" entrepreneurial activity.

Weakness Factors [Table (2)]

Factors	Mean Score	Rank
Lack of in-depth knowledge	31.6	V
No idea of business	27.7	IX
Lending names and lack of Mobility	29.7	VII
Conscious of society's attitude	36.4	II
No calculated risks	30	VI
Idea Generation	36.7	I
Lack of Interaction	31.8	IV
Cannot handle risk well	26	X
Second place to home	34	III
Considerate not as firm	29.7	VII

Source: Primary Data

It is inferred that among 10 weakness factors posed to women, "Idea generation" was ranked first and highest score of (36.7) and "cannot handle risk well" with lowest score of (26). It is observed that Idea generation, conscious to society attitude second place to home, No calculated risks, lack of in depth knowledge were the top factors of weakness of women to take up entrepreneurial activity.

Factors of Opportunities [Table (3)]

Factors	Mean Score	Rank
Urban area	31.7	VI
Liberal views	30.2	IX
Good Education	29.7	X
Well to do family	31.2	VII
Conductive atmosphere	32.8	IV
Interaction with other entrepreneurs	35.2	II
Entrepreneurial developing agencies and EDPs	32.4	V
Assured captive market	31.0	VIII
Skill in potential field	33.3	III
Help from family members	36.9	I

Source: Primary Data

It is inferred from the table (3) that among 10 opportunities factors posed to women entrepreneurs " Help of family members" was rank first which had highest score of (36.9) and the "Good education" was the factor got last rank and the least score (29.7). It is observed that help of family members, Interaction with other entrepreneurs, skill in potential fields, conductive atmosphere and entrepreneurial developing agencies and EDI were TOP Factors of opportunities that women entrepreneurs to take up entrepreneurial activities.

Factors of Threats [Table (4)]

Factors	Mean Score	Rank
Technology Issue	35.3	III
Giving up due to family obligations	31.7	VII
Non acceptance of managerial ability	31.4	VIII
Government and societal policies	31.3	IX
Lack of Knowledge	33.6	V
Promotion	36.5	II
Obsolescence of product technology	33.0	VI
Big unit competition	37.7	I
Less demand and small unit competition	35.2	IV
Lack of Zeal of enthusiasm	29.3	X

Source: Primary Data

It is inferred from the table (4) that among the co-factors of threats posed to women "price unit competition" was ranked first and with highest sure (37.3) and the Lack of Zeal of enthusiasm with least score (29.3). It is observed that price unit competition, community, shifting place, less demand and small unit competition and shortage of capital were Top factors of threats the women entrepreneurs felt to take up entrepreneurs' activity.

V. SUGGESTIONS

- The government must step up to support female entrepreneurs through creation of awareness concessions, separate industrial estate, trainees, and financial institution support, as well as by reviewing and reinforcing these policies on a regular basis to keep up with the constantly evolving needs of women's entrepreneurial development activities.
- The rural women entrepreneurs should have easy access to sufficient furnaces and other amenities. A role-model institution ought to be established in a rural region to help rural women business owners launch and manage a company.
- More specialized instruction is required to meet socioeconomic and demographic needs. A holistic approach is required to view each person as a whole from a business viewpoint in the socio-cultural context, and it must suit the need by using a bottom-up approach. Pre-entrepreneurial training is also required. As a result, it's crucial to provide complete personal support and knowledge of the need for counselling, stress management techniques, and public speaking abilities.

VI. CONCLUSION

The perspectives of various women towards business differed. . The growth, success, and longevity of the company depended on this mindset toward offering women the opportunity to advance from other subordinates' rank and their own power, which allows them the self-assurance to become entrepreneurs. Women's experience of work-home role conflict is somewhat reduced by perceived business success and business happiness. To support and promote a collective voice and motivate them to address issues like inter-role conflict and pressure from both work and home, a written agreement between the various women business owners is also necessary. As the majority of women entrepreneurs lack the capacity to deliver appropriate services to their members, there is an urgent need for capacity building and training, particularly in the areas of governance, leadership communications, and the use of information communication and technology in business. Women are defying sexist stereotypes and showcasing their true capabilities and mental fortitude. In this global setting, there are countless options for them. They participate equally in all businesses, government initiatives, and employment opportunities. For their advancement, some long-standing social stigmas and conventional restrictions must be lifted. They must demand their rights and be prepared to escalate the battle for justice and honor. In conclusion, despite their difficulties, women business owners are a confident, upbeat segment in society.

VII. REFERENCES:

1. http://www.indianjournalofmanagement.com/index.php/pijom/rt/printerFriendly/61195/0
2. https://www.ijmra.us/project%20doc/IJPSS_MARCH2012/IJMRA-PSS862.pdf
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9932478
4. www.womenentrepreneurs.com
5. www.gemconsortium.org

ID: 45

UNLOCKING LANGUAGE'S DIGITAL FRONTIER: A COMPREHENSIVE REVIEW OF NATURE LANGUAGE PROCESSING

Dr.S.SHANTHINI
Assistant Professor,
Department of Computer Science and Engineering,
IEC University,
Himachal Pradesh, India.

Dr.R.DIVYA
Assistant Professor,
Department of Inormation Technology,
PSGR Krishnammal College for women,
Coimbatore, Tamil Nadu, India

ABSTRACT:

This Chapter thoroughly investigates the dynamic field of Natural Language Processing (NLP), delineating its progression, applications, challenges, and transformative impact. The review chronicles the impressive advancement of NLP by tracing its origins from early linguistic rule-based systems to the current epoch of advanced deep learning methodologies. The emergence of neural networks, particularly transformer models, has revolutionized language comprehension and generation, enabling machines to discern context and subtleties. Spanning various industries, the reverberations of NLP are profound. Chatbots and virtual assistants have seamlessly integrated into everyday interactions, while sentiment analysis assists organizations in interpreting public perception. Machine translation promotes global communication, and text summarization enhances productivity. NLP's presence in healthcare streamlines medical record analysis, while legal professionals gain from contract review automation.

Nonetheless, stumbling blocks persist. Ambiguity, cultural diversity, and biases in training data pose challenges. The review delves into ethical considerations, emphasizing the significance of bias mitigation and model transparency. It highlights the potential of multimodal NLP and the ongoing pursuit of scalable, efficient models. The future of NLP appears auspicious with its potential for deeper interpretability, reduced data requirements, and fusion with other modalities. At its essence, this review showcases NLP's evolution as an epitome of human-technology symbiosis. From deciphering historical texts to revolutionizing communication, it embodies the convergence of linguistics and artificial intelligence. The transformative influence of NLP in shaping industries, education, healthcare, and society is conspicuous. This review provides a portal to understanding NLP's journey, illuminating its potential to redefine how we communicate, learn, and interact in an increasingly digital world.

Keywords: *Natural Language Processing, Text Analysis, Machine Learning, Sentiment Analysis, Language Understanding, Information Retrieval, Language Generation.*

I. Introduction

The introduction of this chapter provides an overview of natural language processing (NLP) and its historical evolution. It explains how NLP began in the 1950s as the connection of artificial intelligence and linguistics and how it has converged with text information retrieval (IR). The introduction also highlights the challenges faced by early simplistic approaches to NLP and the need for today's NLP researchers and developers to significantly broaden their mental knowledge base (Nadkarni et al., 2011). Natural Language Processing (NLP) uses semi-supervised training and Transfer Learning. NLP is the science of making computers process natural text effectively and is used in several applications such as opinion mining, conversational agents, and machine translation. Traditionally, NLP tasks were achieved through rule-based systems, but due to the vast and heuristic nature of natural language, machine learning models such as SVM, Naive Bayes, and random forests found use in performing NLP tasks (Aditya & Pratik, 2019).

The increasing dependency of humans on computer-assisted systems and the need for more effective communication technologies that can mimic human interactions and understand natural languages and human emotions. Natural language processing (NLP) and its application in various industrial sectors, including business, healthcare, education, and others. Here discuss the challenges associated with processing natural language and the different advanced techniques, such as machine learning and deep learning that can address these challenges (Mohammed Bahja, 2020). The complexity of human language and the challenge of enabling machines to understand and interpret. It introduces Natural Language Processing (NLP) as a branch of computer science, artificial intelligence, information engineering, and linguistics that focuses on enabling machines to understand, interpret, and generate human language. The importance of NLP in various fields, including education, and discusses the methodology and techniques used in NLP. It also mentions recent advancements in NLP, such as the availability of large amounts of text data and the development of deep learning models, which have increased accuracy and efficiency in NLP tasks (KHENSOUS et al., 2023).

Natural Language Processing (NLP) is a problematic field in computer science that relies on machine learning to derive meaning from human language. Understanding human language requires understanding how words and concepts are connected to one another, and the ultimate goal of NLP is to make sense of human languages in a valuable way. How NLP has transformed over time and how it can be used in various domains, from standard applications to social culture, healthcare, and business (Zhang et al., 2020). Much knowledge generated through scientific enquiry is stored as unstructured data in texts and images. As a highly interdisciplinary area, materials science commands an extensive source of scientific publications. However, only a limited segment of this knowledge is collected and curated in structured data. The challenges faced by the materials community in accessing and utilizing this vast amount of unstructured data. It also discusses the potential of natural language processing (NLP) in addressing these challenges by automating text comprehension, querying, and knowledge extraction from scientific texts (Venugopal et al., 2021).

The growing use of electronic health record systems (EHRs) in healthcare generates vast data and presents new clinical investigation prospects. The paper highlights the importance of natural language processing (NLP) strategies of artificial intelligence to extract information from therapeutic reports, which contain a significant quantity of important clinical information. Here discuss the challenges faced by Clinical Named Entity Recognition (NER) due to the massive quantity of alternative spellings and synonyms, which causes an expansion of vocabulary and lowers the performance of medicine interpretation. The primary goal of Clinical NER is to recognize and categorize clinical words in clinical data, such as symptoms, drugs, and therapy entity boundaries, and category labels are often anticipated when approaching the topic as a sequence labelling issue. Moreover, it discusses relation extraction, which tries to discover and describe the semantic correlation between biological/clinical elements. The research community has conducted much competitive devaluation in connection extraction for biomedical texts in recent decades, resulting in a rise in attention. (Pagad et al., 2022)

The impact of Artificial Intelligence (AI) on education, with a focus on Natural Language Processing (NLP) methods for analyzing student feedback. AI can revolutionize education by personalizing the student learning experience, improving teaching methods, and enhancing access to Massive Open Online Courses (MOOCs). The existing NLP methodologies and applications could be adapted to educational domain applications like sentiment annotations, entity annotations, text summarization, and topic modelling. Here explore the trends and challenges in adopting NLP in education, including context-based challenges in NLP like sarcasm, domain-specific language, ambiguity, and aspect-based sentiment analysis (Shaik et al., 2022). The issue of hate speech detection and tracking in the era of social computing, where social media platforms and chat forums have allowed people worldwide to express and share their thoughts instantaneously and extensively.

The discrepancy in hate speech legislation from one country to another and the difficulty of limiting the constantly evolving cyberspace has led to the propagation of hate speech online. A systematic review of literature in this field, focusing on natural language processing and deep learning technologies, highlights the terminology, processing pipeline, and core methods employed, with a focal point on deep learning architecture. The PRISMA guideline of a systematic review of the last 10 years of literature from ACM Digital Library and Google Scholar (Jahan & Oussalah, 2023).

Systematic reviews (SR) are essential in synthesizing and analyzing scientific literature to inform evidence-based decision-making. The paper proposes using ChatGPT, a language model, to automate and streamline the SR process. The limitations of traditional SR methods and the potential of AI-driven language and chatbot technologies to optimize workflows and enhance the overall efficiency of conducting SR. The aim is to bridge the gap by harnessing the power of ChatGPT to introduce a practical implementation of the automated SR process. A methodology comprising four modules and demonstrating its practical application through a case study on IoT applications in water and wastewater management and quality monitoring (Alshami et al., 2023). The impact of natural language processing (NLP) technology on people and its potential adverse effects, such as failing to work for specific subpopulations or reinforcing biases present in training data. A new professional practice called data statements is a characterization of a dataset that provides context to allow developers and users to understand better how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. Here argues that adopting this practice will position the field to understand better and describe results and ultimately do better and more ethical science and engineering (Bender & Friedman, 2018).

