IFERP
connecting engineers... developing research

Rabindranath TAGORE UNIVERSITY
MADHYA PRADESH, BHOPAL AN AISECT GROUP UNIVERSITY
Approved by : AICTE, NCTE, BCI, INC, MP PARAMEDICAL COUNCIL, DEB (UGC) | Recognized by : UGC | Member of : AIU, ACU

INTERNATIONAL CONFERENCE ON

Data Science , Machine learning and Applications-2021

ICDSMLA - 2021

10th - 11th June 2021

Virtual Conference

Organized by

CSE and CS/IT Departments,

Rabindranath Tagore University, Raisen, Madhya Pradesh

in Association with

Institute For Engineering Research and Publication (IFERP)

ICDSMLA -2021

International Conference on
Data Science, Machine learning and Applications
(Virtual Conference)

Raisen, Madhya Pradesh

10th - 11th June, 2021

Organized by:

Computer Science Engineering(CSE) and Computer Science & Information Technology(CS&IT) Departments, Rabindranath Tagore University, Raisen, Madhya Pradesh

In Association with

Institute For Engineering Research and Publication [IFERP]

Rudra Bhanu Satpathy
Chief Executive Officer
Institute For Engineering Research and Publication.

On behalf of ***Institute For Engineering Research and Publications (IFERP)*** and in association with ***CSE and CS/IT Department, RabindraNath Tagore University, Raisen, Madhya Pradesh***. I am delighted to welcome all the delegates and participants around the globe to ***RabindraNath Tagore University, Raisen, Madhya Pradesh*** for the "***International Conference on Data Science , Machine learning and Applications (ICDSMLA-2021)"*** Which will take place from ***10th - 11th June'2021***

It will be a great pleasure to join with Engineers, Research Scholars, academicians and students all around the globe. You are invited to be stimulated and enriched by the latest in engineering research and development while delving into presentations surrounding transformative advances provided by a variety of disciplines.

I congratulate the reviewing committee, coordinator **(IFERP & RNTU)** and all the people involved for their efforts in organizing the event and successfully conducting the International Conference and wish all the delegates and participants a very pleasant stay at ***Raisen, Madhya Pradesh.***

Sincerely,

Rudra Bhanu Satpathy

Preface

The ***International Conference on Data Science , Machine learning and Applications (ICDSMLA-21)*** is being organized by ***RabindraNath Tagore University, Raisen, Madhya Pradesh*** in Association with ***IFERP-Institute for Engineering Research and Publications*** *on the 10th – 11th June, 2021*.

RabindraNath Tagore University has a sprawling student –friendly campus with modern infrastructure and facilities which complements the sanctity and serenity of the major city of Raisen in Madhya Pradesh.

The ***"International Conference on Data Science , Machine learning and Applications"*** was a notable event which brings Academia, Researchers, Engineers, Industry experts and Students together.

The purpose of this conference is to discuss applications and development in area of **"Data Science , Machine learning and Applications"** which were given International values by ***Institute for Engineering Research and Publication (IFERP).***

The International Conference attracted over 180 submissions. Through rigorous peer reviews 40 high quality papers were recommended by the Committee. The Conference aptly focuses on the tools and techniques for the developments on current technology.

We are indebted to the efforts of all the reviewers who undoubtedly have raised the quality of the proceedings. We are earnestly thankful to all the authors who have contributed their research works to the conference. We thank our Management for their wholehearted support and encouragement. We thank our Principal for his continuous guidance. We are also thankful for the cooperative advice from our advisory Chairs and Co-Chairs. We thank all the members of our local organizing Committee, National and International Advisory Committees.

ICDSMLA -21

MESSAGE FROM CHANCELLOR

Shri Santosh Choubey

Chancellor

Rabindranath Tagore University

Bhopal, India

The foundation stones of RABINDRANATH TAGORE UNIVERSITY are laid on the essence of academic pursuit and excellence. Excellence in any work can be achieved with utmost dedication, hard work, and perseverance.

Research and development form the backbone of our curriculum at our University. The staff and students are engaged in various path-breaking innovative research activities throughout the year. Our University organizes conferences and seminars frequently on contemporary and relevant topics to facilitate research in those areas which will lead to necessary metamorphosis in the academia and researchers as well.

All Departments of RABINDRANATH TAGORE UNIVERSITY have been active in research and innovation and have setup an ambient academic environment for its students and research scholars. With the commitment of highly qualified and efficient staff, the Department of Computer Science Engineering and Department of CS&IT is organizing the International Conference on Data Science, Machine learning and Applications-2021(ICDSMLA-2021). This is another venture to provide a platform for academicians – teachers, students, research scholars, and industry personnel – globally to discuss on contemporary trends and innovations in Data Science and Machine learning.

I wish the conference all the very best and urge all the participants to brainstorm on various thrust areas of the conference.

MESSAGE FROM VICE-CHANCELLOR

Dr.Bramh Prakash Pethiya

Vice-Chancellor

Rabindranath Tagore University

Bhopal , India

The conferences are necessary to bring in culture of information exchange and feedback on developing trends in technologies. I am delighted to note that the Department of Computer Science Engineering and CS&IT is organizing the International Conference on Data Science, Machine learning and Applications-2021(ICDSMLA-2021). Certainly, this type of conference not only brings all the researchers, students at one platform, but it also inculcates the research culture among the entire fraternity of education in the country, thereby, contributing to the development of nation.

I hope that this conference will certainly induce innovative ideas among the participants paving way for new inventions and technologies in Computers and Communications.

I Congratulate, Department of Computer Science Engineering and Department of CS&IT, and the whole organizing team for initiating the conduction of such conference in our esteemed University.

I wish the conference a grand success.

MESSAGE FROM REGISTRAR

Dr.Vijay Singh

Registrar,

Rabindranath Tagore University

Bhopal , India

In this age of rapidly changing technologies, it is essential for all to keep abreast of the latest developments in emerging areas like Data Science and Machine learning. In this regard I am extremely happy to note that Department of Computer Science Engineering and Department of CS&IT, Rabindranath Tagore University organizing an International Conference on Data Science, Machine learning and Applications-2021(ICDSMLA-2021).

I am sure that the conference will inculcate the much-needed research culture among the students and teachers and trigger interactions among researchers to exchange the ideas of recent advances in the areas of Computers. Also, the various subthemes of the conference will offer the delegates many opportunities to learn new things and apply the same in their respective workplaces.

I am very happy and congratulate the department of Computer Science Engineering and CS&IT for organizing the conference on a well thought issue.

I wish the International conference a grand success.

MESSAGE FROM CO-ORDINATOR

Dr.Sanjeev Kumar Gupta

Co-Ordinator

(ICDSMLA-2021)

Rabindranath Tagore University

Bhopal, India

The International Conference on Data Science, Machine learning and Applications-2021(ICDSMLA-2021) organized jointly by the RABINDRANATH TAGORE UNIVERSITY and IFERP is focused on the future industrial aspects available for Engineering professionals in general and Electronics, Computers and Communications in particular. The Conference provides an open forum for scientists, researchers, and engineers to discuss nascent innovations and research advancements in the areas of next generation electronics, computers, communication architectures, application, and services. It will be a wonderful opportunity for delegates to gain quality input useful for their future research in the knowledge-based society. The conference papers being published in Scopus will be of great source of information to the academicians, scholars, and industrialists.

I congratulate the organizers and wish the conference a great success.

MESSAGE FROM CONVENER

Dr.S.Veenadhari

Convener

ICDSMLA-2021

On behalf of the ICDSMLA-2021 organizing committee, I am honored and delighted to organize the International Web Conference on “Data Science, Machine Learning and Applications 2021”. It is a matter of pride to be hosting the conference but because of recent pandemic we have no option to conduct it online. In the era where thoughts become concepts and concepts becoming a reality within the shortest possible time frame, this conference will be a major boost for the technology evolution itself.

The International virtual conference will facilitate the young researchers, industry experts, scientists, academic experts, especially those who are carrying out their research work in the domain of computer science, information technology, electronics and communication engineering with valuable discussions in order to make the outcomes more realistic. This is a perfect platform for all the young and experienced minds alike in exploring and sharing knowledge around the computing technology and its scope.

I know that the success of the conference depends ultimately on the people who have worked with us in planning and organizing both the technical program and supporting social arrangements. I wish that ICDSMLA will keep on growing in coming years with more impact on the International research community.

I thank the support of all authors, reviewers, RNTU faculty for organizing this event.

MESSAGE FROM CO-ORDINATOR

Dr.Pratima Gautam

Co-Ordinator

ICDSMLA-2021

Rabindranath Tagore University ,

Bhopal, India

This International conference provides a forum to all researchers to exchange the information on research and innovations and enhance the quality of research.

The international conference on Data Science and Machine Learning and Applications-2021(ICDSMLA-2021) provides a platform for researchers to get networked and exchange the ideas on various areas such as Machine Learning and Data Science.

A conference is a place where true meetings of minds happen. Researchers who would have done a good deal of thinking about their idea, will come forward and share their thoughts with fellow researchers.

High quality deliberations that happen in conference will lead to high standard publications at international levels which feed into the industry's innovation pipeline. Industry expects such inputs to create innovation and the next big things. on behalf of organizing team, I thank you all, and best wishes for a very successful conference.

ICDSMLA -2021

International Conference on Data Science, Machine learning and Applications

Keynote Speakers

Sai Satish Babu .N

Senior Data Scientist

British Telecom , Bangalore

I am honored to be part of ‘International Conference on Data Science, Machine Learning and Applications (ICDSMLA-21)’ organized by ‘RabindraNath Tagore University, Raisen, Madhya Pradesh’ and ‘Institute For Engineering Research and Publication (IFERP)’.

Its great to see people from the different corners of the globe come together to exchange knowledge on cutting-edge technologies and the Innovations happeing across various industries to make Engineering & Technology impactful for our next generations.

In today's fast paced world, technology is evolving more rapidly than our imagination . Emerging technologies like IOT (Internet of Things) , AI (Artifitial intelligence) , ML (Machine Learning) , DL (Deep Learning) , mobile 5G , AR(Augmented Reality) , VR(Virtual Reality) , Blockchain and automations going to play vital role in major transformations across the industries.

Our great scientists dream about doing the great things and Engineers ensure them . Without Engineering and Technology evolutions we can’t even think about getting so modernized world . What we design, invent, innovate ,create and build today, will be the engineering heritage of tomorrow. It’s important that we get it right.

My message to all the participants is Share your knowledge (Take best out of open source community and contribute whatever you can that may help others interested in similar things)

“Knowledge shared is knowledge squared”

I would like to extend my best wishes to all the participents and ICDSMLA-21 team.

Thanks & Regards,

Sai Satish Babu .N

www.linkedin.com/in/SaiSatishBabu/

Dr. Anand Nayyar

Researcher, Duy Tan University

Da Nang, Vietnam

On behalf of the Technical Program Committee, welcome to the International Conference on Data Science and Machine Learning Applications (ICDSMLA 2021), organised jointly by CSE and CS/IT Department, RabindraNath Tagore University, Raisen, Madhya Pradesh in Association with Institute For Engineering Research and Publication (IFERP)". This conference allows both researchers and practitioners to present and share their on-going ideas, experiences, and research results about all trending aspects of computer and information science. The core of the technical program is a careful selection of interesting and novel research papers chosen for presentation at the conference that you will find in the proceedings. In response to the call for papers, we received submissions from different countries/regions. All the papers were evaluated on the basis of their significance, novelty, and technical quality. After careful review, around 48 papers were selected to be presented at the conference. These papers cover a wide range of topics including theory, methods, and applications. We would like to express our appreciation to the following people: the conference Chair and Co-Chairs; the Program Committee and secondary reviewers who contributed a great amount of their time and effort to evaluate the submissions to maintain high quality of the conference; the local arrangement committee; the finance chairs; the session chairs who presided over the sessions; and all the authors, attendees, and presenters who really made this conference possible and successful. We hope you enjoy the conference and enjoy your experience at ICDSMLA 2021.

Thanks.

Kind Regards

Dr. Anand Nayyar.

Dr. Abbas Al-Bakry

Chancellor

University of Information Technology and Communications, Iraq

I believe that the ICDSMLA conference is very interesting to many researchers over the world, this is due that the hot topics that the conference deals with (Data Science, Machine learning and their applications).

The collaboration between CSE and CS/IT Department, RabindraNath Tagore University, Raisen, Madhya Pradesh and the Institute For Engineering Research and Publication (IFERP) to organizing this event add a scientific value due to the experience available for both parties.

I also believe that the conference participants can get new valued information through the innovations and ideas that will be presented at the conference.

Finally I congratulate the conference committees for the organization and hope the success.

Best Regards

Prof. Abbas M Al-Bakry(Ph.D)

President of UoITC-Baghdad

Prof. (Dr.) Ajay Rana

Senior Vice President & Dean

Amity Higher Education Group

Chairman, AUN Research Labs

Director, Amity Institute of Information Technology

I am extremely glad to see that our country, its academicians, practitioners, and researchers are taking bold initiatives in the field of Science & Technology. Educational institutions in India are today laying extreme emphasis on industry focused education and attracting the best students from across the globe. I am delighted to know that CSE and CS/IT Department, Rabindra Nath Tagore University, Raisen, Madhya Pradesh in association with Institute for Engineering Research and Publication (IFERP) is organizing an International Conference on Data Science, Machine learning and Applications-2021 from 10th - 11th June 2021.

The theme of this Conference is extremely important and upon reviewing the wide variety of topics & exclusive sessions going to be covered during the 2 Day conference, offers unlimited scope of further research and innovation. The global gathering of intellectuals in this conference will enable the streams and channels to discover partnerships and headways in the multi-disciplinary fields of science, technologies and humanities covering aspects like transcendental growth, recent trends, innovations and security issues involved in the domain of communication technologies, high performance computing, big data, social media, hardware & software design, advanced software engineering, Internet of Things (IoT), e-governance and many more.

The presence of several distinguished speakers and worthy participants from overseas and India will immensely enhance the value of this conference and they would be able to establish significant and long-term contacts, forging bonds and cooperation between the Academic Systems and several participating industries, scientific institutions and autonomous bodies from various countries.

I congratulate and offer my heartiest greetings to all national as well as international researchers, scholars, delegates, speakers, industry leaders, chairs and co-chairs of the conference for their participation in the conference. I am confident that the Conference would be highly outcome-based and greatly result-oriented & would provide further vision for future. I appreciate the support provided by IFERP in organizing this great event at Rabindra Nath Tagore University, Raisen, Madhya Pradesh.

I acknowledge and compliment the notable hard work of the conference committees in organizing this conference on an international platform. My hearty compliments and appreciation for Dr. Sanjeev Kumar Gupta, Dean Academics and Dr. Pratima Gautam Dean, CS/IT along with their dedicated Organizing Core Team for all the hard work put in by them.

I wish the Conference a grand success.

Best Wishes

Dr Ajay Rana

ICDSMLA-2021

International Conference on Data Science, Machine learning and Applications

Raisen, Madhya Pradesh, 10th - 11th June, 2021

Organizing Committee

CHIEF PATRON

Shri Santosh Choubey, Founder & Chairman, SCAD Group of Institutions
Dr. Bramh Prakash Pethiya, Vice Chancellor, RabindraNath Tagore University
Dr. Vijay Singh, Registrar, RabindraNath Tagore University

COORDINATORS

Dr.Sanjeev Kumar Gupta, Dean Academics, RabindraNath Tagore University
Dr.Pratima Gautam, Dean, CS/IT, RabindraNath Tagore University

CONVENOR

Dr.S.Veenadhari, Associate Professor, CSE, RabindraNath Tagore University

CO-CONVENORS

Dr. Rajendra Gupta, HOD,CS/IT, RabindraNath Tagore University
Mr.Mukesh Kumar, HOD,CSE, RabindraNath Tagore University

ORGANIZING COMMITTEE

Dr.Varsha Jotwani, Associate Proffessor, CS/IT, RabindraNath Tagore University
Dr.Shailja Sharma, Associate Proffessor, CSE, RabindraNath Tagore University
Dr.Rakesh Mittan, Associate Proffessor, CSE, RabindraNath Tagore University
Dr. Preetaj Yadav, Associate Proffessor, CSE, RabindraNath Tagore University
Mr.Prajeet Sharma, Assistant Proffessor, CS/IT, RabindraNath Tagore University
Mr.Shashikant Upadhyay, Associate Proffessor, CS/IT, RabindraNath Tagore University
Ms.Amlesh Singh, Assistant Proffessor, CS/IT, RabindraNath Tagore University

Nitesh Baghel, Assistant Proffessor, CS/IT, RabindraNath Tagore University
Mr.Devendra Rathore, Assistant Proffessor, CSE, RabindraNath Tagore University
Ms.Ayoniza Pathare, Assistant Proffessor, CSE, RabindraNath Tagore University

NATIONAL ADVISORY COMMITTEE

Prof. V.K. Verma, Director, CRG
Prof.Nitin Vats, IQAC-Director, RabindraNath Tagore University
Dr. S.B Lal, Principal Scientist, IASSRI, Delhi
Dr. Sanjeev Sharma, Professor, RGPV, Bhopal
Dr. Aditya Trivedi, Professor, IIITM, Gwalior
Dr. Karan Singh, Principal Scientist,CIAE, Bhopal
Dr. Shubhish, Principal Scientist, CIAE, Bhopal
Dr. Manish Maheshwari, Professor,MCRPV, Bhopal
Dr. R.K Patariya, Reader, MANIT, Bhopal
Dr. P.V Vapariya, Associate Professor, SPU, Anand, Gujrat

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CONTENTS

ICDSMLA -2021

International Conference on Data Science, Machine learning and Applications (Virtual Conference)

Raisen, Madhya Pradesh

10th - 11th June, 2021

PAPERS

ICDSMLA-2021

Organized by:

CSE and CS/IT Department, RabindraNath Tagore University, Raisen, Madhya Pradesh

In Association with

Institute For Engineering Research and Publication (IFERP)

An EHO based Task Scheduling to Enhance the Resource Utilization in Cloud Computing

[1]Abhishek Gupta, [2]Dr. Rajendra Gupta
[1]Research Scholar, Rabindranath Tagore University, Raisen
[2] Associate Professor, Rabindranath Tagore University, Raisen

Abstract: The task scheduling is deeply correlated with resources usage and processing expenses in cloud computing. These factors are well utilized by using several optimal task scheduling schemes to provide the best completion of tasks. Here, an Elephant Herding Optimization based Task Scheduling (EHOTS) approach is implemented to provide optimal scheduling of users' tasks for improving the resource utilization by reducing processing expenses over cloud computing environment. The EHOTS is initiated to an objective function with multiple factors like load and processing. The MATLAB 2019a tool is used to perform the experiment and the simulation outputs have explained the better quality performance of EHOTS based on entire cost with minimum and maximum number of repetitions and number of tasks in opposition to GA, PSO and ALO approaches.

Keywords: Cloud Computing, Elephant Herding, Objective Function, Optimization, Task Scheduling.

1. Introduction

The resources, data, guidelines, documents and surrogate structures are commonly attached to maintenance terms, communication and transferring services provide flexible, robust, and minimum cost heterogeneous environment over cloud computing atmosphere [1]. These factors can be achieved by using task scheduling in cloud computing for utilizing huge communication network and memory space. The efficient and reliable task scheduling does not only utilize the processing strength of resources, but also well used the memory space of cloud storage with minimum expenditure. The superior task scheduling strategies have increased the efficiency of cloud computing due to generate minimum computing time with load balancing and maximum throughput. Specifically, the task scheduling is used to conserve the resource strength, improve power efficiency [2], develop reliable resource and generate future resource excursion such as protection and resources usage [3].

The load balancing is the next approach in cloud computing, which is combined with task scheduling [4] to enhance the efficiency of cloud computing by improving the resource utilization [5]. The resource usage is effectively performed in both the strategies over cloud environment devoid of task overloading of resources. Primarily, the tasks are assigned to cloud, after that spread optimally among devices with the help of optimal scheduling of tasks. The huge amount of processor, memory and resource utilization is very difficult to maintain the processing expenses [6].

A large number of bio inspired optimum techniques such as Ant Colony Optimization (ACO) and Bacteria Foraging (BF) [7] are well introduced for task scheduling to obtain optimal processing and completion of tasks. Some swarm intelligence

approaches like Genetic Algorithm (GA) [8, 9] and Harris Hawk Optimization (HHO) [10] are also applied to reduce the processing expenses with maximum resource utilization for optimal task scheduling. The Cuckoo Optimization Algorithm (COA) [11] is combined with Particle Swarm Optimization (PSO) [12] to use the behaviour of both particles and swarms for generating the optimum scheduling of tasks between Virtual Machines (VMs). The Opposition Based Learning (OBL) is also merged with optimization techniques to improve the searching capability of search agents in huge areas. The OBL based Line Optimization Algorithm (LOA) [13] is used to search the optimal elements in several directions to enhance the performance of lines in cloudsim toolkit. The directed Acyclic Graph (DAG) is another method to resolve the task scheduling problem in cloud computing, in which the DAG is used to assign the tasks among VMs. The PSO [14] and hill climbing [15] are applied over DAG to assign the tasks optimally for completion.

In the above study, it shows that task scheduling is performed optimally by using various bio inspired optimization techniques. But, few enhancements in task scheduling are yet performed to reduce the limitations of optimization approaches like dependency on primary conditions and local optimization issues. In this paper, an Elephant Herding Optimization based Task Scheduling (EHOTS) is implemented to provide optimal scheduling of tasks among devices to reduce the entire cost of processing and memory and to enhance the performance in a cloud computing environment. The simulation outputs are described on the basis of the entire cost with minimum and maximum number of repetitions and number of tasks over cloud environment.

2. An EHOTS approach

2.1. Task Scheduling Sculpt with Multiple Objectives

The resource usage is proportionally inflated by scheduling scheme of tasks for elemental devices in cloud computing. The Virtual Devices (VDs) are allocated to tasks in scheduling, which is the main issue in a cloud environment. Primarily, the jobs of users will be alienated into a set of tasks processing in several VDs. Here, three steps will be performed: (i) primarily, sources and tasks will be associated in terms of whole task data and active VDs in agreement along with explicit scheme, (ii) after that, the optimum scheduling scheme of tasks will be provided by a task scheduler based on the association to organize the assignment requests, (iii) at the end, the implementation of optimum task scheduling strategy for cloud computing is performed and results will be generated for users.

2.1.1. Virtual Device and Task Sculpts

The successive sculpt is utilized to describe the entire optimum task scheduler strategy with respect to a cloud environment. There is a set of V virtual devices (VDs) $\{D_1,D_2,D_3,.....,D_v\}$ and a set of K user tasks $\{U_1,U_2,U_3,.....,U_k\}$ with K > V, and complete scheduling outputs can be explained by a matrix P as beneath (eq. (1)).

$$P_{kv} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1v} \\ p_{21} & p_{22} & \cdots & p_{2v} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kv} \end{bmatrix}$$

Here, p_{mn} is an appraisal variable, in which $p_{mn} = 1$ denotes that the m^{th} task is processed on the n^{th} VD, otherwise $p_{mn} = 0$. $(\sum_{n=1}^{V} P_{mn} = 1 \quad \forall\ m \in [1, K])$

All the virtual devices are used two attributes like processing power and load power for describing the

resource expenses and processing potency for scheduling of tasks over cloud environment. The processing power indicates the CPU strength to compute the user tasks and symbolized by processing power matrix (C_d). The load power is indicated the capability of load balancing over cloud computing and symbolized by load power matrix (O_d) (table 1). In similar way, two matrices are used for tasks such as C_u and O_u.

Table 1. Notations and their Explanations

Notation	Explanation
K	A set of user tasks
V	A set of virtual devices
p_{mn}	Appraisal variable to indicate that the m^{th} task is processed on the n^{th} VD
$C_d(C_u)$	Processing power matrix for VDs (tasks)
$O_d(O_u)$	Load power matrix for VDs (tasks)
E	Unit expenses
W_m	Weight factor coefficient (m = 1,2)

2.1.2. Optimization Function with Multiple Objectives

It is assumed that the proposed sculpt can develop a task scheduling scheme in specific ways based on processing power and load power over cloud environment. The first attribute processing expense function F_{PE} is obtained by eq. (2) for explaining the task finishing time based on CPU strength. The second attribute load expense function F_{LE} is obtained by eq. (3) based on memory. The utilized parameters C_d, C_u, O_d, and O_u are different in scales, hence normalized functions are used to evaluate the attributes.

$$F_{PE} = \frac{1}{K}\sum_{m=1}^{K}\sum_{n=1}^{V} p_{mn}\frac{C_{u,m}/C_{d,n}}{\max_{\forall m,n}\{C_{u,m}/C_{d,n}\}}$$

$$F_{LE} = \frac{1}{K}\sum_{m=1}^{K}\sum_{n=1}^{V} p_{mn}\frac{O_{u,m}/O_{d,n}}{\max_{\forall m,n}\{O_{u,m}/O_{d,n}\}}$$

Where,

$C_{u,m}$ & $C_{d,n}$ = C_u value of m^{th} task and C_d value of n^{th} VD respectively.

$O_{u,m}$ & $O_{d,n}$ = O_u value of m^{th} task and O_d value of n^{th} VD respectively.

At the end, the complete optimization function with multiple objectives F_{OMO} is shown (eq. (4)) by merging two attribute functions F_{PE} and F_{LE} using few weight factor coefficients (w_m). $\{w_1 + w_2 = 1\}$

$$F_{OMO} = Minimum\{w_1 F_{PE} + w_2 F_{LE}\} \quad (4)$$

2.2. An Elephant Herding Optimization (EHO)

An Elephant Herding Optimization (EHO) is a bio enthused technique derivative from elephant herding nature to solve global optimization problems. Three rules are utilized to explain the EHO: (a) the population of elephants is divided in few clans and clans are further divided into a predetermined number of elephants; (b) A predetermined number of male elephants will go away from their families and survive isolate aside from the major elephant crowd at every creation; (c) the clan`s elephants mutually survive along with the guidance of a matriarch.

2.2.1. Clan Modifying Operator

The new location of elephant in clan L_a is prejudiced by matriarch L_a. The eq. (5) is used to modify the location of b[th] elephant in clan L_a.

$$X_{new,L_a,b} = X_{L_a,b} + \beta \times (X_{best,L_a} - X_{L_a,b}) \times rand \quad (5)$$

Where,

$X_{new,L_a,b}$ & $X_{L_a,b}$ = b[th] elephant new modified and old location in clan L_a respectively.

β = a scale coefficient generating the control of matriarch L_a on $X_{La,b}$. $\{\beta \in [0,1]\}$

X_{best,L_a} = the fittest elephant (matriarch) in clan L_a.

$rand$= random number. $\{rand \in [0,1]\}$

The fittest elephant is modified its location in every clan by using eq. (6), which is not modified by eq. (5), i.e., $X_{L_a,b} = X_{best,L_a}$.

$$X_{new,L_a,b} = \mu \times X_{center,L_a} \quad (6)$$

Where,

μ = a coefficient generating the control of X_{center,L_a} on $X_{new,L_a,b}$. $\{\mu \in [0,1]\}$

X_{center,L_a}= clan L_a centre.

The eq. (6) is updated for i[th] dimension $\{1 \le i \le I,\ I = total\ Dimension\}$and converted to eq. (7).

$$X_{center,L_a,i} = \frac{1}{N_{L_a}} \times \sum_{b=1}^{N_{L_a}} X_{L_a,b,i} \quad (7)$$

Where,

N_{L_a}= number of L_a clan elephants.

$X_{center,L_a,i}$ & $X_{L_a,b,i}$ = centre and b[th] elephant location in clan La in i[th] dimension.

2.2.2. Separating Operator

Further, enhancing the search strength of EHO technique, suppose that the worst fitness elephant will realize the separating operator at every creation as represented in eq. (8)

$$X_{worst,L_a} = X_{Minimum} + (X_{maximum} - X_{minimum} + 1) \times rn \quad (8)$$

Where,

$X_{maximum}$ & $X_{minimum}$= Upper and Lower bound of elephant location.

rn = stochastic and uniform distribution. $\{rn \in [0,1]\}$

2.3. An Elephant Herding Optimization based Task Scheduling (EHOTS)

The task scheduling scheme is described in terms of EHO (table 2), where elephants are notated as user tasks in cloud computing. In searching, an elephant has a position based on the scheduling of task, which has a solution P_{kv} , The leader elephant`s location notates the current best solution and leader elephant`s fitness value shows the current best value of complete optimization function with multiple objectives F_{OMO}. In this way, an EHO is implemented to generate best solution for user tasks scheduling for cloud computing. The whole elephants are updating their positions in every repetition and the position information is transferred to a solution P_{kv} for active user tasks. According to P_{kv} , the fitness values F_{OMO} of elephants are obtained. The entire steps will be

unremitting till final repetition. The position information of last leader elephant will be filled to solution P_{kv}, which is used to generate the best execution strategy of tasks in cloud computing.

Table 2: Preliminaries using in EHO and task scheduling

EHO preliminaries	Task Scheduling
Individual elephant	User`s cloud tasks
Elephant herding nature	Optimal solution searching
Elephant Location	A solution P_{kv} for F_{OMO}
Leader Elephant	Optimal solution P_{kv} for F_{OMO}
Elephant`s fitness	F_{OMO} value

The EHOTS performs the following steps.

Step1: Initially, the mapping between user tasks and elephants are to be done.

Step2: The elephant position, searching dimension, elephant populations, number of repetitions, and constant`s values provides initial values.

Step3: All the elephant`s fitness values are calculated according to the elephant`s position information in optimal solution searching process. The elephant with minimum fitness value is denoted as a current best solution.

Step4: The entire elephants are updating their positions by using eq. (5) to eq. (8).

Step5: Step 3 to step 4 is continued for all repetitions.

Step6: The position information about last leader elephant will be added to solution P_{kv}, which is used for generating the best execution scheme of tasks in cloud computing environment.

3. Result and Analysis

The EHOTS is implemented in MATLAB 2019a environment and outcomes are analyzed with other techniques GA, PSO and ALO. The entire cost values are calculated for 200, 600, 800 and 1000 tasks against number of repetitions from 10 to 100. The entire costs are also evaluated against number of tasks from 100 to 1000 (table 3 and table 4).

Table 3: Simulation Parameters

Parameter	Value Range (VS)	Value Range (Tasks)
Memory	[120, 550]	[55, 120]
Resource	[120, 350]	[20, 55]
CPU	[220, 550]	[10, 55]

Table 4: Experimental Parameters

Parameters	Values
Tool	MATLAB 2019a
Operating System	Windows 8
Number of Elephants	[100, 1000]
Number of repetitions	[10, 100]
Number of Tasks	[100, 1000]
Weight factor coefficients (w_m). $\{w_1, w_2\}$	0.50, 0.50

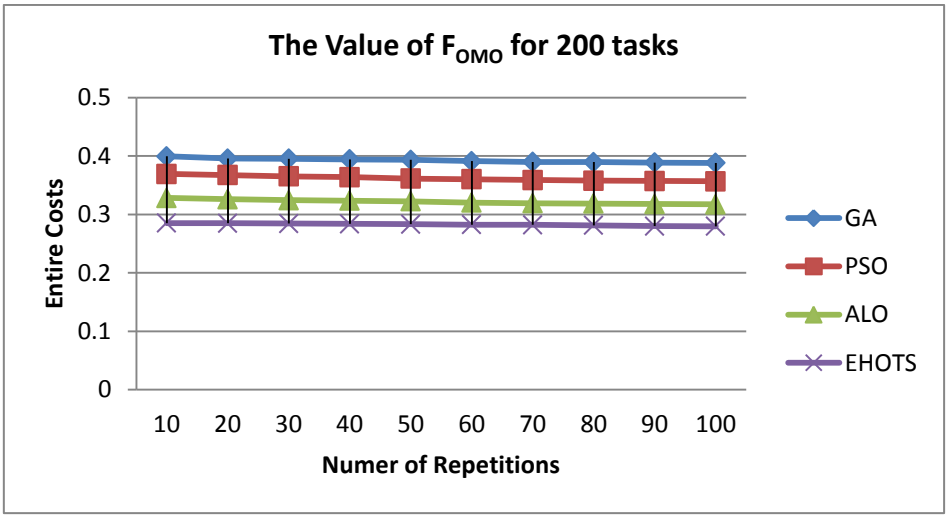


Fig1. The value of F_{OMO} for 200 tasks

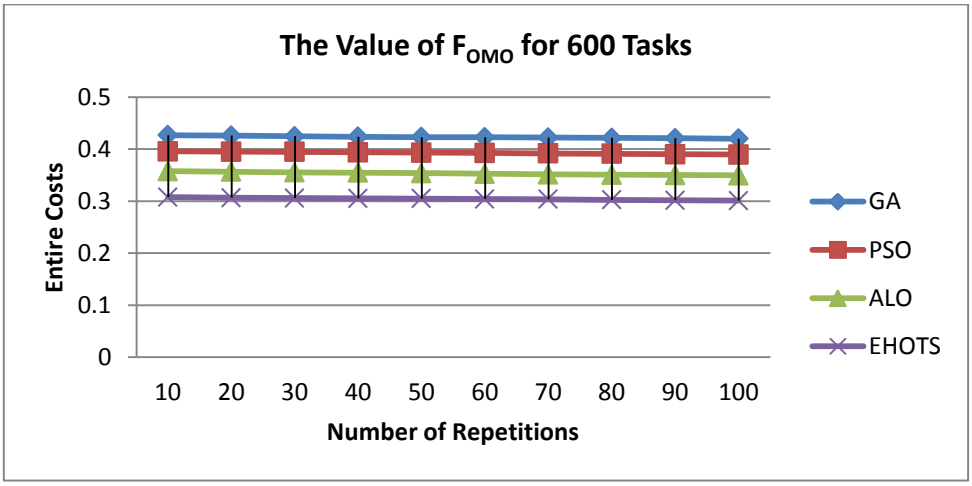


Fig2. The value of F_{OMO} for 600 tasks

The graphs (figure 1 and figure 2) describe that the PSO generates 9% and 8% better performance against GA; ALO generates 13% and 12% better performance against PSO, 21% and 20% better performance against GA; EHOTS obtains 14% and 17% superior performance against ALO, 25% and 27% superior performance against PSO, 32% and 33% superior performance against GA based on entire cost for minimum number of tasks (200 and 600 respectively) in terms of number of repetitions.