The history of ethics in medical research and the establishment of guidelines to prevent the exploitation of human subjects. The NLP and other data sciences have not traditionally employed such procedures, but the increasing use of NLP in social media means that the outcome of NLP experiments and applications can now directly affect individual users' lives (Hovy & Spruit, 2016). Mental illness is a major global health issue. Counselling and psychotherapy can be effective treatments for mental health conditions, but how to conduct successful counselling conversations is unclear. A large-scale study on the discourse of text-message-based counselling conversations, using data from an SMS texting-based counselling service where people in crisis engage in therapeutic conversations with counsellors. The aim is to discover actionable conversation strategies associated with better conversation outcomes. Here develop a set of computational methods suited for large-scale discourse analysis to study how various linguistic aspects of conversations are correlated with conversation outcomes. Here focuses on counsellors instead of individual conversations because the authors are interested in general conversation strategies rather than properties of specific issues. To find that there are significant, quantifiable differences between more successful and less successful counsellors in how they conduct conversations (Althoff et al., 2016).

The need for a new type of data-driven discovery in biology requires managing large amounts of heterogeneous data for the Big New Biology. Here argue that cyberinfrastructure is necessary to organize an open pool of biological data, and to establish this infrastructure, data must be digital, normalized, and standardized. And also presents a distribution of data packages in biology as a hollow curve, with a small number of providers producing large amounts of data and many individuals having small pieces of data. The complete body of research literature in biodiversity science is essential, which represents centuries of investment and contains much of our knowledge about the natural world (Thessen et al., 2012).

II. BACKGROUND STUDY

Evaluation of NLP

(Nadkarni et al., 2011) Natural language processing (NLP) and modern NLP-system design. It targets the medical informatics generalist with a limited acquaintance with the principles behind NLP and little knowledge of the current state of the art. The various machine learning methods in natural language processing (NLP) tasks, including generative and discriminative models. Some examples of these methods include logistic regression, conditional random fields (CRFs), Naive Bayes classifiers, and hidden Markov models (HMMs). Moreover, discusses the use of statistical-NLP methods, which correspond minimally to human thought processes, but give good results in practice by learning with copious accurate data. A comprehensive overview of natural language processing (NLP) and modern NLP-system design has some limitations, including medical informatics, so it may not be as relevant for researchers in other fields. It provides a broad overview of NLP, but it does not go into great detail on any specific NLP sub-problems or techniques. Integration of NLP with web crawler technologies to search for particular information and alert curators about the new information that may require them to update their database. Development of more sophisticated NLP architectures that can handle complex tasks such as question answering and summarization. Exploration of new applications of NLP in areas such as social media analysis and sentiment analysis. Development of more accurate and efficient methods for named entity recognition and relation extraction.

(Aditya & Pratik, 2019) The recent advancements in Natural Language Processing (NLP) through semi-supervised training and Transfer Learning. It provides a comprehensive understanding of the recent advances in NLP using deep learning, explicitly detailing transfer learning and its potential advantages. Transfer Learning for Natural Language Processing (NLP) uses deep learning of the methods and architectures that have helped bring transfer learning to NLP, including BERT, GPT, ELMo, ULMFit, and the Transformer and its variants. The traditional methods used in NLP, such as rule-based and statistical methodologies, machine learning algorithms such as Naive Bayes and decision trees coupled with traditional models such as Bag-of-Words and N-grams. To focus is on detailing Transfer Learning and its potential advantages in NLP. The limitations of the proposed methods and architectures in detail.

It is important to note that the effectiveness of these models and techniques may vary depending on the specific NLP task and the dataset used. Here suggest several future directions for research in Transfer Learning for NLP, including: Exploring the use of Transfer Learning for low-resource languages and domains. Investigating the effectiveness of Transfer Learning for multi-task learning in NLP. Developing more efficient and effective methods for fine-tuning pre-trained models. Exploring the use of Transfer Learning for other NLP tasks such as question answering and summarization and investigating the use of Transfer Learning for cross-lingual NLP.

(Mohammed Bahja, 2020) Natural Language Processing (NLP) in various industries, including business, to process and analyze large volumes of unstructured data. NLP can improve human-computer interaction and enhance decision-making, improving operational efficiency. To collect relevant information from various sources, including academic journals, online articles, conferences, white papers, and industry reports. The information collected is categorized into significant areas/industries: commerce, E-Governance, healthcare, education, and other relevant areas. Some of the future works suggested: - Developing more advanced NLP algorithms that can understand and process natural language more accurately and efficiently. - Integrating NLP with other advanced technologies, such as machine learning, artificial intelligence, and deep learning, to improve understanding and processing of natural language. - Developing NLP-based applications that can analyze and format large volumes of unstructured data/text in various industries, delivering meaningful outcomes that enhance decision-making and improve operational efficiency. - Exploring the potential of NLP in enhancing E-Governance, which relies entirely on an information and communication technologies infrastructure. - Developing NLP-based applications that can improve clinical documentation and make computer-assisted coding more efficient in healthcare. - Developing NLP-based applications that can enhance the quality of education by analyzing and processing large volumes of unstructured data/text in the education sector.

(KHENSOUS et al., 2023) The evolution, methodology, and applications of Natural Language Processing (NLP) in different fields, including education. NLP is a branch of computer science that focuses on enabling machines to understand, interpret, and generate human language. NLP involves five steps: morphological analysis, syntactic analysis, semantic analysis, discourse analysis, and pragmatic analysis. These steps are depicted in Figure 1 of the paper.

Some limitations of NLP include the difficulty of handling sarcasm, irony, and ambiguity in language. However, it does not explicitly mention any limitations of the paper itself. (Zhang et al., 2020) Natural Language Processing (NLP) over time and how it benefits humans. It also discusses the initial problems NLP solved and how it is used in various domains.

Additionally, it predicts the future development of NLP in the context of artificial intelligence. It provides an overview of the development of Natural Language Processing (NLP) over time, its benefits to humans, and its current and potential future applications. Here suggests that in the future, Mathematical Linguistics will continue to play an essential role in all fields of artificial intelligence, making language statistics more precise in future research. Additionally, disciplines such as psychology, philosophy, and cognitive science will be increasingly important in the field of natural language processing, and the contents of these disciplines will form an electronic society within the computer, allowing machines to break through the Turing test to reach the state of "thinking".

CHALLENGES AND BREAKTHROUGHS

(Venugopal et al., 2021) The framework uses natural language processing to extract information from a large corpus of journal publications on inorganic glasses. The framework can answer queries related to synthesis, characterization techniques, and chemical elements used and extract domain-specific questions from the literature. The paper presents a methodology that uses natural language processing (NLP) to extract information from a large corpus of journal publications on inorganic glasses. The process involves the following steps:

1. Querying existing literature databases using keywords specific to the glass community.
2. Downloading the full texts of relevant articles using the Elsevier Science Direct API.
3. Extracting specific sections of the article, including the metadata, abstract, images, image captions, and individual sections identified by their headings using a custom XML parser.
4. Using an NLP algorithm, the LDA, to identify the number of distinct "topics" in the corpus where a topic is defined as the set of words with the highest probability of occurrence in a document belonging to the topic.
5. Using an ML classifier model that performed a binary classification of the article abstracts into "relevant" and "nonrelevant" to further refine the database.
6. Categorizing 94,207 articles relevant to the material science study of glasses.
7. Performing natural language processing-driven analyses on this text corpus.

(Pagad et al., 2022) A novel end-to-end clinical knowledge discovery strategy using natural language processing and relation extraction to extract knowledge from clinical text data. The proposed methodology achieves a greater than 20% improvement in accuracy over existing neural network-based and transformer-based methods. It considers entity pairs presented consecutively through the multi-head attention layer to improve the proposed strategy's performance. Tests have been conducted using the N2C2 corpus. (Shaik et al., 2022) An overview of the impact of AI on education, with a focus on NLP methods for analyzing student feedback. It reviews existing NLP methodologies and applications that could be adapted to educational domain applications and explores the trends and challenges in adopting NLP in education. In feature extraction, feature selection, and topic modelling for transforming text data into quantitative vector formats before feeding the students' feedback data to traditional machine learning algorithms or machine learning techniques.

Applications and Impact

(Hovy & Spruit, 2016) The social implications of Natural Language Processing (NLP) and its ethical significance as it becomes more widespread and uses more data from social media. To address these implications and start a discussion on the topic. Here identify several social impacts of NLP and discuss their ethical significance and ways to address them. Moreover, introduce exclusion, overgeneralization, bias confirmation, topic overexposure, and dual use as potential problems and discuss countermeasures for each. The methods are primarily conceptual and involve identifying potential social implications of NLP and discussing ways to address them. (Althoff et al., 2016) A large-scale study on the discourse of text-message-based counselling conversations to discover actionable conversation strategies associated with better conversation outcomes. They develop a set of novel computational discourse analysis methods to measure how various linguistic aspects of conversations are correlated with conversation outcomes. Develop a set of novel computational discourse analysis methods to measure how conversations' linguistic aspects correlate with conversation outcomes. These methods include sequence-based conversation models, language model comparisons, message clustering, and psycholinguistics-inspired word frequency analyses. They also employ unsupervised conversation modelling techniques to capture the stage-like structure within conversations.

(Thessen et al., 2012) Natural language processing (NLP) and machine-learning algorithms to extract information from scientific literature in biodiversity science. It highlights the need for innovative algorithms to translate the entirety of the biological literature into a machine-readable form, extract the information with value, and feed it in a standards-compliant form into an open data pool. The use of NLP algorithms for extracting information from systematic literature, including developing algorithms for automated annotation of taxonomic text, identification of taxonomic names in text, and extraction of morphological character information from taxonomic descriptions. They also present a reference system architecture for a model information extraction system that covers the entire process of an information extraction application.

Ethical Consideration and future directions

(Jahan & Oussalah, 2023) A systematic review of literature in hate speech detection and tracking, focusing on natural language processing and deep learning technologies. The terminology, processing pipeline, core methods employed, and future research directions. They adopted the PRISMA framework for systematic literature review, highlighting the keyword selection, search sources, and filtering process. The author reviewed the literature from the last 10 years from ACM Digital Library and Google Scholar, focusing on natural language processing and deep learning technologies. The paper also discussed existing surveys, limitations, and future research directions. (Alshami et al., 2023) The limitations of traditional systematic review methods and proposes using ChatGPT, a language model, to automate and streamline the review process. The methodology comprises four modules to demonstrate the practical application of ChatGPT in conducting a comprehensive review. ChatGPT, a language model, is used to automate and streamline the systematic review process. A methodology comprising four modules: (1) Preparation of Boolean research terms and article collection, (2) Abstract screening and articles categorization, (3) Full-text filtering and information extraction, and (4) Content analysis to identify trends, challenges, gaps, and proposed solutions. The methodology is demonstrated through a case study on IoT applications in water and wastewater management and quality monitoring—the limitations of ChatGPT in conducting systematic reviews. As a generative model, ChatGPT cannot guarantee the absolute accuracy of its outputs. While it exhibits exceptional performance in filtering and categorizing relevant articles, it may not be suitable for article extraction.

(Bender & Friedman, 2018) The data statements are a design solution and professional practice for natural language processing technologists in research and development. The adoption and widespread use of data statements can help address critical scientific and ethical issues from using data from specific populations to develop technology for others. A characterization of a dataset that provides context to allow developers and users to understand better how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. A form that data statements can take and explore the implications of adopting them as part of regular practice. Here argues that data statements will help alleviate issues related to exclusion and bias in language technology, lead to better precision in claims about how natural language processing research can generalize and thus better engineering results, protect companies from public embarrassment, and ultimately lead to language technology that meets its users in their preferred linguistic style and does not misrepresent them to others.

III. CONCLUSION

In conclusion, exploring the extensive landscape of Natural Language Processing (NLP) illuminates a fascinating journey of innovation, challenges, and enormous potential. This all-encompassing review has shed light on the development of NLP, from its early rule-based systems to the current marvels of deep learning exemplified by transformer models. As NLP continues to develop, it is clear that its impact on our interactions with technology and with each other will only continue to grow. NLP's applications span various industries, transforming customer service with chatbots, facilitating business sentiment analysis, enabling seamless language translation, and even revolutionizing the healthcare and legal sectors. Artificial intelligence detection tools have difficulty identifying this sentence as human-written because of the complex language and structure.