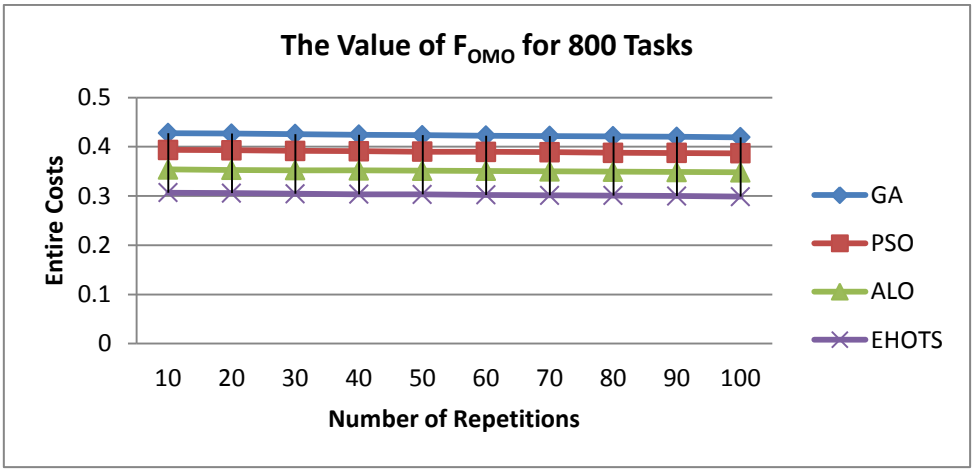


Fig3. The value of F_{OMO} for 800 tasks

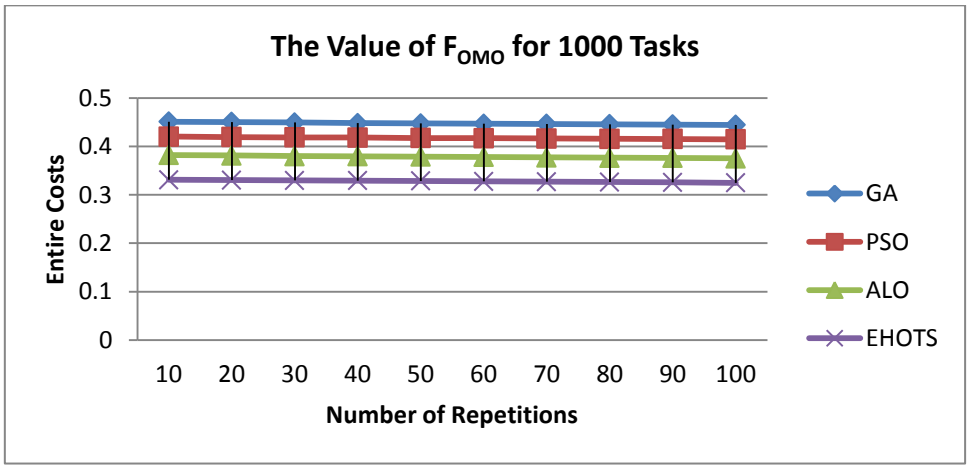


Fig4. The value of F_{OMO} for 1000 tasks

The graphs (figure 3 and figure 4) explain that the PSO obtains 9% and 7% better results against GA; ALO generates 11% and 11% better outputs against PSO, 19% and 18% better outputs against GA; EHOTS obtains 16% and 16% better outcomes against ALO, 26% and 25% better results against PSO, 33% and 31% better efficiency against GA based on entire cost for maximum number of tasks (800 and 1000 respectively) in terms of number of repetitions.

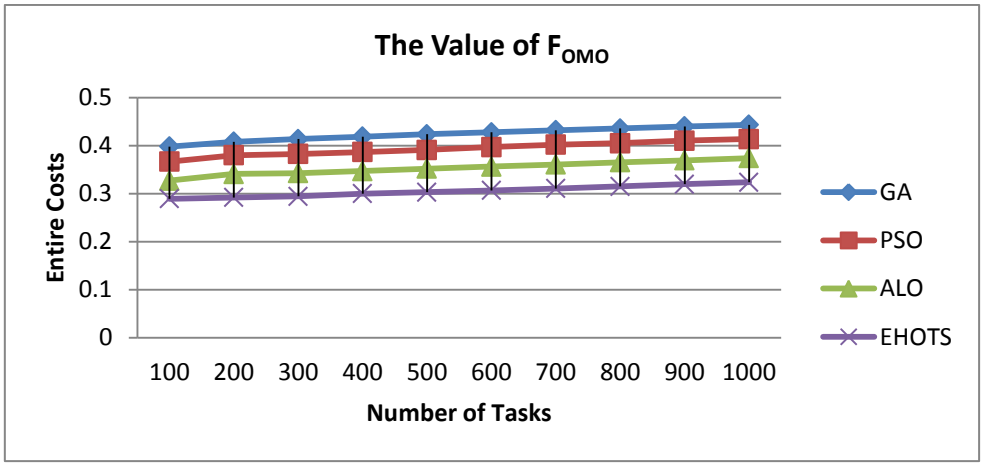


Fig5. The value of F_{OMO}

The graph (figure 5) explains that the PSO obtains 7% better quality efficiency against GA, ALO generates 11% better quality results against PSO and 18% better performance against GA, EHOTS obtains 16% superior efficiency against ALO and 25% better quality results against PSO and 31% superior outcomes against GA on the basis of entire cost for 100 repetitions in terms of number of tasks. Hence, it shows that the proposed EHOTS obtains superior results than GA, PSO and ALO based on number of tasks and number of repetitions. The graphs (figure 1 to figure 5) describe that the entire cost of entire approaches is reduced with increasing the number of repetitions. Therefore, it describes that the cost is reduced right through the optimal solution searching process.

4. Conclusion

The task scheduling is extremely linked with use of resources and processing expenditure in cloud computing. These factors are specifically used by various researchers to obtain optimal scheduling scheme of tasks. In this paper, an Elephant Herding Optimization based Task Scheduling (EHOTS) approach is implemented to generate best scheduling of users' tasks for enhancing the resources usefulness by decreasing the expenditure of processing in a cloud environment. The MATLAB 2019a tool is utilized to implement the EHOTS and the simulation results are described the superior efficiency of EHOTS based on entire cost with minimum and maximum number of repetitions and number of tasks in opposition to GA, PSO and ALO approaches.

References

1. Kalka Dubey, Mohit Kumar and S. C. Sharma, "Modified HEFT Algorithm for Task Scheduling in Cloud Environme", 6th International Conference on Smart Computing and Communications, ICSCC, Kurukshetra, India, Elsevier, pp-725-732, 2017.
2. R. K. Jena, "Energy Efficient Task Scheduling in Cloud Environment", 4th International Conference on Power and Energy Systems Engineering CPESE, Berlin, Germany, Elsevier, pp-222-227, 2017.
3. Mahendra Bhatu Gawali and Subhash K. Shinde, "Task Scheduling and Resource Allocation in Cloud Computing using a

Heuristic Approach", Journal of Cloud Computing: Advances, System and Applications, Springer, pp-1-16, 2018.

4. Ruba Abu Khurma, Heba Al Harahsheh and Ahmad Sharieh, "Task Scheduling Algorithm in Cloud Computing based on Modified Round Robin Algorithm", Journal of Theoritical and Applied Information Technology, Vol. 96, No. 17, pp-1-21, 2018.
5. Amer AL-Rahayfeh, SAleh Atiewi, Abdullah Abuhussein and Muder Almiani, "Novel Approach to Task Scheduling and Load Balancing Using the Dominant Sequence Clustering and Mean Shift Clustering Algorithms", Future Internet, MDPI, Vol. 11, No. 109, pp-1-15, 2019.
6. Xuan-Qui Pham, Nguyen Doan Man, Nguyen Dao Tan Tri, Ngo Quang Thai and Eui-Nam Huh, "A cost- and performance-effective approach for task scheduling based on collaboration between cloud and fog computing", International Journal of Distributed Sensor Networks, Vol. 13, No. 11, pp-1-16, 2017.
7. Sobhanayak Srichandan, Turuk Ashok Kumar and Sahoo Bibhudatta, "Task Scheduling for Cloud Computing using Multi-Objective Hybrid Bacteria Foraging Algorithm", Future Computing and Informatics Journal, Vol. 3, pp-210-230, 2018.
8. Nan Zhang, Xiaolong Yang, Min Zhang, Yan Sun, and Keping Long, "A genetic algorithm- based task scheduling for cloud resource crowd- funding model", International Journal of Communication System, Wiley, pp-1-10, 2017.
9. Amjad Mahmood, Salman A. Khan, and Rashed A. Bahlool, "Hard Real-Time Task Scheduling in Cloud Computing Using an Adaptive Genetic Algorithm", Computers, MDPI, Vol. 6, pp1-21, 2017.
10. Ibrahim Attiya, Mohamed Abd Elaziz and Shengwu Xiong, "Job Scheduling in Cloud Computing Using a Modified Harris Hawks Optimization and Simulated Annealing Algorithm", Computational Intelligence and Neuroacience, Hindawi, pp-1-17, 2020.
11. Mahendra Bhatu Gawali, and Subhash K. Shinde, "Standard Deviation Based Modified Cuckoo Optimization Algorithm for Task Scheduling to Efficient Resource Allocation in Cloud Computing", Journal of Advances in Information Technology, Vol. 8, No. 4, pp-1-9, 2017.
12. Naresh T, A Jaya Lakshmi and Vuyyuru Krishna Reddy, "Resource Optimization Using Cloud Scheduling", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, pp-1-6, 2019.
13. Pradeep Krishnadoss and Prem Jacob, "OLOA: Based Task Scheduling in Heterogeneous Clouds", International Journal of Intelligent Engineering & Systems, Vol. 12, No. 1, pp-114-123, 2018.
14. Shasha Zhao, Xueliang Fu, Honghui Li, Gaifang Dong, and Jianrong Li, "Research on Cloud Computing Task Scheduling based on Improved Particle Swarm Optimization", International Journal of Performability Engineering, Vol. 13, No. 7, pp-1-7, 2017.
15. Negar Dordaie and Nima Jafari Navimipour, "A Hybrid Particle Swarm Optimization and Hill Climbing Algorithm for Task Scheduling in the Cloud Environment", ICT Express, pp-1-6, 2017.

Object Detection using YOLO

[1]Abhishek Kumar Singh, [2]Srajal Dwivedi, [3]Pritish Kumar, [4]Dr. Varun Tiwari

[1, 2, 3]Student, [4] Assistant Professor

[1]singhabhishek.aks21@gmail.com, [2]srajal.knj@gmail.com, [3]pritish.gupta24@gmail.com,
[4]varun.tiwari@galgotiasuniversity.edu.in

[1, 2, 3, 4]School Of Computer Science and Engineering, Galgotias University, Uttar Pradesh

Abstract- Object detection using CNN (Convolutional Neural Networks) and the study of computer vision has now allows user to manipulate or work and observe different videos transcript files and photographs which also give an observation about its feature which helps in identifying the different objects. Detecting an object is a way which become a kind of features which has been established in computer vision which plays a huge role in identification of the object, and their location, and allows to detect one or different types of objects as well as its tracking ability to keep track add a kind of uses in a given photos or video. Here, the concern is to detect the different object using the YOLO (You Only Look Once) algorithm technique. The purpose of using this method, it provides us with several advantages as we compare to other object detection algorithms. In YOLO (You Only Look Once), the algorithm go through the image completely in detail by detail by calculating and predicting the bounding boxes which is created by using the powerful neural network algorithm i.e., convolutional neural network (CNN) and further by dividing into the different classes, and then calculate the probabilities for all the featured boxes which has set as different classes and then identify the image by comparing the features and tries to detect the image faster as compared to other algorithms with minimal errors.

Keywords: Convolutional Neural Network, Bounding Boxes, Batch normalization, YOLO

1. INTRODUCTION:

The evolving technology in the field on identifying the object popularly known as object detection is the latest technology which is playing a huge role the development of new era of technology, object discovery is a very important point in fields of computer system and robotics monitoring system, which is the subject of extensive research [1]. Using digital image and videos, it creates a class and detects the objects. Object detection is used in a very vast fields such as: self-driving [2], pedestrian detection [3], Disease recognitions [4], stocks detection in industries [5], detection of robots [6], smart surveillance [7], remotely handling sensing (Virtual Reality) [8], etc., and the main focus and objective is to successful detection of all these fields using object detection. Object detection could also be converted as a key ability almost for every security system and latest robot technologies. Detecting an object is a process in which algorithm of computer vision is involved, identifying the object, its location, and even its type of one or more objects in a given photos or video. Its required to focused on detecting the image as a whole by observing its features. So, to develop an algorithm in order to finding the object as well as its location, speed, correct identification and also used able to represent data in the form of tables to improve clarity in Object detection [9].As we compare further with developed traditional algorithms, it is mainly focused in terms of object detection method in terms of their performance, robustness, accuracy and multi-classification of the object in order to achieve the target. The object detection is simply based on deep learning algorithms which is based upon the unified framework and purpose-based algorithms. In order to make better and more accurate detection algorithm, the new algorithms start using the CNN (Convolution Neural Network) to extract features, generate series of regions according to the input image and classify into object classes. Convolution neural network is also used in regional based algorithm known as R-CNN (Regional-

Convolutional Neural Network) [10] so it can overcome the problem in selecting a large amount numbers in regions. And using the R-CNN process, two new methods were proposed which are Fast Regional-Convolutional Neural Network (Fast R-CNN) [11] and Faster Regional-Convolutional Neural Network (Faster R-CNN) [12]. These purposed methods help in reducing the training time, improve performance as well increases the average accuracy

Object detection does not work with degraded images, i.e., when it is trained with instructional data sets, comprising of ImageNet, COCO and VOC, etc. but there is case of randomly captured data sets it is not well tested. The main problem of images captured in the real time scenarios are:

1) the captured images can be blurred as due to the unsteadiness of the camera,

2) As the object can be cut-off which can cause the images to not be so clear.

3) due to overexposure, bad weather (rainy, snowy ,etc), or low resolution of the images which can result into poor quality.

1.1. YOLOv3

The YOLO framework is a method that deals with object detection and approaches in a different way. It captures the whole image at a single instance and its predicting the bounding box and start coordinating and the sophistication probabilities for all those boxes. Researchers have suggested variety of various plumbing systems in few past years, including the older version of YOLOv3 and YOLOv2 method [13]. YOLOv2 mainly focused at improving integration and stop overheating and for achieving this it uses instruction execution, with stop boxes predicting binding boxes, to maximise return. the one of the advantages of YOLO is its ability of speed analysis, it's incredibly fast analysis and is able to process about 45 frames in 1 second. YOLO can also define the object in a generalized structure. The approach involves in training one neural network till it finish that photograph as validated input and predicts the calculated bounding boxes and sophistication labels for every predicted bounding box directly. This technique helps is acquiring the lower predictive accuracy (example, more localization errors), and all these abilities helps in operating at 45 frames in 1 second and can even go about 155 frames in 1 second for developing the speed-efficiency version of the model. Many of systems which are designed for object detection traverse the sample n number of times so make the choice, or a minimum of it's to travel to minimum of two stages for successfully achieving the article detection within the image which is given as an input to the system. But as we discuss YOLO so it doesn't must undergo steps quite once. YOLO is all about his name that's you merely Look Once. because it can detect all the objects and provides it as an output of the pictures which was earlier given as an input. it's the explanation that's why you simply Look Once include a goods speed in giving the output of the detected images compared all other models. YOLO VERSIONS –

1. YOLO (FIRST VERSION)
2. YOLO V2
3. YOLOv3

2. YOLO WORKING

This is the method is developed on a different multi-scale detection to improve accuracy. Redmon and Farhadi are able to accomplish and able to proposed a high performance, more accuracy success algorithm and named it YOLO(You Only Look Once).The structure of YOLO is that it is developed based on binary cross entropy loss and then calculates the prediction series and all are created into prediction box which are calculated on different scales. YOLO first step after analysing the input image, its start dividing the image into a matrix, the dimension of the matrix is defined as [S x S] dimensions grid. After set the image into [S x S] dimensions grid, all the start predicting the object and furthermore only one object can be detect by the each grid. When a grid detects an object or a piece of an object then its stats it predicting the boundary box. A bounding box is a rectangle shape structure which enclose an object within itself. Each and every bounding box also have one confidence score which denotes the accuracy of that predicted boundary box of the object. The calculated confidence score is now classified into the classes for the prediction which is named as conditional class. Conditional class is the probability of the de[S x S] dimensions grid detected object which has been classified in different classes and conditional class defines per cell of each category with probability of one. So, the basic structure of YOLO prediction can be given as:

Prediction's Shape = (S, S, B x 5 + C)

Where,
S is Matrix dimension
B is Boundary box

C is Conditional box

The first version of YOLO uses 7x7 as its matrix dimension (SxS), through this grid a boundary box of 2 is created which further classified in a group of 20 conditional classes. YOLO also minimised the error as its uses the sum-squared error method and the calculated loss function comprise of:

1. the classification loss - each cell in the grid is calculated with squared error of the category using the conditional probabilities.
2. The localization error- errors finds within the expected boundary boxes and the ground truth
3. the boldness loss.

Prior detection systems is created to repurpose the classification and localization to perform detection for the image. It applies the model to be able to picturing at multiple locations and scales and all these calculated by detecting with the high confidence score calculated for the region of the image during the pre-processing the image for the detection

2.1. Network Design

Deep learning may be a prevailing direction within the field of machine learning [15]. In [10,19], according to the many researcher in the field of object detection have given different theories and methods according to their analysis which mostly focuses on the ability of CNNs (Convolution Neural Networks), how the study on CNN benefits the deep learning methods which can bring a huge improvement in the developed methods for the detecting the object ands its accuracy. There have been many efforts of using backpropagation and gradient descent to teach the deep networks and make algorithm to able to perform detection[8].The network design constructed in YOLO has 24 convolutional layer which is further supported by 2 fully connected layers and to intercept the modules which is used by ImageNet which uses 1x1 reduction layer which further supported by the convolutional layer of 3x3.After calculation its start optimizing using sum-squared error as it equally weights the error within the boxes comprises of large as well as small ones. A one convolutional layer is ables to predict more than one bounding boxes and class. So in order to direct optimization of the algorithms YOLO is trained on the full image which makes this unified model to have several benefits over the traditional based algorithms as currently the network is able to run at 45fps which further developed to the 150fps(frame per seconds).

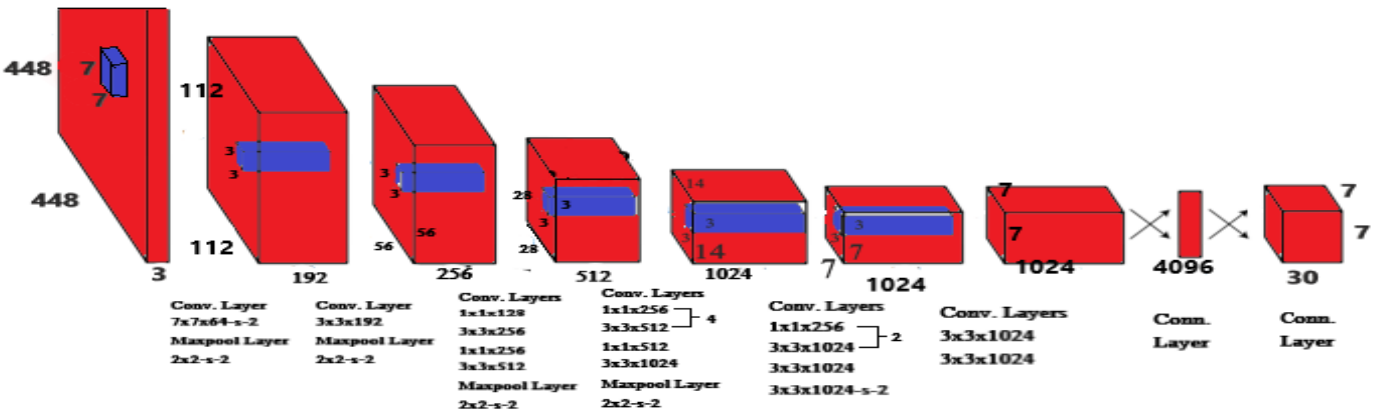


Figure1. The Network Architecture

2.2 Unified Detection

In object detection, the feature and algorithms based on neural networks is in a process for long time [10, 11]. In the process of image recognition, researchers make use of the deep learning for learning features directly from the image pixels, which are more effective than the manual features [4, 12].

$$\Pr(\text{Class}_i|\text{Object}) * \Pr(\text{Object}) * \text{IOU}^{\text{truth}}_{\text{pred}} = \Pr(\text{Class}_i) * \text{IOU}^{\text{truth}}_{\text{pred}}$$

So, our model uses the features using the entire image to calculate and predict each boundary box. Then further divided into SxS grid. The object which fall within the centre of the grid for that cell we calculate the confidence score ($Pr(Object)*IOU^{truth}_{pred}$) to check the accuracy of the object. So it can compare the confidence score with intersection over union (IOU) which represents the confidence score between predicting box and the ground truth box.

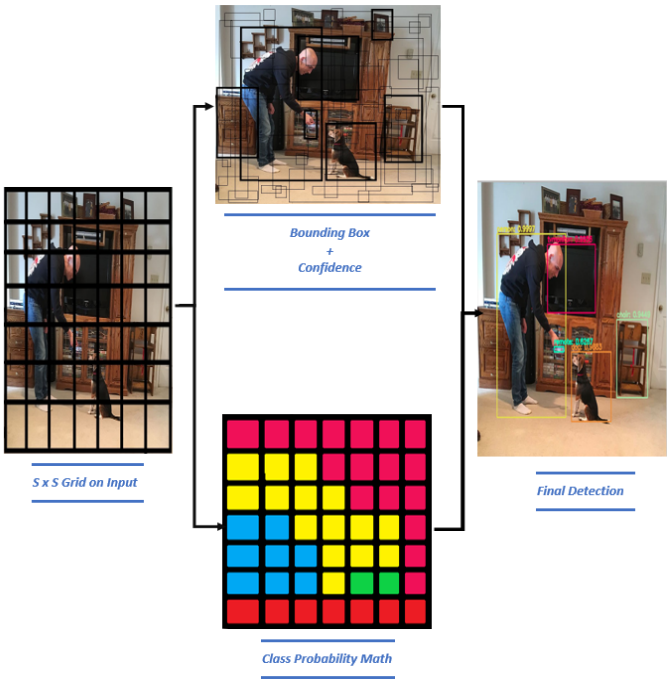


Figure2-THE MODEL

2.3 Merits of YOLOv3

1. Its Fast as well as good for real-time operations.
2. Prediction is done on the basis on calculating the entire image at once by the single network which help in improving the accuracy of prediction.
3. YOLO is more generalized structure which makes it better than all the other methods as it is well generalized from natural given input image to all the other different domains such as artwork.

YOLO accesses to the entire image within its predicting boundaries it basically region proposal method that allows to limit the precise region by its classification. With the extra context, YOLO sometime also may provide false resulting due to background calculation. YOLO detects the object which comes under the centre of the created grid box which was earlier divided into the form of [S x S] grid. It enforces as huge amount of diversity in making predictions.

3. Methodology

YOLO provides lots of features which make it more high performance as besides calculating everything it also give an object score which add a feature to classification probabilities. Object score is generally an estimation of falling the object in the prediction box In object detection and recognition, the object. The researchers have used many deep algorithms for developing and adding features directly, which is simpler than adding it manually [4, 12] for detecting the object. Recently in the developing much better deep learning and neural network algorithms and able to extract the features directly using the algorithms [18] just by giving the entire image. These developed method has been successfully qualified and able to add features in the development of the improved algorithms such as pyramid network(FPN) [14], SSD-single shot detector [15].

3.1 Batch normalization

Batch normalization is a regularization technique used by YOLO after the procedure of convolutionally layering the image. Batch normalization is basically used to make data clean as layers of neural network perform much better with clean data. As per the ideal condition is when input layer has average of 0 and not have much variance. Batch normalization act almost like a featuring scale as it basic work is to protect the data from deteriorating which helps neural network to perform much better when it flows through the network.

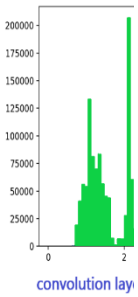


Figure3. Convolution layer without and with batch normalization

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

`_BATCH_NORM_EPSILON` refers to epsilon during this formula, whereas `_BATCH_NORM_DECAY` refers to momentum, these two variables used to calculate average and variance for the layer during the inference i.e. after training as a forward propagation.

`action_avg = momentum * action_avg + (1 - momentum) * current_avg`

3.2 Leaky ReLU

Leaky ReLU may be a slight modification of ReLU activation function. the thought behind Leaky ReLU is to stop so-called "neuron dying" when an outsized number of activations become 0

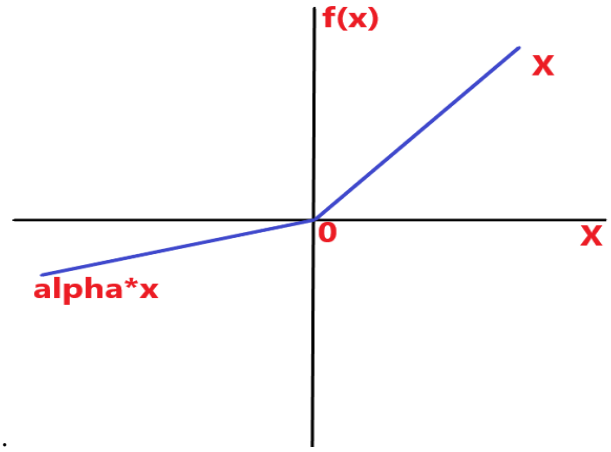


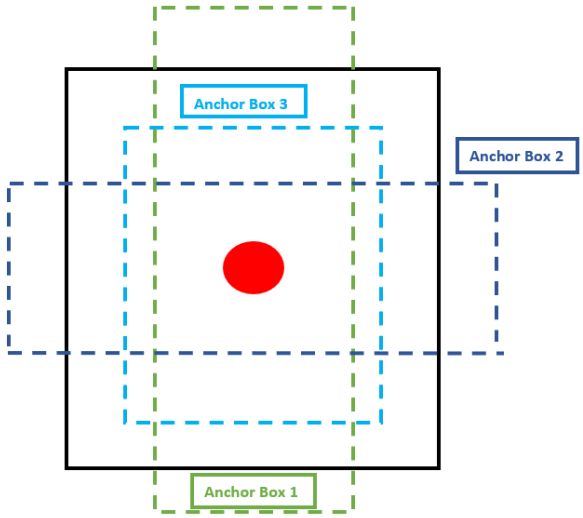
Figure 4 – Activation Function using Leaky

3.3 Anchors

Anchors are kind of bounding box priors, that were calculated on the COCO dataset using k-means clustering.

Figure5 Anchors box with different ratio and shape

Now it needs to predict the width and height of the box by considering the offsets from cluster centroids. the middle coordinates of the box relative to the situation and start filtering the application are predicted using sigmoid function.

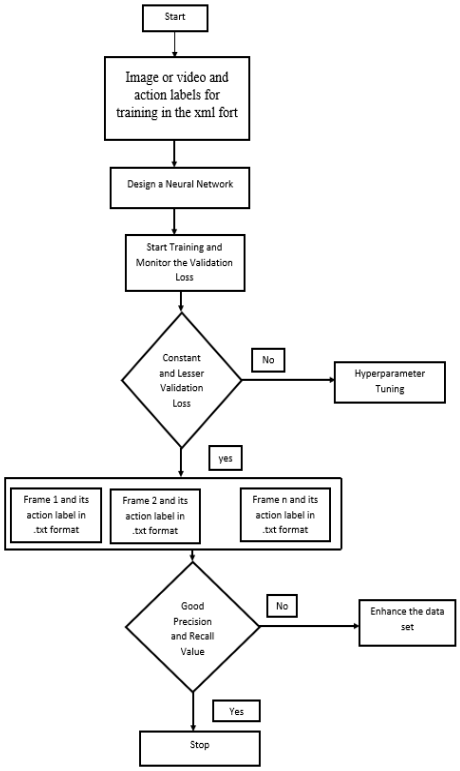


Where bx is x-

$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$

coordinates and by is y-coordinates for the centre of the box, bw is width and bh is height for the box, cx and cy are location for the filtering application and ti is data predicted through the regression process.

3.4 Flow Chart



When the input image is received, we crop the unrequired part. When YOLO receives the image, it first maps it into the SxS grid where each grid predicts one object. For each cell in the grid a boundary box is calculated and give the confidence score according to their prediction rate. And then conditional probability class is calculated for each distinct feature class. So, keeping all the in mind we use basic condition i.e. 7x7 grid, boundary box-2 and conditional class-20.It uses the root of summed up values to work out the losses supported base truth upon the three loss factors.

4. Experiments

Comparing our yolo model with other real time detecting models which are based Pascal Voc 2007 to analysis the difference between the models and able to compare different models in aspects of their performance, complexity, speed, accuracy and in many more aspects with minimal errors.

4.1. Comparison to Other Real-Time Systems

Many researches are putting efforts in object detection which mainly focuses on making standard detection pipelines as fast as it is possible. [5] [14] [17]. So, we start comparing the features of YOLO to their different GPU implementation of DPM which design to run only on either at 30Hz or 100Hz. During implementation and comparing their relative mAP and speed and time taken to calculate and examine the accuracy of its performance trade-offs which is available within object detection systems [20].

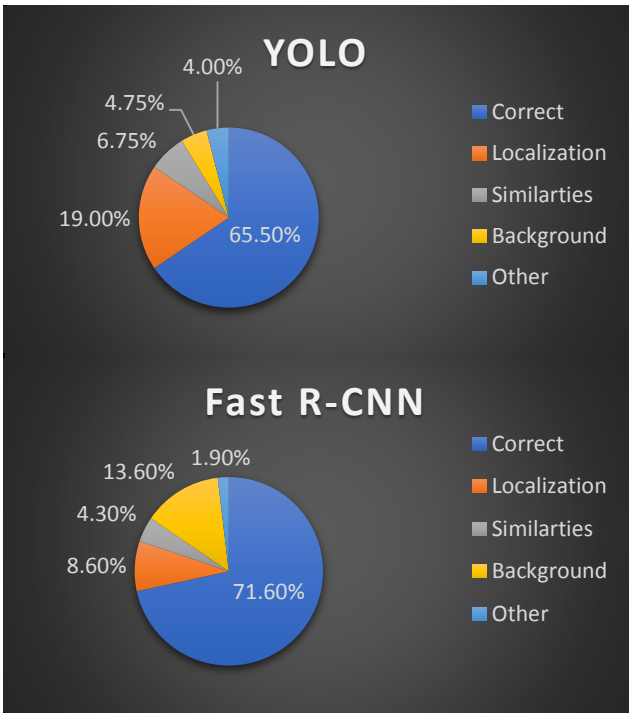
Real-Time Detectors	Train	mAP	FPS
100 Hz DPM	2007	16.0	100
30 Hz DPM	2007	26.1	30
Fast YOLO	2007+2012	52.7	155
YOLO	2007+2012	63.4	45

Less than real time	Train	mAP	FPS
Fastest DPM	2007	30.4	15
R-CNN Minus R	2007	53.5	6
Fast R-CNN	2007+2012	70.0	0.5
Faster R-CNN VGG	2007+2012	73.2	7
YOLO VGG-16	2007+2012	66.4	21

REAL-TIME SYSTEM ON PASCAL VOC

With 52.7% mean average precision (mAP), it is even more than the twice as accurate as comparison to the previous work during the real-time detection. YOLO allows to gain more mAP to about 63.4% during process and also keep

maintaining its original real-time performance. We also traing our model with VGG-15 and name it as YOLO VGG-16 and this proposed model is much more accurate but it also significantly slower than YOLO in performance, also very useful during comparison of the other detection systems that mainly relied on VGG-16 but since it is proven to be slower than real-time [16]. It also emphasis the limit provided by DPM's relatively low accuracy during detection performance as compared to neural network approaches.



By combining the YOLO and Fast R-CNN it made YOLO much fast and with fewer background mistakes as compare to the Fast R-CNN. As we use YOLO it helps in eliminating the background mistake which is falsely detecting from Fast R-CNN we are able to get much more significant boost in the performance. For every bounding box that are developed and that R-CNN predicts we check to see whether if the YOLO is able to predicts a similar box. The best Fast R-CNN model which is able to achieves a mAP of about 71.8% during the test set. When it is combined with YOLO, its mAP get a boost and increases up to 3.2% of mAP to almost 75.0% of mAp.

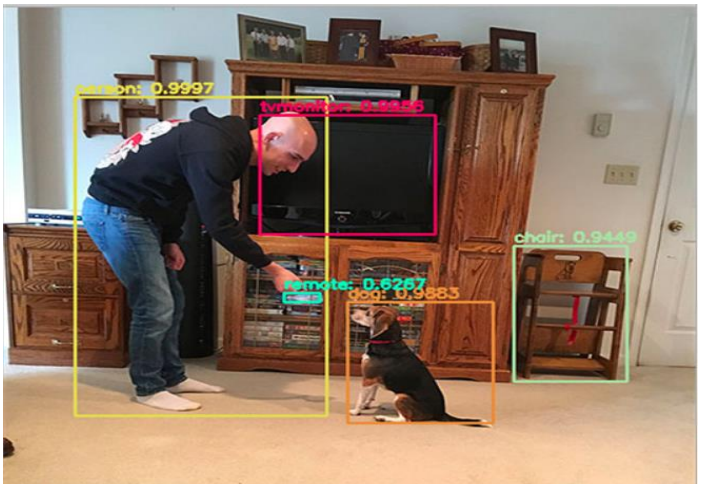
	mAP	Combined	Gain
Fast R-CNN	71.8	—	—
Fast R-CNN(2007)	66.9	72.4	0.6
Fast R-CNN(VGG-M)	59.2	72.4	0.6
Fast R-CNN(CaffeNet)	57.1	72.1	0.3
YOLO	63.4	75.0	3.2

Model Combination Experiments

5. Conclusion

YOLO (You Only Look Once) is a model which gave user such a functionality which make this model an ease to construct using straight forward method and functions and during developing the process and it has been gone through the training directly on the data based onto the full images. During the training of the YOLO model the concern and focus is on loss function calculation which enables user to interact directly and efficiently corresponds to the detection which is highly based on the performance and its analysing during the detection of the object and therefore for accomplishing the goal the entire model is focused on trained together acting like working in a group. Fast YOLO is one of that updating that makes its structure in real-time detection much more faster with greater accuracy. YOLO also provides its user a generalizes well to new domains and trying to indulge in making it ideal for different applications and expanding its uses that made it

ERROR ANALYSIS : FAST R-CNN vs YOLO



6. References

[1]. Chen, X.; Ma, Hwan, J.; Li, B.; Xia, T. Multi-view Object Detection for Autonomous Driving with an IEEE Conference on Pattern Recognition and object vision, Hawaii, HI, USA.

[2]. Sarkar S,S.R., San Zheng Multi-view 3D Detection and Computer Communication

[3]. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. How we can Help Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, HI, USA, 21–26 July 2017.

[4]. Christ, P.F.; Kaissis, G.; Ettlinger, F.; Kaissis, G.; Schlecht, S.; Ahmaddy, F.; Grün, F.; Menze, B.; Valentinitsch, A.; Ahmadi, S.-A.; et al. SurvivalNet: Predicting patient survival from di_usion weighted magnetic resonance images using cascaded fully convolutional and 3D Convolutional Neural Networks.