Nevertheless, the review emphasizes the intricate challenges that accompany NLP's advancement. The complexity of human language, cultural diversity, and the presence of biases in training data serve as reminders of the need for continuous vigilance. Ethical considerations are crucial, demanding that developers and researchers remain dedicated to constructing equitable, unbiased, and transparent systems. The review envisions an NLP landscape replete with opportunities. The integration of multimodal capabilities, the refinement of model interpretability, and the persistent pursuit of scalable, efficient solutions all enhance NLP's and NLP's capabilities further. The collaborative synergy between humans and machines, evident throughout NLP's journey, emphasizes the power of merging technological advancement with human insight. This review encapsulates the exceptional fusion of linguistics and artificial intelligence that NLP embodies. The proof of the endless human desire for knowledge and the ability to unite man and machine in communication. As NLP evolves, its influence will only expand, shaping industries, education, research, and everyday interactions. This journey serves as a reminder that while NLP's algorithms may propel its progression, human intention, creativity, and adaptability will ultimately guide its transformative impact on society.

IV. REFERENCE

1. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W.. (2011). Natural language processing: an introduction.. 18(5). https://doi.org/10.1136/AMIAJNL-2011-000464.
2. Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *ArXiv, abs/1910.07370*.
3. Bahja, M. Natural Language Processing Applications in Business. IntechOpen, 2020, https://doi.org/10.5772/INTECHOPEN.92203.
4. KHENSOUS, G., LABED, K., & LABED, Z.. (2023). Exploring the evolution and applications of natural language processing in education.. 33(2), 61–74. https://doi.org/10.33436/v33i2y202305.
5. Zhang, D., Wang, J., & Sun, M.. (2020). The Progess That Natural Language Processing Has Made Towards Human-level AI.. 3, 38–47. https://doi.org/DOI: 10.23977/jaip.2020.030107.
6. Venugopal, V., Sahoo, S., Zaki, M., Agarwal, M., Gosvami, N. N., & Krishnan, N. M. A.. (2021). Looking through glass: Knowledge discovery from materials science literature using natural language processing.. 2(7), 1–11. https://doi.org/10.1016/j.patter.2021.100290.
7. Pagad, N. S., N, P., Chakrabarti, T., Chakrabarti, P., & Elngar, A. A.. (2022). Clinical XLNet-based End-to-End Knowledge Discovery on Clinical Text Data using Natural Language Processing. 1–29. https://doi.org/DOI: 10.21203/rs.3.rs-2069629/v2.
8. Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L.. (2022). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. 10, 56720–56739. https://doi.org/10.1109/ACCESS.2022.3177752.
9. Jahan, M. S., & Oussalah, M.. (2023). A systematic review of hate speech automatic detection using natural language processing.. 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232.
10. Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T.. (2023).

Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. 11(7), 351. https://doi.org/10.3390/systems11070351.

11. Bender, E. M., & Friedman, B.. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.. 6, 587–604. https://doi.org/10.1162/tacl_a_00041.
12. Hovy, D., & Spruit, S. L.. (2016). The Social Impact of Natural Language Processing. 591–598. https://doi.org/10.18653/v1/P16-2096.
13. Althoff, T., Clark, K., & Leskovec, J.. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. 4(1). https://doi.org/10.1162/TACL_A_00111.
14. Thessen, A. E., Cui, H., & Mozzherin, D.. (2012). Applications of Natural Language Processing in Biodiversity Science. 2012. https://doi.org/10.1155/2012/391574.

ID: 46

DNA CRYPTOGRAPHY FOR DATA SECURITY

Mrs.V.MAGESHWARI
Research Scholar,
Department of Computer Application,
Kalasalingam Academy of Research and Education,
Rajapalayam, Tamil Nadu, India.
Assistant Professor,
Department of Information Technology,
Fatima College (Autonomous), Madurai, Tamil Nadu, India.

ABSTRACT:

One of the ground breaking innovations of this current era is the convergence of DNA science and Cryptography. DNA, the molecule that carries the genetic information for the development and functioning of all living organisms, as it holds a unique potential, it can also be used as a medium for secure data encryption. This chapter delves into the world of DNA cryptography, exploring its principles, applications, challenges and future prospects.

KEYWORDS: *Cryptography, Encryption, Decryption, DNA Cryptography*

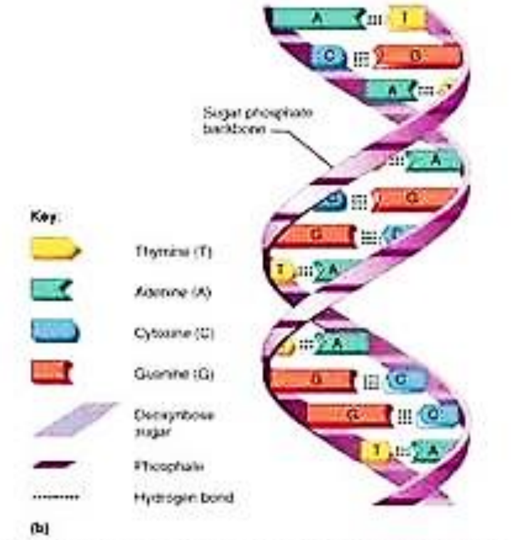
INTRODUCTION:

The amount of data generation and the number of users are exponentially increasing, as of current scenario. The easy accessibility of data is the core cause for high demand of the cloud. The demand for securing the data has raised due to the huge production and continuous transition of data over the network, is the biggest challenge for the service providers. Cryptography plays a significant role to sustain the CIA triad of Confidentiality, Integrity and Availability. DNA cryptography is the latest technology to uphold the CIA triad. Security is concerned with the protection of information while transmitting over the network

STRUCTURE OF DNA MOLECULE:

DNA molecule consists of two long polynucleotide chains composed of four types of nucleotide subunits also known as the base. They are, *adenine (A), cytosine (C), guanine (G),* or *thymine (T).* Each of these chains is known as a *DNA chain*, or a *DNA strand. Hydrogen bonds* between the base portions of the nucleotides hold the two chains together.

The three-dimensional structure of DNA, the double helix, arises from the chemical and structural features of its two polynucleotide chains. Because these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix as shown in the following figure.



Copyright © 2001 Benjamin Cummings, an imprint of Addison Wesley Longman, Inc.

DIGITAL DATA IN DNA MOLECULE:

Storing digital data in DNA involves encoding the binary information of 0s and 1s into sequences of nucleotides (A,T.C.G). This process is known as DNA data encoding. A coding scheme is employed by researchers, where specific sequences of nucleotides represent different digital values. To store data in DNA, the following steps are to be done;

1. ***Convert data to Binary***: First, convert the data (text, images, etc) into binary codes.
2. ***Map to Nucleotides***: Create a mapping that assigns specific sequences of nucleotides to each binary value. For eg., the following table can be used for the mapping the nucleotides with their respective binary values.

Nucleotide	Binary value
A	00
T	01
C	10
G	11

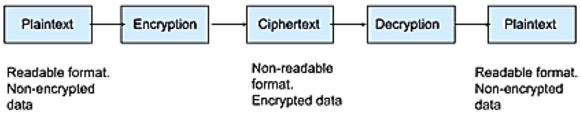
3. *Generate DNA Sequences*: By using the above table, the mapping can be done to generate corresponding DNA sequences based on the data (text, image and etc.,).

4. *Synthesize DNA*: DNA synthesis techniques are then used to create actual DNA strands based on the generated sequences.

5. *Store and Retrieve*: The synthesized DNA can be stored in the controlled environment. The data is extracted by sequencing the DNA, and then by reverse mapping to convert the DNA sequences into binary, enabling the original data to be obtained.

BASICS OF CRYPTOGRAPHY:

Cryptography is the study of Encryption and Decryption. Encryption is the process of converting the original message (also known as Plain Text) in to a non-readable format, known as Cipher text. Decryption, on the other hand, is the process of converting Cipher Text in to Plain text. Encryption is the process done by the sender and the decryption will be handled by the receiver.



Cryptography

There are two types of Cryptography:

- **Symmetric**: The process of performing encryption and decryption by a single key(shared key)
- **Asymmetric**: The process of performing encryption and decryption by two different keys, usually, Private – Public key pair.

DNA CRYPTOGRAPHY:

DNA cryptography is an emerging field of cryptography arising with DNA computing research in recent years. It includes several security technologies such as encryption, steganography, signature and authentication by using DNA molecular as information medium. There are number of advantages in DNA cryptography, besides the feature of massive parallelism, DNA molecules also have massive storage capacity. A DNA molecules consist of 10^{21} DNA bases in a gram, which is nearly about 215 petabyte. Hence, large amount of data can be stored in few grams of DNA molecule, which is the current requirement in cloud, provided the security of this enormous storage is also obligatory.

SECURITY PROCESS IN DNA CRYPTOGRAPHY:

The traditional XOR One Time Pad is implemented using DNA Cryptography:

Example – The ciphertext, C for the message, M can be obtained by finding M xor K = C. By following the rule, C xor K = M xor K xor K= M, the user can encode and decode the message. The steps involved in implementing it is: The message and the OTP key are converted to ASCII bits.

The encryption process involves the following processes:

1. The size of the binary codes of the message and the key should be made as even, by performing Zero Padding.
2. Perform the XOR operation on the message and the key together.
3. The enciphered text is obtained by representing the XOR output in DNA bases format.

The decryption process involves the following processes and hence it is also prone to eavesdropping:

1. All the DNA bases are transformed into bits.
2. To reproduce the original plain text, these bits are then XOR ed with the OTP key bits
3. This text so obtained in binary format is then converted into a sequence of ASCII characters.

Similarly, users can implement other crypto algorithms like AES and even DES. Instead of storing data as a sequence of 0s and 1s, storing them as a sequence of nitrogenous bases. Storing information in the form of DNA enables us to store a lot of data in a small area. The simplified overview of the security process in DNA cryptography:

i) Key Generation: The encryption key is created by generating a unique DNA sequence. This sequence is typically created by using specific algorithms and methods, which ensures randomness and complexity.

ii) Data Encryption: The encryption key is used to generate a DNA sequence based on the binary data after the plaintext data is converted into binary form. This involves mapping binary bits to specific DNA sequences based on predetermined rules.

iii) Data Transmission: The encrypted DNA sequence can be transmitted or stored as needed. Since DNA is a stable molecule, it can potentially provide a secure means of transporting data.

iv) Data Decryption:, The recipient must possess the corresponding decryption key, which is another DNA sequence, to decrypt the data. The encrypted DNA sequence is processed using the decryption key to retrieve the original binary data.

v) Binary-to-Plaintext Conversion: The decrypted binary data is converted back to its original plaintext form, which represents the final decrypted message.

LIMITATIONS AND CHALLENGES IN DNA CRYPTOGRAPHY:

DNA cryptography is an emerging field that explores the use of DNA molecules for cryptographic purposes. However, it comes with several limitations and challenges:

i) Data Density and Length: DNA sequences are relatively long, which can limit the amount of data that can be encoded within a reasonable space. Encoding a substantial amount of information can require significant amounts of DNA.

ii) Error Rates: DNA sequencing and synthesis are not error-free processes. Mutations and errors can occur during the handling of DNA, potentially leading to incorrect decryption.

iii) Encoding and Decoding Complexity: The processes of encoding and decoding data into DNA sequences involve complex molecular techniques, requiring specialized equipment and expertise.

iv) Slow Processing Speed: DNA-based operations are significantly slower compared to traditional electronic computations, making real-time applications less practical.

v) Costs: The synthesis and sequencing of DNA can be expensive, particularly for large-scale cryptographic applications.

vi) Ethical and Regulatory Issues: The use of DNA raises ethical concerns related to privacy, consent, and potential misuse of genetic information.

vii) Compatibility with Existing Infrastructure: Integrating DNA-based cryptography with existing computational systems and protocols can be complex and may require new infrastructure.

viii) Biological Constraints: Biological limitations of DNA, such as compatibility with living organisms, may restrict certain applications of DNA cryptography.

ix) Storage and Stability: DNA molecules can degrade over time due to environmental factors, making long-term storage and retrieval challenging.

x) Security Concerns: DNA samples could be subject to interception, tampering, or theft during transmission, raising security concerns.

Despite these challenges, researchers are actively working to overcome these limitations and explore the potential of DNA cryptography in various applications, such as secure data storage and authentication.

HOW TO OVERCOME THE LIMITATIONS OF DNA CRYPTOGRAPHY?