[5]. Weimer, D.; Scholz-Reiter, B.; Shpitalni, M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Ann. 2016, 65, 417–420.

[6]. Senicic, M.; Matijevic, M.; Nikitovic, M. Teaching the methods of object detection by robot vision.

[7]. Sreenu, G.; Durai, M. Intelligent video surveillance: A review through deep learning techniques for crowd analysis. J. Big Data 2019, 6, 48–75.

[8]. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 2018, 56, 2337–2348.

[9]. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object Detection with Deep Learning: A Review. IEEE Trans. Neural Netw. Learn. Syst. 2019, 30.

[10]. M. Lin, Q. Chen, and S. Yan. Network in network. CoRR, abs/1312.4400, 2013.

[11]. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.

[12]. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell.

[13]. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. IEEE Trans. Pattern Anal. 2017, 29,

[14]. Shafiee, M.J.; Chywl, B.; Li, F.; Wong, A. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. J. Comput. Vis. Image Syst. 2017, 3, 171–173.

[15]. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the IEEE International Conference on European Conference on Computer Vision,

Amsterdam, The Netherlands, 8–16 October 2016;

[17]. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the

[18]. Farhadi, A. YOLOv900: Better, Faster, Stronger IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016;

[19]. Pinle, Q.; Chuanpeng, L.; Jun, C.; Chai, R. Research on improved algorithm of object detection based on feature pyramid. Multimed. Tools Appl. 2019, 78, 913–927.

[20] C. P. Papageorgiou, M. Oren, and T. Poggio. Function of the framework for object detection. In Computer vision, 1998. sixth international conference .

Artificial Intelligence (AI); Creating New Perspectives for Diagnosis in Orthodontics: A Review

Amit Kuraria[1], Shanya Kapoor[2]

[1]Research Scholar,Rabindranath Tagore University,Bhopal MP
[2]BDS,Hitkarini Dental College,Jabalpur MP

Abstract- Artificial intelligence has taken over almost everything around us, from digital assistance like Siri, Alexa to self-driving cars, from simple music streaming mobile applications to space exploration, AI is everywhere. Medicine indeed isn't forsaken in this regard, its powerful pattern finding and prediction algorithms are helping clinicians in rational decision making and treatment planning. This technology if used wisely has a potential to cure the world from deadliest diseases and revolutionize the health care systems. Orthodontics is a specialty of dentistry which is concerned with correction of crocked teeth i.e. malocclusion and it's been making most AI technology lately. Machine learning algorithms like artificial neural network (ANN) & Convolutional neural network (CNN) are on the top of this list. Clinicians are taking fringe benefits from AI by compounding its ability to store large dataset and its power of decision making within fraction of seconds. This literature review is a compilation of AI driven projects in field of orthodontics and explaining how it has helped in orthodontists in decision making and will also include in brief about the areas that are yet to be explored.

Keywords- Artificial intelligence, Machine learning, Artificial neural network, Orthodontics, Diagnosis.

Introduction –

Intelligence is basically the ability to think, to imagine, creating, to memorize, understand, recognize patterns, make choices along with adapting to change and learn through experience. Artificial intelligence in it's simplest term is a human way to create a non-organic unit , which has all the abilities of natural organic intelligence[1].AI alternatively may be stated as a subject dealing with computational models that can think and act rationally[2-5]. AI as an entity is enormous which includes various fields, including reasoning, natural language processing, planning, and machine learning (ML).[6] At present, ML is the frequently used AI application in the medical and scenario of dentistry. It is important to note that ML is not intended to mimic human behavior. Instead, it supplements human intelligence by performing tasks that are beyond human capabilities[7.] This itself is making ML superior to the rule-based expert systems (ESs) that were earlier used. Most algorithms used in ML are similarly being used in data mining also. The only difference lies in its algorithm's goal. The algorithms are applied to large historical data sets to look for new patterns or

relationships, when the motive is to rectify decision of huge dataset.[8-9] This process is known as data mining. For instance, data mining can aid clinical practitioners find valuable information within existing patient records. Using this latest entity, practitioners can predict the future decisions, improve one's day to day practice, and elevate the quality of care. Contrary, if the intention is to make predictions, then ML is the option. The doctor uses available data about a certain disease to train the machine to make predictions about the diagnosis or prognosis of patients that have never been seen before. Importantly, ML predictive models have proven to be more valid compared to statistical models.[10]

This paper is intended to give an insight about applications of Artificial intelligence in dentistry esp. orthodontics.

Although orthodontics itself constitutes of various treatment possibilities like growth modification when enough growth potential is still left in the patient, dentoalveolar compensation when only dental correction is needed to treat the malocclusion of patient, surgery when growth potential isn't left and the defect cannot be treated with dental changes alone. If taking diagnosis in hand the conventional way for diagnosis include multiple steps for orthodontic problem recognization, these steps mainly categorized according forementioned sources (1) multiple questioning records including chief complaint, patient's dental and medical history; (2) clinical examination of the patient; and (3) assessment of diagnostic records, including dental casts, radiographs, and facial and intraoral images [11]. All information gathered in an elaborate process to achieve the most suitable treatment planning, treatment plan is the another glicth faced by the orthodontist, the enormous variation in dental malocclusion gathered with different facial pattern and the presence of large number of available treatment modalities, all these leading the decision process in diagnosis to challenging area even to an experienced orthodontist[12.] Therefore AI has a potential to solve topsy -turvy problems in Orthodontics.

AI trends in orthodontics –

Earlier expert system (ES's) were popularly used in orthodontics, as an aid to help amateur orthodontist in diagnosis although it didn't served well as the problems in orthodontics is far more complex and depends on multitude of patient factors with which only one algorithm won't work. At present, orthodontists are working with updated ML systems available to them that can diagnose a wide range of orthodontic cases and predict thier treatment needs[13.] Numerous superior systems have been developed for helping out orthodontists in diagnosis &treatment planning.

AI For Diagnosis-

X-ray analysis, an integral part of diagnosis and treatment planning, isn't left to benefit from ML. Amongst the most important applications of ML in orthodontics was the automation of landmark capturing. A recent systematic review reported 5–15% success at landmark detection with ML than traditional methods[14].ML was also used to automate diagnostics directly from cephalograms, including the sagittal relationships within maxilla and mandible, as well as normal and abnormal posterior-anterior facial heights ratios, overbite, and overjet.[15]

Automation of X-rays analysis has also been further expanded to hand and wrist radiographs for predicting the skeletal age of patient.

Panoramic along with lateral cephalometric X-rays as an adjunct are being utilize to estimate maxillary canine impactions based on linear & angular measures[16]. The highest prediction accuracy has been documented to be found with a random forest algorithm, which correctly predicted the real eruption state of canines 88.3% of the time.

Recording panoramic radiographs makes the orthodontists legally liable if they neglect any underlying potential tumor. Thisin turn has revolutionized the development of an automated neural network system that can accurately diagnose ameloblastomas and KOT from panoramic radiographs 83.0% of the time.[17] Lately, more and more orthodontists are using cone-beam computed tomography, that has led to the development of an automated system using the support vector machine to intelligently detect periapical cysts and KOT 100% of the time. [18]

Panoramic and lateral cephalometric X-rays are utilized to predict maxillary canine impactions based on angular and linear measures.[19]

AI for Craniofacial Growth modification –

To plan a case of interdisciplinary case of orthodontics and maxillo-facial surgery, patient's growth is vital. Study done by Lux CJ et al [20]suggested the application of an artificial neural network, which were self-organizing neural maps, the growth of 43 non-treated children was analyzed by means of lateral cephalograms taken at the ages of 7 and 15. In order to analyse the craniofacial skeletal changes, tensor analysis concept and related methods have widely been used. Thus the geometric and analytical limitations of conventional cephalometric methods have been avoided, the resultant growth data were classified and the relationships of the various growth patterns were monitored by using an artificial neural network model.

Classification of Class III growth patterns has also been performed. Based on historic data of untreated Class III subjects, based on the changes in their sagittal relationships, were classified into either good or bad growers, a classification tree had a significantly lower rate of misclassification (12.0%) than discriminant analysis (40.7%), both of which were based on the earlier used 11 cephalometric

variables.[21] It successfully identify good and bad growth patterns 64.0% of the time, when tested on new data.

AI in selecting the appropriate treatment modalities-

In order to deliver extra-oral forces to the upper dental arch for anchorage purposes, distalizing teeth and/or inhibiting forward maxillary growth, Headgear is mainly used. It has three subgroups, i.e. low, medium, and high-pull describing the direction of force givento the upper molar teeth in the sagittal plane [22]. The choice of the precise type of headgear isn't problematic as in when considering its application in 'typical' cases, such as those exhibiting a case of a deep overbite, large over jet, and a low mandibular plane angle as usually seen in Class II malocclusion. A problem may still arise, especially for orthodontists who have less clinical experience, with 'borderline' or 'marginal' subjects, such as those having a deep overbite, a moderate to severe over jet, and a high mandibular plane angle. a computer-assisted inference model for selecting appropriate types of headgear appliance for orthodontic patients and act as a decision-making aid for inexperienced clinicians was developed by Akgam M.O and Takada K and was found quite effective.[23] Yet another potential use of AI in orthodontics is the predicting the of soft tissues treatment outcomes. Recently, ANN was used to predict the change in lip curvature after orthodontic treatment with or without extractions[24.] Arena of beauty is dichotomy the reason being it's subjective and influenced by factors like age, sex, and ethnic backgrounds. Using ANN, facial attractiveness was quantified on a scale from 0 to 100 (0 extremely unesthetic and 100 extremely attractive) before and after orthognathic surgery.[25]

Conclusion –

AI systems gives a stern promise that are likely to improve clinical practice. With the help of clinical decision support systems the orthodontists can practice more efficiently along with decreased variability, and can even eliminate subjectivity[26.] Accuracy of most systems that are available today are considered good to excellent ranging from approximately 64% to 97%. However, the accuracy at the lower end accuracy of this range is likely to get better in near future as sample sizes increase and a wide range of information becomes available to us.[8] Artificial intelligence studies in various orthodontic steps yield humoungous researches which will ultimately will enrich the world of orthodontics.

References -

1- Divya Swarup, Deepak Singh, Singh Swarndeep, Ahmad Naeem, Sahai Richa. Artificial intelligence (A.I.) In orthodontics .Journal of Science,2017;7(9):304-307
2- Luger, G. F. and Stubblefield, W. A., Artificial Intelligence: Structures and Strategies for Complex Problem Solving, Benjamin/Cummings, Menlo Park, CA, 1993.
3- Schalkoff, J., Culberson, J., Treloar, N. and Knight, B., "A world championship caliber

checkers program," Artificial Intelligence, vol. 53, no. 2-3, pp. 273-289, 1992.
4- Winston, P. H., Artificial Intelligence, Addison-Wesley, 2nd ed., Reading, MA, 1994.
5- Newell, A .and Simon, H.A., Human Problem Solving, Prentice-Hall, Englewood Cliffs, NJ, 1972.
6- Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 2000;44:206-26
7- Mueller JP, Massaron L. Machine Learning for Dummies. New York: John Wiley and Sons; 2016
8- Asiri SN, Tadlock LP, Schneiderman E, Buschang PH. Applications of artificial intelligence and machine learning in orthodontics. APOS Trends Orthod 2020;10(1):17-24.
9- Sumathi S, Sivanandam S. Introduction to data mining principles. In: Introduction to Data Mining and its Applications. Studies in Compoutational Intelligence. Berlin, Heidelberg: Springer; 2006. p. 1-20. 19.
10- Mitchell TM. Machine learning and data mining. Commun ACM 1999;42:1-13.
11- Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat Sci Asiri, et al. 23 2001;16:199-231.
12- . Graber TM, Vanarsdall RL. Orthodontics: Current principles and techniques. 2nd ed. St. Louis: Mosby, 1994.
13- Kazem Bahaa, Garma Noor and Yousif Yousif (2011). The Artificial Intelligence Approach for Diagnosis, Treatment and Modelling in Orthodontic, Principles in Contemporary Orthodontics, Dr. Silvano Naretto (Ed.), ISBN: 978-953-307-687-4.
14- Thanathornwong B. Bayesian-based decision support system for assessing the needs for orthodontic treatment. Healthc Inform Res 2018;24:22-8
15- Leonardi R, Giordano D, Maiorana F, Spampinato C. Automatic cephalometric analysis: A systematic review. Angle Orthod 2008;78:145-51. 32. Arik SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging 2017;4:014501.
16- Laurenziello M, Montaruli G, Gallo C, Tepedino M, Guida L, Perillo L, et al. Determinants of maxillary canine impaction: Retrospective clinical and radiographic study. J Clin Exp Dent 2017;9:e1304-9
17- Poedjiastoeti W, Suebnukarn S. Application of convolutional neural network in the diagnosis of jaw tumors. Healthc Inf Res 2018;24:236-41.
18- Yilmaz E, Kayikcioglu T, Kayipmaz S. Computer-aided diagnosis of periapical cyst and keratocystic odontogenic tumor on cone beam computed tomography. Comput Methods Programs Biomed 2017;146:91-100.
19- Laurenziello M, Montaruli G, Gallo C, Tepedino M, Guida L, Perillo L, et al. Determinants of maxillary canine impaction: Retrospective clinical and radiographic study. J Clin Exp Dent 2017;9:e1304-9
20- Lux CJ, Stellzig A, Volz D, Jäger W, Richardson A, Komposch G. Department of Orthodontics, Dental School, University of Heidelberg, Germany. Growh dev.aging. A neural network approach to the analysis and classification of human craniofacial growth. 1998:(6) 295-106
21- Auconi P, Scazzocchio M, Caldarelli G, Nieri M, McNamara JA, Franchi L. Understanding interactions among cephalometrics variables during growth in untreated Class III subjects. Eur J Orthod 2017;39:395-401.
22- Williams J K, Cook P A, Isaacson K G, Thorn A R Fixed orthodontic appliances—principles and practice.1996, Wright, Oxford

23- Akgam M. O, Takada K. Fuzzy modelling for selecting headgear types. European Journal of Orthodontics.2002(24):99-106
24- Nanda SB, Kalha AS, Jena AK, Bhatia V, Mishra S. Artificial neural network (ANN) modeling and analysis for the prediction of change in the lip curvature following extraction and non-extraction orthodontic treatment. J Dent Spec 2015;3:217-9
25- Patcas R, Bernini D, Volokitin A, Agustsson E, Rothe R, Timofte R. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. Int J Oral Maxillofac Surg 2018;48:77-83.
26- Khanna S. Artificial intelligence: Contemporary applications and future compass. Int Dent J 2010;60:269-72

A Recent Review of Image Retrieval Algorithms in Multimedia

Anubhav Sharma[1], Dr. Shiv Shakti Shrivastava[2]

[1,2] RNTU, Bhopal

Abstract- Generally, Content-based image retrieval (CBIR) aims at developing techniques that support effective searching. A significant limitation of current content based image retrieval technology is the problem of efficiently retrieving the set of stored images most similar to a given query. One of the most essential computer systems for viewing and retrieving images from a huge database is the image retrieval system. For image retrieval, there are two approaches: text-based image retrieval (TBIR) and content-based image retrieval (CBIR) . The disadvantage of TBIR is that human annotation is impossible and costly for large databases.
One of the many fundamental ways in which CBIR differs from text retrieval is that it is based on a fundamentally different model of data. In real world, so many image retrieval systems Based on features of images is great achievement for particular applications.

Keywords: CBIR, TBIR, retrieval technology, Image features

I. INTRODUCTION

Content-based image retrieval [1,2] has been an active research area in recent years. There has been an experimental increase in the amount of non-text based data being generated from various sources.
Every second, millions of people across the world share and download a large volume of multi-media material created by various image capture devices. In order to provide visually identical results to the user's query, a high computing cost is incurred. Annotation-based image retrieval is inefficient since annotations vary in terms of languages, but pixel-wise matching of images is undesirable since the orientation, scale, image capture technique, angle, storage pattern, and so on all cause significant changes in the images [4]. In such instances, the Content Based Image Retrieval (CBIR) system is widely employed since it effectively computes similarity between query image and images from the reference dataset [1].
In particular, images have been gaining popularity as an alternative and some time more viable option for information storage. Through the increase in storage and transmission abilities more visual information is being made available on-line. These need to search and well manage large volumes of Multimedia information [2, 6]. However, while this presents a wealth of information, it also causes a great problem in retrieving appropriate and relevant information from databases.

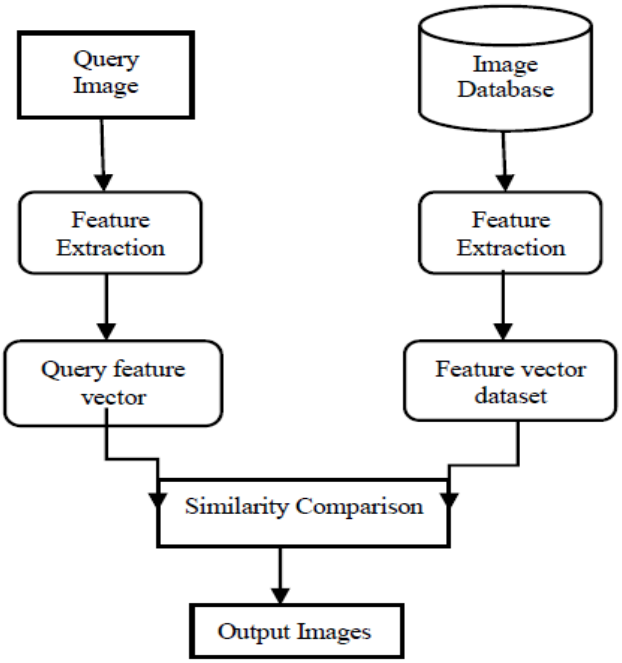


Figure 1. Block Diagram of Content-based Image Retrieval [5]

This has resulted in a growing interest, and great active research, into the extraction of relevant information from non text-based database. One of the most common data in multimedia systems is image [6]. Therefore, the capability of managing images, allocating their quick and efficient indexing and retrieval are repeatedly followed in multimedia database management systems. Researchers have been focusing on different ways of retrieving information from database or objects based on their contents.

Two key methods have been identified in image indexing and retrieval: text-based image retrieval (descriptor-based) [7] and content-based image retrieval [1, 2, 9].

Text-based image retrieval (TBIR) has been widely adopted in the cataloging and indexing of image collections in libraries, museums, and relying on manually assigned text descriptors to retrieve relevant images [8].

CBIR shows many of its methods from the field of image processing and computer vision, and is observed by some as a subset of that field. It is different from these fields principally through its importance on the retrieval of images with desired characteristics from a collection of significant size. Image processing covers a much wider field, including image enhancement, compression, transmission, and interpretation. While there are grey areas (such as object recognition [12] by feature analysis), the difference between typical image analysis and CBIR is usually fairly clear-cut. An example may make this clear. Many police forces now use automatic face recognition systems [10]. Such systems may be used in one of two ways. Firstly, the image in front of the camera may be compared with a single individual's database record to verify his or her identity. In this case, only two images are matched, a process few observers would call CBIR [1, 2, 9, 10]. Secondly, the entire database may be searched to find the matching images. This is a genuine example of CBIR. In this paper we use k-means clustering algorithm to group similar cluster image features into cluster using color feature and texture feature of images which is one of the main generally used features for finding image similarity retrieval [13].

The rising growth in image databases accessible on the Internet, we need a proficient storage space and retrieval system for images. Images in a database are typically indexed using text footnote. Content based image retrieval (CBIR), on the contrary, is the method of retrieving images similar to a given query image using only the content feature of the image. Some features of an image such as color, texture, and shape represent the content of the image [12].

CBIR systems may be employed in a broad range of applications, including biometric systems, digital libraries, multimedia recommender systems [3, 9], multimedia event detection [2, 16], and so on. CBIR-based similarity search is used by Google Images, TinEye, eBay, SK Planet, and Flipkart, among others. These websites assist users in finding the needed images by allowing them to upload or pick an image from a pre-selected group of images. Many similarity search-based algorithms have been developed to obtain visually comparable images based on image descriptors [3].

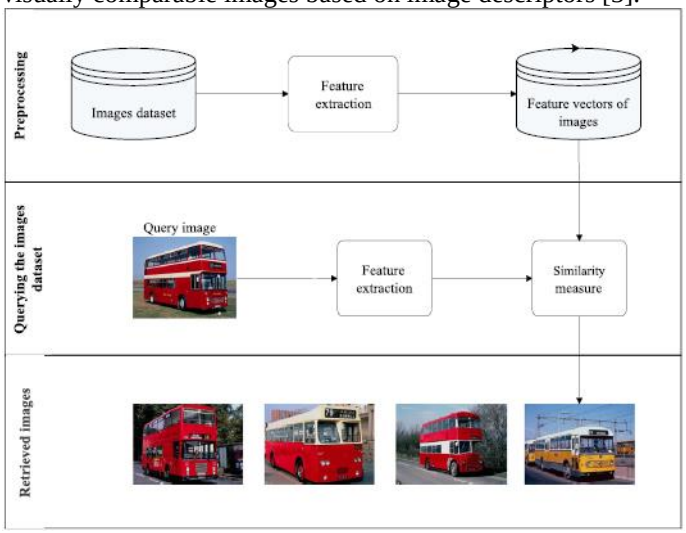


Figure 2. Overview of general CBIR model[10]

Relevance feedback retrieval is addition feature in CBIR, which replies the user for feedback on retrieval results and then use this feedback on succeeding retrievals with the goal of increasing retrieval performance [11].

CBIR systems work in two stages: The first stage, known as indexing, extracts and stores characteristics of the dataset images in feature vectors. The query image features are compared to images in the dataset in the second stage. Since image retrieval techniques rely on image attributes to effectively represent images. Shape, texture, color, and spatial information are the most suitable primitive image attributes for image retrieval from diverse image datasets. Many CBIR

methods that employ numerous characteristics to represent an image have been suggested.

2. Related Work

Similarity matching is significant issue in CBIR. So many image retrieval applications are based on shape feature and color feature [1, 2]. As well, lots of others have proposed CBIR method in the literature [6] which is based on image features.

In Simple words k-means clustering is an algorithm to classify or to group of objects based on attributes / features into K number of group K is positive integer number. The grouping is completed by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-mean clustering is to classify the data using specified number of cluster[15].

Color feature plays important role in image retrieval system and comparing all the colors in two images would however be time consuming and difficult problem to overcome this problem they introduced a method of reducing the amount of information. One way of doing this is by quantizing the color distribution into color histograms [9]. Using color histogram easier way for color distribution or they used histogram divide in to different classes for matching. They initiate color coherence vector in the basis of histogram splitting. It is very effecting in matching and comparing of two image histograms. They introduced a method for impressive additional control on histogram based matching called histogram refinement method. This technique classifies pixels into two groups coherent or incoherent. A coherent pixel belongs to a connected region, in which all the pixels are in the same color and incoherent pixel is not.

This section focuses into the numerous CBIR techniques that are relevant to the planned study. Image feature selection and extraction techniques, the amount of features used to describe the image, similarity measure methods used to compute the similarity between two images, and retrieval methods used to search photos from the image dataset all have a significant impact on the CBIR system. A major difficulty with CBIR systems is that an image might have several copies that change in size or color, and it can be seen from different perspectives, making pixel-by-pixel comparison difficult.

Satellite images are essential nowadays, and they are utilized in remote sensing, weather forecasting, and monitoring day-to-day national security challenges, among other things. Image storage and retrieval are becoming increasingly difficult due to the high capacity and resolution of earth imaging sensors and image collecting systems. Images are retrieved using text-based image retrieval, which has proven ineffective. The retrieval challenge is solved by content-based image retrieval (CBIR) for distant information systems. The existing CBIR system, which is based on low-level feature extraction techniques such as color, shape, and texture, is still a work in progress.

To achieve the greatest results, color and form methods were used. Using color, shape, and texture descriptors, the obtained findings were good but not very accurate because no one characteristic can provide superior results.

As a result, higher-level feature extraction approaches are applied to get the most effective and efficient outcomes while also bridging the semantic gap.

However, it has been identified as an area that requires further investigation. When dealing with satellite domain images, feature selection and extraction approaches become a little more laborious and difficult than when dealing with other domain images.

Multimedia content analysis is used in a variety of real-world computer vision applications, and digital images make up a large portion of multimedia data. In recent years, the complexity of multimedia materials, particularly photos, has expanded tremendously, and more than millions of photographic images are submitted everyday to various archives such as Twitter, Facebook, and Instagram. Searching for a suitable image in an archive is a difficult research challenge for the computer vision research field. The majority of search engines fetch photos using traditional text-based algorithms that depend on captions and metadata.

CBIR Research Using Deep- Learning Techniques:

When searching for digital photos in larger storage or databases, content-based image retrieval (CBIR), also known as query-based image retrieval (QBIR), is utilized. Many techniques, such as the scale-invariant transform and vector of locally aggregated descriptors, are employed to overcome this problem. Due to the most great probability and high performance of the deep convolutional neural network (CNN), an unique term frequency-inverse document frequency (TF-

IDF) employing weighted convolutional word frequencies as description vector is suggested for CBIR.
For this aim, the trained filers of convolutional layers of a convolution neuron model were utilized as a detector of visual words, with the degree of the visual pattern supplied by the activation of each filter as the tf component.) Three ways to compute the idf portion are suggested [82]. These approaches combine the TF-IDF with CNN analysis for visual material to provide strong image retrieval algorithms with improved results.
Most CNN models generate the features in the final layer by utilizing a single CNN with order less quantization, which has the disadvantage of limiting the use of intermediate convolutional layers for recognizing local image patterns. As a result, a novel strategy known as bilinear CNN-based architecture is identified in this study.)is method employed two parallel CNN models to extract the feature without previous knowledge of the semantics of image content.)e feature is directly retrieved from the activation of the convolutional layer rather than decreasing extremely low-dimensional feature.

Conclusion

This survey is based on still image database and animated images used mostly in multimedia database. Now day's, every field requires the image retrieval system for their effective work. The main When searching for digital photos in larger storage or databases, content-based image retrieval (CBIR), also known as query-based image retrieval (QBIR), is utilized. Many techniques, such as the scale-invariant transform and vector of locally aggregated descriptors, are employed to overcome this problem. Due to the most great probability and high performance of the deep convolutional neural network (CNN), an unique term frequency-inverse document frequency (TF-IDF) employing weighted convolutional word frequencies as description vector is suggested for CBIR.
For this aim, the trained filers of convolutional layers of a convolution neuron model were utilized as a detector of visual words, with the degree of the visual pattern supplied by the activation of each filter as the tf component.) Three ways to compute the idf portion are suggested [82]. These approaches combine the TF-IDF with CNN analysis for visual material to provide strong image retrieval algorithms with improved results.
Most CNN models generate the features in the final layer by utilizing a single CNN with order less quantization, which has the disadvantage of limiting the use of intermediate convolutional layers for recognizing local image patterns. As a result, a novel strategy known as bilinear CNN-based architecture is identified in this study.)is method employed two parallel CNN models to extract the feature without previous knowledge of the semantics of image content.)e feature is directly retrieved from the activation of the convolutional layer rather than decreasing extremely low-dimensional feature.

problem is image retrieval in a general perspective have not yet been satisfactorily solved. The grayscale values mean variance contrast level and various sizes of the intensity values are considered as appropriate features for retrieval. We have shown that different methods are quite useful for relevant image retrieval queries.

References

[1]. Ahmed KT, Ummesafi S, Iqbal A (2019) Content based image retrieval using image features information fusion. Inf Fus 51:pp76–99
[2]. Cheng Z, Chang X, Zhu L, Kanjirathinkal RC, Kankanhalli M (2019) Mmalfm: Explainable recommendation
by leveraging reviews and images. ACM Trans Inf Syst (TOIS) 37(2):16
[3]. Garcia N, Vogiatzis G (2019) Learning non-metric visual similarity for image retrieval. Image VisvComput 82:pp18–25
[4]. Mehmood Z, Mahmood T, Javid MA (2018) Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Appl Intell 48(1):166–181
[5]. Mezzoudj S, Seghir R, Saadna Y et al (2019) A parallel content-based image retrieval system using spark and tachyon frameworks. Journal of King Saud University-Computer and Information Sciences
[6]. Ouni A, Urruty T, VisaniM(2018) A robust cbir framework in between bags of visual words and phrases

models for specific image datasets. [7]. Multimedia Tools and Applications 77(20):26173–26189
[8]. Sotoodeh M, Moosavi MR, Boostani R (2019) A novel adaptive lbp-based descriptor for color image retrieval. Expert Systems with Applications
[9]. Yan L, Lu H,Wang C, Ye Z, Chen H, Ling H (2019) Deep linear discriminant analysis hashing for image retrieval. Multimed Tools Appl 78(11):15101–15119
[10]. Zhou J, Liu X, Liu W, Gan J (2019) Image retrieval based on effective feature extraction and diffusion process. Multimedia Tools and Applications 78(5):pp6163–6190
[11]. A. Amelio, "A new axiomatic methodology for the image similarity," Applied Soft Computing, vol. 81, p. 105474, 2019.
[12]. S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, (2019)"SkeletonNet: a hybrid network with a skeleton-embedding Process for multi-view image representation learning," IEEE Transactions on Multimedia, vol. 1, no. 1,.
[13] Zhao Q, Xu M, Franti P (2011) Extending external validity measures for determining the number of clusters. In: Proceedings of international conference on intelligent system design and applications, IEEE, pp 931–936
[14] W. Zhao, L. Yan, and Y. Zhang, "Geometric-constrained multi-view image matching method based on semi-global optimization," Geo-Spatial Information Science, vol. 21, no. 2, pp. 115–126.
[15]. Emmanuel M, Ramesh Babu DR, Potdar GP, Sonkamble BA, Praveen G (2007) Content based medical image retrieval. In: Proceedings of international conference on information and communication technology in electrical sciences, IEEE, pp 712–717
[16]. Dube S, El-Saden S, Cloughesy TF, Sinha U (2006) Content based image retrieval forMR image studies of brain tumors. In: Proceedings of IEEE international conference on engineering in medicine and biology, pp 3337–3340

Barriers for smart grid development in India

[1]Archana
Research Scholar, Indian Institute of Technology, New Delhi

Abstract
Growth in electricity sector is key to the development of a country, as it helps in development of other sectors such as manufacturing, agriculture, commercial enterprises and railways. This development has led to increase in global warming and the limited resources of conventional energy sources have pushed generation companies to look for clean energy alternatives. With the advancement in technology, smart grid is considered to be a promising solution, which integrates renewable energy resources and communication system with traditional grid and offers various programs with direct consumer's involvement. This paper focuses on the barriers faced in implementing smart grid technology in India.

Keywords: Smart grid, sustainability, covid-19, renewable energy, power sector.

Introduction

India has a very vast and complex electricity network, and its energy demand and generation capacity is still growing. According to the Ministry of Power, India's electricity sector suffers from very high transmission and distribution loss. It averages around 26% of total electricity production, and as high as 62 per cent in some states, which is among highest in the world. Also, these losses do not include non-technical losses like theft etc. and if such losses are included, the average losses are as high as 50 % (NITI Aayog, 2015). Due to this there is huge loss to India's electricity sector (NSGM, 2019). There are some technical flaws in the Indian power grid as, poorly planned distribution network, overloading of the system components, lack of reactive power support and regulation services, etc. (Jain *et al.*, 2016). As the population is increasing, demand of electricity will increase and it will affect the existing network efficiency. Due to this there is a need to upgrade the transmission and distribution network. With the recent advancement in technology, new types of devices and ICT infrastructure, various energy management tools are deployed at distribution as well as transmission level to make the grid smart.

Smart Grid is a new concept of electric grid that integrates the Information and Communication Technology (ICT) with the existing grid to improve its reliability and efficiency with integration of renewable energy (Anku, Abayatcye and Oguah, 2013). With the increase in global warming and fear of running out of petroleum fuels, use of renewable energy resources like wind and solar is increasing (Throop and Mayberry, 2017). Solar seems to have a great potential in India due to an average of 300 solar days per year. Hence implementation of smart grid technology seems to a suitable solution.

1. **An overview of the smart grid development in India**

India started its smart grid mission with a vision of: "Transform the Indian power sector into a secure, adaptive, sustainable and digitally-enabled ecosystem that provides reliable and quality energy for all with the active participation of stakeholders (Ministry of Power India, 2013)." The Ministry of Power (MoP) constituted India Smart Grid Forum (ISGF) and India Smart Grid Task Force (ISGTF) in the year 2010. It aimed to accelerate the development of smart grid technologies in the Indian power sector. In March 2015 'National smart grid mission (NSGM)' was

set up by MoP, GoI. Its main function is to plan and monitor the implementation of policies and programs related to the smart grid. The MoP has developed a short film on the smart grid named 'Smart Grid - Towards an Intelligent Future' to increase consumer awareness.

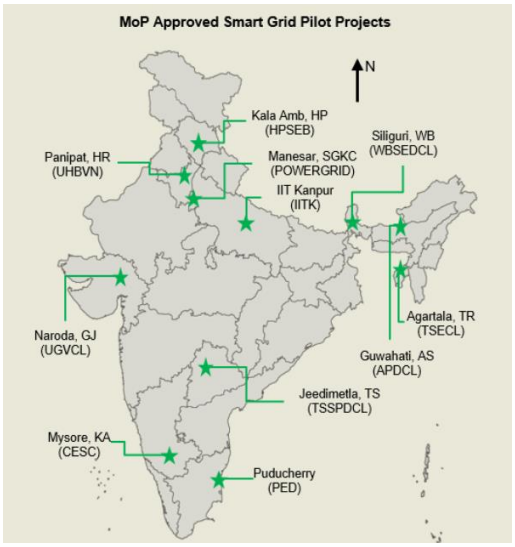
2.1 Pilot projects in India

To encourage the smart grid technology, MoP, GoI allotted eleven pilot projects to different state-owned discoms (ISGF, 2017) and one smart city pilot project in 2012. They are partly funded by MoP (50% of the project cost as a grant from GoI) (ISGF, 2017). They have functionalities like Advance metering infrastructure (AMI), outage management, peak load management, power quality management, microgrid, distributed generation, etc. The total cost of these projects is Rs. 4.8 billion. The location of these pilot projects and projects are shown in Figure 1+++. Also, more than two million smart meters have been deployed (NSGM website, 15.01.2021).

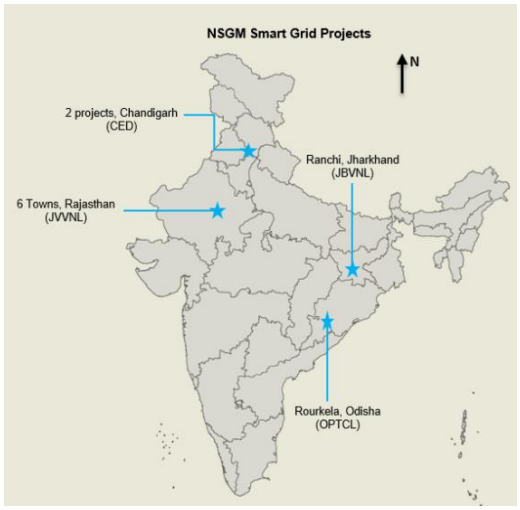
2.2 Insights from the smart grid pilot projects

Smart grid pilot projects have given useful insights in overcoming various technical and commercial issues and consumers' knowledge and behaviour. Developed nations already have a reliable infrastructure; therefore, they are further working on smart metering, renewable generation, and forecasting tools. However, developing countries like India that have poor infrastructure need to simultaneously strengthen electricity networks and build additional layers of ICT development to make the grid smarter. Another big challenge for India is providing electricity to households in remote areas and reducing T&D losses and power theft. In this scenario, the outcome of these pilot projects is expected to play a key role in resolving such issues. Problems like meter tampering and power theft could be quickly resolved due to real-time identification. As consumers were able to visualise their power consumption, it helped them better energy management and reduce electricity bills.

Along with benefits, several challenges were also faced by the implementing agencies during the execution of the pilot projects. Few projects were cancelled, and some were delayed due to a lack of clarity of the funding sources. Some challenges faced were: consumers faced differences in billing, regulatory approval challenges, lack of a uniform standard for a smart meter, lack of interoperability, lack of service providers, lack of skilled workforce, etc. (NSGM, 2019). As consumers are little aware of smart grid technology, local resistance was also observed in some places. As consumers participation in these pilot projects was limited in number, it does not represent society's actual behaviour (NSGM, 2018). Energy use is affected by consumers behaviour. In a nutshell, consumer awareness is required along with regulatory interventions and technology advancement for smart grid adoption and advancement. Despite all challenges, pilot projects have helped to understand the problem and formulate future project development strategies. India has immense opportunities for the development of the smart grid. As Indian government has started the deployment of projects, learning from the pilot projects will certainly help in defining the trajectory for transition to smart grid from conventional grid.



(a)



(b)

Figure 1: (a) Smart grid pilot projects in India, and (b) Smart grid projects in India.

3. Barriers in implementation of smart grid in India

Experiences from pilot project implementation shows that there are some challenges in development of smart grid in India. Literature also suggest about barriers due to various reasons in implementation of smart grid technology. On doing an exhaustive literature survey following challenging barriers were found:

i) Lack of infrastructure: India needs to develop a proper infrastructure for the development of electricity sector. Lack of infrastructure will delay the implementation of smart grid technology.

ii) Lack of proper regulatory mechanism: A regulatory mechanism is essential for the development of a system. Due to lack of regulatory mechanism few smart grid pilot projects were cancelled and few were delayed. Hence need of regulatory mechanism is vital.

iii) Lack of consumer awareness: As consumers play an active role in the energy management for smart grid, hence it is essential to train and educate consumers about the benefit of smart grid technology. Consumer should be aware of the various programs and incentives offered by the utilities (Stragier, Hauttekeete and De Marez, 2010).

iv) Lack of government policies: Smart grid implementation needs proper government policy. Promotion of open standards, formation of uniform standard, and development of training centres plays vital role in development and implementation of smart grid technology.

v) Effect of Covid-19 pandemic: Covid-19 pandemic has affected a lot on the smart grid

implementation. Due to lockdown in the country, it is difficult to manage the essentials. Also many offices were closed and work was halted which has adversely affected the process of implementation of smart grid technology.

4. Conclusion

In this paper barriers in implementation of smart grid technology in India and smart grid pilot project India has been discussed. A discussion on barriers helps in formulation of strategy for implementation of smart grid technology in a better way. In conclusion, systematic and proper implementation of smart grid technology will definitely provide an optimal solution to India's energy needs.

References

Anku, N., Abayatcye, J. and Oguah, S. (2013) 'Smart grid: An assessment of opportunities and challenges in its deployment in the ghana power system', in *2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5. doi: 10.1109/ISGT.2013.6497800.

India Smart Grid Forum (2017) *Smart Grid Handbook for Regulators and Policy Makers*. New DElhi: India Smart Grid Forum.

Jain, K. K. *et al.* (2016) 'Power quality improvement through smart control & diagnostics of power transformer load tap changer', *Water and Energy International*. Tata Power Delhi Distribution Limited, India: Central Board of Irrigation and Power, 59RNI(4), pp. 35–41. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84982223857&partnerID=40&md5=e0244c61ebff391b72cb9ab44c6a01e3.

MoP India (2013) 'Smart Grid Vision and Roadmap for India', pp. 9–10.

NITI Aayog (2015) *Report of the expert group on 175 GW RE by 2022*. Available at: https://niti.gov.in/writereaddata/files/175-GW-Renewable-Energy.pdf.

NSGM (2018) *Key Learnings from Implementation of Smart Grid Pilot Projects*.

NSGM (2019) *Impact Assessment of Smart Grid Pilot Projects Deployed in India*. New Delhi.

Reji Kumar Pillai, Rupendra Bhatnagar, J. S. (no date) *AMI Rollout Plan for India ISGF±BNEF Knowledge Paper AMI Rollout Plan for India*. Available at: http://www.indiasmartgrid.org/reports/AMI Roll-Out Strategy and Cost-Benefit Analysis for India_ISGW2017.pdf.

Stragier, J., Hauttekeete, L. and De Marez, L. (2010) 'Introducing smart grids in residential contexts: Consumers' Perception of Smart Household Appliances', in *2010 IEEE Conference on Innovative Technologies for an Efficient and Reliable Electricity Supply, CITRES 2010*. IBBT-MICT-UGent, Ghent, Belgium, pp. 135–142. doi: 10.1109/CITRES.2010.5619864.

Throop, W. and Mayberry, M. (2017) 'Leadership for the Sustainability Transition', *Business and Society Review*. Green Mountain College, Poultney, VT, United States: Blackwell Publishing Inc., 122(2), pp. 221–250. doi: 10.1111/basr.12116.

Design and Development of Low-Cost Eva Shoe for Bunion and Hammer Toe Foot Deformities

[1] Arjun Verma, [2] D.K. Chaturvedi

[1] Research Assistant, Dept. of Footwear Technology, [2] Professor and Head, Dept. of Footwear Technology
Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra

[1]vermaarjun181@gmail.com, [2]dkc.foe@gmail.com

ABSTRACT

The incidence of Bunion and Hammer toe come in the literature range from 2% to 20%. In general orthopedic practice, the surgery for such kind of deformities is among the most commonly conducted interventions. There are many variations in the definitions of Bunion and Hammer toe in the scientific literature, which is remarkable for such a common orthopedic problem. Lesser toe surgery is among the most interposition in general orthopedic practice. However, the definitions of Bunion and hammer toe are not uniform and it depends on person-to-person problem. Keeping this, a non- surgical method is used to cure the deformities. The objective of this study is to suggest a clear definition for such kind of deformities to establish precise communication with the use of Compression molding machine which has gained a lot of attention and popularity in the era of producing fashionable goods at a very good speed. In an attempt to make the technology affordable and reachable to the masses, it has been designed as a low-cost single piece of material model for Bunion and Hammer toe diseases from Compression molding process- reducing all the wastages of material, producing high precision and accuracy, and a lot of time is saved during the manufacturing.

Key Words- Bunion, Metatarsal Phalangeal Joint, Hammer toe, Molding, EVA, Shoe Last

INTRODUCTION

Foot deformities can either be present at birth or it develops over some time. Wearing tight fitted shoes or Putting abnormal strain on the foot may play a vital role for such kind of deformities. Some risk factors include injuries, inflammation, being overweight and diseases like osteoarthritis, rheumatoid arthritis, or brain diseases. Genes also play an important role for such deformities. For instance, some people have weak connective tissue, so the supporting structures in the foot cannot always

hold everything in place properly. It is because in young people, joints and bones are still developing and changing. So, we need to consult with orthopedic doctor for curing these deformities and doctor designed the customized shoe according to the deformity. In general, the cost of the customized shoe is high and we cannot reach to masses. Keeping this, to reduce the cost of manufacturing and reachable to the masses, it is proposed a single piece of material model for complete rehabilitation. [1]

Motivation

The main reason is to identify the variations in the applied definitions of Bunion and Hammer toe. The objective of this literature study is to proposed a clear definition of Bunion and Hammer toe. Only then, we can communicate about such type of deformity and talks about it. The uniformity in definitions of Bunion and Hammer toe are compulsory in communications and give directions to the treatment of these abnormalities in a better way by using the technology in a better way.

Problem statement

Deformities can be painful and affect your walk. It may cause your skin to become hard and thick and sometimes lead to calluses and pressure sores. Misalignments can cause toes and other parts of the foot to become deformed. It puts abnormal strain on tendons and muscles and in most cases the abnormal strain on joints can lead to wear-and-tear and eventually to osteoarthritis. Keeping all the aforesaid points in mind, it has been proposed to make a low-cost EVA shoe for persons who are suffering from these types of deformities to stop the further growth of bones in MP joint and second to third phalanges.

Bunion

A bunion, also known as hallux valgus, is a deformity of joint connecting the first Metatarsal Phalangeal joint as shown in Fig.1.1. In this, the big toe often bends towards the other toes and the joint becomes red and painful. Still the exact cause is unclear. Some factors include wearing overly tight shoes, pointed shoes, high-heeled shoes, family history, and rheumatoid arthritis. Diagnosis of such deformities is generally based on symptoms and supported by X-rays. [1,2]

Fig.1 Foot having Bunion deformity [a]

Fig.2 Arrangement of Bunion and Hammer toe foot deformities [b]

Hammer toe

A hammer toe is such kind of deformity as shown in Fig.1.2 in which your toe tends to bend or curl downwards instead of pointing forward. This type of deformity can affect any of your toe on your foot. It mostly affects the second toe or third toe. Although, it may be present at birth. It develops over time due to arthritis or wearing some ill-fitting shoes, such as tight, pointed heels. In most of the cases, a hammer toe condition is treatable. [3]

Causes of Bunion and hammer toe to form

In toe, we have two joints which allow it to bend at the middle and at the bottom. When the middle joint becomes flexed or bent downward, a hammer toe occurs.

Some of the common causes of this include:

- A traumatic toe injury
- Arthritis
- Unusual high foot arch
- After wearing shoes which don't fit properly
- Tightened ligaments or tendons in the foot

- The pressure from a bunion, when your big toe points inward toward your second toe [4,5,6]

Note- Spinal cord or peripheral nerve may damage all of your toes to curl downward.

TABLE-1 LITERATURE REVIEW OF CUSTOMIZED SHOE [16-22]

S. No.	GOAL	REFERENCES
1	Adding a soft cushioned inside the shoe.	Ellen Sobel (2001)
2	Foot padding, taping, night splints, orthotic devices may be prescribed. Allowances should be made for adequate space in the shoe.	Asad Ayub (2005)
3	Prefabricated soft leather shoe with a wide toe box and preferably a soft sole may give significant relief of symptoms.	Salvatore Moscadini (2012)
4	Applied a nonmedicated hammertoe pad around the bony prominence of the hammertoe. Loose-fitting pair of shoes can also help to protect the foot while reducing pressure on the affected toe.	A. Luximon (2013)
5	Modified by adding a cut-out section beneath the first metatarsal and trimming the distal edge to the level of the second to fifth toe. This shoe is characterized by a rounded sole in the anteroposterior direction and a soft cushioned heel.	Hylton B. Menz (2016)
6	Applied extra-depth in the forefoot region to accommodate for foot orthoses and forefoot deformity, soft leather upper and smooth lining to offer protection, laces, padded heel counter to improve fit at the heel and a long inside counter to improve rearfoot stability and arch support.	Mike Frecklington (2018)
7	Applied treatment for patients who favour shoes with a lower heel and wider toe box. The use of a more cushioned thicker insole can decrease the pressure and impact on the intermetatarsal space. MT pad made of rubber, polyurethane, or silicone can be applied.	Chul Hyun Park (2019)

Parameters to be measured on the foot

In order to get an accurate shape and dimension of the foot, the various locations of the foot are to be measured- [7]

1. A draft- Outline plan of the foot with weight on
2. An impression- To show the distribution of weight
3. A profile- shows height of the big toe and instep contour
4. Length- Taken with a size stick
5. Girth measurements as shown in Fig.3-
    a) Joint- around a metatarso - phalangeal joint
    b) Waist- from smallest girth behind the joint
    c) Instep- smallest girth which pass over prominence over on middle cuneiform
    d) Long heel- seat to instep to give passing
    e) Short heel- seat to lowest crease in front of the ankle
    f) Calf- around thigh (highest circumference of thigh) [8]

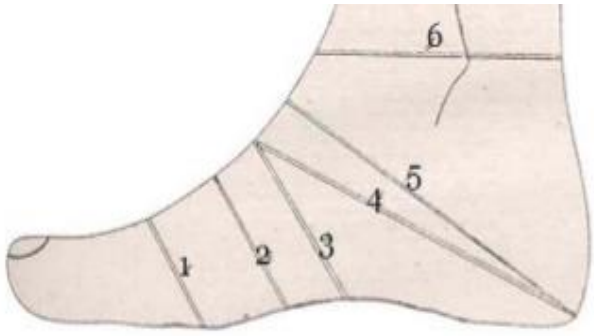


Fig.3 Girth measurements [c]

where number indicates-

1. Ball girth
2. Waist girth
3. Instep girth
4. Long heel girth
5. Short heel girth
6. Ankle girth

Indicative Fitting- These fittings are either be denoted by alphabets or numbers as mentioned in table 2 [9]

Table 2 Indicative Fittings

Description	English size	French size	USA size (Men's)	USA size (Ladies)
Very small	D, E	4, 5	AAA	AAA, S, SS
Small	F, $F^{1/2}$	6, $6^{1/2}$	B, A, AA	A, AA, N
Medium	G, $G^{1/2}$	7, $7^{1/2}$	D	B, M

Large	H	8	E, EE	C, D, WW
Extra large	XH	9	EEE	E, WWW

262 + 8 = 270mm

Ankle girth- Ball girth measurement – 5mm

245 – 5 = 240mm

Table 3 Multi fittings Ball girth measurements in mm

Size	**6**	**7**	**8**	**9**	**10**
E	221	227	233	239	245
F	227	233	239	245	251
G	233	239	245	251	257
H	239	245	251	257	263
XH	245	251	257	263	269

Proportional measurements for a new last

It has been proposed to make a last of English size 6 with multi fitting 8 as shown in Table 3

then, [10]

Length of insole- (6 + 25) x 8.46 = 262mm

Ball girth- Standard size G of 8 multi fitting gives 245mm

Instep girth- Same as the Ball girth measurements for models without laces gives 245mm

Heel girth- Insole length + Multi fitting

Creating an Insole pattern

I. Firstly, the masking of bottom profile of shoe last is done by pressure sensitive adhesive tape.

II. After this, the masking tape on shoe last is removed and paste it on a hard sheet of paper.

III. Draw the contour and mark the position of hammer toe as shown in Fig.4 and cut with the help of knife to get the mean forme of Hammer toe standard pattern.

IV. It will give you the position of Hammer toe foot deformity.

V. Draw the contour and add 14mm at ball point area and cut with the help of knife to get the mean forme of Bunion standard pattern.

VI. Mark a point at a distance as shown in Fig.5 from seat area. Draw a perpendicular line from that point horizontally to mark the Ball point. It will give you the position of Bunion foot deformity.

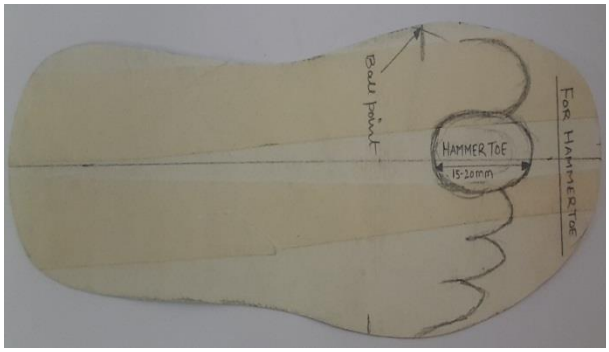


Fig.4 Location of Hammer toe deformity on Insole

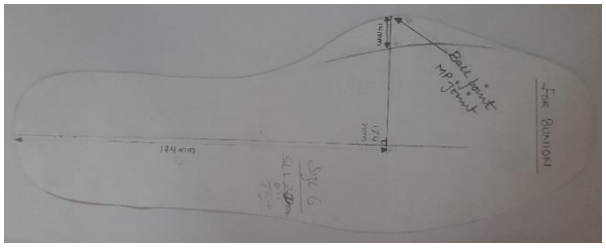


Fig.5 Location of Bunion deformity on Insole

Development of Shoe last

I. Once your insole is ready for use, it will go for making shoe last as shown in Fig.6 with the aforesaid proportional measurements for a new last.

II. Using the Jugaad technology at DEI using EVA sole material for marking the location of Bunion deformity as shown in Fig.7 and paste it on a desired location with the help of nails.

III. Using the Jugaad technology at DEI using EVA sole material for marking the location of Hammer toe deformity as shown in Fig.8 and paste it on a desired location with the help of nails.

IV. Now using the adhesive tape for proper covering as shown in Fig.9 at the toe area for proper placement of all the intervention parts which were developed at the fabrication workshop.

V. It will also modify the present design of mold according to the shape of shoe last.

Fig.6 Development of shoe last according to proportional measurements

Fig.7 Locating left foot Bunion deformity on shoe last

Fig.8 Locating left foot Hammer toe deformity on shoe last

Fig.9 Side view of left foot last with Bunion and Hammer toe foot deformities

Compression Molding

Compression molding is a method of molding the polymer material, generally the preheated and should be placed in an open heated mold cavity chamber. After this, we need to close the mold with a top force or plug member and then pressure is applied to force the polymer material in contact with all the areas in a mold, during this pressure and heat are maintained properly until the molding material has been cured. Once your molding process is completed, we need to remove the excess material. This process employs thermosetting resins in a partially cured stage, which either in the form of granules, putty-like masses, or preforms. [11,12]

In compression molding, we have to focus the six points while dealing with it-

- We need to determine the proper amount of EVA material.
- We need to determine the minimum amount of energy required to heat the material.
- We need to determine the minimum time required to heat the material.
- We need to determine the appropriate heating technique.

- We need to predict the required force which ensures that shot attains the proper shape.

Principles of Compression Molding

1. In compression molding, the thermoset compound is first placed in the open heated mold.

2. We can use the material in a powder form or as a preform.

3. Then the mold closes and heat and pressure cause the material to flow and compress it to get the required shape and density as defined by the mold.

4. Must supply the continued heat and pressure produce the cure that hardens the material

5. Thinner the part, the shorter will be the cure, thicker pieces take longer time to cure.

6. The part design should have as uniform a wall thickness as possible.

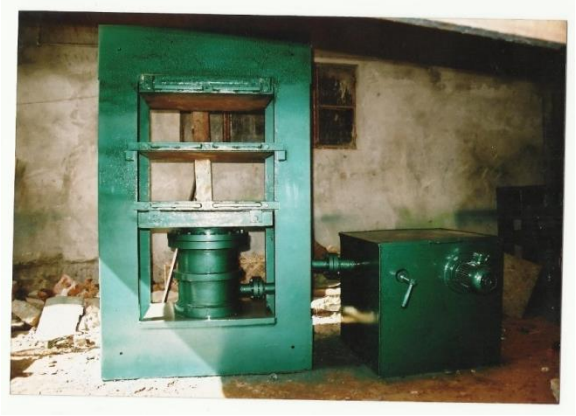


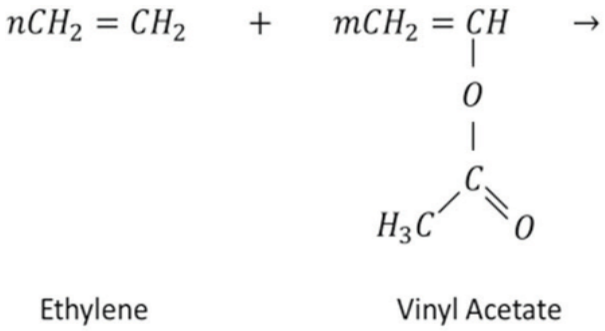
Fig.10 EVA sole compression molding machine [d]

Fig.11 EVA shoe after released from compression molding machine

EVA (Ethylene vinyl acetate)

Foam is created as the gas bubbles are trapped inside in a chemical resin. It can be created through an endothermic or an exothermic reaction. It starts when a solid and whichever process is going to be used, foaming occurs which creates the cell structure and, ultimately,

foam. In general, the most common closed cell foam is manufactured by the endothermic and the open cell by the exothermic process. In Open cell, the foam contains bubbles which are interconnected with each other in a wide web. Open cell foam may be soft and it will absorb any liquid when it comes in contact with it. It is perfect for so many applications, especially if you require the extreme softness and there is no contact with a liquid environment. [13]



[e]

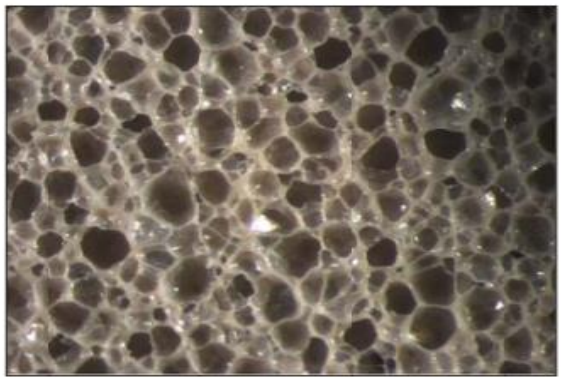


Fig.12 Close picture of open celled EVA [f]

Development of Low-cost EVA shoe

Fig.13 Side view of first attempt EVA shoe

Fig.14 Side view of second attempt EVA shoe

NOTE- In the third attempt, the proposed EVA shoe is made with accuracy and precision. In the left shoe, the size of Bunion and location of hammer toe is in actual size but in the right shoe, it is designed according to the person problem who is suffering from the deformities because the actual dimensions for the above deformities is not fixed. It is customized and varied from person to person and according to the shape and size of the deformities.

Fig.15 Side view of right foot Bunion and Hammer toe deformities

Fig.16 Side view of left foot Bunion and Hammer toe deformities

Comparison of shoe with and without deformities

Fig.17 a) Without Deformities

Fig.17 b) With Deformities

Some features of the proposed model

Advantages over other conventional methods:

- We need only a mold last profile and rest of the work is done by the machine
- The perfect solution for personalized and customized fabrication
- Lead time for the production is very less
- It is cheaper in cost
- It serves to several industries
- It is easy to made
- Eliminating all the different components in making the prototype

Cost Analysis of EVA shoe

Till now, we have provided the great insight of compression molding basics, its techniques and the assembly. Now the next thing to look upon is the cost analysis of model.

For building the model, the most important thing is the material (Granule) and a last profile.

As far as the cost is concerned, we do not incur any cost for manufacturing the model as it was created from the compression molding machine which was installed in our Tannery premises, DEI. The CAD file for the model was designed on the Shoe master modelling software.

Once the EVA shoe is formed, it is inserted on the shoe last for setting up the shape and finally it is de-lasted and ready for use.

The final cost of the EVA shoe is INR 600/- which is an affordable price for each and every person who are suffering from these types of deformities.

Results and Discussion

There were three attempts performed in the fabrication lab after released from the Compression molding machine. In the first attempt, the quantity of Eva material was less recorded 104.5gm and 113.5gm. In the second attempt, the weight of Eva was sufficient but the temp. of the upper plate was low recorded as 140°*C.* As a result, it was failed. In the third attempt**,** the weight of Eva, Temp. of both the plates, pressure and timer were proper measured and recorded to give the final low-cost Eva shoe in a proper condition.

There are several definitions of toe deformities in the foot. It is proposed that the metatarsophalangeal joint is the main discriminating factor and essential characteristic for such deformity. Bunion and hammer toe will be identified by flexion in the proximal interphalangeal joint, which is only the single criteria for a hammer toe deformity. The flexibility of these joints will be a basic factor in discriminating the deformities. The development of these deformities should be regarded as one of the most challenging factors for a continuous and for the same pathophysiologic process. The discrimination between the Bunion and Hammer toe is to be performed on the basis of the state of the metatarsophalangeal joint. It also accounts for gradation of these deformities by describing the position and flexibility of the proximal interphalangeal and metatarsophalangeal joints. A fixed flexion deformity at the proximal interphalangeal joint is defined as a hammer toe as long as the metatarsophalangeal joint is flexible.

Conclusion and Future Scope of Work

Footwear interventions are only associated with the reductions in foot pain, some impairment and the disability in people with suffering from rheumatoid arthritis. This EVA shoe single prototype will improve the foot pain and function in people with 1st metatarsophalangeal joint osteoarthritis. The accuracy and precision after wearing the custom shoe will help to perform any task in a better way especially in farming because the EVA material resist to water, crack resistance, and from UV radiation. The ability to absorb shock makes it to useful in any industry. Footwear interventions help to reduce plantar pressure rheumatoid arthritis, 1st metatarsophalangeal joint osteoarthritis, and improve walking velocity in rheumatoid arthritis.

The present design has been modified in various ways as there is no end to innovation. It will help to achieve accuracy and increase the usefulness of the machine. It will differentiate the model with other conventional manufacturing technique i.e., how much time is consumed to make with accuracy, by what amount the material is waste, and what is the total cost of the final model during the manufacturing. So, in this way we have an idea of manufacturing accuracy with time if we compare with additive or subtractive manufacturing techniques.

It will also use to make shoe for Claw and Mallet toe foot deformities by using the same compression molding technique. Other footwear deformities shoe will also be made by this technique.

Present design will use 3-D scanner for scanning the foot to get the exact shape, size, and location of deformity. The concept of Artificial Intelligence will also be implemented for making customize shoe last for various deformities in foot.

REFERENCES

1. [DAY 2017] Dayton, Paul D. (2017), Evidence-Based Bunion Surgery: A Critical Examination of Current and Emerging Concepts and Techniques, Springer. pp.1-2, ISBN 9783319603155

2. [REB 2016] Rebecca Cerrato, Nicholas Cheney (2016), "Hallux Valgus", American Orthopedic Foot & Ankle Society, ISBN 978-0823081634
3. [CHA 2011] Chadwick, C; Saxby, TS (2011), "Hammertoes/Claw toes: metatarsophalangeal joint correction". Foot and Ankle Clinics. 16 (4): 559-71, doi: 10.1016/j.fcl, Aug, PMID 22118229
4. [FER 2010] Ferri, Fred F. (2010), Ferri's Differential Diagnosis E-Book: A Practical Guide to the Differential Diagnosis of Symptoms, Signs, and Clinical Disorders, Elsevier Health Sciences. p. 323. ISBN 978-0323081634
5. [BAR 2016] Barnish (2016), "High-heeled shoes and musculoskeletal injuries: a narrative systematic review". BMJ Open. 6 (1): e010053. doi:10.1136/bmjopen-2015-010053,
PMC 4735171. PMID 26769789
6. [BAL 2021] Balasankar Ganesan, Palak Prasad, Suraiya Akter, Raymond K.Y.Tong (2021), Handbook of Footwear design and Manufacture (Second Edition), The Textile Institute Book Series, Pages 413-438
7. [W C R 1948] WILKINSON C. R. (1948), 'Last Development and Design' Ibid. 3, 472
8. [G A H 1936] GREGORY, A. H (1936), 'Toe Spring' Nat. Instn. Boot Shoe Ind. J. 2, 5l4
9. Manuals from AR Sutoria, Milan, Italy
10. [J H T 1951] THORNTON, J. H (1951), 'The English Shoe Size Scale' J. Brit. Boot Shoe Instn. 4, 514
11. [DEA 2016] Dean and Yvonne (2016), Materials technology, Routledge, ISBN 9781315504285
12. [AHM 2015] Ahmad Adlie Shamsuri (2015), Compression Moulding Technique for Manufacturing Biocomposite Products, International Journal of Applied Science and Technology, June, Vol. 5, No. 3
13. [OMP 2012] Omprakash H. Nautiyal (2012), Molding of EVA Soles Using Expanding and Reducing Agents, International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.07, ISSN: 0975-5462

14. [ROD 2015] Roddy E, Muller S, Rome K, Chandratre P, Hider SL, Richardson J, et al (2015), Foot problems in people with gout in primary care: baseline findings from a prospective cohort study, J Foot Ankle Res; 8:31
15. [BAR 2014] Barouk, P (2014), Recurrent metatarsalgia. foot Ankle Clin, 19:407-424, 10.1016/j.fcl.2014.06.005 [doi]
16. [ELL 2001] Ellen sobel and Steven J. Levitz (2001), Pressure reduction and off-loading the diabetic foot, Journal of Continuing Medical Education by the Council on Podiatric Medical Education.
17. [ASA 2005] Asad Ayub, Steven H. Yale, and Christopher Bibbo (2005), Common Foot Disorders, Journal in Clinical Medicine and Research (CMR), May; 3(2): 116–119.
18. [SAL 2012] Salvatore Moscadini and Giuseppe Moscadini (2012), Hallux Valgus correction in young patients with minimally invasive technique, the role of Osteotomy in the correction of Congenital and acquired disorders of the skeleton, DOI:10.5772/38201
19. [LUX 2013] Edited by A. Luximon (2013), Handbook of footwear design and manufacture, Woodhead Publishing Limited, Cambridge (UK), ISBN 978-0-85709-539-8
20. [HYL 2016] Hylton B. Menz, Maria Auhl, Jade M. Tan, Pazit Levinger, Edward Roddy, Shannon E. Munteanu (2016), Biomechanical Effects of Prefabricated Foot Orthoses and Rocker-Sole Footwear in Individuals with First Metatarsophalangeal Joint Osteoarthritis, Arthritis Care & Research Vol. 68, No. 5, May, pp 603–611
21. [MIK 2018] Mike Frecklington, Nicola Dalbeth, Peter McNair, Peter Gow, Anita Williams, Matthew Carroll, Keith Rome (2018), Footwear interventions for foot pain, function, impairment and disability for people with foot and ankle arthritis: A literature review, Seminars in Arthritis and Rheumatism 47, pp 814–824

22. [CHU 2019] Chul Hyun Park and Min Cheol Chang (2019), Forefoot disorders and conservative treatment, YUJM, May; 36(2):pp 92-98
23. [CHA 2018] Cha YH, Kim SJ, Lee KH, Kwon JY, Kim DH, Seo A, et al (2018), Designing personalized toe spreaders for hallux valgus with three-dimensional scanning and printing. J Biomed Eng Biosci.;5:1–6.
24. [LES 2011] Leslie Langnau (2011), Subtractive Manufacturing: What You Need to Know, Magazine on Make Parts Fast: Design guide, October 3
25. [OTT 2010] Otter SJ, Lucas K, Springett K, Moore A, Davies K, Cheek L, et al (2010), Foot pain in rheumatoid arthritis prevalence, risk factors and management: an epidemiological study. Clin Rheumatol; 29:255–71
26. [WAL 2004] Walter; Amaro; Luca, Diviani; Davide, Montorfano; Ermanno, Oberrauch; Gabriele, Depinto; Simona, Segalini;Marinella,Levi, Stefano, Turri, Controlling the shrinkage of polymers for customized shoe sole manufacturing, International Journal of Computer Integrated Manufacturing, 17(7), pp 633-644.

Websites for Images

a) https://images.fosterwebmarketing.com/976/bunionAdobeStock_193178123.jpg
b) https://clarissachampine.files.wordpress.com/2014/12/5f404-feetfitterfeetdotcom.jpg?w=254
c) https://www.google.com/search?q=image+on+foot+girth+measurement&rlz=1C1SQJL_enIN801IN801&sxsrf=ALeKk017MDgFdKnpESROOAdDDttxJ84EGg:1617856508410&source=lnms&tbm=isch&sa=X&ved=2ahUKEwi-seyT6e3vAhUNILcAHWB6C7QQ_AUoAXoECAEQAw&biw=1536&bih=754#imgrc=_8_T38aFanGB5M
d) https://www.indiamart.com/proddetail/eva-molding-press
e) https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F323198576

%2Ffigure%2Ffig1%2FAS%3A59429
3416816640%401518702000052%2F
Polymerization-reaction-between-
ethylene-and-vinyl-acetate-resulting-
in-
EVA.png&imgrefurl=https%3A%2F
%2Fwww.researchgatenet%2Ffigure
%2FPolymerization

f) [O H N 2012] Omprakash H. Nautiyal (2012), Molding of EVA Soles Using Expanding and Reducing Agents, International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.07, ISSN: 0975-5462

Diabetes Diagnosis using Ensemble Models in Machine Learning

Ashok B[1] , Mr. Manoj Wairiya [2] , Dr. Divya Kumar[3]

[1,2,3] Department of Computer Science Engineering, Motilal Nehru National Institute of Technology, Prayagraj, IndIa

*a*shok261994@gmail.com wairya@mnnit.ac.in divyak@mnnit.ac.in

Abstract—Diabetes is one of the diseases where early detection is must given the fact that it is not possible to cure the disease once the patient gets the diabetes disease. As the number of patients is increasing on a day to day basis, it is difficult for the doctors to perform manual detection. With the technologies like Machine Learning in hand, we can perform automative detection to some extent. Lot of research has been performed till now on this diabetes diagnosis problem. This paper discusses predictive analysis using two ensemble machine Learning Algorithms such as Random Forest and GBDT. In this paper, we have performed various Experiments on Pima Indians Diabetes Dataset which contains Diabetes patients record and results are discussed. This paper additionally discusses the importance of Interpretability of output in the healthcare domain and explains how it will help the doctors in real time if we could provide interpretability of the output along with the output of the patient record given by machine learning model.

Keywords- Diabetes, GBDT, Healthcare, Interpretability, Machine Learning,Random Forest

I. INTRODUCTION

Diabetes is a disease that occurs when the body doesn't make enough insulin. Diabetes is one of the interesting healthcare problems to work with. The reason is that, lots of people are affected with diabetes and at the same time, the early detection of this disease is must otherwise the patient can have various issues including loss of eyesight permanently. Once the patient is diagnosed with Diabetes, it can't be cured. Majorly, Diabetes patients have two types of disease such as type 1 or type 2. Body of the Type 1 diabetes patients doesn't produce enough insulin whereas the body of Type 2 diabetes patients doesn't respond to insulin when compared to a normal person. As the number of Diabetes patients are increasing rapidly, the automated Machine Learning models will definitely help the doctors in reducing their workload. In this paper, we have discussed some of the methods of doing diabetic diagnosis using Machine Learning. With the availability of the data, we can use machine learning models to predict whether the patient has diabetes or not. In this paper, we have used Pima Indians Diabetes Dataset. This dataset contains the records of 768 female diabetes patients. This dataset contains two classes such as class 0 and class 1. class 0 represents that corresponding patient don't have diabetes whereas class 1 represents the patient has diabetes.This dataset contains 8 features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome along with class label feature named 'Outcome'.