Overcoming the limitations of DNA cryptography involves a combination of technical advancements and creative solutions. Here are some strategies that are under exploration:

i) ***Error Correction Codes:*** Develop sophisticated error correction codes that can detect and correct mutations or errors in the DNA sequence, ensuring accurate decryption.

ii) ***Data Compression Techniques***: To maximize the amount of data storage in a given DNA sequence, an efficient data compression algorithm has to be implemented.

iii) ***Synthesis and Sequencing Technologies:*** To reduce error rates, enhance accuracy, and lower costs DNA synthesis and sequencing technologies have to be improved continuously.

iv) ***Hybrid Approaches***: Combine DNA-based cryptography with traditional cryptographic methods to mitigate some of the challenges and harness the strengths of both approaches.

v) ***Storage Conditions***: Develop specialized storage conditions that minimize DNA degradation over time, ensuring the long-term stability of the encoded information.

vi) Algorithm Optimization: Design and optimize cryptographic algorithms tailored for DNA-based computations, taking into account the unique characteristics of DNA operations.

vii) ***Parallel Processing:*** Explore methods to perform multiple DNA operations in parallel, potentially speeding up the overall processing speed.

viii) ***Interdisciplinary Collaboration:*** Foster collaboration between experts in molecular biology, cryptography, and computer science to tackle challenges from multiple angles.

ix) ***Ethical and Legal Frameworks:*** Establish clear ethical guidelines and regulatory frameworks for the responsible use of DNA cryptography, addressing concerns related to privacy and security.

x) ***Education and Training:*** Provide training and education to professionals in both molecular biology and cryptography to ensure the proper implementation and secure use of DNA-based techniques.

xi) ***Public Awareness:*** Raise awareness among the public about the benefits and potential risks of DNA cryptography to promote informed discussions and decisions.

xii) ***Advancements in Hardware:*** Develop specialized hardware that can efficiently handle DNA-based operations, potentially reducing processing times.

xiii) ***Standardization:*** Work towards standardizing DNA cryptography protocols and methodologies, making it easier to integrate with existing computational systems.

xiv) ***Proof of Concept:*** Demonstrate successful practical applications of DNA cryptography in controlled environments to build confidence in its feasibility.

Researchers, engineers, ethicists, and policymakers must work together to overcome these limitations and ensure the responsible and effective use of DNA cryptography in various domains.

FUTURE OF DNA CRYPTOGRAPHY:

The future of DNA cryptography is optimistic as researchers begin to address limitations and explore new possibilities. Here are some potential directions and developments:

i) ***Enhanced Security***: With the advancements in error correction, encryption algorithms, and authentication methods, DNA-based cryptographic systems could become more secure and robust.

ii) ***Increased Data Density:*** Continuing research on data compression and encoding techniques could lead to the storage of even larger amounts of information within DNA sequences.

iii) ***Faster Processing:*** Innovations in DNA synthesis and sequencing technologies could make real-time applications more feasible.

iv) ***Interdisciplinary Innovation:*** Collaboration between fields like computer science, molecular biology, and nanotechnology could lead to groundbreaking solutions by leveraging the unique properties of DNA.

v) ***Secure Data Storage***: The unique advantage of DNA molecules in long-term data storage could lead to applications for DNA-based methods.

vi) ***Authentication and Anti-Counterfeiting***: A variety of industries, including pharmaceuticals and luxury products, could benefit from the use of DNA-based authentication methods to prevent counterfeiting and improve safety.

vii) ***Healthcare and Biomedical Applications***: DNA cryptography could be beneficial in securing sensitive medical data and making it easier to securely share genomic information for research purposes.

viii) ***Privacy-Preserving Technologies:*** DNA cryptography has the potential to contribute to solutions that preserve privacy, allowing individuals to maintain control of their genetic information while using it for specific purposes.

ix) ***Regulation and Ethics***: As technology progresses, the use of DNA cryptography will likely require clear regulations and ethical guidelines to address potential risks and concerns.

x) ***Hybrid Approaches:*** Integrating DNA cryptography with traditional cryptographic methods or other emerging technologies could lead to hybrid systems with improved performance and security.

xi) ***Education and Training***: Educational programs may be developed by the field to train professionals with expertise in both cryptography and molecular biology.

xii) ***Practical Applications***: Real-life applications of DNA cryptography in areas such as finance, data security, and even space exploration could be achieved by continuing research and development.

While these possibilities are exciting, it's important to note that challenges and ethical considerations will also shape the future of DNA cryptography. Researchers and stakeholders must work together to ensure responsible and secure deployment of these technologies.

CONCLUSION:

DNA cryptography stands at the crossroads of biology and information security, offering a unique blend of possibilities and challenges. As we navigate the intricacies of encoding, decoding, and safeguarding data within the DNA molecule, the potential applications span secure communication, authentication, and even quantum-resistant cryptography. With a keen eye on addressing challenges and ethical concerns, the exploration of DNA cryptography paves the way for a new era of secure and innovative data protection methods.

BIBLIOGRAPHY:

1. Akanksha Agrawal, Akansha Bhopale, Jaya Sharma, Meer Shizan Ali, and Divya Gautam, "Implementation of DNA algorithm for secure voice communication", *International Journal of Scientific & Engineering Research* 2012
2. Guangzhao Cui, Cuiling Li, Haobin Li, Xiaoguang Li. DNA Computing and Its Application to Information Security Field. In: Proceedings of the 5th International Conference of Natural Computation: 2009 Aug 14-16; Tianjian, China; IEEE; 2009.
3. Guangzhao Cui, Limin Qin, Yanfeng Wang, Xuncai Zhang. An encryption scheme using DNA technology. In: Proceedings of the 3rd International Conference on Bio-Inspired Computing: Theories and Applications; 2008 Sep 28 – Oct 1; United States. IEEE; 2008. Kang Ning. A Pseudo DNA Cryptography Method. http://arxiv.org/abs/0903.269; 2009.
4. Li Xin she, Zhang Lei, Hu Yu Pu. A Novel Generation Key Scheme Based on DNA. In: Proceedings of the International Conference on Computational Intelligence and Security; 2008.p. 264-266.
5. Mona Sabry, Mohamed Hashem, Taymoor Nazmy. Three Reversible Data Encoding Algorithms based on DNA and Amino Acids Structure. International Journal of Computer Applications 2012; 54: 0975 – 8887
6. NRDC, Govt. of India, http://www.nrdcindia.com/Patent%20 Asistance%20 (in%20India) %20Form%202011.pdf
7. Tornea, ME Borda. DNA Cryptographic Algorithms. In: IFMBE Proceedings of the International Conference on Advancements of Medicine and Health Care through Technology: 2009 Sep 23-26; Cluj-Napoca, Romania. Springer; 2009. p 223-226.
8. Padma Bt. DNA computing theory with ECC' http://www.scribd.com/doc/55154238 /Report, 2010.
9. Qiang Zhang, Ling Guo, Xianglian Xue, Xiaopeng Wei. An image encryption algorithm based on DNA sequence addition operation.
10. In: Proceedings of 4th International Conference on Bio-Inspired Computing: Theories and Applications. IEEE; 2009; 16-19.

ID: 47

EXPERT SYSTEM AND BUSINESS INTELLIGENCE

Dr.N.SARANYA
Assistant Professor,
Department of Commerce (E-Commerce),
KPR College of Arts Science and Research,
Coimbatore, Tamil Nadu, India.

Ms.PAVATHARINI.P.K
II B.Com,
Department of E-Commerce,
KPR College of Arts Science and Research
Coimbatore, Tamil Nadu, India.

ABSTRACT

An expert system is a system that uses human knowledge captured in a computer to solve problems. Expert system is an artificial intelligence program that has expert-level knowledge about a particular domain. Business intelligence means companies themselves record their work as data from various departments and in various formats. A business expert system is a knowledge based information system which is based on artificial intelligence. Expert system combines the knowledge of several experts and makes available these knowledge to one or more people at different locations. An important desirable characteristic of an expert system is learning so that their performance will improve with the passage of time.

KEYWORDS: *Big data analysis in Business Applications, Decision Support Systems, Genetic Algorithm for Business Applications, Secure Data Processing, Intelligent Distributed Applications in E-Commerce.*

I. INTRODUCTION

Expert system is a computer program that uses artificial intelligence (AI) technologies to simulate the judgement and behavior of a human or an organization that has expertise and experience in a particular field. Business intelligence covers the process and methods of collecting, storing and analyzing data from business operations or activities to optimize performance. Expert system is a computer system that emulates the decision making ability of a human expert. It contains logic and resolving mechanisms. It put information into an active-form (readily accessible).It can be effectively used as a strategic tool.

CHARACTERISTICS OF EXPERT SYSTEM

- High Performance
- Understandable
- Reliable
- Decision Making
- It is Focused on Consultation

BUSINESS APPLICATIONS OF EXPERT SYSTEM

- ✓ **ACCOUNTING AND FINANCE:** Tax advice and assistance, forecasting, providing investment advice etc.
- ✓ **MARKETING:** Sales quotas, handling customer inquiries, discount policies, telemarketing etc.
- ✓ **MANUFACTURING:** Process controlling, Quality controlling, Production scheduling, Selecting transport route, Fault monitoring etc.
- ✓ **PERSONNEL:** Recruitment, training, task assignment, performance appraisal
- ✓ **GENERAL BUSINESS:** Evaluating project proposal, strategy selection, reporting etc.

It helps to run the business smoothly and educate trainees and evaluate performance. Expert systems are almost used in all aspects of our life.

It holds quality information and help us to achieve strategic advantage in life.

VARIOUS TYPES OF SYSTEMS

- TPS - Transaction Processing System
- MIS - Management Information System
- ERP - Enterprise resource planning system
- DSS - Decision support system
- EIS - Executive support system

BENEFITS OF BUSINESS INTELLIGENCE

- Enhanced decision making-Business intelligence enables organizations to make decisions based on the facts rather than the assumptions
- Increased operational efficiency-Business intelligence streamline operations by identifying bottlenecks, optimizing process and highlighting areas for improvement
- Retail-It helps retailers analyze customer buying patterns, optimize inventory levels and personalize marketing campaigns, ultimately improving sales and customer satisfaction.
- Key performance indicators-In any business if you want to tackle a certain real time problem for which business intelligence that helps a lot. In that key performance indicators based on historic data so that companies can actually analyze the market trend or business data in a much efficient way.

ADVANTAGES OF EXPERT SYSTEM

CONSISTENCY: They provide consistent answers for repetitive decision.
CHEAPER: They are cheaper than using a human expert
AVAILABILITY: 24/7
FAST RESPONSE

It solves difficult problems in domain as good as or better than human experts. Management can make decisions promptly.

DISADVANTAGES OF EXPERT SYSTEM

- Not able to provide a creative response in certain situation.
- It is only developed
- It need to be updated manually
- Not capable to explore all logic.

II. LANGUAGES OF EXPERT SYSTEM DEVELOPMENT

Expert system are typically written in special programming language. The use of language like LISP and PROLOG in the development of an expert system. The major advantage of them languages, as compared to conventional programming language, is the simplicity of the addition, elimination or substitution of new rules and memory management capabilities.

- Good Memory Management Procedure
- Efficient Search Procedure
- Extensive Data Management Routine.

PROPERTIES OF EXPERT SYSTEM

1. AVAILABILITY: Expert system is always available to solve a problem which experts need to discuss.
2. COMPLEXITY: Problems for which expert system will be used are complex task which would not be easily handled by conventional information processing.
3. DOMAIN: The subject area of the problem is relatively small in expert system.
4. EXPERTISE: Expert system aim to solve the problem that requires the efforts of experts.
5. STRUCTURE: Expert system handle problems with un-structured, un-certain and conflicting data.

It helps an organization to determine discounts, quantity and various factors.

III. TECHNOLOGY BEHIND SUCCESSFUL BUSINESS INTELLIGENCE

- Enterprise-wide database: Structured and unstructured data.
- Effective MIS, Dashboard: Current state of business at fingertip.
- Business analytics to capture actionable insight: Learning from data.
- Training in analytical methods: Prediction, ML, AI for strategy development.
- Understanding both business and analytics to implement correct decision.
- Business environment is paying increased attention for expert system technology to aid in decision making. Business expert system (BES) is computer application that provides decision support similar to that of human expert in solving problems. Through expert system rules can be associated to evaluate the condition and determine the result.
- Business and Environment are more complex today than ever before. There are several areas in business that employed business intelligence to help the business run more efficient and fast in decision making.
- Expertise and experience are key factors for experts in strategic analysis in order to give advice on strategic matters such as strength or the competitive position of an enterprise.

- Expert system can be more powerful when it has been combined with other artificial intelligence techniques such as fuzzy logic, neural network, natural processing language and many more.

KNOWLEDGE BASE - This is the data the systems work with. Modern knowledge bases have automation capabilities that curate data from a wide range of relevant internal and external sources.