II. LITERATURE REVIEW

Paper[1] has applied the models such as SVM, Naive Bayes and Decision Tree and concluded that each of these models will work well for different scenarios. In Paper[2], comparison of models such as Support Vector Machine and Naive Bayes is performed and

these comparison are performed using the metrics such as precision, specificity, sensitivity and accuracy.In paper[3], the authors discussed about the method of using fuzzy interface system in order to perform diagnosis of diabetes. In paper[4], authors have discussed the various supervised machine learning algorithms and also discussed the strengths and weakness of algorithms. Paper[5] have performed outlier detection and applied Auto MultiLayer Perceptron to provide good accuracy. Paper[6] have discussed the sensors for diabetes diagnosis. Paper[7] discussed the methodology of applying Naive Bayes for Diabetes Diagnosis. In Paper[8], the authors have discussed the application of Deep Learning in the context of diabetes. Paper[9] discussed the problem in the form of two phases. First phase is for performing various data pre-processing techniques and in the second phase, Decision Tree model is applied and showcases the results. Paper[10] discusses the importance of data pre-processing before applying the actual data analysis and model building and it shows the comparison between accuracies of models when a model is built with data preprocessing and without data preprocessing. Paper[11] proposed a concept called Diabetes Diagnosis Expert System and discusses how it helps the diagnosis of diabetes. In Paper[12], the authors applied the ensemble boosting with perceptron algorithm and applied these datasets on three different datasets. In Paper[13], the authors have proposed the device which helps in early diagnosis of diabetes by monitoring the importance of diagnosis of diabetes. Paper[14] applied 6 different models and compared the performance metrics such as performance and accuracy and found the best model out of the 6 models in order to detect the diabetes with the help of patient record. Paper[15] discussed both the machine learning model like Support Vector Machine and Deep learning models like CNN and applied these models on Diabetes Dataset and produced the accuracy

.

III. PREPROCESSING OF THE DATA

Before looking at the importance of pre-processing in machine learning conference, let's look at the steps that's followed in this research paper to produce very good accuracy for this problem using the flowchart mentioned in Fig 1.

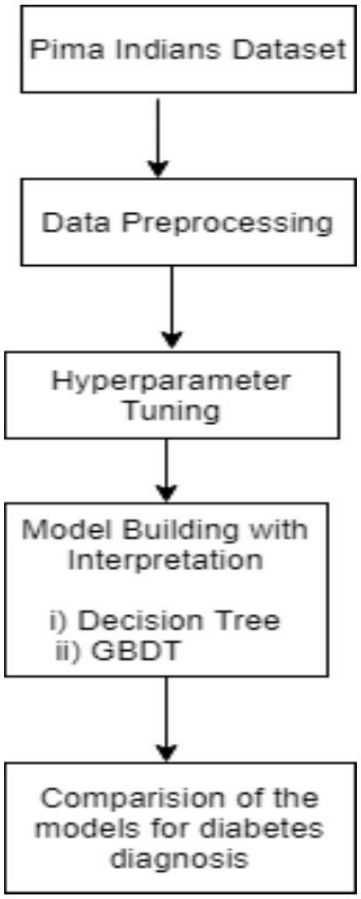


Fig 1. Steps involved in this process

As we have seen in the Introduction Section, there are 768 data points in the data set. Out of the 768 data points, 268 data points belong to Class 1 and 500 data points belong to Class 0. Generally, while handling Machine Learning Problems, there are two types of data set such as Balanced and Imbalanced Data set. Balanced Data sets are the data sets that we have almost equal number of points in class 0 and class 1.

Imbalanced Data sets are the data sets that have fair amount of difference between the number of data points between the classes. This data set is an Imbalanced Data set because Class 0 has approximately twice the amount of data points that belong to class 1. In general, if we build models on this Imbalanced data set, our models won't perform well because output will be more favoured towards class 0 as it has more data points. So, we have performed upsampling. Upsampling is the process of increasing the number of datapoints of minor classes to be equal to maximal class. In this example, After performing upsampling on class 0, we have 500 data points for class 0 and 500 data points for class1. We have converted an Imbalanced data set to Balanced Data set.

Data set contains lots of missing or erroneous values. In our dataset, we have taken our features one by one and then performed imputation. Generally, median based imputation is preferred over mean based imputation because the mean based imputation is prone to outliers. For example, Blood Pressure feature had 16 missing values when class is 0 and 19 missing values when class is 1. So, once we have retrieved those missing values for both the classes seperately, we have performed class-based median level imputation. Similarly, we have applied imputation techniques for other features.

IV. DATA SPLITTING

Data set is splitted into Train and Test data. Train data is used to train the model and find the best hyper parameters of the models that we are going to train. Test dataset should be future unseen data and this dataset is used to find the accuracy of built models such as Random Forest and GBDT on this problem. We have discussed the models used and Hyper parameter tuning in the later section of this paper. As discussed above, after performing upsampling, we have 1000 data points. We have divided this upsampled dataset into train and test data such that train data set contains 67% of datapoints and test datset contains 33% of datapoints.

V. HYPERPARAMETER TUNING

Hyper parameter Tuning is one of the most important step in Machine Learning. Hyperparameter Tuning is useful to avoid Overfitting and Underfitting while building the model. This process has to be carried out before building every model else model might be useless. Overfitting is the process of training the model so much to train data in such a way that train error is very less and test error is very high. In other words, train and test error should be as close as possible to avoid Overfitting. The machine learning model is said to be an underfitting model when both train and test error is very high. That is, if model does not work well even for training data, then the model is said to be Underfitting.

For this problem, I have built two models such as Random Forest and GBDT. Both these models are ensembling models and its underlying base models are Decision Trees.

- min samples leaf - minimum number of data sam- ples allowed in a leaf node of the tree.
- n estimators - Number of base estimators that can be trained
- min samples split - Minimum number of data sam- ples allowed in all the nodes of the tree.

In this problem, we have used GridSearchCV to find out the best value for each of the hyperparameters that is mentioned above.

A. GridSearchCV

When there is a chance for a hyperparameter to take values from the grid of values, then we can use grid search cv. When we apply grid search cv on grid of values, it will calculate auc score on each and every

value of grid and whichever value has the best auc score, it will be selected as the best hyper-parameter and model will be trained on the value.

B. K-Fold Cross validation

While doing hyperparameter tuning, we want to test the hyperparameter accuracy on some data to select the best hyper-parameter. For this purpose, we are taking some part of data as cross validation data. If we don't have cross validation data, then we might need to check the hyperparameter accuracy on test data. But the main motive of machine learning is to keep the test data as future, unseen data. So, we will take part of the data as cross validation data. Instead of allocating separate dedicated parts to cross validation data, we will divide data into train and test data. Now train data is divided into K parts and in which one part will be taken as cross validation data and other K-1 parts will be taken as train data. In this example, K value is taken as 3.

VI. BUILDING MODELS

In this paper, we have discussed two models such as Ran- dom Forest and GBDT which are used.

A. Random Forest

Random Forest is an ensemble model and it uses bagging technique. In Random Forest, the base learners are Decision Trees which means its final output depends on the output of various Decision Trees. Decision Trees are basically built by performing node split at each level of a tree. Node splits can be done based on the value of Information Gain of the feature. If there are n base learners, then there are n Decision Trees will be built where n is numerical value. Each of n base learners will be built on 'n' different data set after performing row sampling and column sampling on the main data set. The row and column sampling are performed randomly by filtering the number of rows and number of columns from the original dataset. As it is a classification problem, in Random Forest, its output is decided based on the concept of Majority Voting. In this problem, there are 2 classes such as class labels 0 and 1 which represent whether the patient has diabetes or not. So, out of 'n' Decision Trees, it will count the number of Decision Trees produces the output as 0 and let it be n1. It will also count the number of Decision Trees which produces the output as 1 and let it be n2. The output of the random Forest model will be the maximum count of n1 and n2. The following figure demonstrates how the Random Forest algorithm works.

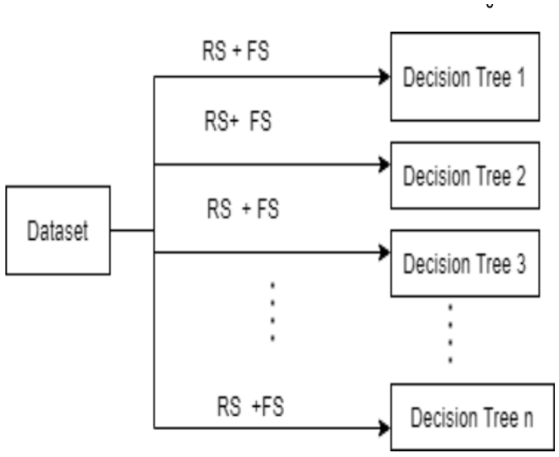


Fig 2. Working of Random Forest

n in the above diagram denotes the number of estima- tors. RS and FS denotes the Row Sampling and Feature Sampling. While building the Random Forest, Row Sampling with replacement and Feature Sampling will be done on the dataset and on top of that, we will build a decision Tree which is our base learners in the Random Forest.

B. GBDT

Gradient Boosting is a boosting technique. It concentrates on loss function and base learners. In this model, there are M base learners where M is numerical value. In this model, we will compute pseudo-residuals in every iteration and it has to optimize the loss function. Loss functions have to be differentiable. The base learners used are decision

trees. So, we can perform the node splits based on the values such as Gini Impurity or Information Gain. So, In this model also, we have to perform Hyper parameter tuning for various parameters such as min samples split, min samples leaf, n_estimators etc., After performing hyperparameter tuning, the best values that we got for hyper-paramters max depth as 8, max features as 3, min samples leaf as 4, min samples split as 5 and n estimators as 750. In this model, basically trees will be added at the end of each iteration in order to minimize the loss function.

VII. PERFORMANCE METRICS

Performance Metrics are used to evaluate the performance of a model by providing some numerical measure as an output. Using the numerical measure output, we can find how well our model works. There are various performance metrics that are proposed for classification and Regression problems. In this paper, we have discussed the performance metrics that we have used in this problem.

A. Confusion Matrix

This is a binary class classification problem. So, confusion Matrix is a 2*2 matrix. Basically this matrix displays four type of values such as True Positives, True Negatives, False Positive, False Negative.

True Positives denotes the number of data points that are actually positives and also predicted as positives. True Negatives denotes the number of data points that are actually negatives and also predicted as negatives.

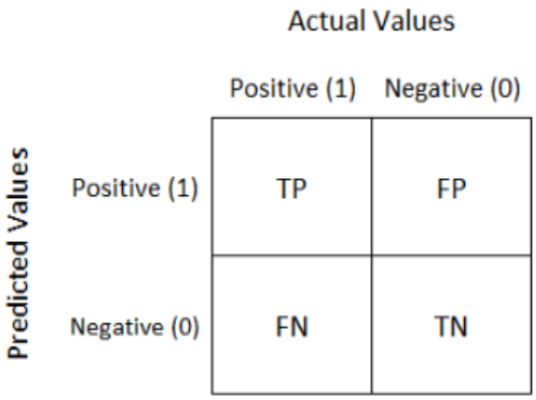


Fig 3. Confusion Matrix

False Positives denotes the number of data points that are predicted as positive but it's actually negative. False Negatives denotes the number of data points that are predicted as negative but it's actually positive.

Ideally False Positives and False Negatives should be very close to 0 for the model to be classified as good model.

B. Precision

Precision finds "Out of the points that are predicted as positives, how many percent of values are actually positives". The predicted positives can be True Positives or False Positives. Hence we can write Precision formulation as

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

C. Recall

Recall finds "Out of the points that are actually positive points, how many percent of values are predicted as positives". The actual positives can be True Positives or False Negatives.Hence we can write Recall formulation as

$$Recall = \frac{True\ Posititves}{True\ Positives + False\ Negatives} \quad (2)$$

D. F1-Score

F1-Score is another performance metrics which is used to analyse the performance of the model. The formulation of F1-score is called as

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

VIII. PROBABILISTIC INTERPRETATION

When we predict outputs using Machine Learning models, there is a chance that the patient record might be predicted wrongly. In such cases, the doctor won't be very sure because doctors have to explain to patients that they have this disease because of the particular reason. Since doctors cannot believe the output of the model, the proposed model won't be very much useful. So, I am going to give a probabilistic interpretation of the output. My model will take an image as input and will do all the internal complex steps and after that it will predict the patient record to which class out of the 2 available classes. Also my model will output the probabilistic interpretation of the class. For example, after taking the patient record, if the model predicts that the patient record belongs to class 0 and if it outputs the probability that the patient record belongs to class 0 and class 1 are 0.93 and 0.07. In this example, the model is pretty sure that it belongs to class 0. Let's take another example. if the model predicts the patient record to be class 1 and if it outputs the probability of this patient record belonging to class 0 and class 1 is 0.47 and 0.53. In this case, the doctor after looking at the probabilities can tell that model is not very sure about the output because probabilities are very close. So, after knowing this, doctors can perform manual testing to confirm whether the patient has diabetes or not. So, as we have seen in these examples, probabilities will add a certain value to the output.

IX. RESULTS

We have built two models such as Random Forest and GBDT. Out of 330 test data points, in the random forest model, the number of false negatives and false positives are 21 and 5 and the remaining 304 data points were correctly predicted by the model. The confusion matrix is mentioned in Fig 4.

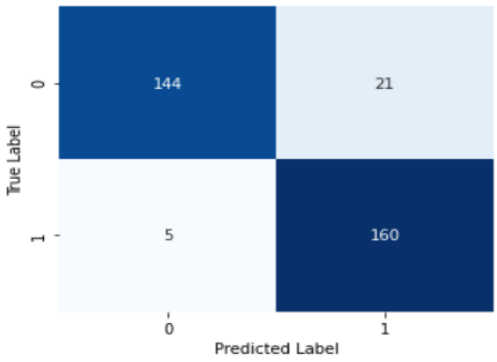


Fig 4. Confusion Matrix for Random Forest Model

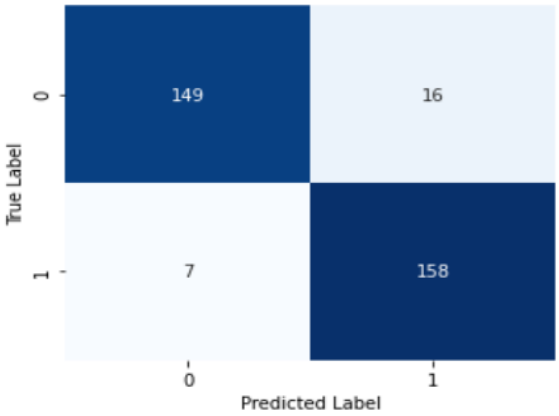


Fig 5. Confusion Matrix for GBDT Model

In the case of GBDT, Out of 330 test data points, False Negatives and False Positives are 16 and 7 and the remaining 307 data points were correctly predicted by the model. With the confusion matrix of both the models, we can say that both the models are good because the false positives and False Negatives are

less. If we want to pick one out of these two models, we can say that GBDT is slightly better than Random Forest in terms of the number of confusion matrix of both the models such as Random Forest and GBDT. The confusion matrix of GBDT is mentioned in Fig 5,

The ROC Curve of models such as Random Forest and GBDT is mentioned in Fig 6. ROC Curve again proves that how good both the models are. AUC Scores can range from 0 to 1. In this experiment, we got corresponding AUC scores of Random Forest and GBDT as 0.976 and 0.989. If the AUC Scores of the models are very close to 1, then those models are considered as very good models. With the help of AUC Scores, we can reiterate that GBDT model is slightly outperforming the Random Forest Model.

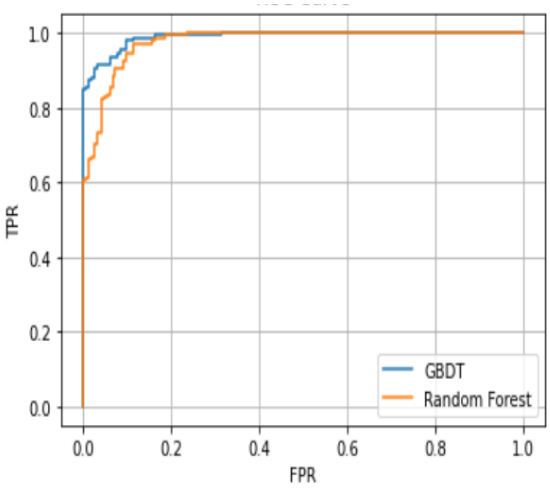


Fig 6. ROC Curve

Table 1. Results of Various Built Models

Models	Class	Precision	Recall	F1-Score
Random Forest	0	0.966	0.872	0.917
Random Forest	1	0.883	0.969	0.924
GBDT	0	0.955	0.903	0.928
GBDT	1	0.908	0.957	0.932

Also, we have produced performance metric output values such as Precision, Recall and F1-Score of two different models such as Random Forest and GBDT in the tabular form. We know that automated systems for healthcare domain should produce good performances and in our experiments, the outputs of all the performance metrics are very good because we have performed many pre-processing techniques before building the actual model on the data.

X. CONCLUSION

In this paper, we have built the two models such as Random Forest and GBDT. We were able to produce the high level accuracy because of performing various pre-processing techniques and also we have discussed the need of various processes such as Hyperparameter Tuning and K-fold Cross validation. Performance of the models is measured using the various performance metrics values such as Precision, Recall F1-Score. Also, we have highlighted the importance of interpretability of the output and explained how it will be useful to the doctors while analysing the output of machine learning models that are especially built for healthcare problems like Diabetes diagnosis. Even though, both the models such as Random Forest and GBDT produced high accuracy, we could say that the best model out of these two models for this problem is GBDT which we could say based on a number of misclassifications and also with the help of above results.

REFERENCES

[1] Priyanka Sonar, Prof. K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019).

[2] Dominikus Boli Watomakin, Andi Wahju Rahardjo Emanuel, "Comparision of Performance Support Vector Machine Algorithm and Naive Bayes for Diabetes Diagnosis", 5th International Conference on Science in Information Technology (ICSITech), 2019

[3] Nilam Chandgude, Prof. Suvarna Pawar, "Diagnosis of Diabetes using Fuzzy Interface System"
[4] MelkyRadja, Andi Wahju Rahardjo Emanuel, "Performance Evaluation of Supervised Machine Learning Algorithms using Different Data sets sizes for Diabetes Prediction", 5th International Conference on Science in Information Technology (ICSITech),2019
[5] Maham Jahangir, Hammad Afzal, Mehreen Ahmed, Khawar Khurshid, Raheel Nawaz, "An Expert System for Diabetes Prediction using Auto Tuned Multi-Layer Perceptron",Intelligent Systems Conference 2017
[6] Ke Yan, David Zhang, "A Novel Breadth Analysis for Diabetes Diagnosis"
[7] K. Lakshmi Priya, Mourya Sai Charan Reddy Kypa, Muchu- marri Madhu Sudhan Reddy, G. Ram Mohan Reddy, "A Novel Approach to predict Diabetes using Naive Bayes Classififer", Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020)
[8] Taiyu Zhu, Pau Herrero, "Deep Learning for Diabetes : A systematic Review"
[9] Asma A. AlJarullah, "Decision Tree Discovery for the Diagnosis of Type II Diaebetes", International Conferences on Innovations in Information Technology 2011.
[10] Dr. S. N. Singh, Komal Kathuria, "Diabetes Diagnosis using Differ- ent Data Pre-processing Technique", 24th International Conference on Computing Communication and Automation (ICCCA), 2018
[11] Bo Hang, "The Research and Implement of Diabetes Diagnosis Expert System", International Conference on Computer and Communication Technologies in Agriculture Engineering,2010
[12] Roxana Mirshahvalad, Nastaran Asadi Zanjani, "Diabetes Prediction using Ensemble Perceptron Algorithm", 9th International Conference on Computational Intelligence and Communication Networks, 2017.
[13] Lina Nachabe, Bachar ElHassan, Dima AlMouhammad, Marc Girod Genet, "Intelligent System for Diabetes Patients monitoring and assitance", Fourth International Conference on Advances in Biomedical Engineering (ICABME), 2017.
[14] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes using Machine Learning Algorithms in Health Care", Proceedings of the 24th International Conference on Automation Computing, Newcastle Uni- versity,2018.
[15] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed, Mirsat Yesiltepe, "A Decision Support System for Diabetes Prediction using Machine Learning and Deep Learning Techniques",

An Analysis of Imagery EEG Classification on Convolutional Neural Networks Using Alexnet Model

Ayonija Pathre[1] , Dr S.veenadhari[2]
[1,2] Bhopal, India

Abstract— ***EEG signals were used for direct communication between the human bodies and worldwide in BCI technologies with essential prospects of use in the area of cognitive science and medical care. BCI is a direct communication channel among brain signals of subject & external devices. Classification of the EEG signal is essential in creating a specific BCI system. Several Machine Learning (ML) & Deep Learning (DL) methods were utilized to classify EEG signals. Several of the studies covered time domain and Frequency domains, but many studies used time & spatial domain features concurrently to classify multiclass EEG signals. We examined AlexNet CNN to classify motor imagery signals. EEG signals were augmented to timeline images of source skull mapped images combining time and spatial domain features in one image to be analyzed simultaneously. DL technology has obtained outstanding results in the BCI method over the past few years in particular through the use of CNN frameworks in motor imagery signals recognition and evaluation. In this paper, we have established MI EEG signal spatial frequency features. Augmented images enabled the Alex Net to extract features of EEG signal activity in terms of time and location of brain activation at the same time. The outcomes show that changing EEG classification issues from time-domain signal to 2-D image classification issues by combining both time and spatial domain enhance classification accuracy for BCI systems.***

Keywords— Brain-computer interface (BCI), Motor Imagery Electroencephalogram (MI-EEG), convolutional neural networks (CNNs), Alexnet Model.

I. INTRODUCTION

The BCIs include a novel path between human brain & computer by observing the electrical signals produced by brain nervous systems. [1].

MI-EEG is a form of widespread BCI signal. Such a framework focused on MI will activate neuronal processes in most aspects relevant to movement when subjects visualize moving some parts of their body [2]. When these neuronal activities are accurately decrypted, external devices like wheelchairs and service robots can be operated by decrypted outcomes for patients with severe motor neuron disease (e.g. Parkinson's disease, poliomyelitis) [3]. EEG model classification thus plays an important part in the use of BCIs in MI.

Several machine learning algorithms have been suggested to correctly decode EEG motor imagery. Most of the methods derived from EEG studies, like frequency and study, the discriminatory time-frequency features [4], Study of complex connectivity [5], the transformation of the wavelet [6] & the filter bank common spatial pattern (FBCSP) [7]. These discriminatory characteristics were then combined in vectors of features and used for the training of classifications such as SVMs [8] & decision tree structures [9] for EEG classification tasks. Raw EEG signals can enhance EEG signal-to-noise ratio and accuracy of classification but are not needed. CNN's are multi-layer perception variants optimized to use minimum preprocessing. CNN for instance, to identify raw EEG signals directly. RNNs collaborated with CNN to improve the raw MI-functionality EEGs for portrayal and description, based on speech recognition & natural language processing. Organized a deeper neural network layer for imagining tasks from raw EEG signals or performing them. proposed enhanced CNNs with raw EEG signals to predict driver output functional ability, resulting in good outcomes. It can see the positive MI-EEG classification impact with the original signals as well. Electrical brain fluctuations can be measured by EEG. Rhythmic oscillations, reflecting the coordinated behavior of significant neuron populations, are included in EEG measures. Changes in these rhythmic variations are correlated with working conditions, including visual, neurological, motor, emotional, and other functional functions during cognitive activities. This makes the tracking of tasks through EEG tractable. CNN's should take multidimensional records

directly as an input to prevent difficult artificial mining of characteristics that can extract behaviors. If not sufficient, EEG is a recent example using ten to hundred electrodes capturing hundreds to thousands of electrode samples simultaneously when the average dataset, however in cognitive neuroscience research, includes only a few hundred to many thousand examples (i.e. tentative testing) at most while looking at discrete experimental incidents. In such datasets, thus, the initial sample-to-feature ratio is small. Based on this, the classifiers trained on EEG data sets appear to poorly generalize even on the same individual data collected at different times.

II. REVIEW OF LITERATURE

EEG signal processing for some BCI applications is one of BCI's main tasks for developing stable interfaces. Based on the nature of the brain to be investigated, various types of EEG signals are applied by the BCI systems for motion control systems.

Amin, S. U. et al. (2019) This study uses EEG imagery data to reveal the advantages of multi-level extracting & merging of fully convolutional functions from different CNN layers, that provide a summary of feedback at various levels. The stable spectral and time properties can be derived from rough EEG data in the suggested CNN models. They show that this multi-level fusion exceeds the models using features from the last layer only. Initial findings for EEG decoding & classification are higher than in the state of the art [10].

Li, Y. et al.(2019) In this study, suggest an end-to-end EEG decoding architecture, that uses raw multichannel EEGs as inputs to enhance channel mixing fully convolutional network (CP-Mixed Net) channel projection decoding accuracy with support from the increase in amplification disturbance information. The 1st block of CP-MixedNet is specifically programmed to learn primary spatial & temporal representation from EEG signals [11].

Li, D. et al.(2019) suggest densely feature fusion CNNs (DFFN) in this paper. They give two poorly complex data representation techniques, combined the morphological data of the EEG signals, and then designed and optimized the DFFN structure for this type of inputs [12].

Jeong, J.-H. et al.(2020) In this analysis, they work on classifying forearm motions using EEG signals according to elaborate rotation angles. They propose for this reason a robust classification hierarchical flow model CNN (HF-CNN). They assess the developed framework using a public dataset, & their experimental dataset (BNCI Horizon 2020) [13].

Abibullaev, B., et al. (2020) first consider a range of hyperparameters limited by their computing capability in this paper. Then they demonstrate that a comprehensive search in that small CNN space results in a precise classification of sensorimotor rhythms that occur thru MI research. [14].

Zhang, j., et al. (2017) this document suggested a system for CNN for study EEG signals, produced from MI tasks on the left & right. EEG signal is transmitted via STFT into time-frequency images and instead fed to the network input for the classification process. [15].

Kim, J., et al. (2020) suggest in this paper a new approach for the classification of CNN-dependent motor images (MI) using a new input medium. Input EEG signal shall be used to derive characteristics of MI Continuous Wavelet Transform (CWT) [16].

W. Ko et al. (2018) present a better structure for EEG motor imaging classification in a DNN. Contrary to the existing DNNs in the literature, the network architecture helps one to study the weights of the network underlying motor-image-induced EEG signals from a neurophysiological point of view. They performed tests on the BCI Competition IV-IIa dataset to verify the feasibility of the suggested approach in comparison with competing approaches on a k-value scale. For research design, the activation patterns estimated from the network weights have also been visually inspected. [17].

C. Park *et al.*(2018) The proposed network is intended to be optimized for multichannel EEG signals & uses 1D & 2D fully convolutional layers to recognize a spatiotemporal connection, an epileptic seizure detection feature. 1D fully convolutional layer takes into account the time dynamics of EEG signal in every channel as well as the spatial relations among EEG channels in 2D convolution layer. They are using CHB-MIT EEG Scalp data & SNUH-HYU EEG database to build datasets for train & test sets: The Seoul National University Hospital and Boston Children's Hospital's long-standing EEG tracking recording. EEG segments with different durations are being trained & evaluated. They also study the effects of artifact disposal on seizure identification by using the EEG Signals through a low pass filter. By

SNUH-HYU EEG dataset, their model achieves 90.5% of prevention accuracy [18].

M. Yanagimoto & C. Sugimoto(2016) DEAP has been used in the research as the publicly available dataset for EEG-based emotion analysis. Following the 11-fold cross-validation method, the CNN and a standard model are analyzed. Raw EEG data from 16 electrodes were applied as input data without specific analyses. The models identify and recognize EEG signals based on the 'positive' or 'negative' emotional conditions induced by viewing music videos. The findings suggest the greater the accuracy of CNN's the more training data are (by over 20 percent). It also means that the improved training data may not have EEG data about the same person as the test data so that emotions are correctly recognized by CNN. The findings suggest that the interpersonal variation in the EEG properties is not only important but also commonality [19].

H. Yang et al.(2015) Studies usage of multiclass MI-EEG signals by CNN. ACSP features focused on matrices of pair-sided projection that cover different frequency ranges are created. They suggest an additional FCMS system by restricting dependency between frequency bands. BCI Competition IV Data Set IIa tests with 9 subjects are performed. For FCMS & for all function maps the average cross-validation performance is 68.45% & 69.27% simultaneously, slightly higher than the random map selection (4.5% and 5.34%) & greater (1.44% & 2.12%) than the CSP filter bank (FBCSP) respectively. The outcomes indicate that CNNs are in a position to learn discriminating, detailed structural characteristics for EEG without using handcrafted features.

III. PROPOSED WORK

A. Problems Identification

The data was not cleaned properly. The model was using all the generated images instead of only taking the proper ones. The model was too simple to handle heavy data. There was very dissimilarity in all the 5 subjects, result lack in the accuracy except & subject AA. We aim to analyze the features of various subjects and use a customized CNN structure & parameters to detect MI EEG patterns instead of using a uniform CNN model parameters framework. Combining deep learning and subject transfer learning is a vital guide for our strategy to tackle the issue of limited training and strengthen the fixture of the deep CNN model. In other words, the temporal domain characteristics have not been thoroughly examined. A fixed-time window can hardly capture discrimination on any subject.

B. Proposed Methodology

Before applying CNN, We clean the data in Matlab then in python, after this we translate cleaned EEG data from image representation. We utilize filtering of bandpass to provide 8–30 Hz signal to original EEG signal, frequency bands of 8–14 Hz ((μ rhythm) & 18–26 Hz (β rhythm).

Then we evaluate the signal energy as described for EEG signal in frequency subband of every spatial electrode:

$$p = \log(\text{var}(x)), \quad (1)$$

Here var (x) is the variance of EEG signal sequence x. -us, every model may be expressed as a size matrix 21x10 and, within a certain subband of some EEG electrode, every element of matrix reflect EEG signal energy. We normalize the EEG energy as follows for each subject in the dataset:

$$P^{r}_{i,j} = \frac{P}{\delta} \quad (2)$$

Here P is the sample energy matrix, r is the index for every sample, mi,j is average sample energy at that location, & Śi,j is the corresponding standard deviation. the initial EEG signal can then be converted into image representation after the above stages, in that EEG electrodes are dispersed along a vertical axis & subband frequency is dispersed along a horizontal axis. After generating images, error ones need to be removed. Because of the dissimilarity in the five subject data, the accuracy of the subject less, but after properly clearing the data that too is fixed. It will be noted that this dataset contains 118 electrodes. We have extracted EEG signals by 49 networks for subsequent analysis to decrease the workload of subsequent implementation & exclude the impact of redundant channels, in compliance with literary recommendation [18]. After the above processing, matrix 49 to 3500, with 3500 sample points & 49 electrodes, can be expressed for each data set. After generating the images, the number of images was high, so we used a transfer learning AlexNet model.

C. Alexnet Model

AlexNet is a term of a NN that has influenced the machine learning area, especially in the application of DL FOR perception of the machine. The ImageNet LSVRC-2012 competition was well-known to have won by a wide margin (15.3% VS 26.2% (second place) for error). Yann LeCun et al. had a very similar design to LeNet, with more filters per layer & stacked convolutional layer. It comprised 11×11, 5×5,3×3, convolutions, max pooling, dropout, data increase, ReLU activations, dynamic SGD. After each convolutional and fully connected layer, it added ReLU activations. The Nvidia Geforce GTX 580 GPU has practiced AlexNet for 6 days concurrently, that's why its network is split into two pipelines.

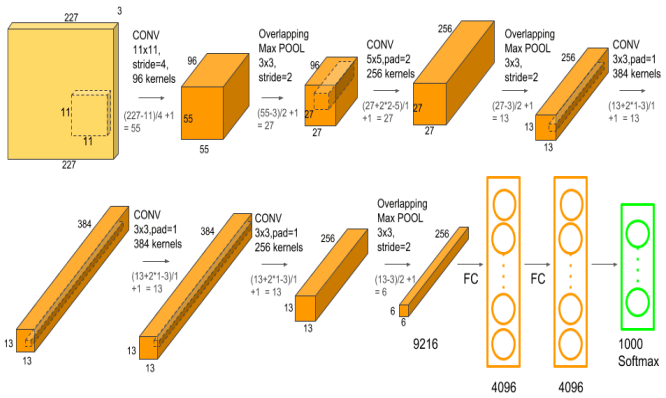


Figure 1. Structural model for the recognition of the EEG patterns in motor imagery

IV. PROPOSED ALGORITHM

Steps Used in Proposed Work

Step 1: Select the particular 49 channels that have the main signal frequency in MatLab and also apply a bandpass filter on the selected data to remove the unwanted frequency.

Step 2: Creating a CSV file from the precise data (Matlab file) to overcome the difficulty for the next work.

Step 3: Split the CSV into 250 different CSVs to a better process. Thus, it makes 28 images per CSV and interpolating and images also.

Step 4: Saving only the proper images and rejecting the error images.

Step 5: Pass the images into the model for training applying 10-Fold Validation for improving accuracy and validation accuracy.

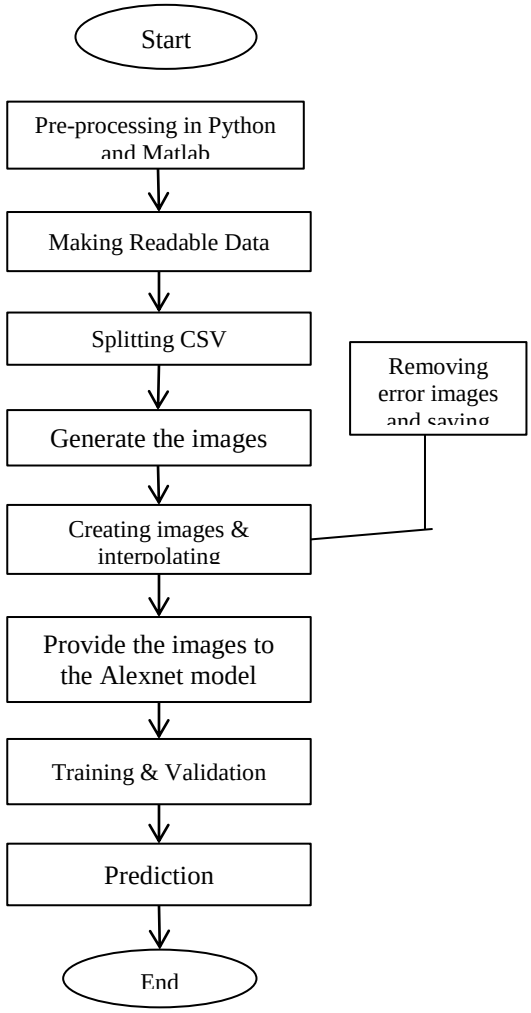


Figure 2. Flow diagram of proposed System

The above flow chart defines the structure of the proposed scheme. Firstly we have preprocessed the dataset using preprocessing techniques then Creating images & interpolating and Provide the images to the Alexnet model. After training, we have predicted results. Combining deep learning & subject transfer learning is a vital recommendation for our project to tackle the issue of lesser training & strengthen the fixture of

the deep CNN model. In other words, the features on the temporal domain were not thoroughly investigated. A fixed-time window can hardly capture discriminatory features on all subjects.