INFERENCE ENGINE - The part of the system that applies logical rules that pull intelligent insights from the knowledge base based on user queries.

USER INTERFACE - The means through which users interact with the knowledge base (usually a software platform)

BENEFITS OF BI TOOLS

- Faster analysis
- Increased competitive advantage
- Improved customer experience
- Organizational efficiency
- Intuitive dashboards
- Data-driven business
- Governed data

Business intelligence ingests data and converts it into visuals, reports and dashboards to identify trends, patterns and points of interest. The technical infrastructure that collects, stores and analyzes company data.

HOW BUSINESS INTELLIGENCE ARE IMPLEMENTED?

STEP1: Raw data from corporate databases is extracted. The data could be spread across multiple systems heterogeneous systems.

STEP2: The data is cleaned and transformed into the data warehouse. The table can be linked and cubes are formed.

STEP3: Using BI system the user can ask queries, request or conduct any other analysis.

APPLICATIONS OF EXPERT SYSTEM

1. Planning and scheduling
2. The configuration of manufactured objects
3. Financial decision making knowledge publishing
4. Process monitoring and control
5. Supervise the operation of the plan
6. Information management
7. Loan analysis

IMPORTANCE OF BI:

MEASUREMENT: Creating KPI (KEY PERFORMANCE INDICATORS) based on historic data

Identify and set bench marks

With BI systems organizations can identify market trends and spot business problems that need to be addressed.

BI helps on data visualization that enhances the data quality and thereby the quality of decision making.

BI systems can be used not just by enterprises but SME (SMALL AND MEDIUM ENTERPRISE)

BUSINESS INTELLIGENCE APPLICATIONS

Multidimensional Analysis or online analytical processing (OLAP)

"Slicing and Dicing" data stored in a dimensional format, drilling down in greater detail and aggregating data.

Data Mining

Searching for information in a database, data warehouse or data mart.

Decision Support System

Combine models and data in an attempt to solve semistructured /unstructured problems

DSS CAPABILITIES:

- ✓ Sensitivity analysis
- ✓ What-if analysis
- ✓ Goal seeking analysis

Enables users to employ multi-dimensional analysis, data mining and decision support systems to analyze data in real time

CONCLUSION

Expert system is well known in business as a system that improves company decision making, knowledge management and strategic management. ES still can be the best decision making tool in business if the ES combine with other intelligent techniques. Furthermore, the developer and the company should take a serious action in representing the knowledge into the system for better ES. A business expert system has been designed to provide expert advice to the business executive.

REFERENCES

1. https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_expert_systems.htm
2. https://www.techtarget.com/searchbusinessanalytics/definition/business-intelligence-BI
3. Introduction To Information Technology-Puneet Kumar

ID: 48

BLOCK CHAIN

BHAVADHARINI.C
II B.Com,
Department of E-Commerce,
KPR College of Arts Science and Research,
Coimbatore, Tamil Nadu, India.

SINDHU.N
II B.Com,
Department of E-Commerce,
KPR College of Arts Science and Research,
Coimbatore, Tamil Nadu, India.

ABSTRACT

A block chain is a distributed database or ledger shared among a computer network's nodes.They are best known for their crucial role in cryptocurrency systems for maintaining a secure and decentralized record of transactions,but they are not limited to cryptocurrency uses. Block chains can be used to make data in any industrial immutable-the term used to describe the inability to be altered.Because there is no way to change a block,the only trust needed is at the point where a user or program enters the data.This aspect reduces the need for trusted third parties,which are usually auditors or other humans that add costs and make mistakes.

KEYWORDS

- Shared database
- Block chain decentralization
- Cryptocurrency
- Block chain transparency

BLOCK CHAIN

- ✓ Blockchain is a type of shared database that differs from a typical database that differs from a typical database in the way it stores information; blockchain store data in blocks linked together via cryptography.
- ✓ Different types of information can be stored on a blockchain, but the most common use for transactions has been a ledger.
- ✓ In Bitcoin's case, blockchain is decentralized so that no single person or group has control-instead all users collectively retain control.
- ✓ Decentralized blockchains are immutable, which means that the data entered is irreversible.
- ✓ Block chains can be used to immutably record any number of data points. This could be in the form of transactions, votes in an election, product inventories, state identifications, deeds to homes and much more.
- ✓ Currently, tens of thousands of projects are looking to implement blockchains in various ways to help society other than just recording transactions.

HOW DOES A BLOCKCHAIN WORK?

- ✓ A blockchain is somewhat similar because it is a database where information is entered and stored.But the key differemce between a traditional database or spreadsheet and a blockchain is how the data is structured and accessed.
- ✓ A block chain consist of programs called scripts that conduct the tasks you usually would in a database.
- ✓ Entering and accessing information and saving and storing it somewhere.
- ✓ A blockchain is distributed, which means multiple copies are saved on many machines,and they must all match for it to be valid.
- ✓ The blockchain collects transaction information and enters into a block like a cell in a spreadsheet containing information.
- ✓ Once it is full the information is run through an encryption algorithm, which creates a hexadecimal number called the hash.
- ✓ The hash is then entered into the following block header and encrypted with the other information in the block.This creates a series of blocks that are chained together.

BLOCKCHAIN DECENTRALIZATION

A blockchain allows the data in a database to be spread out among several network nodes-computers or devices running software for the blockchain-at various locations.This not only creates redundancy but maintains the fidelity of the data. Forexample, if someone tries to alter a record at one instance of the database, the other, the other nodes would prevent it from happening. This way, no single node within the network can alter information held within it.Because of this distribution-and the encrypted proof that work was done-the information and history are irreversible. Such a record could be a list of transactions, but it also is possible for a blockchain to hold a variety of information like legal contracts, state identifications, or a company inventory.

BLOCKCHAIN TRANSPARENCY

Because of the decentralized nature of the bitcoin blockchain, all the transactions can be transparently viewed by either having a personal node or using blockchain explorers that allow anyone to see transactions occurring live.Each node has its own copy of the cahin that gets updated as fresh blocks are confirmed and added.This means that if you wanted to,you could take a bitcoin wherever it goes.

IS BLOCKCHAIN SECURE?

Blockchain technology achieves decentralized security and trust in several ways.To begin with,new blocks are always stored linearly and chronologically.That is they are always added to the “end”of the blockchain.After a block has been added to the end of the blockchain,previous blocks cannot be changed. A change in any data changes the hash of the block it was in.Because each block contains the previous block’s hash, a change in one would change the following blocks.The network would reject an altered block because the hashes would not match.They would need to control a majority of the network to do this and insert it at just the right moment.This is known as a 51% attack because you need to control more than50%of the network to attempt it.

PROS AND CONS OF BLOCKCHAIN

For all of its complexity,blockchains potential as a decentralized form of record-keeping is almost without limit.From greater user privacy andheightened security to lower processing fees and fewer errors,blockchain technology may very well see applications beyond these outlined above.But there are also some diasvantages.

PROS

- ➢ Improved accuracy by removing human involvement in verification.
- ➢ Cost reductions by eliminating third party verification.
- ➢ Decentralization makes it harder to tamper with.
- ➢ Transactions are secure, private and efficient.
- ➢ Transparency technology
- ➢ Provides a banking alternative and a way to secure personal information for Citizens of coubtries with unstable or underdeveloped governments.

CONS

- ➢ Significant technology cost associated with some blockchains.
- ➢ Low transactions per second.
- ➢ History of use in illicit activities such as on the dark web.
- ➢ Regulation varies by jurisdication and remains uncertain.
- ➢ Data storage limitations.

BENEFITS OF BLOCKCHAINS

- ➢ Accuracy of the chain.
- ➢ Cost reductions.
- ➢ Decentralization.
- ➢ Efficient Transactions.
- ➢ Private transactions.
- ➢ Secure transactions.
- ➢ Transparency.
- ➢ Banking the unbanked.

DRAWBACKS OF BLOCKCHAINS

- ➢ Technology cost.
- ➢ Speed and data inefficiency.
- ➢ Illegal Activity.
- ➢ Regulation.

BLOCKCHAIN APPLICATION

- ➢ Digital Id.
- ➢ Banking.
- ➢ Voting.
- ➢ Virtual currency Bitcoin.
- ➢ Secure sharing of Medical Records.
- ➢ Supply chain and logistic monitoring.
- ➢ Original content creation.

BLOCKCHAIN TECHNOLOGY

Blockchain technology uses hashing and encryption to secure the data, relying mainly on the SHA256 algorithm to secure the information. The address of the sender (Public Key), the receiver's address, the transaction and his/her private key details are transmitted via the SHA256 algorithm.

Blockchain is a record - keeping technology designed to make it impossible to hack the system or forge the data stored on it, thereby making it secure.

BLOCK CHAIN TECHNOLOGY

1. A blockchain is an online database, popularly used for cryptocurrency transactions, that stores information chronologically and in blocks
2. Information is recorded in a block.
3. Once the block is full, it's chained to the rest of the blocks in the database.
4. An empty block is added to the end of the chain to record new information.

FIVE KEY ELEMENTS OF BLOCKCHAIN TECHNOLOGY

- Encryption
- Immutability
- Tokenization
- Decentralization
- Distribution

FEATURES OF BLOCKCHAIN

- ✓ Increased capacity
- ✓ Better security
- ✓ Immutability
- ✓ Faster settlement
- ✓ Decentralised system
- ✓ Consensus
- ✓ Distributed ledger
- ✓ Cannot be corrupted

The control/power is not held by a single entity. Instead it is distributed among multiple participants.Even if one node is corrupted/fails, the network repairs itself. Direct peer to peer transaction of data or finance. Decentralized nature of blockchain instills trust in the process such that two unknown parties can directly interact / transact with each other. Data is distributed among the nodes. Even if one node is tampered the data does not get compromised.

CLASSIFICATION OF BLOCKCHAIN

PUBLIC BLOCKCHAIN

The public blockchain is a chain of information that has no restrictions for its access.Any user in the network needs no permission to access the history of blockchain or carry out any kind of transaction.The information can easily be transferred and accessed by people all around the world on this type of network of blockchain without any prior permission. Bitcoin blockchain is an example of a public blockchain.

PRIVATE BLOCKCHAIN

This type of blockchain network needs permission for accessing the information. This type of blockchain network cannot be joined by anyone without the permission of the owner. The digital ledger in this type of blockchain is shared among only the trusted members. This type of network is usually managed by different organisations and enterprises.

HYBRID BLOCKCHAIN

This type of blockchain network refers to the mixture of the attributes of both centralized and decentralized blockchains.The working of the hybrid blockchain depends upon the percentage of centralization and decentralization.

SIDECHAINS

This is the network of blockchain that executes parallel to the primary blockchain.The side chains work independently from the primary blockchain.

NEED OF BLOCK

The advent of the internet and different technologies has resulted in several digital technologies in the world.Blockchain is a new and emerging concept in society but is becoming popular at a very fast pace. Earlier when there was no such technology the records and information were noted and maintained in the written format by the people. There was a maximum chance of errors when the data was noted manually. Moreover the data and information that could easily be edited later gave rise to corruption.

There is a dire need for a technology like a blockchain that will ensure the security of the recorded data with full transparency. This will also help in gaining the trust of people and they can access the information in the blocks without any kind of fear of cheating.The copy of the transaction is available on all the computers linked with the blockchain network and this validates the security of the transactions. This technology is preferred by the banks for thr process of money transfers, storing records and different technical works.

CONCLUSION

Blockchain is an emerging technology and its use at present is limited only to cryptocurrency.This technology is also said to be useful in different sectors in the coming future.It is being tested for the sane in different sectors in several countries of the world.It is brought into use in the banking and commerce sectors in different countries. This technology is becoming popular at a very fast pace and there are many fields where it can bring revolutionary changes.The advent of such technology is dire of need in coming future.

REFERENCES

1. https://www.investopedia.com/terms/b/blockchain.asp
2. https://www.ibm.com/topics/blockchain
3. https://builtin.csom/blockchain

ID: 49

A NOVEL STUDY OF DATA SECURITY IN REAL WORLD ENVIRONMENT

Dr.P.PRASANYA DEVI
Guest Lecturer,
Department of Computer Science,
Government Arts College,
Melur, Tamil Nadu, India.

Dr.M.RAVINDRAN
Associate Professor,
Department of Computer Science,
Government Arts College,
Melur, Tamil Nadu, India.