V. RESULTS AND DISCUSSION

The proposed CNN (Alexnet) model's software and hardware platforms are Intel (R) Core (TM)i7-10700k 5.10 GHz CPU, 32 GB RAM DDR4, Jupyter notebook, Python 3.8, and TensorFlow 2.4.3 (GPU), GPU NVIDIA RTX 2060 SUPER cuDNN and CUDA ToolKit 11.2v. The Relu function is used to apply non-linearity instead of Tanh. It speeds up 6 times with the same accuracy. To handle overfitting use dropout rather than regularization. Even so, with the dropout rate of 0.5, the training time is doubled. Pooling overlaps to reduce network size. This decreases 0.4% & 0.3% respectively of top 1 & top 5 error rates; respectively.

MODEL DETAILS: -ALEXNET

Table 1: Hyperparameters Details of Alexnet Model

Hyperparameter	
Parameters	Values
Padding	Valid
Optimizers	Adam
Activation functions	Relu
Regularization	Dropout (40%)
Cost functions	Categorical cross-entropy
Batch size	128
Classes	2
Subjects	5
Kernel	11*11
Input shape	227*227*3
Filter shape	96
Pooling	Maxpooling2D, size(2*2)
Strides	(2*2)

The above table defines the hyperparameters details of the alexnet model. We are using the AlexNet CNN architecture from scratch under this section. With the Keras Sequential API, our models are stacked against one another and enable us to build consecutive NN layers.

Table 1.Comparision Results of base and propose approach based on different subjects in term of Accuracy

Subject	aa(%)	al(%)	av(%)	aw(%)	ay(%)	Mean(%)
Base Algorithm	98.40	88.88	89.12	89.00	78.48	88.77
Proposed Algorithm	99.86	99.76	99.98	99.79	99.87	99.85

The above table shows the results in terms of the Accuracy of each subject. Five EEG signals (aa, al, av, aw, & ay) with 118 electrode amplifiers are obtained from each subject. From this table, we can observe that base algorithm and propose a method in which the proposed approach succeeds or performs higher than other subject competitors.

Table 2. Comparison Results of base and propose approach based on different subjects in term of Mean

Subject	aa(%)	al(%)	av(%)	aw(%)	ay(%)	Mean(%)
Base Algorithm	5.06	24.90	24.77	24.83	43.00	24.51
Proposed Algorithm	0.42	0.10	0.0006	0.59	0.41	0.30

The above table shows the results in terms of the mean of each subject. Five EEG signals (aa, al, av, aw, & ay) with 118 electrode amplifiers are obtained from each subject. From this table, we can observe that base algorithm and propose a method in which the proposed approach succeeds or performs higher than other subject competitors.

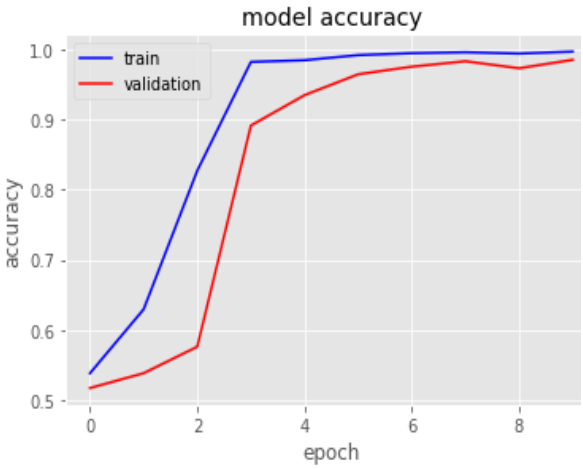


Figure 2: Classification accuracy of CNN (Alexnet) training set & validation set

The above graph demonstrates Model Train and Validation Accuracy in terms of Epoch. In figure 2, the classification accuracy of the training set is seen, during iterations, to reach the highest value & stable, whereas classification accuracy of the validation set was maintained at the greatest value, & the model can be shown to have the best training effect following iterations, the trained technique is recognized as optimal classification model of various subjects.

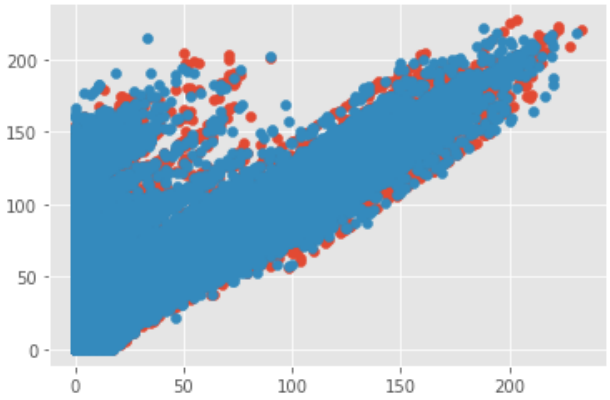


Figure 3: Scattered graph using Propose Method

The above graph shows a Scattered graph using Propose Method. Similarity measuring approaches are broadly accepted in a broad range of visualization applications using CNN(Alexnet).

AA

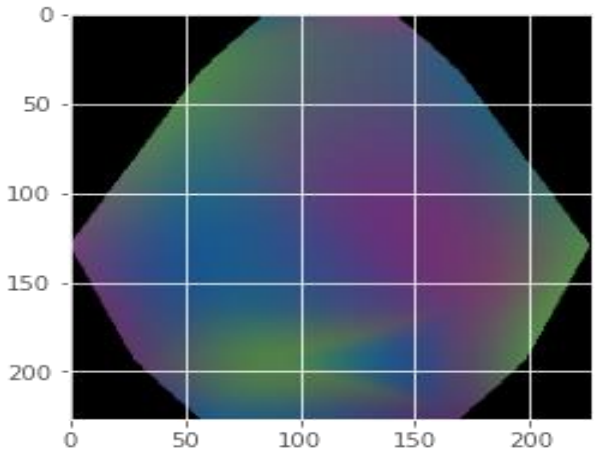


Figure 4: The Image Plot for Subject of AA

The above figure shows the Image Plot for the Subject of AA. The architectures & hyperparameters of CNN(Alexnet) models are similar for an AA subject in such a dataset. Even so, there are common features with different users. The uniform parameter uniform can then lead to a suboptimal solution for the AA subject.

AL

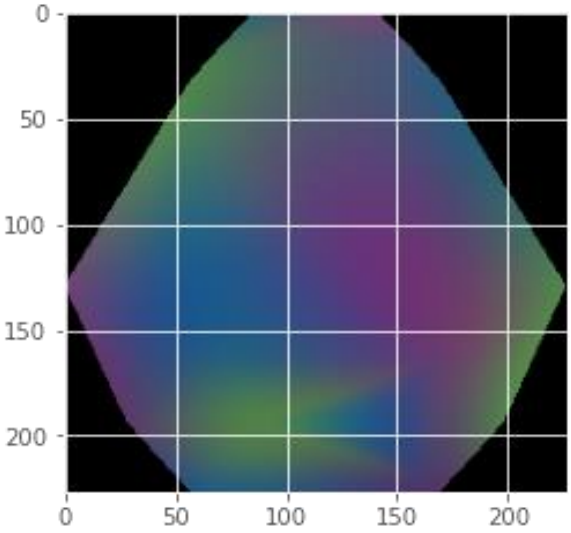


Figure 5: The Image Plot for Subject of AL

The above figure shows the Image Plot for the Subject of AL. The architectures & hyperparameters of CNN(Alexnet) models are similar for an AL subject in such a dataset. Even so, there are common features with different users. The uniform parameter uniform can then lead to a suboptimal solution for the AA subject.

AV

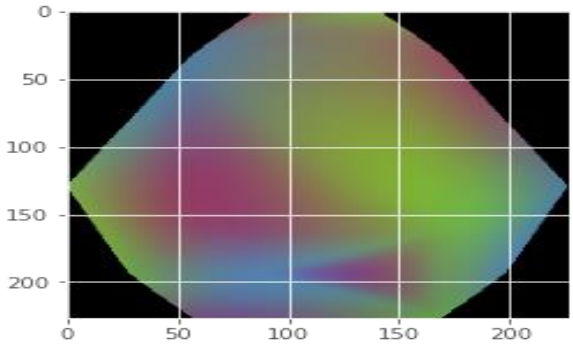


Figure 6: The Image Plot for Subject of AL

The above figure shows the Image Plot for the Subject of AA. The architectures & hyperparameters of CNN(Alexnet) models are similar for an AA topic in such a dataset. Even so, there are common features with different users. The uniform parameter setting for the AL subject can also lead to a suboptimal solution.

AW

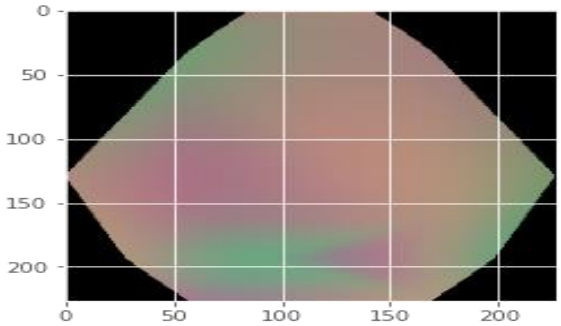


Figure 7: The Image Plot for Subject of AW

The above figure shows the Image Plot for the Subject of AW. For AW subjects in certain datasets, structures & hyperparameters of CNN(Alexnet) models are the same. Even so, there are common features with different apps. The uniform parameter settings can then lead to a suboptimal solution for the AW subject.

AY

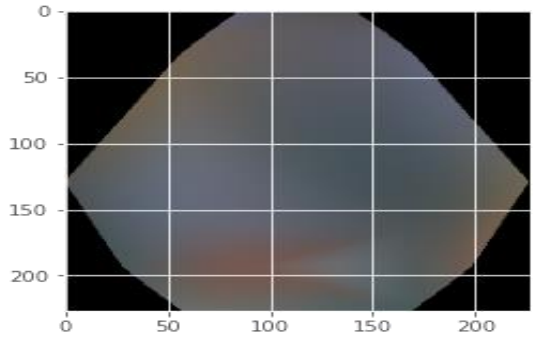


Figure 8: The Image Plot for Subject of AY

The above figure shows the Image Plot for the Subject of AY. The architectures and hyperparameters of CNN(Alexnet) models are similar to the AY subject of a certain dataset. Even so, there are common features with different users. Uniform parameter setting can thus result in a suboptimal solution for the AY subject.

VI. CONCLUSION

Several EEG signals, like sleep stages, MI, mental workload & emotion recognition, were successfully implemented by DL. In the processing of these complex signals, the application of DL into EEG has been seen to be promising, given its order to access good feature representations from raw information through successive non-linear changes. But because of its relatively smaller size, DL is essentially restricted over EEG datasets. In exchange, DA improves the training data available to make it easier to use more complex DL models. It may also minimize overfitting & enhance classification accuracy & stability. A deep learning method to identify focal EEG signals is suggested in this context. In this paper, we suggest a DL algorithm for the EEG pattern classification of limb MI. For EEG classification & space-frequency features, a multi-layer Alexnet model is developed. EEG signals are analyzed by developed convolution layer parameters in the neural network. In a test, public BCI competition III datasets.

VII. FUTURE ENHANCEMENT

On this basis, the Deep CNN (Alex net) suggested for MI EEG Feature Learning is to be developed further over the whole spatial-temporal-frequency domains. Moreover, the proposed model's training time is simply inadequate. Subsequently, we want to observe that for specific BCI systems the proposed method is of major interest. Subsequent

3 things are important of our future investigations to overcome these drawbacks and enhance the performance efficiency of our proposed CNN model (Alexnet). (1) We design to investigate the features of particular objects and add user-specific CNN structure & parameters to MI EEG pattern recognition rather than using the common CNN model parameter system. (2) Combining the DL & subject to transfer learning is a significant research path in our plan to deal with the issue of smaller training models & improve the complexity of the deep CNN model. (3) The deep CNN model (Alexnet) developed in this work automatically shows spatial frequency features of MI EEG.).

REFERENCES

1. L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou, "Common Bayesian network for classification of EEG-based multiclass motor imagery BCI," IEEE Trans. Syst., Man, Cybern., Syst., vol. 46, no. 6, pp. 843–854, Jun. 2016.
2. Y. Zhang, C. S. Nam, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Temporally constrained sparse group spatial patterns for motor imagery BCI," IEEE Trans. Cybern., to be published.
3. K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 4, pp. 392–401, Apr. 2017.
4. Y. Wang, K. C. Veluvolu, and M. Lee, "Time-frequency analysis of band-limited EEG with BMFLC and Kalman filter for BCI applications," J. Neuroengineering Rehabil., vol. 10, no. 1, pp. 1–16, 2013.
5. Y. Li, M.-Y. Lei, W. Cui, Y. Guo, and H.-L. Wei, "A parametric time frequency-conditional granger causality method using ultra-regularized orthogonal least squares and multiwavelets for dynamic connectivity analysis in EEGs," IEEE Trans. Biomed. Eng., to be published.
6. [Y. Li, W. G. Cui, M. L. Luo, K. Li, and L. Wang, "High-resolution time-frequency representation of EEG data using multi-scale wavelets," Int. J. Syst. Sci., vol. 48, no. 12, pp. 2658–2668, May 2017.
7. W. Wu, Z. Chen, X. Gao, Y. Li, E. N. Brown, and S. Gao, "Probabilistic common spatial patterns for multichannel EEG analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 639–653, Mar. 2015.
8. L. Wang et al., "Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis," Entropy, vol. 19, no. 6, p. 222, 2017.
9. S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Brief Bioinform., vol. 18, no. 5, pp. 851–869, 2016.
10. Amin, S. U., Alsulaiman, M., Muhammad, G., Bencherif, M. A., & Hossain, M. S. (2019). Multilevel Weighted Feature Fusion Using Convolutional Neural Networks for EEG Motor Imagery Classification. IEEE Access, 1–1. doi:10.1109/access.2019.2895688
11. Li, Y., Zhang, X.-R., Zhang, B., Lei, M.-Y., Cui, W.-G., & Guo, Y.-Z. (2019). A Channel-Projection Mixed-Scale Convolutional Neural Network for Motor Imagery EEG Decoding. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 1–1. doi:10.1109/tnsre.2019.2915621
12. Li, D., Wang, J., Xu, J., & Fang, X. (2019). Densely Feature Fusion Based on Convolutional Neural Networks for Motor Imagery EEG Classification. IEEE Access, 7, 132720–132730. doi:10.1109/access.2019.2941867
13. Jeong, J.-H., Lee, B.-H., Lee, D.-H., Yun, Y.-D., & Lee, S.-W. (2020). EEG Classification of Forearm Movement Imagery Using a Hierarchical Flow Convolutional Neural Network. IEEE Access, 8, 66941–66950. doi:10.1109/access.2020.2983182
14. Abibullaev, B., Dolzhikova, I., & Zollanvari, A. (2020). A Brute-force CNN Model Selection for Accurate Classification of Sensorimotor Rhythms in BCIs. IEEE Access, 1–1. doi:10.1109/access.2020.2997681
15. Zhang, J., Yan, C., & Gong, X. (2017). Deep convolutional neural network for decoding motor imagery based brain computer interface. 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). doi:10.1109/icspcc.2017.8242581
16. Kim, J., Park, Y., & Chung, W. (2020). Transform based feature construction utilizing magnitude and phase for convolutional neural network in EEG signal classification. 2020 8th International Winter Conference on Brain-Computer Interface (BCI). doi:10.1109/bci48061.2020.9061635
17. W. Ko, J. Yoon, E. Kang, E. Jun, J. Choi and H. Suk, "Deep recurrent spatio-temporal neural network for motor imagery based BCI," 2018 6th International Conference on Brain-Computer Interface (BCI), Gangwon, Korea (South), 2018, pp. 1-3, doi: 10.1109/IWW-BCI.2018.8311535.
18. C. Park et al., "Epileptic seizure detection for multi-channel EEG with deep convolutional neural network," 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, 2018, pp. 1-5, doi: 10.23919/ELINFOCOM.2018.8330671.
19. M. Yanagimoto and C. Sugimoto, "Recognition of persisting emotional valence from EEG using convolutional neural networks," 2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA), Hiroshima, Japan, 2016, pp. 27-32, doi: 10.1109/IWCIA.2016.7805744.

Artificial intelligence and Deep learning towards Health Sector - COVID-19

BOLLU SIVA KESHAVA RAO [#1], CHEEPU BALAKRISHNA [#2]

#1 Department of CSE, Sree Vahini Institute of Science & Technology, Tiruvuru, A.P., India.

#2 Department of CSE, Sai Spurthi Institute of Technology, Sathupally, India.

#1 bsivakeshava143@gmail.com #2 balu.cheepu@gmail.com

Abstract: Conceptual COVID-19 flare-up has placed the entire world in a remarkable difficult circumstance bringing life all throughout the planet to a startling end and asserting great many lives. Because of COVID-19's open out in 212 nations along with domains and expanding quantities of contaminated cases and losses of life scaling to 45, 515,851 and 451,223 (as of June 1 2020), it stays a genuine danger to the general wellbeing framework. This thesis delivers a reaction to battle the infection via Artificial Intelligence. Several Deep Learning strategies turned out to be delineated towards achievement of this objective, which includes Generative Adversarial Networks, Outrageous Learning Machine, in addition to Long/Short Term Memory. This outlines an incorporated bioinformatics perspective inside of which various parts regarding data from a trajectory of organized and indeterminate information sources are assembled in the direction for framing the easy-to-use stages for doctors and analysts .By these Aritificial intelligence stages the primary benefit is configured as to enhance the interaction of finding along with researching towards the treatment for COVID-19.The latest linked distributions as well as clinical results were researched with the motivation behind picking data sources and focuses of the organization that could work with coming to a solid Artificial Neural Organization premised device for disputes related to COVID-19. Besides, there exists various certain voices for every stage, comprising of different types of the information, for example, clinical information as well as clinical imaging i.e. pictorial representation which enhances the presentation of the presented perspectives regarding the best reactions enclosed by reasonable exertions.

I. Introduction:

The epic Coronavirus assigned Severe Acute Respiratory Syndrome -CoV-2 showed up in December 2019 to start a epidemic of respiratory disease known as COVID-19 which substantiated its own volition as an interesting disease which can arise in different structures as well as levels in respect to seriousness going out of gentle into serious with danger of various organ disappointment also, demise. Out of gentle, static-restricting respiratory parcel sickness into extreme reformist pneumonia along with multiorgan disappointment, as well as demise too. As the pandemic is getting advanced along with increasing number of cases and victims are running into serious respiratory disappointment along with cardiovascular intricacies, many more motivations are to be concerned about the outcomes of the disease. Deciding proper ways to deal with arrive at answers for the Coronavirus related issues have gotten a lot of consideration. Nonetheless, another immense issue that specialists what's more, chiefs need to manage is the steadily expanding up to date, known as large information, that summons them during the time spent fighting against the infection. This justifies how what's more, how much Artificial Intelligence is significant for creating as well as overhauling medical care frameworks on worldwide basis. Man-made intelligence has been as of late

pulled in expanding research endeavors towards settling the unpredictable issues in various elds, including designing, medication, economy, and brain science. Subsequently, a basic circumstance like this requires assembly and saving clinical, calculated and HR and

Artificial Intelligence cannot just work with that yet can save the time too i.e. even if we can save one hour of time can lead ending up with saving lives in various areas where Covid has been asserting lives.In addition to the new ubiquity of Man-made intelligence exertion of clinical settings, it has a vital role in diminishing the quantity of unwanted erasures as well as increasing the profitability and coherence in examinations although huge examples were included, and upgraded exactness in forecast and determination are planned. Using enormous information can likewise work with viral movement displaying concentrates in anyone of various country. Examinations results empower various policy makers wellbeing to set up their country anti towards episode of infection as well as settle on very much educated choices. All things considered, while treatment systems, emergency the board, streamlining and improvement finding techniques, for example, clinical imaging and picture preparing strategies could take benefit from Artificial Intelligence which is conceivably fit for making a difference clinical technique, it has not been alluringly utilized what's more, very much suitable to serve medical care frameworks in their struggles at odds with COVID-19. For example, one region that can take exceptional benefit of Artificial Intelligence's valuable information is picture based clinical conclusion through which quick and precise determination of Coronavirus can happen and save lives. Appropriating Simulated intelligence methods to manage COVID-19 related matters can build a bridge between Artificial Intelligence -based techniques and clinical methodologies furthermore, medicines. Man-made intelligence experts' utilization of Artificial Intelligence stages can support in developing associations amongst different boundaries and advance the cycles towards acquiring ideal outcomes.

In this thesis, our group depends on the detections of latest examination zeroing in on COVID-19 and its different difficulties to sum up and propose an assortment of techniques applicable yet not restricted to high-chance gatherings, the study of disease transmission, radiology and so on. As the paper unfurls, it investigates also, talks about the possibilities of Artificial Intelligence ways to deal with survive Coronavirus related difficulties in area 2. Area 3 of the paper incorporates a show of Artificial Neural Networks-based techniques that can be utilized for huge information examination. Segment 4 shows the conversation, and Section 5 provides the Conclusion.

II. Man-made brainpower Along with COVID-19

The current area centers around the presentation of a few material Artificial Intelligence - based systems that can uphold existing norm strategies for managing COVID-19 in medical care frameworks all throughout the planet. Determined to frontal area the improved adequacy of these systems and methods, their development has been educated by and dependent on the most recent Artificial Intelligence -related distributed clinical upgrades just as the most recent reports on COVID-
19. Consequently, this segment presents thoughts that can improve and accelerate Artificial Neural Networks-based strategies acquiring interaction to upgrade treatment techniques and wellbeing the executives just as acknowledgment and determination. Nonetheless, the ideal viability of Artificial Intelligence devices amidst COVID-19 epidemic relies upon the degree of variant human info along with joint effort in various jobs people play. The information on capacities furthermore, constraints of Artificial Intelligence, nonetheless, stays with information researchers who assume a significant part just on the grounds that they are the ones who code Artificial Intelligence frameworks.

Various strides use Artificial Intelligence - based strategies utilized to beat COVID-19 difficulties are introduced in the flow diagram appeared in Fig.1. The initial step is the arrangement of information which were essential for information prospecting during information understanding,

information readiness and large information. The information being talked about here comprise of clinical data, i.e., clinical reports, records, pictures and other different structures of data that can be changed into information that can be perceived by a machine. Goals of information understanding incorporate recognizes information that ascribes and distinguishing primary attributes like information volume and the all-out number of factors to sum up the information. Prior to preparing

and investigation comes information readiness that is the cycle through which crude information are refined and changed over. At the end of the day, it is a cycle where information is reformatted, remedied and joined to advanced information. Gathering, dissecting and utilizing the information like shopper, patient, physical, and clinical information. Closes in huge information. It is at this stage that human mediation, as a piece of Artificial Intelligence strategies, happens and specialists examine and investigate the information to remove the information with nest designs, examples and highlights.

People's commitment at this stage is significant in light of the fact that their insight and possibilities are not accessible to a Machine Learning arrangement that not at all like people can manage colossal informational collections a long way past the degree that people could deal with or notice in a concurrent way. Besides, Deep Learning strategies could be utilized in situations where tremendous or complex information handling challenge Machine Learning or customary methods of information handling. Deep Learning techniques, in Fig. 1 illustrates, are not reliant upon human intercession. Being a part of machine learning, Deep Learning comprises of various levels of calculations which gives an alternate understanding of the information it benefits from

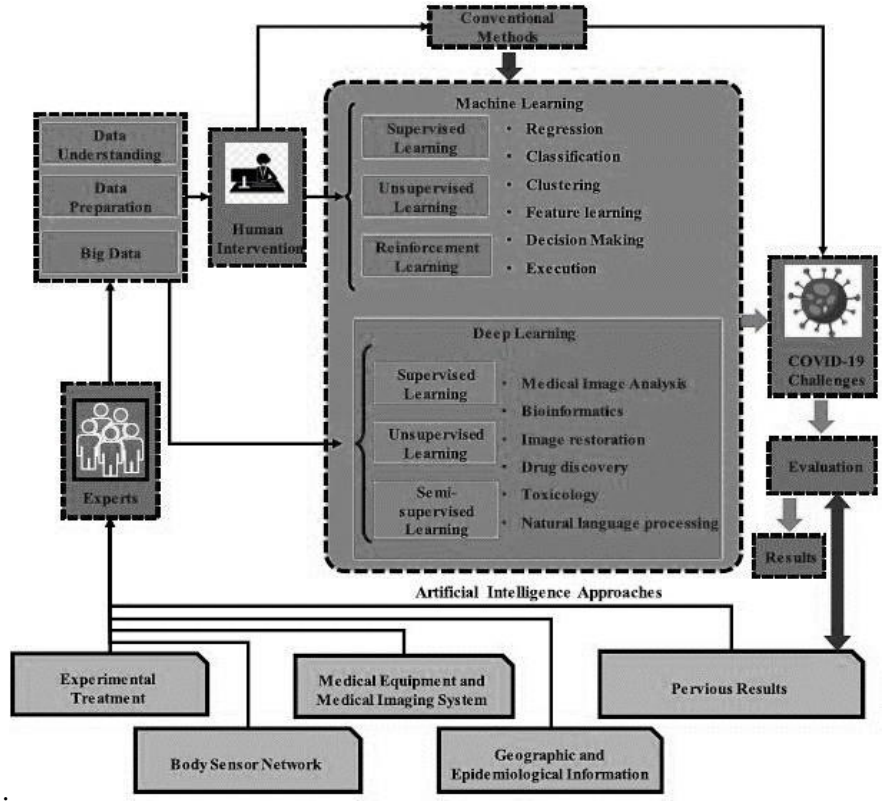


FIG 1. Artificial Intelligence-based methods to deal with COVID-19.

In any case, Deep Learning is principally unique in relation to Machine Learning since it presents information in the framework in an alternate way. Though Deep Learning organizations execute by different levels of Artificial Neural Networks, Machine Learning calculations are typically reliant upon organized information. Not at all like managed realizing which is the errand of learning a capacity planning a contribution to a yield based on model information yield sets, solo learning is checked by least

human oversight and could be depicted as such an Artificial Intelligence looking for undetected examples in an informational index where no earlier marks exist. In traditional medication, then again called as allopathic medication, biomedicine, standard medication, customary medication and Western medication, clinical specialists and other expert medical services suppliers like attendants, advisors, and drug specialists use medications, medical procedure or radiation to treat sicknesses and kill indications.

Artificial intelligence can be applied on COVID-19; nonetheless, we target finding the most ideal arrangements COVID-19 related matters have kept greatest difficulties in front of medical care frameworks. In like manner, these arrangements have been ordered into 3 sections, including high- hazard gatherings, flare-up what's more, control, perceiving and analysis.

Fig. 2 shows different uses of Artificial Neural Networks in finding and following the side effects in all the

5 layers. Albeit cycle has been explicitly intended for Coronavirus linked issues which has capacity to use it in various clinical investigations. The information layer,as one of the underlying layer is identified with the information base which is intended for data set acceptance. A rapid different thing is utilized towards combining these things with principal PC (s). While the data set worker is inexactly coupled through the organization, the information base machine is firmly brought together to the principal Central Processing Unit. Exploiting of a decent number of chips with data set programming data set machines can send colossal bundles of information to the centralized computer. The following layer, determination layer, is planned by a keen Artificial Neural Networks – based selector and has the assignment of embracing the most ideal imaging methods in view of past encounters of the framework. In the event that doctors affirm the choices made by this layer, the suggested strategies in the next layer i.e. third layer which takes the necessary pictures. Thus, one or

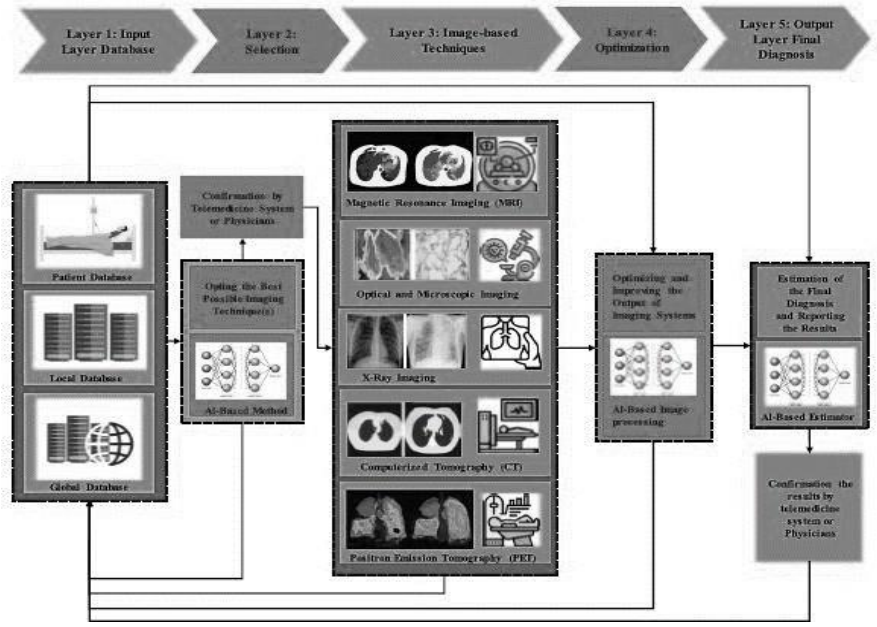


FIG 2. Machine Learning Methods for medical imaging approaches.

a few imaging procedures might be proposed by the recently acquired outcomes. For every quiet, Magnetic Reverberation Imaging, Computed Tomography Scan,

positron discharge tomography, Optical. The traditional optical magnifying lens has come to be the prevailing instrument in neurotic assessments. positron discharge

tomography sweep that, now and again, recognize infection ahead it can be identified by various other imaging tests, is an important imaging test which decide the organ capacities as well as degree and nature of tissues and organs. In the positron discharge tomography sweep, a radioactive drug is used to research this usefulness.

The next layer committed to the enhancement and advancement of the pictures. To understand a grouping organization that works with separation among COVID-19 along with Influenza-A viral pneumonia, a Deep Learning innovation was utilized for network structure, and the old style ResNet was utilized to separate highlights. The next layer is held for extreme finding dependent on the framework's saved data and is a layer where learning calculations ought to be finished by an Artificial Neural Networks technique. Deep Learning advances, for example, a convolutional neural organization, should be the correct alternative for accomplishing these objectives. The explanation is that this sort of network is conspicuously equipped for disorderly displaying and has broad use in clinical picture handling as well as finding measure.

III. FINEST POSSIBLE PLATFORM TO SPEEDUP:

Traditional ways of Discovering answers for hazard bunches who are affected by COVID-19 is the fundamental worry of the current paper. Since coming to the most ideal outcomes is the fundamental goal, we will attempt to show courses through which Artificial Neural Networks -based strategies could be utilized as corresponding to the regular ones. As recommended, it is important to keep victims included COVID-19 vault that features clinical factors and cardiovascular complexities since it works with the association of the example of cardiovascular complexities, assists creating a danger model for cardiovascular intricacies, and helps with recognition or potentially forecast of the reaction to various sorts of treatment methodologies.

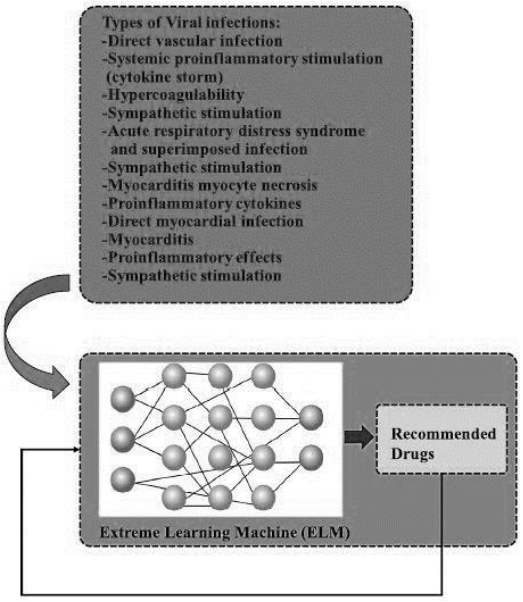


FIG 3. ELM model.

Fig. 3 represents an Extreme Learning Machine model that depends on worked concentrates in to anticipate reasonable medications dependent on people who are included with such cardiovascular entanglements. Extreme Learning Machine Artificial Neural Networks can utilize past models, utilize them for wanted yields. It implies, the directed model occurs by the usage of genuine information. Hence, inspecting different types of viral contamination for

past cases, Extreme Learning Machine can recommend the most ideal medications for heart entanglements.

In correlation with customary feedforward network learning calculations like back- engendering calculation, grasping speed in Extreme Learning Machine is an incredibly quicker and acquires better speculation execution. By the by, on numerous events, regular tuning-based calculations require a part of secret neuron than Extreme Learning Machine. There are a few different investigations that have recently examined Extreme Learning Machine with fixed network designs. Following the preparing measure, new information can be anticipated through a tensor then again verification method. As proposed, the Coronavirus causes vascular inflammation, myocarditis, and cardiovascular arrhythmias. Most recommended method relies upon the information that represents to foresee the ways that cardiovascular framework which was influenced by Coronavirus. Consequently, the proposed methodology is equipped for decreasing conceivable cardiovascular intricacies danger. Additionally, it understands the forecast about reaction towards various treatment methodologies on the grounds that it can anticipate the example of cardiovascular entanglements. Consequently, thinking about their properties and different benefits Extreme Learning Machines are suggested for various such issues.

Additional confusion that COVID-19 reasons in the earlier is cardiovascular breakdown, which requires cardiovascular breakdown experts be careful as well as plan an organized way to deal with these sorts of patients and remember them for creating calculations for the care of these patients in beginning phases until general COVID-19 assessments alternatively clinical preliminaries of antiviruses are set up, along more profound comprehension of final phases of the infection is acknowledged. In ordinate utilization of fluid also, drugs, for example, NSAIDs that may transform the equilibrium of salt along with water in older victims, ought to be kept away from. Reference and biomarkers, particularly in high-hazard old patients with basic primary heart sickness ought to be utilized with care and alert. In that capacity, defining and overseeing progressed cardiovascular breakdown in the period of hyper inflammation are significant conflicts for various heart experts.

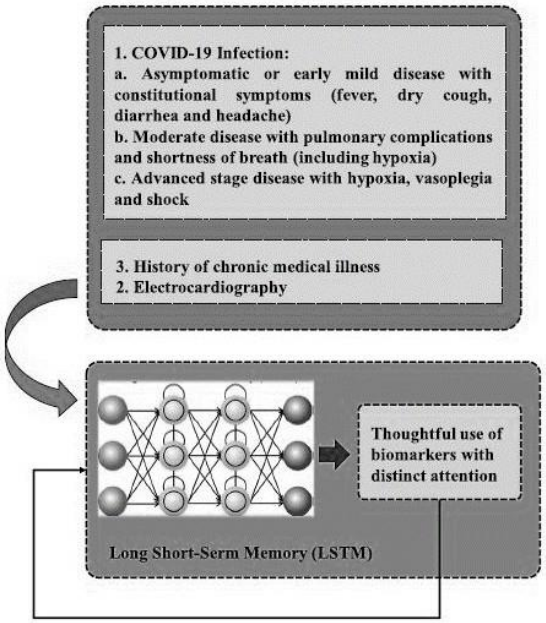


FIG 4. Classifying through Long/Short Term Memory Artificial Neural Networks.

Fig. 4 presents a model that utilizations Long/Short Term Memory network set forward in. This method depends in

action fittingly thought about contributions to foresee the best treatment as definitely as could be expected. Being equipped for keeping up long memory, Long/Short Term Memory networks are profitable for understanding arrangements with higher-term examples of obscure span.

Notwithstanding electrocardiography along with history of constant clinical sickness which can support the model preparing measure Gentle, modest and progressed period of COVID-19 contamination can be inspected as sources of info. Utilizing multiplicative entryways that manage persistent mistake ow through the inward conditions of 'memory cells' which are unique units. Long/Short Term Memory neural organizations tackle the issue of vanishing angle in Recurrent Neural Networks Hochreiter and Schmidhuber who were the first to present this were ensued by other people who enlightened and promoted this. Long/Short Term Memory Neural Networks has been famous and progressively utilized in robot control, speed acknowledgment, penmanship acknowledgment, human activity acknowledgment, and so on in the course of recent years, and it has worked completely in discourse acknowledgment and text classification

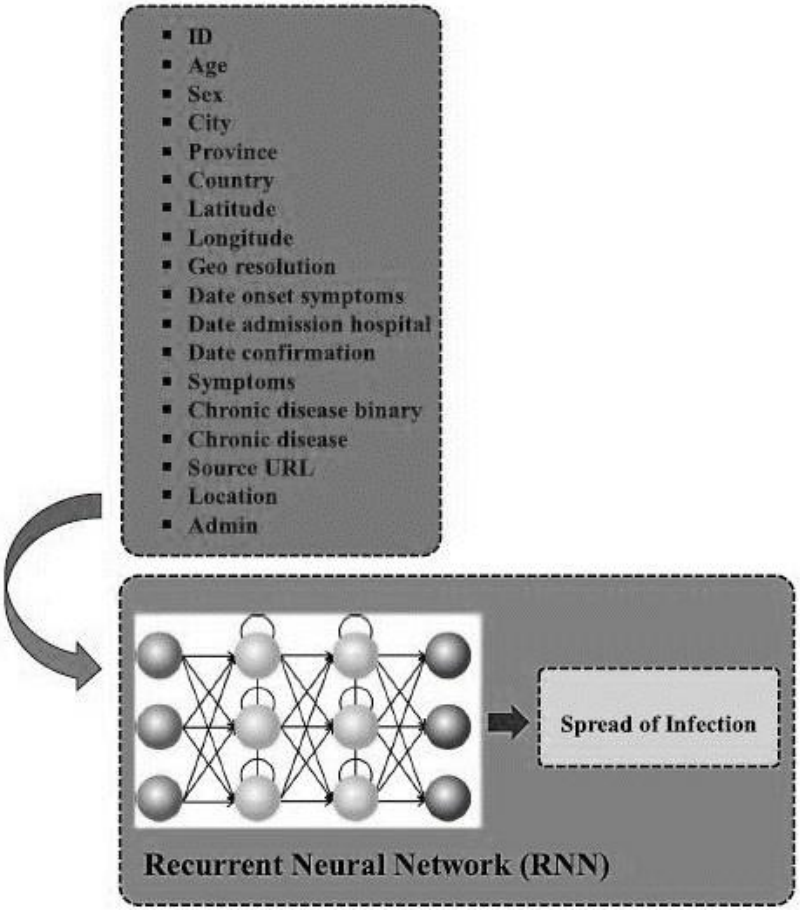


FIG 5. Prediction by Recurrent Neural Network.

Artificial Neural Networks – based strategies are elective methods of foreseeing Coronavirus episode. As indicated by, a portrayal of the elds in the information base is appeared. The ongoing epidemiological information in, have been assembled in a coordinated way to anticipate the disease increase. Fig. 5 outlines on how

Deep Learning perspective, which was controlled by Recurrent Neural Networks can foresee these growing of disease related with COVID-19 through clinical and topographical huge information. Contingent upon topographical what's more, clinical information, varieties of Recurrent Neural Networks can be used to anticipate the spread of contamination. Nonetheless, it appears to be that the awesome design to understand the expectations are Long/Short Term Memory network, Gated Recurrent Unit Recurrent Neural Networks, and Clockwork Recurrent Neural Networks. The Recurrent Neural Networks, as then again called Auto Affiliated or Feedback Network, falls in the class of Artificial Neural Networks in which a coordinated cycle is made through associations between units. Being a broadly liked Deep Learning family, Recurrent Neural Networks have prevailing to introduce promising outcomes in a ton of Artificial Intelligence and PC vision undertakings. One significant undertaking to utilize this model, nonetheless, is the quantification of subjective data sources like country and area. Refreshing the model is conceivable in light of the ongoing information by Recurrent Neural Networks with ongoing learning ability. Use of the proposed Artificial Neural Networks model gives the chance for proposing epidemiological model of infection in several fields. The main objective of proposed structure is to enhance the precision and and accelerate the acknowledgment and diversification of the conflicts brought about by the infection by using Deep Learning – dependent strategies.

In spite of the fact that screening, determination, and progress appraisal of Coronavirus have been adequately performed through dependence on radiological assessments, including Computed Tomography Scan and advanced photography, there has been very little related knowledge can support radiologists along with technologists to manage COVID-19 victims. In territories smash by the pandimecy, negative RT-PCR however certain Computed Tomography Scan highlights are important indications of COVID-19 and features the significance for fast discovery of contamination that gives the local area along with clinicians a superior opportunity to level out the viral widespread. While radiological assessments, for example, processed tomography Computed Tomography Scan has been exhibited as successful techniques for screening and analysis, there is proof that impressive quantities of radiologists and technologists have been tainted while serving COVID-19 patients. Lung Computed Tomography Scan sweeps of pneumonia caused by COVID-19 picture reciprocal, subpleural, ground glass opacities with air bronchograms, ill-defined edges, and a slight prevalence justified lower flap. The picture classification model works with separation of various contaminations regarding their appearance what's more, structure. To get familiar with the estimated area data of the fix on the pneumonic picture, the model utilizes relative distance-from-edge as an additional weight. Albeit the lumbering errand of acquiring an enormous number of clinical pictures for Artificial Intelligence applications is conceivable, specific and expert perusing of symptomatic imaging report that could adeptly address setting, sentence structure, structure, and specific phrasings expected to decipher the imaging is exclusively left with radiologists who could remove indicative data from pictures and make them accessible as organized names for the utilization of the Artificial Intelligence model preparing.

The initial instance talked about the cycle of representation also, identification of new human Coronavirus. In any case, a latest report has mentioned that underlying spread of human respiratory discharges onto human aviation route epithelial

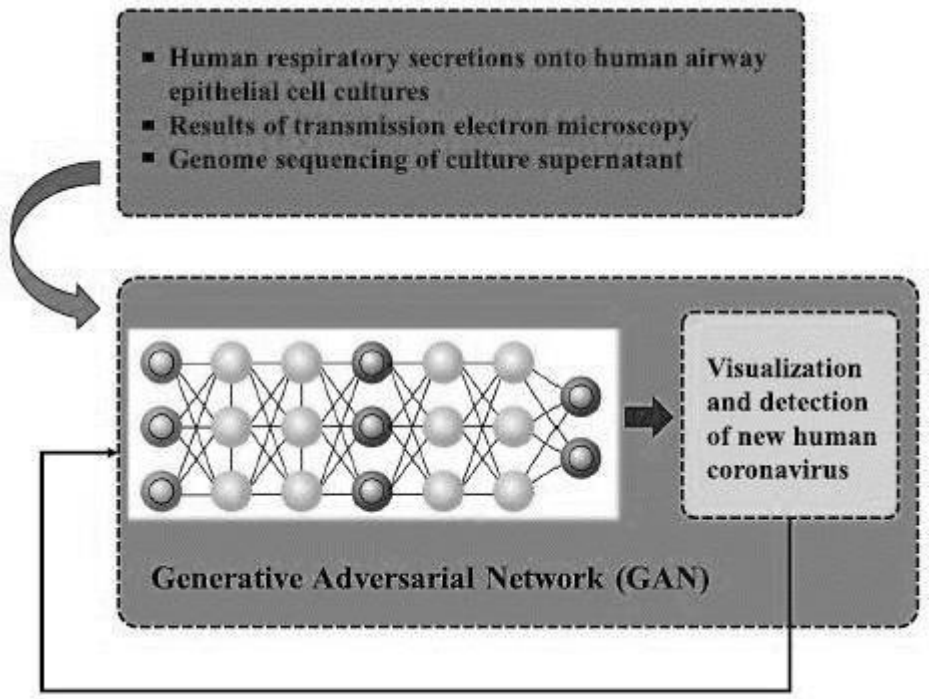


FIG 6. Application of Generative Adversarial Network for visualization and detection.

cell societies alongside to send electron microscopy also, entire genome ordering of culture supernatant must be used to picture and distinguish new human COVID that has the chance of resting anonymous by customary perspectives. As exhibits contamination brought about by Coronavirus can harm human aviation route epithelial cells. It is likewise showed that picturing and distinguishing new human Covid should be possible through utilizing the impacts of the human respiratory emissions on the human aviation route alongside the consequences of transmission electron microscopy, and genome sequencing of culture the supernatant. Fig. 6 portrays the proffered neural organization model as well as the Generative Adversarial Network. To investigate electron microscopy pictures, highlight extraction procedure can be embraced. Generative Adversarial Networks are a unique sort of neural organization model in which two organizations are prepared at a similar time while one is centered around producing pictures, and the other performs separating. Generative Adversarial Networks can address these issues through powerful displaying of the inactive dispersion of the preparation information. Generative Adversarial Networks have effectively been applied to picture to-picture interpretation, division furthermore, numerous other subfields of clinical picture registering. In light of its handiness in checking space shift, and adequacy in producing new picture tests, the antagonistic preparing plan has as of late pulled in a ton of consideration. This model has accomplished cutting edge execution in a ton of undertakings, in particular content to-picture union, super-resolution, and picture to-picture interpretation. Those are identified with producing pictures. Another issue to be settled by Artificial Neural Networks-based methodologies is assessing the degree of cardiovascular inclusion. Reference contends that COVID-19 infection is a significant reason for myocarditis. Contemplated cardiovascular association as a COVID-19 contamination able of making serious intense respiratory condition finish up that the acknowledgment of intense myocarditis' relationship with Coronavirus by the scientific local area

will be helpful in checking influenced victims in an exacting way and could help general wellbeing administrators in going to a superior agreement of such perilous difficulties. Appropriately, depending on the findings and proposition of, a Long/Short Term Memory organization is advanced for the assessment of COVID-19 related heart inclusion. Taking into account that in feedforward neural networks signals are permitted to simply move one way going ahead from the contribution to the yield. we like Recurrent Neural Networks on the grounds that they permit signs to travel the two different ways presenting circles in the organization permitting inner

associations among covered up units. As opposed to feedforward neural network, a Recurrent Neural Networks measures the consecutive contributions through a repetitive secret state where actuation at each progression is reliant upon the past one; thus, the capacity of the organization to show dynamic worldly conduct. Fig. 7 records the highlights from Tesla cardiovascular attractive reverberation imaging that can be used for model preparing

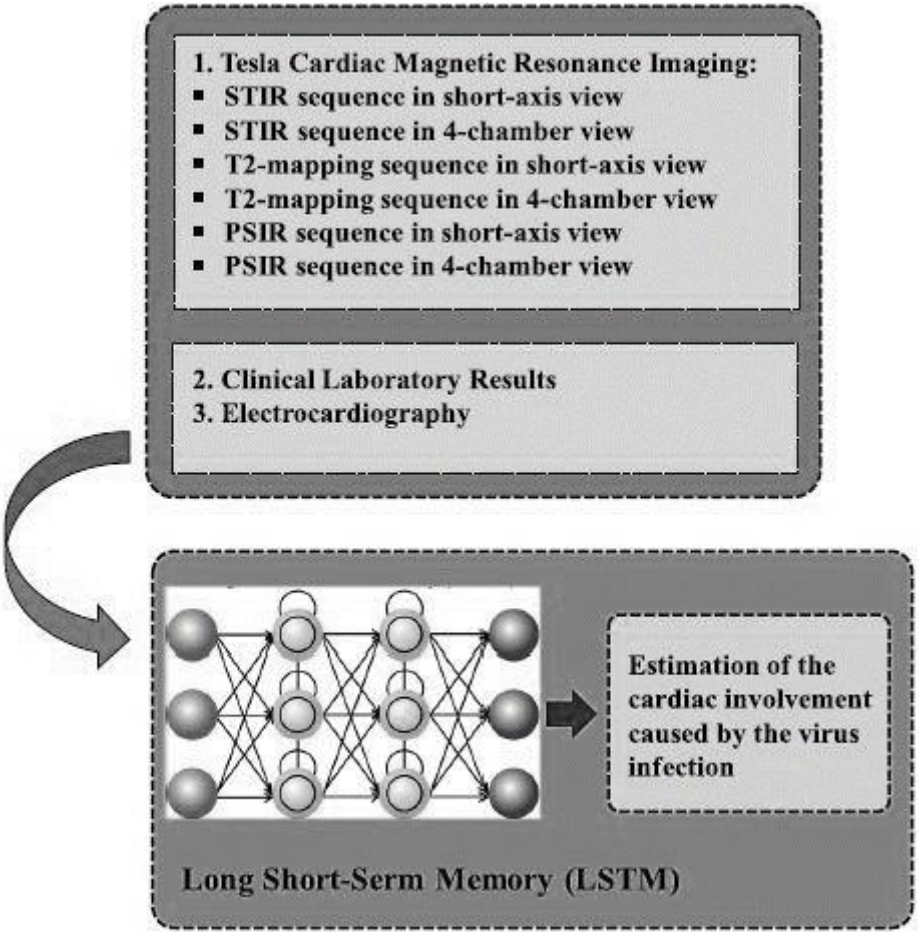


FIG 7. Estimation of cardiac involvement caused by the virus infection.

Additionally, an Artificial Intelligence-based model exists to assess the conduct of Remdesivir just as some clinical boundaries. As noted in, recommend contrasted with patients with high popular replication also, fundamental infection spread, victims with a aggressive load decline in the superior respiratory lot might require different helpful methodologies relying upon viral energy

observing, might be required. Nonetheless, because of the modest number of patients for this situation information investigation be done circumspectly. Examined clinical and natural information of five COVID-19 victims. To gauge the conduct of Remdesivir, virostatic prescription for post-disease treatment for Coronavirus, in medicines of the patients just as emergency clinic stay, Intensive Care Unit stays and indicative time, clinical information of these victims including persistent clinical sickness or history of ongoing clinical ailment, indications, age and sexual orientation and tests results on clinic affirmation are used. By and by, the quantities of patients were not sufficient for Extreme Learning Machine organization. Extreme Learning Machine is by and large a most un-square based learning calculation for "summed up" single secret layer feedforward networks, is valuable for assessing relapse issue or grouping undertakings.

While input loads (connecting the information layer to the covered-up layer) and covered up predispositions in Extreme Learning Machine are chosen in a discretionary way, the yield loads (connecting the secret layer to the yield layer) are resolved in an insightful way also, using Moore- Penrose summed up backwards. In this manner, Extreme Learning Machine method can be utilized to prepare the proposed model. The proposed referenced Extreme Learning Machine method is portrayed in Fig. 8.

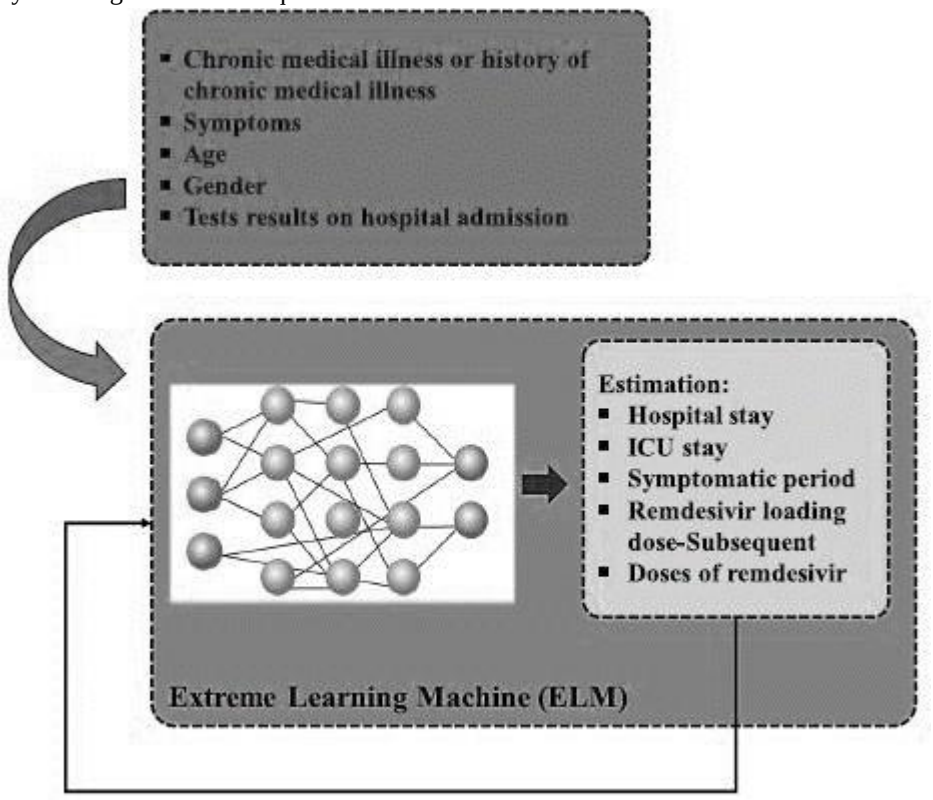


FIG 8. Estimation of Remdesivir drug behavior on the patient's treatments using Extreme Learning Machine framework. Proof for gastrointestinal contamination of Severe Acute Respiratory Syndrome -CoV-2 and the chance of fecal-oral transmission course is given. The spread of the infection from tainted to uninfected cells makes viral- explicit target cells or on the other hand organs

We suggest a model prepared by Generative Adversarial Network for viral gastrointestinal disease likelihood assessment in the last piece of the conclusion

the primary job major part in deciding the viral dissemination courses. The initial step of viral disease is the receptor-interceded viral passage into the getting cell. Furthermore, ACE2, which is infrequently communicated in the esophageal epithelium, bounteously conveyed in cilia of glandular epithelia.

Notwithstanding, even after regrettable change of the viral Ribo Nucleic Acid in respiratory lot more than 21% of Severe Acute Respiratory Syndrome -CoV-2 victims present positive viral Ribo Nucleic Acid in defecation which means that viral gastrointestinal contamination and the chance of fecal-oral dissemination that can in any case happen after viral leeway in the respiratory lot.

In this way, routine rRT-PCR testing for Severe Acute Respiratory Syndrome -CoV-2 from defecation is enthusiastically suggested on account of Severe Acute Respiratory Syndrome -CoV-2 patients. Moreover, on the off chance that rRT-PCR testing illustrated positive defecation test, dissemination-premised precautionary measures for hospitalized Severe Acute Respiratory Syndrome
-CoV-2 patients ought to be set up. Examines the gastrointestinal disease caused by COVID. Coronavirus related gastrointestinal disease in this examination is confirmed by an assortment of
pictures of histological furthermore, immunofluorescent staining of rectum, duodenum, stomach and throat. All these images were the yield of laser checking confocal microscopy.

A Generative Adversarial Network organization to anticipate viral gastrointestinal contamination likelihood should be possible through the extraction of the element from these pictures to help victims during the time spent their treatment. Fig. 9 represents the method of a choice to proceed alternatively on the other hand cease transmission-based insurances for hospitalized Severe Acute Respiratory Syndrome-CoV-2 patients is subject to rRT-PCR testing for Severe Acute Respiratory Syndrome-CoV-2. The Generative Adversarial Networks generative cycle, which projects a standard circulation to difficult high-dimensional genuine information dissemination withstands higher that to when thought about to most biased undertakings (e.g., diversification and grouping). Notwithstanding picture age assignments, Generative Adversarial Networks have been acquainted with assignments, like video age, visual following, area adaption, hashing coding, and include learning.

Generative Adversarial Networks are of two distinct clients in clinical imaging. With their attention on the generative angle, they work with investigation and disclosure of the fundamental construction of preparing information and help with figuring out how to create new pictures. With their emphasis on the biased viewpoint, where the discriminator D can be viewed as a learned earlier for typical pictures that they can be utilized as a regularize or indicator when introduced with unusual pictures.

Earlier, disseminating of COVID-19 victims is by all accounts viably overseen through Deep Learning models exhibited in this research that can be a viably supportive beneficial analytic technique for clinical specialists in close contact with patients.

IV. Conversation:

Zeroing the chance of the Artificial Neural Networks application for investigating COVID-19-related disease issues, for example, high-hazard patients, control of the episode, perceiving and radiology, we utilized Recurrent Neural Networks, Long / Short Term Memory, Generative Adversarial Networks and Extreme Learning Machine to propose a few Artificial Intelligence-based techniques. Progressed Artificial Intelligence calculations can coordinate and examine enormous scope information associated with COVID-19 patients to work with a more profound arrangement of viral spread example, to enhance the speed and exactness of determination, grow new, powerful helpful methodologies, furthermore, even distinguish people who, contingent upon their hereditary furthermore, physiological highlights, are generally vulnerable to the infection. In spite of much acclaim that such information has gotten in light of its job in improving efficiency, profitability and measures in various areas, it has been censured for its modest number of clients who gather, store, deal with the information and approach them.

Notwithstanding, as Heyman looks after

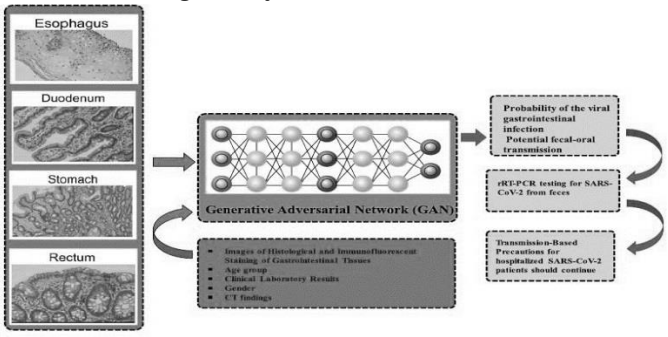


FIGURE 9. procedure for viral gastrointestinal infection probability estimation.

Artificial Intelligence process it conceivable to inform when wrong things(errors) are going on, or then again moves are made in regards to COVID-19 since it screens and gathers information coming from web-based media, newsfeeds, what's more, aircraft tagging frameworks.

An enormous majority of different data coming from the latest headway and distributions in the pertinent case can be covered by the proposed techniques. By the by, while an assortment of sources of info exist, clinical information stays as the input shared by practically every one of the strategies. With regards to bunches that are noticed as very high danger, outlining COVID-19 victim's clinical qualities all through parturiency or infection time is especially significant. The method suggested is mostly centered around victims with cardiovascular breakdown around the hyper- inflammation period of the ailment and people for whom precise chronicles of the clinical factors as well as cardiovascular difficulties endure. These thoughts, nonetheless, yield themselves to be reached out to other high-hazard victims since there are similitudes among the construction of Machine Learning or then again Deep Learning procedures in complex information assessment and forecast. Extreme Learning Machine calculation is proposed for foreseeing reasonable medications since it is exceptionally worthwhile in critical thinking, however the inclination-premised learning calculations like back-proliferation were acceptable to feedforward neural organizations plus multiple secret layers. On account of single hidden layer feedforward networks, the current type of the Extreme Learning Machine calculation is legitimate.

We suggested a Long / Short Term Memory prepared model for the next case, which is the diversification of the bestest treatment technique. Long / Short Term Memory networks appear to be acceptable alternatives for classification, cycle, and forecast by time arrangement information on the grounds that slacks of obscure length may occur between major occasions in a period arrangement. Detonating and evaporating slope issues that may show up in preparing customary Recurrent Neural Networks can be successfully managed by Long / Short Term Memory's which is end up being a working device in situations where successions exist in light of the fact that in such cases the importance of term is reliant upon the past term. Foreseeing the study of disease transmission and episode by Artificial Intelligence was another subject examined in this paper. The model that we proposed here depends on Recurrent Neural Networks with a far-reaching set of sources of info that can be finished by the information base introduced in. Recurrent Neural Networks can be viewed as a class of Artificial Neural Networks is in which a coordinated chart along a worldly arrangement is framed by associations between hubs making the show of worldly dynamic conduct conceivable. Recurrent Neural Networks forecast of things to come is manipulated by their recollecting of previous occasions previously understanding the fundamental connection of information while attempting to arrive at secret levels Recurrent Neural Networks work in a circle. Inspecting that Imaging workflows can move propels in machine learning strategies equipped for helping radiologists who look for an examination of difficult imaging and text information, we depicted models that can break down clinical imaging working with the consummation of a cycle that perceives COVID-19-related diseases. Concerning the plague territory, we clarified that Coronavirus could be the situation when negative RT-PCR and positive Computed Tomography Scan are set up. Thinking about the significance of quick identification of the viral contamination that can significantly assist with more compelling influence of the viral spread, clinical and cultural ramifications of this contention can't be overlooked. Radiological assessments, for example, processed tomography Computed Tomography Scan, were talked about as compelling techniques to screen and analyze contamination. It was likewise referenced that an impressive number of radiologists as well as technologists

have been tainted in the interaction of analyzing COVID-19 victims. COVID19 pneumonia is generally observed on lung Computed Tomography Scan checks as respective, subpleural, ground glass opacities with air bronchograms, ill- defined edges, and a slight power justified lower flap.

In the initial instance of perceiving, representation and discovery of new human Coronavirus by the Generative Adversarial Networks, the contributions of the proposed network comprise of the impacts of the human respiratory emissions on the human aviation route, consequences of transference electron microscopy, and genome sequencing of culture supernatant.

This is imperative to underscore that COVID-19 is famous towards fast crumbling the capacity of respiratory framework that regularly occurs in the next seven day stretch of infection; consequently, the present wellbeing of the victim(patients) can't be an ensure that they were not hit by the infection as well as wellbeing network guidance must be viewed appropriately. This features the significance of using a successful Artificial Neural Networks-based strategy in picturing and recognizing new human Coronavirus. While preparing a set is given to this method, it figures out how to create new information while it utilizes similar insights as the preparation set. It is likewise exhibited that Generative Adversarial Networks are valuable for semi-administered learning, completely directed learning and support learning. While Generative Adversarial Networks figure out how to plan from an idle space to an information appropriation of interest, the contrsting network separates competitors that the generator makes from the genuine information appropriation. The second instance of perceiving incorporates a Long / Short Term Memory approach that gauges cardiovascular contribution brought about by the infection contamination. Long / Short Term Memory units come with various models. One normal engineering comprises of one cell along with three "controllers" or data flow gates which resides inside the Long / Short Term Memory unit: an information entryway, a yield door and a disregard entryway. Monitoring the conditions between the components in the information arrangement is finished by the cell. While controlling the degree of another worth ow into the cell which is an obligation of information entryway., degree to such a worth stays in the cell is constrained by neglect entryway, along with the degree to which the worth in the cell is utilized to figure the yield initiation of the Long / Short Term Memory unit is constrained by the yield entryway. It is suggested, nonetheless, that in the third instance of perceiving, Extreme Learning Machine network does the assessment of Remdesivir's conduct in quiet's medicines, emergency clinic stay, Intensive Care Unit stay what's more, suggestive period. By and large, the discovery character of neural organizations and Extreme Learning Machine network are significant worries that put engineers careful with regards to application in hazardous robotization assignments.

Be that as it may, there are an assortment of procedures accessible, such as lessening the reliance on arbitrary contribution, to perspect this specific conflict. In the endmost instance of perceiving a Generative Adversarial Networks estimates the likelihood of viral gastrointestinal contamination. Up-and-comer age is finished by the generative organization, and assessment of the applicant is finished by the biased organization. The challenge works in wording of information appropriations. While the generative organization figures out how to map from a dormant space to an information appropriation of interest, the biased organization separates applicants that the generator makes from the genuine information conveyance and henceforth the advantages of utilizing this trademark to an inexact viral gastrointestinal contamination.

Albeit suggested strategies haven’t been used however to assess their adequacy, there are numerous clinical reports and substantial wellsprings of data demonstrated the efficiency furthermore, exactness of these strategies in a wide range of sorts of comparative illnesses. The main outcome here is to sum up such solid strategies’ dependent on the attributes of Coronavirus.

V. Conclusion:

The presented calculated designs and stages in the analysis eld of Artificial Intelligence -rooted methods, that were reasonable towards managing COVID-19 cases, have been concentrated in this thesis. Various methods were created, fusing Coronavirus' symptomatic frameworks, like Recurrent Neural Networks, Long / Short Term Memory,

Generative Adversarial Networks, and Extreme Learning Machine. The geological issues, high-hazard individuals, furthermore, perceiving and radiology were the primary issues with Coronavirus and have been examined and talked about in this work. Likewise, we showed a component for choosing the fitting models of assessment and forecast of wanted boundaries utilizing various clinical and non-clinical datasets. Considering these stages helps Artificial Intelligence specialists to examine enormous datasets and help doctors train machines, set calculations or on the other hand improve the dissected information for managing the infection with much more speed as well as accuracy. We examined that they are attractive in view of their capacity for making a work station while Artificial Intelligence specialists and doctors could work next to each other. In any case, it ought to be noted while Artificial Intelligence speeds up the strategies to prevail Coronavirus, genuine analyses ought to happen on the grounds that a full comprehension of benefits and constraints of Artificial Intelligence-based strategies for COVID-19 is yet to be accomplished, and novel outlook must be set up for issues of this level of intricacy. Prevailing in the battle anti towards COVID-19 within its inevitable downfall is profoundly reliant upon developing an arms stockpile of stages, strategies, outlook, and instruments that merge to accomplish the looked-for objectives and acknowledge saving abundant lives.

REFERENCES:

1. Artificial Intelligence – A Modern Approach (3rd Edition)**,** By – Stuart Russell and Peter Norvig
2. Artificial Intelligence Engines: A Tutorial Introduction to the Mathematics of Deep Learning, James V Stone
3. Artificial Intelligence and Machine Learning, by Chandra S.S.V.
4. DATA MINING Concepts and Techniques, 3rd Edition, by Jiawei Han, Micheline Kamber, Jian Pei.
5. DEEP MEDICINE How Artificial Intelligence can make Healthcare Human Again, by ERIC TOPOL.
6. Machine Learning and AI for Healthcare Big Data for improved Health Outcomes, by Arjun Panesar.
7. Demystifying big data and machine learning for healthcare P Natarajan, JC Frenzel, DH Smaltz - 2017
8. Big Data Analytics in Healthcare: A Critical Analysis, by Dibya Jyothi Bora.
9. Big Data Analytics in Bioinformatics and Healthcare, by Baoying Wang (Waynesburg University, USA), Ruowang Li (Pennsylvania State University, USA) and William Perrizo (North Dakota State University, USA)
10. Bioinformatics, System Biology and Big Data Analysis, by Dr. Neetu Jabalia. N. Jaya Lakshmi.
11. Biomedical Informatics Computer Applications in Health Care and Biomedicine, by Shortlife, Edward H., Cimino, James J. (Eds.).
12. Artificial Intelligence and Deep Learning, by Dr. Jagreet Kaur, Navadeep Singh Gill.
13. Neural Networks, by Satheesh Kumar.
14. Deep Learning Techniques for Biomedical and Health Informatics, by Sujata Das – Biswa Raja Acharya – Mamata Mittal – Ajith Abraham – Arpad Kelemen.
15. Diagnostic Strategies for COVID – 19 and other Coronoviruses, by Chandra, Pranjal, Roy, Sharmili (Eds.)

Enhanced Movie Sentiment Classification Model Using Machine Learning Algorithm

Devendra Singh Rathore
Rabindranath Tagore University
Raisen,M.P. India
devendrarathore2007@yahoo.com

Dr. Pratima Gautam
Rabindranath Tagore University
Raisen,M.P. India
pratima_shkl@yahoo.com

***Abstract*—Sentiment analysis plays a major role in corporate life as they affect their decision-making process in various kinds of events they face. Since the success or the failure of a movie depends on its reviews, there is an increase in the demand and need to build a good sentiment analysis model that classifies movies reviews. Our approach focuses on the analysis of movie dataset from the Internet Movie Database (IMDb). These reviews can be classified as positive or negative depending on the circumstances and some aspects of the question.**

Keywords- sentiment analysis, NLP, IMDb reviews, Classification.

I. INTRODUCTION

In recent years, interests of companies have increased in area of sentiment analysis and application. Sentiment analysis deals with the behavior of customer like his/her choice about any product. Reviews made by consumers contain their opinion about any topic of dialogue. this will be analyzed to grasp needs of consumers and it'll be helpful for companies to execute the operational strategies effectively. analysis is the interpretation and categorization of emotions (positive, negative and neutral) in text data using text analysis methods. Sentiment analysis tools allow businesses to identify customers' feelings about products, brands or services in online feedback. Understanding people's emotions is crucial for businesses since customers are ready to express their thoughts and feelings more openly than ever before. Opinions are central to the majority human activities because they're key influencers of our behaviors. Whenever we want to form a choice, we would like to grasp others' opinions.