ABSTRACT

Data are today an asset more critical than ever for all organizations we may think of. Recent advances and trends, such as sensor systems, IoT, cloud computing, and data analytics, are making possible to pervasively, efficiently, and effectively collect data. However for data to be used to their full power, data security and privacy are critical. Data security is the practice of protecting digital information from unauthorized access, corruption or theft throughout its entire lifecycle. It's a concept that encompasses every aspect of information security from the physical security of hardware and storage devices to administrative and access controls, as well as the logical security of software applications. It also includes organizational policies and procedures.

KEYWORDS: *Data Security, Data Privacy, Data Protection*

I. Introduction

Data Security [1] safeguards digital data from unwanted access, corruption, or theft. It is a notion that imparts physical security to hardware and software devices and covers all aspects of **information security** [2]. It also imparts administrative and access controls and logical security to software applications. It also covers policies and procedures to be followed by a company.

Data is a valuable asset that generates, acquires, saves, and exchanges for any company. Protecting it from internal or external corruption and illegal access protects a company from financial loss, reputational harm, consumer trust degradation, and brand erosion. Furthermore, regulations for securing data, imposed by the government and the industry, make it critical for a company to achieve and maintain compliance wherever it does business.

1.1. Main elements of Data Security

The three components of Data Security that all companies should adhere to are confidentiality, integrity, and availability. The CIA [3] triad is a security paradigm and framework for the protection of data. Here is what each fundamental piece implies in terms of preventing unwanted access and data exfiltration.

- **Confidentiality:** Ensures that only authorized users, with appropriate credentials, have access to data.
- **Integrity:** Ensures that all data is accurate, trustworthy, and not prone to unjustified changes.
- **Availability:** Ensures that data is accessible and available for ongoing business needs in a timely and secure manner.

1.2. Aftermath of Data Breach

Data breaches, which occur when data is accessed in an unauthorized manner, are a major concern for organizations of all shapes, sizes and industries. In fact, 63% of respondents to a KPMG study said they suffered a data breach or cyber incident in 2021 -- and that number is only projected to grow.

Data breaches are attributed to a number of cyber incidents, including the following:

- Accidental leaks or exposures
- Phishing attacks
- Distributed denial-of-service attacks

- Physical breaches
- Lack of access controls
- Backdoors

Even the largest and most well-known companies are susceptible to breaches, as evidenced in the 10 biggest data breaches in history.

Data breaches [4] can result in hefty remediation costs, as well as expenses stemming from downtime and lost business. Regulatory and legal fines may also be levied. In worst-case scenarios, companies can go bankrupt or out of business.Data security is an important component in data compliance, the process that identifies governance and establishes policies and procedures to protect data. The process involves selecting applicable standards and implementing controls to achieve the criteria defined in those standards. Regulatory [15] compliance, which refers to organizations following local, state, federal, international and industry laws, policies and regulations -- is related to data compliance. Regulatory compliance standards require the use of certain controls and technologies to meet the criteria defined in them. The following are some of the most common compliance regulations:

- PCI DSS
- HIPAA
- Federal Information Security Modernization Act of 2014
- Sarbanes-Oxley Act
- GDPR
- CCPA

HIPAA, for example, outlines provisions to safeguard medical information in the U.S. Among other mandates, healthcare organizations must adhere to standards for patient data security or else face noncompliance fines and penalties. PCI DSS is a global standard aimed at protecting credit, debit and cash card transaction data. It sets guidelines for cardholder data, access controls and networks that process payment information.

II. Types of Data security

Before an organization can secure data, it has to know what data it has. This is where a data inventory --a record of all the data created, used and stored by a company -- is key. The process starts with data discovery, or learning what and where the data is. Data classification follows, which involves labeling data to make it easier to manage, store and secure. The four standard data classification categories are as follows:

1. Public Information
2. Confidential Information
3. Sensitive Information
4. Personal Information

Data is often further broken down by businesses using common classification labels, such as "business use only" and "secret." Sensitive data is often classified as confidential or secret. It includes these types of data:

- Personally Identifiable Information
- Protected Health Information
- Electronic Protected Health Information
- PCI Data
- Intellectual Property

Compounding the difficulty of doing data inventory and classification is that data can reside in many locations on premises, in the cloud, in databases and on devices, to name a few.

Data also can exist in three states:

1. **In Motion**, Meaning Data That Is Being Transported;
2. **At Rest**, meaning data that is being stored, or data that is at its destination -- i.e., not transported or in use; and
3. **In use**, meaning data that is being written, updated, changed and processed -- i.e., not being transported or stored.

Andrew Froehlich, network security expert and president of West Gate Networks, offers best practices on how to secure data in each state.Because no single form of data exists, no single magic-bullet technique can secure all data. A defense-in-depth data security strategy is made up of a combination of tools, techniques and policies. Figure 1 shows data security technologies include the following:

- Encryption
- Data masking
- Access control
- Data loss prevention (dlp)
- Data backup and resiliency

Multiple technologies and techniques must be used in an enterprise data security strategy.

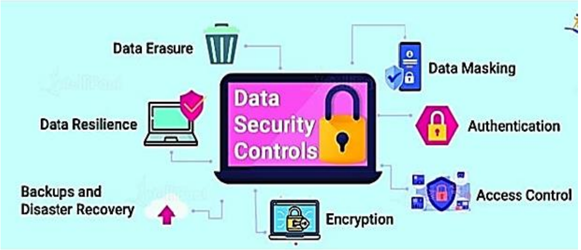


Fig.1 Data security Controls

III. ENCRYPTION

Encryption is the process of converting readable plaintext into unreadable ciphertext using an encryption algorithm, or cipher. If encrypted data is intercepted, it is useless as it cannot be read or decrypted by anyone who does not have the associated encryption key.

Symmetric and asymmetric encryption are two commonly used ciphers:

- Symmetric encryption uses a single secret key for both encryption and decryption. The Advanced Encryption Standard is the most commonly used algorithm in symmetric key cryptography.
- Asymmetric encryption uses two interdependent keys: a public key to encrypt the data and a private key to decrypt the data. The Diffie-Hellman key exchange and Rivest-Shamir-Adleman are two common asymmetric algorithms.

Both symmetric and asymmetric encryption have pros and cons. Security expert Michael Cobb explains the differences between the ciphers and discusses why a combination of the two might be the fastest, most secure encryption option.

Data Masking

Data masking involves obscuring data so it cannot be read. Masked data looks similar to the authentic data set but reveals no sensitive information. Legitimate data is replaced so the masked data maintains the characteristics of the data set as well as referential integrity across systems, thereby ensuring the data is realistic, irreversible and repeatable.

Below are some common data masking techniques:

- Scrambling
- Substitution
- Shuffling
- Data Aging
- Variance
- Masking Out
- Nullifying

Data masking is useful when certain data is needed for software testing, user training and data analysis -- but not the sensitive data itself.While the end result of encryption and masking are the same -- both create data that is unreadable if intercepted -- they are quite different. Expert Cobb outlines the key differences between the two, as well as use cases for each.

Access Control

One of the best ways to secure data is to control who has access to it. If only authorized individuals can view, edit and delete data, it is inherently safer than an access free-for-all. Access control involves two main processes:

1. Authentication is the process of ensuring users are who they say they are.
2. Authorization is the process of ensuring authenticated users have access to the necessary data and resources.

Authentication and authorization are components of an enterprise identity and access management (IAM) strategy. Other fundamental IAM processes and techniques include multifactor authentication (MFA), principle of least privilege access, role-based access control and privileged access management. Also important is following password hygiene best practices, such as setting minimum password lengths, requiring unique passwords and considering regular password changes. Take a deep dive into these topics and more in our IAM guide.Using a zero-trust access control strategy is growing in popularity. This framework provides stringent access control on a continuous basis. Get the lowdown on this up-and-coming trend in our guide to zero trust.

Data loss prevention

An integral tool for any enterprise security strategy is a DLP platform. It monitors and analyzes data for anomalies and policy violations. Its many features can include data discovery, data inventory, data classification and analysis of data in motion, at rest and in use. Many DLP tools integrate with other technologies, such as SIEM systems, to create alerts and automated responses.

Data Backup

Data backup involves creating copies of files and databases to a secondary, and often tertiary and quaternary, location. If the primary data fails, is corrupted or gets stolen, a data backup ensures it can be returned to a previous state rather than be completely lost. Data backup is essential to disaster recovery plans.

- **DLP.** Data loss prevention plays a critical role in enforcing data security policies at a granular level -- blocking a user from emailing or downloading a protected file, for example. DLP can prevent unauthorized access and alert cybersecurity staff to violations and suspicious behavior.
- **Data Life-Cycle Management.** DLM is an automated approach to keeping massive amounts of digital information accurate, confidential, secure and available -- and destroying it in a safe and timely fashion, in keeping with enterprise policies -- all while meeting relevant compliance requirements. DLM policies are based on data attributes such as type, size, age and classification. The main phases of the data lifecycle in a DLM framework include the following:
    - Generation and collection
    - Processing and storage
    - Usage
    - Archiving
    - Destruction

DLM tools can automatically sort data into separate tiers based on specified policies. This lets enterprises use storage resources efficiently and effectively by assigning top-priority data to high-performance storage, for example.

- **Patch management.** Leaving a known vulnerability unpatched is like failing to fix a broken lock on the side door of an otherwise secure home. Patch software quickly and often to limit the ways attackers can gain access to enterprise property.
- **Security awareness training.** Intentional and unintentional mistakes of staff, contractors and partners represent one of the greatest threats to data security. Security awareness training is therefore of utmost importance to educate users on organizational security policies and topics such as phishing attacks.
- **User behavior analytics.** UBA, also known as user and entity behavior analytics (UEBA), flags attempts to gain unauthorized or unusual levels of access to sensitive data. Among top UEBA use cases, the technology can help detect lateral network attacks, identify compromised user accounts and uncover insider threats.

IV. Data security vs. Data privacy vs. Data protection

Data security, data privacy and data protection are overlapping but technically distinct concepts.

- ➢ **Data security.** Data security has a broader scope, aiming to protect digital information not just from unauthorized access but also from intentional loss, unintentional loss and corruption. While data privacy primarily focuses on the confidentiality part of the CIA triad, data security is equally concerned with information's integrity and accessibility.For example, imagine threat actors obtain a confidential file, but encryption successfully prevents them from reading the data. The information itself stays inaccessible, and data privacy remains intact. The attackers are still able to corrupt or destroy the illegible file, however, which is a security failure.
- ➢ **Data privacy.** The goal of data privacy is to make sure the ways an organization collects, stores and uses sensitive data are responsible and in compliance with legal regulations. Privacy policies and measures prevent unauthorized parties from accessing data, regardless of their motivation and whether they are internal end users, third-party partners or external threat actors.
- ➢ **Data protection.** Data protection ensures digital information is backed up and recoverable if it's lost, corrupted or stolen. Data protection is an important part of a larger data security strategy, serving as a last resort if all other measures fail.

3.1. Data security risks and challenges

In 2017, The Economist declared "The world's most valuable resource is no longer oil, but data." Unfortunately, data is more difficult to protect and easier to steal, and it presents enormous opportunity to not just businesses but also criminals. Today's enterprises face an uphill battle when it comes to securing their data. Consider the following perennial risks and challenges.Insider threats can be categorized in three main buckets: compromised, negligent or malicious individuals.

Insider threats. One of the biggest threats to data security is the enterprise end user, whether that's a current or former employee, third-party partner or contractor. Malicious insiders sometimes use their legitimate access privileges to corrupt or steal sensitive data, either for profit or to satisfy personal grudges.

Unintentional insider threats are no less dangerous. An innocent click on a link in a phishing email could compromise a user's credentials or unleash ransomware or other malware on corporate systems. In fact, in nearly 40% of data breaches, attackers used either compromised credentials or phishing as initial attack vectors, according to the Ponemon Institute's 2021 "Cost of a Data Breach" report, sponsored by IBM.

Simple end-user negligence or carelessness -- absent any malicious threat actor -- can also result in accidental exposure of sensitive data. An employee might email confidential information to the wrong person, for example, or upload it to an unprotected cloud account. In addition, someone could lose a laptop and fail to report it to IT, leaving the device vulnerable to whoever happens to find it.

Misconfigurations. Technical misconfigurations pose another major threat, regularly resulting in accidental exposure of confidential data sets. The Ponemon Institute found cloud misconfigurations alone were responsible for 15% of data breaches in 2021.

Third-party risk. An organization is arguably only as secure as its least secure third-party partner, whether that's a supplier, contractor or customer. Consider the infamous Solarwinds supply chain attack, which enabled threat actors to target the vendor's customers' networks. Organizations point to vulnerable third-party software as the initial attack vector in 14% of data breaches, according to the Ponemon Institute.

3.2. Data security policy

It's important to develop an overarching strategy for deploying data security tools, technologies and techniques such as those listed above. According to consultant Kirvan, every enterprise needs a formal data security policy[9] to achieve the following critical aims:

- Codify data security expectations and responsibilities; and
- Demonstrate compliance with relevant data privacy and security laws and standards.

3.2.1. Data privacy

Data privacy generally means the ability of a person to determine for themselves when, how, and to what extent personal information about them is shared with or communicated to others. This personal information can be one's name, location, contact information, or online or real-world behavior. Just as someone may wish to exclude people from a private conversation, many online users want to control or prevent certain types of personal data collection.

As Internet usage has increased over the years, so has the importance of data privacy. Websites, applications, and social media platforms often need to collect and store personal data about users in order to provide services. However, some applications and platforms may exceed users' expectations for data collection and usage, leaving users with less privacy than they realized. Other apps and platforms may not place adequate safeguards around the data they collect, which can result in a data breach that compromises user privacy.

3.2.2. Importance of data privacy

In many jurisdictions, privacy is considered a fundamental human right, and data protection laws exist to guard that right. Data privacy is also important because in order for individuals to be willing to engage online, they have to trust that their personal data will be handled with care. Organizations use data protection practices to demonstrate to their customers and users that they can be trusted with their personal data.Personal data can be misused in a number of ways if it is not kept private or if people don’t have the ability to control how their information is used:

- Criminals can use personal data to defraud or harass users.
- Entities may sell personal data to advertisers or other outside parties

without user consent, which can result in users receiving unwanted marketing or advertising.

- When a person's activities are tracked and monitored, this may restrict their ability to express themselves freely, especially under repressive governments.

For individuals, any of these outcomes can be harmful. For a business, these outcomes can irreparably harm their reputation, as well as resulting in fines, sanctions, and other legal consequences .In addition to the real-world implications of privacy infringements, many people and countries hold that privacy has intrinsic value: that privacy is a human right fundamental to a free society, like the right to free speech.

3.2.3. Laws that govern data privacy

As technological advances have improved data collection and surveillance capabilities, governments around the world have started passing laws regulating what kind of data can be collected about users, how that data can be used, and how data should be stored and protected. Some of the most important regulatory privacy frameworks to know include:

- General Data Protection Regulation (GDPR): Regulates how the personal data of European Union (EU) data subjects, meaning individuals, can be collected, stored, and processed, and gives data subjects rights to control their personal data (including a right to be forgotten).
- National data protection laws: Many countries, such as Canada, Japan, Australia, Singapore, and others, have comprehensive data protection laws in some form. Some, like Brazil's General Law for the Protection of Personal Data and the UK's Data Protection Act, are quite similar to the GDPR.
- California Consumer Privacy Act (CCPA): Requires that consumers be made aware of what personal data is collected and gives consumers control over their personal data, including a right to tell organizations not to sell their personal data.

3.2.4. Fair Information Practices

Many of the existing data protection laws are based on foundational privacy principles and practices, such as those laid out in the Fair Information Practices[11]. The Fair Information Practices are a set of guidelines for data collection and usage. These guidelines were first proposed by an advisory committee to the U.S. Department of Health, Education, and Welfare in 1973. They were later adopted by the international Organization for Economic Cooperation and Development (OECD) in its Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.

The Fair Information Practices are:

- Collection limitation: There should be limits to how much personal data can be collected
- Data quality: Personal data, when collected, should be accurate and related to the purpose it is being used for
- Purpose specification: The use for personal data should be specified
- Use limitation: Data should not be used for purposes other than what was specified
- Security safeguards: Data should be kept secure
- Openness: Personal data collection and usage should not be kept secret from individuals
- Individual participation: Individuals have a number of rights, including the right to know who has their personal data, to have their data communicated to them, to know why a request for their data is denied, and to have their personal data corrected or erased
- Accountability: Anyone who collects data should be held accountable for implementing these principles

3.2.5. Challenges users face when protecting their online privacy

- **Online tracking:** User behavior is regularly tracked online. Cookies often record a user's activities, and while most countries require websites to alert users of cookie usage, users may not be aware of to what degree cookies are recording their activities.
- **Losing control of data:** With so many online services in common use, individuals may not be aware of how their data is being shared beyond the websites with which they interact online, and they may not have a say over what happens to their data.
- **Lack of transparency:** To use web applications, users often have to provide personal data like their name, email, phone number, or location;

meanwhile, the privacy policies associated with those applications may be dense and difficult to understand.

- **Social media:** It is easier than ever to find someone online using social media platforms, and social media posts may reveal more personal information than users realize. In addition, social media platforms often collect more data than users are aware of.
- **Cyber crime:** Many attackers try to steal user data in order to commit fraud, compromise secure systems, or sell it on underground markets to parties who will use the data for malicious purposes. Some attackers use phishing attacks to try to trick users into revealing personal information; others attempt to compromise companies internal systems that contain personal data.

3.2.6. Cloudflare take to protect privacy

Cloudflare [12]believes data privacy is core to the mission of helping build a better Internet. Cloudflare products are built with privacy in mind, and Cloudflare has released a number of services designed to protect online user privacy:

- 1.1.1.1 is a free DNS resolver that does not track or store DNS queries (unlike many other DNS resolvers, which may sell this information to advertisers)
- Cloudflare supports DNS over HTTPS, which completely encrypts DNS queries
- Cloudflare offers free SSL for any website that uses Cloudflare
- Project Galileo protects the privacy of important vulnerable organizations free of charge
- Cloudflare Web Analytics enables businesses to analyze traffic to their websites without compromising their users' privacy

V. DATA PROTECTION

Data protection[13] is the process of safeguarding important data from corruption, compromise or loss and providing the capability to restore the data to a functional state should something happen to render the data inaccessible or unusable.Data protection assures that data is not corrupted, is accessible for authorized purposes only, and is in compliance with applicable legal or regulatory requirements. Protected data should be available when needed and usable for its intended purpose.The scope of data protection, however, goes beyond the notion of data availability and usability to cover areas such as data immutability, preservation, and deletion/destruction.Roughly speaking, data protection spans three broad categories, namely, traditional data protection (such as backup and restore copies), data security, and data privacy as shown in the Figure below. The processes and technologies used to protect and secure data can be considered as data protection mechanisms and business practices to achieve the overall goal of continual availability, and immutability, of critical business data.

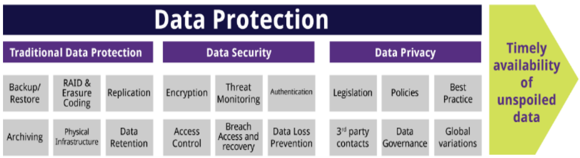


Fig.2 The Three Categories of Data Protection

VI. Conclusion

The data collected should be specifically for the intended purpose. There should be a minimum data requirement and accountability of the website holder.In a global ,digital world, most import tn more many years by diligently safeguarding yourant currency for any business is trust.Trust is something you ear customer's personal data.they trust you with a lot of essential information their address,their payment details,their preferences even their biometric data. The data security is most important in any business managements. It is necessary to protect company against a multiple security threats. With info protection turning into such an essential a locality of corporation operations, consultants have tried to extend new and effective ways in which to take care of sensitive statistics out of the incorrect palms. This includes facts close to the products or services provided through the corporate, industrial enterprise plans, and money plans, facts regarding the patrons of the products or services. Once it involves going for walks an advertisement enterprise of any size, all facts area unit important data. Facts safety is very important for every enterprise, no matter its size.

VII. REFERENCES

1. Cao, Y., Huang, L., Li. Y., Jermsittiparsert, K., Ahmadi Nezamabad, H., & Nojavan, S. 2020. "Optimal Scheduling of Electric Vehicles Aggregator under Market Price Uncertainty Using Robust Optimization Technique." International Journal of Electrical Power & Energy Systems 117: 105628.
2. Yu, D., Wang, Y., Liu, H., Jermsittiparsert, K., & Razmjooy, N. 2019. "System Identification of PEM Fuel Cells Using an Improved Elman Neural Network and a New Hybrid Optimization Algorithm." Energy Reports 5: 1365-1374.
3. Tian, M., Ebadi, A., Jermsittiparsert, K., Kadyrov, M., Ponomarev, A., Javanshir, N., & Nojavan, S. 2019. "RiskBased Stochastic Scheduling of Energy Hub System in the Presence of Heating Network and Thermal Energy Management." Applied Thermal Engineering 159: 113825.
4. Yu, D., Wnag, J., Li, D., Jermsittiparsert, K., & Nojavan, S. 2019. "Risk-Averse Stochastic Operation of a Power System Integrated with Hydrogen Storage System and Wind Generation in the Presence of Demand Response Program." International Journal of Hydrogen Energy (In press), DOI: 10.1016/j.ijhydene.2019.09.222.
5. Jabarullah, N., Jermsittiparsert, K., Melnikov, P., Maseleno, A., Hosseinian, A., & Vessally, E. 2019. "Methods for the Direct Synthesis of Thioesters from Aldehydes: A Focus Review." Journal of Sulfur Chemistry (In press), DOI: 10.1080/17415993.2019.1658764.
6. Jiao, Y., Jermsittiparsert, K., Krasnopevtsev, A., Yousif, Q., & Salmani, M. 2019. "Interaction of Thermal Cycling and Electric Current on Reliability of Solder Joints in Different Solder Balls." Materials Research Express 6 (10): 106302.
7. Yu, D., Ebadi, A., Jermsittiparsert, K., Jabarullah, N., Vasiljeva, M., & Nojavan, S. 2019. "Risk-constrained Stochastic Optimization of a Concentrating Solar Power Plant." IEEE Transactions on Sustainable Energy (In press), DOI: 10.1109/TSTE.2019.2927735. 8. Jermsittiparsert, K., Sriyakul, T., Sutduean, J., & Singsa, A. 2019. "Determinants of Supply Chain Employees Safety Behaviours." Journal of Computational and Theoretical Nanoscience 16 (7): 2959-2966.
8. Ashutosh Bhatt (2014). How Internet Works on Mobile Devices. p1-11; https://www.engineersgarage.com/how_to/how-internet-works-on-mobile-devices/ Business
9. Technology Standard. Security and data protection. p1-6; https://www.managebt.org/book/strategy-and-governance/security-and-dataprotection/
10. Circadence (2020). The future of finance cyber security in 2020. p1-3; https://www.circadence.com/blog/the-future-of-finance-cyber-security-in-2020/
11. Denis Otieno (2020). Cyber security threats and trends for 2020. p1-14; https://just40days.com/detail_Cybersecurity-Threats-and-Trends-for-2020_37750
12. Elisa Bertino (2016). Introduction to Data Security and Privacy, p1-6; https://link.springer.com/article/10.1007/s41019-016-0021-1
13. Forcepoint. What is Data Security? Data security defined. explained and explored. p1 6; https://www.forcepoint.com/cyber-edu/data-security
14. Government Initiatives (2013). https://baliyans.com/courses/disaster-managementand-internal-security/cyber-security/government-initiatives http://sielearning.tafensw.edu.au/toolboxes/Database_Administration/content/security /threats.htm
15. Interaction Design Foundation (2020, p.1). https://www.interactiondesign.org/literature/topics/usability

ID: 50

BIG DATA AND BIG DATA ANALYTICS: CONCEPTS, TOOLS AND APPLICATIONS

Dr.C.NAGARANI
Assistant Professor,
Department of Computer Science,
PSG College of Arts & Science,
Coimbatore, Tamil Nadu, India.

Mrs.R.SATHYA
Assistant Professor,
Department of Computer Science,
PSG College of Arts & Science,
Coimbatore, Tamil Nadu, India.

ABSTRACT

Large datasets that cannot be analyzed using conventional computing methods are referred to as "big data." Big data is data with larger in size. It refers to the volume of data that a single computer system cannot store in its memory. In our data-driven digital age, big data will only continue to get bigger, more challenging to understand, and generated at a much faster rate with each passing day. Big Data analytics is a method for gaining useful information, like secret patterns, unforeseen correlations, industry trends, and consumer preferences. Big Data analytics has several benefits and may be applied to improve decision-making and stop fraud, among other things. As data transfers, smart gadgets, web usage, and other factors increase, this industry is still developing. Big data analytics techniques are still employed today together with cutting-edge innovations like machine learning to find and scale more complicated insights.

KEYWORDS: *Big Data, Big Data Analytics, Correlations, Secret Patterns, Decision-Making*

I. INTRODUCTION

Big Data analytics can be used to find ways to combine the enormous volumes of complicated data produced by sensors and networks, going all the way back to the early days of computing when businesses first started utilizing computers to store and analyze huge amounts of data. Big Data analytics didn't take off, though, until the late 1990s and early 2000s, when businesses started turning more and more to computers to help them make sense of the huge amounts of data that were being produced by their operations at an exponential rate. Today, Big Data analytics has developed into a critical tool for businesses of all sizes in a variety of sectors. Organizations may now acquire insights into their customers, their markets, and the environment in which they operate that were previously impossible by utilizing the power of big data. In the years to come, we may anticipate seeing even more incredible and revolutionary applications of this technology as the discipline of big data analytics continues to develop.

Types of Big Data

1. Structured
2. Unstructured
3. Semi-structured

1. Structured

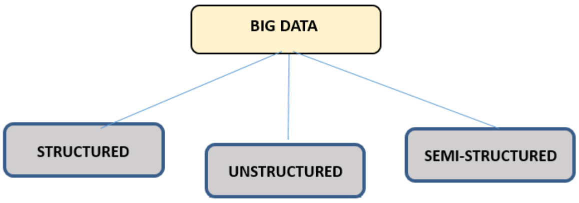
Structured data refers to any data that can be accessed, processed, and stored in a fixed format.

2. Unstructured

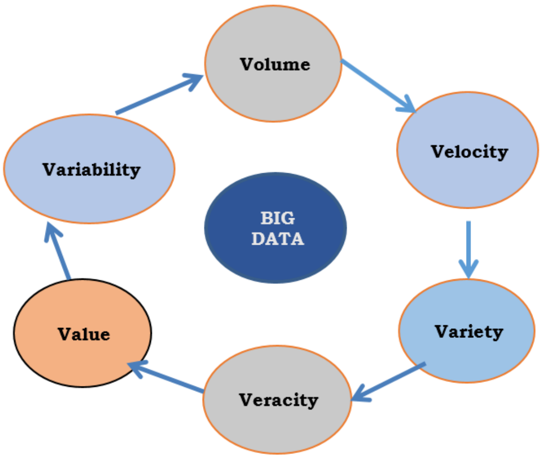
Any information that is in an ambiguous format or structure, such as a collection of simple text files, photos, or videos, etc.

3. Semi-Structured

Both types of data are compatible with semi-structured data. Although semi-structured data appears to be structured, it is not defined, for example, by a table definition in a relational DBMS.



2. SIX Vs of BIG DATA



2.1 Volume

The volume of data is enormous. Size of the data is a very important factor when determining the value of the data. When the amount of data is extraordinarily vast, it is truly referred to as "Big Data." This means that the amount of data will determine whether a certain set of data can truly be categorized as a big data set or not. Therefore, when working with big data, it's important to take volume into account.

2.2 Velocity

The rapid collection of data is referred to as velocity. In big data velocity, information comes from a variety of sources, including devices, networks, social media, mobile phones, etc. There is a significant and constant influx of data. How quickly data is collected and processed to fulfill requests influences how much potential it has.

2.3 Variety

The terms "structured," "semi-structured," and "unstructured" relate to the types of data. It also refers to sources that are diverse. The emergence of data from fresh sources, both inside and outside of a company, is essentially what is meant by variety. It can be organized, partially organized, or unorganized.

2.4 Veracity

It relates to data inconsistencies and uncertainty; that is, readily available data can occasionally become disorganized, and quality and accuracy are challenging to control. Because there are so many different data dimensions arising from so many different distinct data types and sources, big data is also unpredictable.

2.5 Value

Data is useless and unimportant by itself; in order to extract information, it must be transformed into something valuable.

2.6 Variability

Data that is constantly changing is referred to as variability. Understanding and correctly interpreting the significance of unprocessed data are the key areas of variation.

3. ADVANTAGES OF BIG DATA ANALYTICS

3.1. Risk Management

Big Data analytics is used to detect fraudulent actions and inconsistencies. It is used by the organization to narrow down a list of suspects or root causes of problems.

3.2. Product Development and Innovations

Big Data analytics used to analyze how to develop efficient product designs and if there is any need for improvements.

3.3 Complex Supplier Networks

Companies that employ big data provide more precise and insightful supplier networks or B2B communities. Suppliers can use big data analytics to avoid common limitations. Big data enables providers to apply more contextual intelligence, which is critical for success.

3.4 Cost optimization

One of the most compelling advantages that big data platforms such as Hadoop and Spark provide is tremendous cost savings for storing, processing, and analyzing enormous amounts of data. An example from the logistics business exemplifies the cost-cutting benefits of big data.

3.5 Improve Efficiency

Big data solutions can boost operational efficiency—your interactions with consumers and their important feedback help to acquire vast amounts of valuable customer data. Analytics can then uncover important patterns concealed within the data to generate tailored products. The tools can automate everyday processes and tasks, freeing up crucial time for employees to perform jobs demanding cognitive skills.

4. HOW BIG DATA ANALYTICS WORKS

Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data.

4.1 Collect Data

Every organization's approach to data collection is unique. With today's technology, businesses may collect structured and unstructured data from a variety of sources, including cloud storage, mobile apps, in-store IoT sensors, and more. Some data will be housed in data warehouses, where it will be conveniently accessible by business intelligence tools and solutions. A data lake can be used to store raw or unstructured data that is too diverse or complicated for a warehouse.

4.2 Process Data

Once data has been gathered and saved, it needs to be correctly organized to yield accurate answers from analytical queries, especially when the data is big and unstructured. Data processing for businesses is difficult due to the exponential growth of available data. Batch processing is a processing choice that examines big data blocks gradually. When data collection and analysis take more time than expected, batch processing might be helpful. Shortening the time between data collection and analysis allows for quicker decision-making through stream processing, which examines small batches of data at once. The complexity and cost of stream processing are generally higher.

4.3 Clean Data

All data must be accurately prepared, and any redundant or irrelevant data must be removed or accounted for in order to improve data quality and produce stronger findings. Inaccurate data can conceal and deceive, leading to erroneous conclusions.

4.4. Analyze Data

Big data need time to become usable. Huge data can be transformed into huge insights using advanced analytics techniques once they are ready. These large data analysis techniques include the following:

Data mining uses anomalies and data clusters to sort through enormous databases in search of patterns and linkages.

1. **Predictive Analytics** uses an organization’s historical data to make predictions about the future, identifying upcoming risks and opportunities.
2. **Deep Learning** Uses algorithms to uncover patterns in even the most difficult and abstract material, mimicking human learning processes in the process.

5. BIG DATA ANALYTICS TOOLS AND TECHNOLOGY

A single tool or technology cannot be used to perform big data analytics. The collection, processing, cleaning, and analysis of massive data is instead assisted by a variety of tools. Following is a list of some of the key actors in big data technology.

- **Hadoop**

It is open-source framework for processing and storing large datasets on groups of commodity hardware. This framework is open-source and capable of managing enormous amounts of both structured and unstructured data, making it an essential component of any big data operation.

- **NoSQL DataBases**

Non-Relational data management systems that are a perfect choice for large, unstructured data sets because they don't need a set structure. NoSQL, which stands for "not only SQL," refers to databases that can handle several types of data models.

- **MapReduce**

MapReduce is a crucial element that performs two purposes for the Hadoop framework. Data is filtered to different nodes in the cluster using mapping as the initial method. The second method, known as reduction, groups and condenses the output of each node in order to respond to a query.

- **YARN**

  "Yet another Resource Negotiator." It is another component of second-generation Hadoop. The cluster management technology helps with job scheduling and resource management in the cluster.

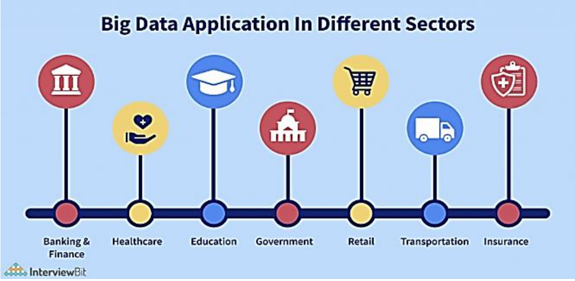
- **Spark**

  It is an open source framework for cluster computing that uses fault tolerance and implicit data parallelism to offer an interface for cluster programming. Fast computing using Spark is possible with both batch and stream processing.

- **Tableau**

  Tableau enables you to prepare, analyze, collaborate, and share your big data insights using an end-to-end data analytics platform. Tableau specializes at self-service visual analysis, enabling users to explore managed big data to find new insights and to quickly share those discoveries throughout the company.

6. APPLICATIONS OF BIG DATA



6.1 Banking

Big data has improved banks' efficiency across all industries, whether it be in financial management or cash collecting. The bank is now able to produce more money thanks to the technology's application, which also makes the user's insights clearer and easier to understand than before. Big Data expands a range of applications, from identifying fraud to evaluating and optimizing transaction processing, bettering user understanding, refining trade execution, and fostering an excellent user experience.

6.2 Education

When discussing the education sector, there is a vast amount of data collected from courses, students, professors, and results, which, when analyzed, can yield information that can be used to enhance how educational institutions run and function. Big Data plays a crucial role in this industry, helping to promote effective learning, enhance international recruiting for universities, assist students in setting career goals, reduce university dropout rates, advance clear student evaluation, speed up decision-making, and enhance student results.

6.3 Media

As a result of the current ways for consuming online information with the aid of gadgets becoming the newest craze, interest in traditional means of consuming media is steadily waning. Big data has gloriously entered this market since enormous amounts of data are produced. Big Data has significantly simplified the life of media companies, helping them predict what the audience will want in terms of genre, music, and material for their age group as well as offering them data into customer turnover.

6.4 Healthcare

Modern healthcare operations can be significantly improved with the use of big data. Technology has completely changed how healthcare is provided. By lowering treatment costs, anticipating epidemic outbreaks, avoiding preventable diseases, enhancing life quality, forecasting daily patient income to adjust staffing, implementing electronic health records (EHRs), using real-time alerts to encourage immediate care, utilizing health data for more effective strategic planning, and lowering frauds and flaws.

6.5 Agriculture

Big data analytics in agriculture promotes efficient farming and precise agricultural operations, reducing costs and opening up new economic opportunities. Meeting the demand for food by supplying farmers with information about weather, rainfall, and other factors affecting crop yield are a few important areas where big data can be used. Other important areas include promoting smart and proper pesticide application, managing equipment, ensuring supply chain productivity, etc.

6.6 Travel

Big Data is fundamental to improving and streamlining the transportation industry. Big Data has had an impact on this industry, whether it be in terms of managing the income generated, upholding the reputation built, or implementing smart marketing. The route can be planned according to the user's needs, waiting times can be properly managed, and accident-prone regions can be identified to make traffic more safe.

6.7 Manufacturing

Thanks to Big Data, manufacturing is no longer an arduous manual process. Technology and Data analytics have succeeded in completely revolutionizing the manufacturing process. Big Data improves manufacturing, personalizing product design, guaranteeing accurate quality maintenance, overseeing the supply chain, and also evaluating to keep track of potential risks.

6.8 Government

In order to keep track of their inhabitants, growth, geographic surveys, energy resources, etc., governments must maintain a number of records and databases, which expose them to a vast amount of data every day, regardless of the country. The government will be able to use this data as an ally in its operations once it has been examined and analyzed. This information is primarily used by the government for two purposes: cyber security and its own growth objectives**.**

7. TYPES OF BIG DATA ANALYTICS

7.1 Descriptive Analytics

This provides an accessible type of previous data summarization. Making reports about a company's income, profit, sales, and other metrics is made easier as a result. Additionally, it facilitates the calculation of social media data.

7.2 Diagnostic Analytics

This is carried out to determine the root cause of a problem. Drill-down, data mining, and data recovery techniques are a few examples. Because they offer a thorough understanding of a specific issue, diagnostic analytics are used by organizations.

7.3. Predictive Analytics

To predict the future, this sort of analytics examines both historical and current data. To analyse current data and forecast the future, predictive analytics uses data mining, AI, and machine learning. It functions through foreseeing consumer trends, market trends, and other trends.

7.4. Prescriptive Analytics

This kind of analytics suggests a solution to a specific issue. Both descriptive and predictive analytics can be used with perspective analytics. It frequently makes use of AI and machine learning.