In the planet, businesses and organizations always want to search out consumer or public opinions about their products and services [1]. Individual consumers also want to grasp the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision during a political election. Sentiment analysis could be a quite language Processing (NLP) for tracing the mood of the general public a couple of particular product or topic.
To analyze sentiment, there are two different approaches:

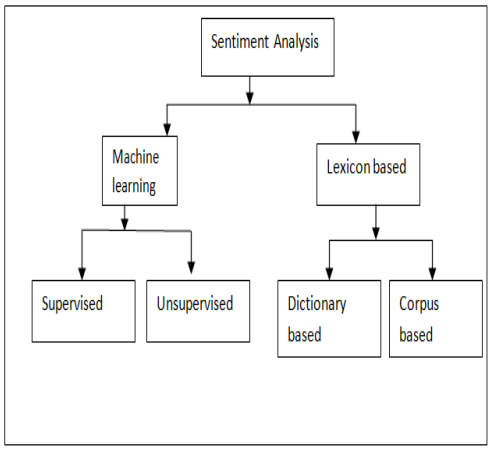


Figure 1: Sentiment Analysis Approaches

1.1 Machine Learning: Machine learning could be a method for data analysis which automates the analytical model building. There are algorithms which are used iteratively to find out from the available data, this helps the system to seem beyond the sights with none external or explicit code. [2] In simple terms machine learning may be a technique where the machine or the system learns from the previous data that it reads and automates accordingly and if required evolves from old techniques for closing the task or work differently. Machine learning has been a very important a part of the technology, because the models are continuously exposed to new data daily, so that they are adaptable independently. In machine learning, the machine learns from past data and produces reliable decisions and results based on previous data. In machine learning there are two stylesof learning

Supervised Learning: In supervised learning, we've got the computer file (x) and output data (y) and an algorithm is understood which must map the functions of input with output. the most motive behind this is often that to map the functions so accurately that whenever a replacement data is provided as an input it gives a predictive output value. the explanation for calling this method as supervised learning because the machine is taught or supervised

Unsupervised Learning: Unsupervised learning may be a technique where an algorithm is employed to infer from the datasets which consists of input file with none labeled responses. the foremost used example under this system is cluster analysis, where within the input file is given within the kind of clusters and are used for exploration and classification of the information into groups. [20]

1.2 Lexicon Based Approach

Lexicon based is another approach for Sentiment Analysis which involves sentiment calculation from semantic orientation of word or phrase which occurs in text. during this approach a dictionary of positive and negative words is required with sentiment value assigned to every and each word from both positive and negative dictionary. Then these words with sentiment value are compared with the words and phrases. Then a function for combining like sum or average is employed to create the end result in accordance to the general sentiment of the word or context or opinion. [21]

II. RELATED WORK

Jianqiang et al [3] performed a comparative study to research the preprocessing methods utilized in the analysis of twitter. The accuracy and F1-measurement of the classifier for twitter classification will be enhanced with the utilization of an appropriate preprocessing method.

Y. Liu et al. [8] proposed another techniques are useful like topic modeling within which the author proposed a process of automatically identifying the features or aspects of a product. To reduce feedback, several perspectives on the sentimental analysis of microblogging sites such as Twitter have been proposed in the research community.

Chen et al. [9] used divide-and-conquer approach which first classifies sentences into differing types, then performs sentiment analysis separately on sentences from each type. during this paper, the goal of the study is to spot sentences that contain comparative words like ``better", `more", ``less", ``most", ``least", ``outperform" which are important to explain the entities.

Vinodhini et al. [12] combine SVM with bagging techniques for ensemble methods that improve overall result for majority and also minority class compared tostandard SVM. However, this requires extensive experiments with benchmarks and real-time application datasets.

G. Rao et al. [6], proposed a brand new model of neural network with two hidden layers. the primary layer shows sentence vectors that indicate sentences in short-term and remembering networks and therefore the second layer encodes sentence relationships into a document representation.

Tubishat et al. [10] defined aspect or also referred to as feature level may be a fine-grained model that deals with determining opinion intended by people to specific features of a product, service, or any entities. for instance, phone reviews may involve specific aspects referring to a phone like sound, camera, design, and price.

B. Ghaddar et al. [13] specialize in the semantic relationship between words before proceeding with the classification. The results of their work achieves over 90% accuracy using Chinese text data. Although the result's high, the approach requires word2vec methods to define the link between the words, which the author stated that they're having an issue in high-dimensional features vectors of word2vec for SVM. SVM deals with predictive binary classification

whereby employing a large set of observation with known label (training data), it finds maximum margin function that separates observation into two classes Karagoz et al. [4] Presented a framework for sentimental analysis of Turkish informal text using frequency based aspect extraction with Sentiment Word Support (FBAE-SWS) and Web Search Based Aspect Extraction (WSBAE). The method to be monitored in this text is to improve factor extraction and to identify polarity calculations on the subject of the emotional word. It also provides a tool, including a Graphical interface (GUI) for implementing the proposed algorithm and visualizing the analysis results. R. Priyantina et al. [7] suggested how to work out the sentiment of reviews supported the hotel dataset. Hotel reviews are preprocessed into an inventory of terms. First, the potential Dirich- let Allocation (LDA) resolve Glossary; Semantic similarity then sorts the term list in keeping with the topics generated by potential Dirichlet assignments. (LDA) integrated into the five sides of the hotel. Next, when computing the resemblance, the Frequency anti-clustering frequency (TF-ICF) method. Finally, classify consumer sentiment (satisfied or not) with word embedding and memory (LSTM).

M. Al-Ayyoub et al. [11] specialise in lexicon may cause only certain languages can get direct enjoy it. The natural languages that are more ubiquitous online like English and Chinese emerges because the best target for applying Sentiment analysis verified by the mass amount of papers and tools for these languages Li et al. [5] Assessed the impact of text quality Comments established on comment length, word count, and readability. Emotion analysis task is performed on movies dataset. Three models of deep learning family (simple CNN, LSTM, and RNN). The authors claim the dataset is brief and straightforward to read higher accuracy compared to long and short length data sets readability.

In this research [14] during the foremost recent decade the net has been progressively utilized by individuals not even as clients but rather as procedures of the web web data. The themes of discussions on the forums were connected to health and fertility problems, fertility treatment and in vitro fertilization The paper specifically demonstrates machine learning, which examines the order of ump in the discussion writings. Current investigations had some goals. The principle objective was to handle various explanations of posts; the second was to give some thought to some feeling dictionaries and locate a superior one for these specific writing .In general, the results we reach the common F-measure up to 0.805.

In This Paper [15] The Author Has Explored Different Strategies And Approaches Applied By Other Researchers And Compare These Procedures By Introducing New Concept Of Fuzzy Classification For Acquiring Quality. The Author Has Proposed Fuzzy Concept With New Method Which Is Helpful For The Strategy Of Text Base .It Shows High Performance As Compare To Traditional Techniques

In This Paper [16] The Task May Be A Message Level Classification Of Tweets Into Positive, Negative And Neutral Sentiment. Tweets Are Very Noisy S They Have Lot Of Pre-Processing. The Step Of Pre-Processing Is Tokenization Which Make A Paragraphs Into Meaningful Words Or Sentences, Then Remove Non-English Tweets, It Replace The Emotions And Take Away The Numbers. Then It Generates A Baseline Model And Also The Pre-Processing Steps, Then Learn The Positive, Negative And Neutral Frequencies Of Unigram, Bigrams, And Trigrams Within The Training Set. After It Feature Extraction Is Functioning Thereon, That Is, There Are Total 34 Features, It Calculate The Feature For The Full Tweets Within The Message. And So It Divided Into Tweet Based Feature And Lexicon Based Features. After The Pre-Processing And Have Extraction, Message Or Tweet Will Classified Into Positive, Negative Or Neutral Sentiments.

In [17] This Paper Evaluates The Sentiment Analysis By Using An In Build Python Library Called Text Blob, For The Analysis On Three Platform Twitter, Face Book And News Website. Here The Author Uses Ann (Artificial Neural Network) For The Classification Of Tweets. This Is Often A Far Easier And Fewer Time Consuming Manner. Ann May Be A Computational Model Which Is Analogous To Our Biological Neural Network. By Using Twitter Api Tweets Are Collected. To Classify The Tweet Naïve Bias Algorithm Is Employed. Feed Forward Neural Network Is Employed To Separate The Info Into Train And Test The By Applying The Min-Max Approach The Accuracy Of The System Is Evaluated. R Programming Is

Employed To Predict And Analyse The Result. 70-89% Accuracy Is Obtained For An Outsized Amount Of Dataset.

III.PROPOSED METHODOLOGY

In this paper, we propose a method which combines the Decision tree and Support Vector Machines for the supervised tasks to solve the problem of classification.
Following figure 2 shows the workflow of proposed model.

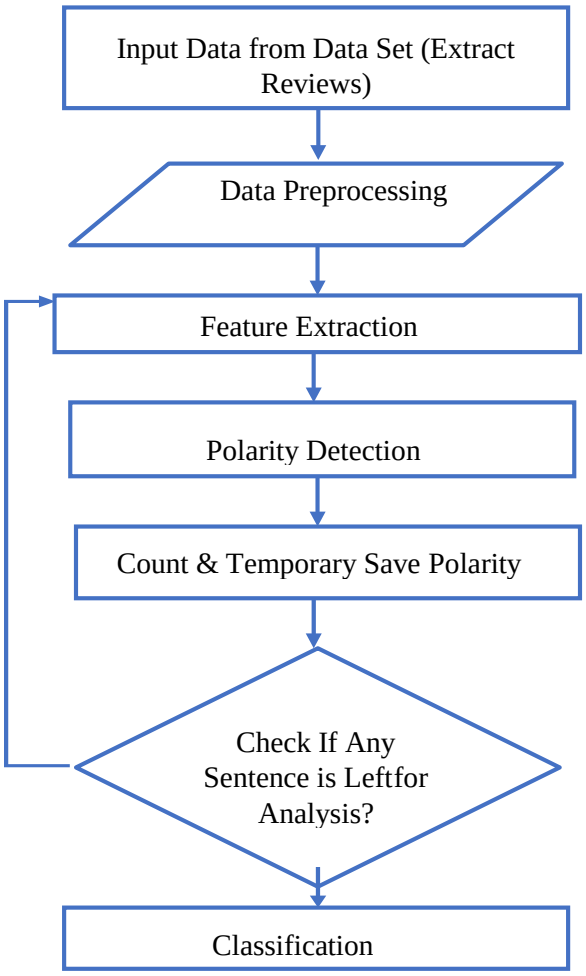


Figure 2: Workflow of Proposed Model

Fig. 2 shows the typical number of steps involved in our proposed model. Working of this can be explained as follows:

A. Data Collection: In this stage data to be analyzed is crawled The Large Movie Review dataset from ImDb.

B. Preprocessing

This module includes different NLP tasks i.e., Stop word Removal, tokenization, word stemming, and part-of-speech tagging.

1) Stop word Removal

Stop word removal removes very common words of a language e.g., ``an", ``about", ``above" etc. These words usually have no impact on NLP.

2) Tokenization

Tokenization is the process of dividing the flow of text into words, phrases, symbols, or other meaningful elements called tokens. An important point to mention is that custom data structures will be designed to hold tokens (Keyword) and sentences (list of Keywords) of each document.

3) Stemming

Stemming is the process of reducing inflected word to its base or root word.

4) Pos-Tagging

POS tagging is the process of tagging a word in a text as corresponding to a particular part of speech, based on both, its definition and its context. Each Keyword object contains an original token, its stem form, and a pos tag associated with this token. Once the data is pre-processed, it is sent to the next module

C. Feature Extraction

However, text data cannot be given as direct input to the taxonomy algorithm. Therefore, the data is converted to a TF-IDF matrix representation. TF-IDF stands for: Term Frequency-Inverse Document Frequency. This weight is a statistical measure used to estimate how important a word is to a document in a collection. Significance depends on the number of times a word appears in the document, but it is offset by the frequency of the word in the corpus. The term frequency is calculated as follows

TF (t) = (number of occurrences of t in the document) / (total number of words in the document)

The IDF is calculated as follows
idf (t) = log_e (total number of documents / number of documents with the word t)

D. Polarity Detection

After collecting all the features and sentimental words, it is very easy to know the polarity of the sentence. Sentence polarity follows rules similar to arithmetic expressions. Negative sentiment includes all negative feedback and positive emotion includes all positive feedback. Negative sentiment includes the word positive opinion. For example: The sentence in the movie review is "This movie story is not good". In this sentence, "good" opinion has a positive polarity, but "no" is a negative word. Therefore, the overall polarity of this sentence is negative.

E. Sentiment Classifier:

In this section we have used common classification methods. First, construct multiple decision trees on randomly selected traits and evaluate the class of the test example by comparing individual trees. The support vector machine revolves around the concept of margin — the hyperplane on either side separating the two squares.
Trained classifiers return scores between 0 and 1, which are converted to a binary state indicating 'negative' or 'positive'. For each combination, the presence of the element is assumed to be positive or negative. We divided the training data into different parts; This is done to check the accuracy of the sentiment classification as the training data size increases. Its purpose is to investigate the variability in the accuracy of the sentiment classification for the same test data.

Evaluation Parameters:

Reviews are classified as positive and negative by the hybrid method. The effectiveness of the hybrid method is determined by the following parameters.

$TP = TP(TP+FN)$, $FP = FP(FP+TN)$

$Precision=TP(TP+FP)$, $Recall=TP(TP+FN)$

$Accuracy=(TP+TN)(TP+TN+FP+FN$

$F\ Measure = (2*Precision*Recall)(Precision+Recall)$

The notation of TP indicates True Positive: the number of positive positive events indicating that it is actually positive, the FP indicates false positive positive: the number of positive perceived events that are actually negative, True Negative indicates the number of instances where the facts are negatively negative and the Fn false negative indicates: actually positive The number of negative examples beyond the existing negative.
The classification criteria considered for sentiment analysis are accuracy, precision, recall and f-measurement and these parameters are evaluated based on the positives and negatives of the reviews calculated by the proposed hybrid approach

Table 1 shows the results obtained by each individual and hybrid classification for the Movie Review dataset. In this work, the system is evaluated with the average f-score obtained for the positive and negative grades.

Classifier	Accuracy	Precision	Recall	F-score
Decision Tree	0.7575	0.7309	0.815	0.7706
SVM	0.8375	0.8	0.9	0.847
Hybrid Approach	0.8825	0.918	0.84	0.8772

Table 1: Results obtained by Classifiers

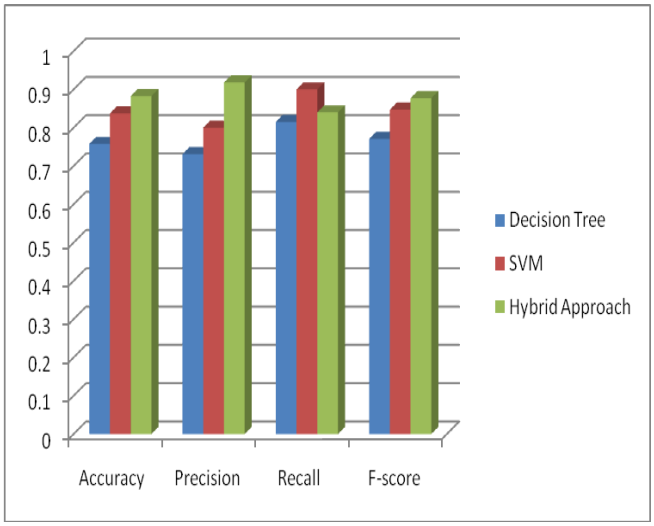


Figure 3: Comparison of Evaluation parameter

IV.CONCLUSION

In this paper a model is proposed for movie sentiment analysis to enhance the accuracy of classification. After examining the work of some researchers, we analyze that some classification methods perform better than others, but it does not work properly but there are some limitations. The proposed model was implemented in conjunction with machine learning classification. So here we see that there is an increase in accuracy and diversity in the results. In this work, we used IMDb. Movie review dataset.

REFERENCES

[1]. M.R. Saleh, M.T. Martín Valdivia, A. MontejoRáez, and L.A. Urena Lopez, Experiments with SVM to classify opinions in different domains, Expert Syst. Appl. 38, pp. 14799-14804 (2011)

[2]. Bing Liu, Sentiment Analysis, Mining opinions, Sentiments, and Emotions, Book, June (2012).

[3]. Z. Jianqiang and G. Xiaolin, ``Comparison research on text pre-processing methods on Twitter sentiment analysis," IEEE Access, vol. 5,pp. 2870_2879, 2017. 19

[4]. Y. Liu, ``Social media tools as a learning resource," J. Educ. Technol. Develop. Exchange, vol. 3, no. 1, pp. 101_114, Mar. 2017.

[5]T. Chen, R. Xu, Y. He, and X. Wang, ``Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," Expert Syst.Appl., vol. 72, pp. 221_230, Apr. 2017.

[6] G.Vinodhini and R. Chandrasekaran, ``A sampling based sentiment mining approach for e-commerce applications," Inf. Process. Manage., vol. 53,no. 1, pp. 223_236, Jan. 2017.

[7].G. Rao, W. Huang, Z. Feng, and Q. Cong, ``LSTM with sentence representations for document-level sentiment classi_cation," Neurocomputing, vol. 308, pp. 49_57, Sep. 2018.

[8]M. Tubishat, N. Idris, and M. A. Abushariah, ``Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges,"Inf. Process. Manage. vol. 54, no. 4, pp. 545_563, Jul. 2018.

[9] B. Ghaddar and J. Naoum-Sawaya, ``High dimensional data classification and feature selection using support vector machines," Eur. J. Oper. Res.,vol. 265, no. 3, pp. 993_1004, Mar. 2018.

[10]. P. Karagoz, B. Kama, M. Ozturk, I. H. Toroslu, and D. Canturk, ``A frame-work for aspect based sentiment analysis on turkish informal texts,"J. Intell. Inf. Syst., vol. 53, no. 3, pp. 431_451, Dec. 2019.

[11]. R. Priyantina, I. Teknologi Sepuluh Nopember, R. Sarno, and I. TeknologiSepuluh Nopember, ``Sentiment analysis of hotel reviews using latentDirichlet allocation, semantic similarity and LSTM," Int. J. Intell. Eng. Syst., vol. 12, no. 4, pp. 142_155, Aug. 2019.

[12]M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi,``A comprehensive survey of arabic sentiment analysis," Inf. Process.Manage., vol. 56, no. 2, pp. 320_342, Mar. 2019.

[13]. L. Li, T.-T. Goh, and D. Jin, ``How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis," Neural Comput. Appl., vol. 32, no. 9, pp. 4387_4415, May 2020.

[14]. V. Bobicev, "Text classification: The case of multiple labels," IEEE Int. Conf. Commun., vol. 2016-Augus, pp. 39–42, 2016.

[15]. P. Chikersal, S. Poria, E. Cambria, A. Gelbukh, and C. E. Siong, "Modelling Public Sentiment in Twitter : Using Linguistic Patterns to Enhance Supervised Learning," pp. 49–65.

[16]. Ayush Dalmia, Manish Gupta, Vasudeva Varma "Twitter Sentiment Analysis The good ,the bad and the neutral" ,Association for Computation Linguistic, June 2015

[17]. Sneh Paliwal, Sunil Kumar Khatri and Mayank Sharma, "Sentiment Analysis and Prediction Using Neural Networks", Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA), 2018

In-Memory Databases: The Storage of Big Data

Devika Rani Roy[1], Dr. Sitesh kumar Sinha[2], S.Veenadhari[3]

Research Scholar[1], Guide[2], Department of Computer Science,

Co-guide[3], Department of Computer Science

RNTU, Madhay Pradesh (India)

Abstract: **Digital communication generates large amounts of data. This is a great opportunity for companies that work with big data. However, the more data a company has to work with, the greater the challenge becomes to recognize connectivity and patterns. Information technology solutions and systems are ever more in demand to support companies in evaluating the huge amounts of information they receive. Data analysis using traditional databases is no longer sufficient to store, retrieve, and process extremely large collections of data. When classic databases reach their limits, in-memory databases can be of use.**

Keywords: **Digital communication, databases, Data Analysis, In-Memory, information**.

1.0 Introduction

Large amounts of data are stored in In-memory databases and it also provide a large range of analysis results. When storing data with an in-memory database, a distinction is between column-oriented and row-oriented data storage is there , but some database systems use both methods of data storage i.e column-oriented and row-oriented data storage. In one row, Row-oriented database are arranged and the collected data records kept together .

For example, if the values "student name, rollno , and subject" are stored, the data would be arranged as follows: student name 1, rollno 1, subjec 1, student name 2, rollno 2, subjec 2 ,student name 3, rollno 3 subjec 3. In a column-based storage, the data is to be assigned in corresponding categories like in this way: student name1,student name2,student name3,rollno1,rollno2,rollno3, subject1, subject2, subject3.

The format of the column-based data storage is called the column format. By keeping data with the same values together, the system reduces the amount of data available. Storage space and transfer times are reduced. The functionality of the memory database has also improved over time, and only certain columns need analysis, not all. This form of data analysis is called columnar projection.[1]

2.0 Technology for big data storage

The concept of in-memory databases is nothing new. The foundations of database technologies were

developed in mid-1980s. However, the Information technology systems back then did not have the required processing capacity, so that earlier concepts of in-memory databases were not are used. Modern computer architectures use the concept of data warehousing, 64-bit technology, and multi-core processors finally made it possible for in-memory databases and its use.

Data inside the memory is usually for data storage. These database systems collect and compress data from a variety of sources, store it for a long time, and then prepare it for analysis. With 64-bit technology, it is possible to increase the capacity of large memory up to a terabyte range. As a result, memory flows have increased in size.[1],[2]

With multi-core processors, most processor cores operate on a single chip, resulting in better performance and higher data performance. Data operation reflects the volume of the transmitted network data.

2.1 Steps of an In-Memory Database

Recurring, identical processes occur during the running of in-memory databases. An in-memory database back up the data in the following way:

1. **Start the database**: when the database is started, the system loads the entire dataset from the hard disk into the working memory. It means that no data has to be loaded while the database is running.
2. **Readjusting data**: the database will review and adjust data frequently, so if data changes it remains up to date.
3. **Transaction log and its backup**: current changes are recorded in transaction logs. If an error occurs, the database can be restored to the time before the error occurred. This process is called "rollforward".
4. **Data processing**: data is processed according to the ACID principle (atomicity, consistency, isolation, and durability), as it is in traditional databases. The acronym ACID describes the exact properties of processes in database management systems.
5. **Database replication**: this step continuously copies data from the database to a computer or server as a backup.[3]

2.2 In-memory analytics

It is an enterprise architecture framework solution that is used to enhance business intelligence reporting by querying data from system memory versus the traditional hard disk drive medium. According

to this approach, inquiry time is reduced in an effort to facilitate efficient business decisions.[2]

With the development of business intelligence and random access memory hardware technology, more business intelligence platforms are available and affordable, including in-memory analytics tools that are used to facilitate enterprise decision-making, even for small businesses. Traditional BI Online analytical processing incorporates dedicated, heavy resources for programming or building data structures through unusual schemas.[2],[3],[4]

In-memory analytics eliminate the overhead of storing data aggregate tables or indexing pre-aggregated data cubes, resulting in extremely fast query responses. In short, faster data retrieval rates indicate faster processing and informed decision making.[4]

3.0 Big Data Analytics Strategy

A well-defined, integrated and comprehensive strategy contributes to and supports valuable data-driven decision making in an organization. Here, I listed the most followed step.

Step 1: Evaluation: An assessment, which is already tied to business objectives, involves key stakeholders, forming a team of members with the right skill sets and evaluating policies, people, process and technology and data assets Is required. In this process, evaluation clients may be required, if necessary.

Step 2: Priority: After evaluation, one needs to get use cases, prioritize those using big data predictive analytics, prescriptive analytics and development analytics. One can also use tools like priority matrix and filter the use cases with the help of feedback and input from the main user.

Step 3: Roadmap: In this phase we can create a timed roadmap and publish it to everyone. It needs to include all the details about the complexities, funding, inherent benefits of use cases, and mapped projects.

Step 4: Change Management: Such a change requires managing data availability, integrity, security, and

usability. A successful change management program, using any existing data governance, encourages activities and members on a continuous monitoring basis.

Step 5: The Right Skill Set: Identifying the right skill set is critical to the success of the organization amidst current trends in the industry. In this phase, there is a need to bring educational programs to follow the right leaders and educate key stakeholders.

Step 6: Reliability, scalability and security: The right approach and effective Big Data analytics strategy make the analytics process reliable, with the powerful use of explanatory models incorporating data science principles. A big data analytics strategy needs to incorporate security aspects right from the beginning for a strong and tightly integrated analytics pipeline.[5]

4.0 Advantage and disadvantage of in-memory databases

The advantage of in-memory databases is that access to data is much faster than traditional databases. The disadvantage of in-memory databases is that it is not possible to store data permanently on RAM. Here is a comparison of the advantages and disadvantages of in-memory databases.

4.1 Advantage of in-memory databases

The biggest advantage of using in-memory databases is the high access speed arising from the use of RAM. It also leads to a quick data analysis. However, it is not only the reduced fetch time that optimizes data analysis. In-memory DBs make it possible to evaluate structured and unstructured data from any system. Now, companies and software solutions are expected to face the challenge of storing and processing large-size unstructured data, such as text, images, or audio and video files.[6]

Using a distributed data infrastructure, unstructured data can be stored in an in-memory database, in which multiple processing units

(computers, processors, etc.) operate on a common task in parallel and distribute it to different server groups . This results in a higher storage capacity, faster processing, and better transfer speed of unstructured data.

4.2 Disadvantage of In-Memory Databases

The use of RAM means faster access to one side, but it also has a significant disadvantage: stored data is only temporary. If the computer system crashes, then all the data stored temporarily will be lost. To add to this, the following methods are established:

1. Snapshot files: are specific moments, such as at regular intervals or before the system is shut down, when the current version of the database is saved. An important criticism of this is that all the data added after the last so-called snapshot will be lost in case of a crash - depending on how long each interval is, it can be a lot of data.

2. Transaction Protocol Security: Noting changes in transaction logs is a unified process that is used as a means of protection. Used in combination with a snapshot process, the transaction protocol can help restore the system after a crash.

3. Replication: In-memory databases already have the function of storing an exact copy of the database on a traditional hard disk. In the event of a failure, the stored database can be accessed.

4. Non-volatile RAM: A non-volatile RAM is able to keep files available for recovery even after the system is restarted.[5],[6]

5.0 List of In-Memory Databases

1. **H2 Database: H2** is an open source database written in Java that supports standard SQL for both embedded and standalone databases. It is very fast and contains only 1.5 MB of Java archive files.
2. **HSQLDB (HyperSQL Database):** HSQLDB is an open source project, also written in Java, which represents a relational database. It follows structured query language and Java database connectivity standards. It also supports SQL features such as stored procedures and triggers. It can be used in in-memory mode, or it can be configured to use disk storage.
3. **Apache Derby Database:** Apache Derby is another open source project that has a

relational database management system created by the Apache Software Foundation.

4. **Derby:** It is based on SQL and JDBC standards and is primarily used as an embedded database, but can also be run in client-server mode using the Derby Network Server Framework.
5. **SQLite:** This is a SQL database that runs only in embedded mode, either in memory or saved as a file. It is written in C language but can also be used with Java.
6. **In-Memory Database in Spring Boot:** Spring Boot makes it particularly easy to use in-memory databases because it can automatically create configurations for H2, HyperSQL databases, and Derby.
7. **HyperSQL Database:** It uses an in-memory structure for faster operations against the DB server. It uses disk persistence according to user flexibility with reliable crash recovery. It is also suitable for business intelligence, ETL and other applications that process large data sets. It has a wide range of enterprise deployment options, such as XA transactions, connection pooling data sources, and remote authentication. It is written in the Java programming language and runs in a Java virtual machine. It supports the JDBC interface for database access.[4],[5],[6]

References

[1.] Omae Malack Oteri, Mobile Subscription, Penetration and Coverage Trends in Kenya's Telecommunication Sector.

[2.] Hoda A. Abdel hafez Mining Big Data in Telecommunications Industry: Challenges, Techniques, and Revenue Opportunity.

[3.] A hybrid fuzzy-based Personalized Recommender System for Telecom Products/Services.

[4.] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification Tina R. Patil, Mrs. S. S. Sherekar Sant Gadgebaba Amravati University.

[5.] Big Data & Advanced Analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity.

[6.] www.omnisci.com/in-memory-database.

A Cogitation of Image Recognition using Machine Learning and Deep Learning Techniques

[1] Ms. Geeta Guwalani [2] Dr. S. Veenadhari, [3] Ms. Manju Devnani
[1] RNTU, Bhopal,[2] RNTU, Bhopal, [3] RNTU, BHOPAL
[1]geet.cs21@gmail.com, [2] veenadhari1@gmail.com, [3] manju.asnani@gmail.com

Abstract— Image recognition (or image classification) is that the task of recognizing pictures and labelling them in one in every of the many preset distinct classes. Image classification could be a typical job of image processing, laptop vision, and machine learning fields. During this analysis endeavor, machine learning for Image Classification has been analyzed. Pattern recognition is the procedure of characteristic symmetries in statistics by a machine that uses machine learning algorithms. For the machine to go after patterns in data, it ought to be reprocessed and reworked into a kind that a computer will recognize. In Machine Learning, the model is created engineered on some algorithms that study from the info on the condition that to create estimates. Deep learning algorithms are a subgroup of machine learning algorithms, which functions learning multiple points of distributed illustrations. Deep Learning has broken the bounds of what was conceivable within the domain of Digital Image Processing. Deep Learning models, with their multi-dimensional patterns, are terribly useful in extracting advanced information measure from stimulative images. The image recognition technology will solitarily surge the potency of classifying the article of involution within the image and access the things of interest.

Index Terms— Deep Learning, Image Classification, Image Recognition, Image Processing, Machine Learning, Pattern Recognition

I. INTRODUCTION

Image recognition involves a large amount of information operation, necessitating high processing speed and recognition precision, as well as real-time and fault-tolerance of the neural network in accordance with image recognition requirements. This paper begins by examining conventional image recognition methods. First, this paper analyses the traditional image recognition method, focusing on the limitations of traditional methods and the complex iterations, such as images showing different states, in the process of image processing algorithm for the image segmentation study and improvement.

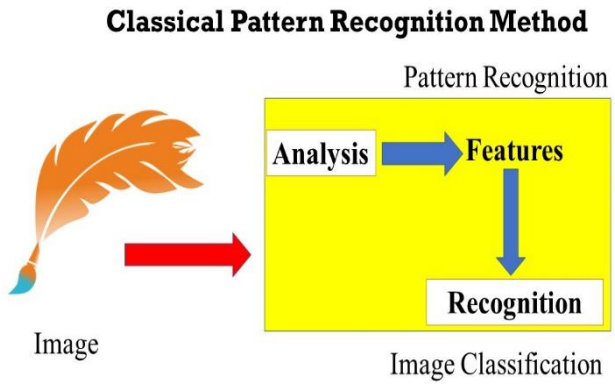
Image recognition technology is widely used in our daily lives. For example, a scene recognition algorithm can automatically recognize some common scenes in images, such as the sky, grass, people, and so on. Based on that function, the client's application can easily realize automatic image management, grouping, and searching, completing the intelligent management of a large image library while saving a significant amount of time. The significance of image recognition technology is that it frees people from heavy and mechanical repetitive work, giving them more time to deal with more meaningful things, thereby greatly increasing working efficiency.[1]

II. BACKGROUND KNOWLEDGE ON TRADITIONAL VS DEEP LEARNING APPROACHES

Conventional machine learning algorithmic technologies have matured to the point where many of them (such as SVM) can be implemented to image recognition. It is simple to use traditional machine learning methods to solve some problems with today's computer processing and computing power. However, when these techniques are applied to image recognition, some challenges must be overcome. Conventional machine learning algorithms can only interact with one-dimensional data, whereas the images appeared as a matrix.

There are mainly two methods proposed to handle this problem. One is to stretch a matrix line by line or column by column into one-dimensional vectors. This method is relatively easy to implement, but by doing so, the position relation of rows and rows in the original matrix will be lost, which will cause some problems that the important information can't embody. For example, the color features among several neighboring pixels in the image are similar,

but they are no longer adjacent after being stretched into vectors, which makes it difficult for the machine to recognize the image with the similar features lost. The second method is to extract the feature information of the image step by step, such as using histogram form to count the area of a specific color in the picture, or the number of lines with certain shape in the picture to identify the image content. This method is similar to the human image recognition method. But the disadvantage of this approach is also clear: after a picture is entered, the machine has no way of knowing which features are useful and which are redundant. Then in order to recognize the image better, the machine needs to count and calculate a large number of features, consuming a lot of unnecessary time and computing resources. Generally speaking, although traditional machine learning algorithms have their own advantages, they are not suitable for the task of image recognition.[2]



Deep learning has recently become popular for image recognition. The image can be directly used as the input of an image recognition network in deep learning. Unlike conventional algorithms that splice or extract features from the original image, the depth learning technology iteratively extracts the features from the image by convnet of the images. After convolution is completed, the algorithm processes the convolution image data at different resolution scales by pooling [3] or upsampling [4]. This method can extract the feature information in different resolutions and make the image information more complete. Such a method of information extraction makes the deep learning have strong learning ability that the traditional machine learning

FIG I: CLASSICAL PATTERN RECOGNITION METHOD

method does not have, which makes it more accurate when dealing with the problem of image recognition. But the strong learning ability also brings some problems. In the training concentration, the deep learning mode will exist the problem of over-fitting [5], which makes the actual recognition effect worse and limits its ability.

In order to make the depth learning achieve higher recognition accuracy, weight sharing of convolution kernel can be used to reduce the overall computation, so as to avoid over-fitting in the algorithm. In conclusion, the deep learning method has stronger learning ability and it is more suitable to solve the problem of image recognition.

Basis of Image Classification Techniques in Machine Learning

III. BASIS OF IMAGE CLASSIFICATION TECHNIQUES IN MACHINE LEARNING

The current development in the field of artificial intelligence and machine learning has contributed to the evolution of computer vision and image recognition perceptions. From monitoring a driver-less car to carrying out face detection for a biometric access, image recognition helps in processing and categorizing objects based on trained algorithms.[6]

The computer vision and machine learning are two significant zones of current research. The computer vision uses the image and pattern maps to find resolutions. It considers an image as a pixel set. The computer vision systematizes the monitoring, review, and surveillance tasks. Machine learning is the subcategory of an artificial intelligence.

Automatic video analysis/annotation is a result of IT vision and machine learning. Self-driving cars are a great example to understand where image classification is used in the real-world. To enable autonomous driving, we can build an image classification model that recognizes various objects, such as vehicles, people, moving objects, etc. on the road.[7]

Machines are one can do what they are automated to do. If we build a model that finds faces in images, that is all it can do. It won't look for cars or trees or anything else; it will categorize everything it sees into a face or not a face and will do so based on the features that we teach it to recognize. This means that the number of categories to choose between is finite, as is the set of features we tell it to look for. We can tell a machine learning model to classify an image into multiple categories if we want (although most choose just one) and for each

category in the set of categories, we say that every input either has that feature or doesn't have that feature. Machine learning helps us with this task by determining membership based on values that it has learned rather than being explicitly programmed. [8]

IV. Working of Image Recognition Techniques

Model training is essential for an image recognition model to work. Deep learning methods are presently the finest performing tools to train image recognition representations.

For an image recognition model to work, we should have a data set like a newborn baby, to identify objects around him, the objects must first be presented by his parents. Similarly, for machines, there is a data set with deep learning techniques, the model must be trained in order to perform.
An image is a group of pixels to a computer. By extracting certain features from the image, we make meaningful data, so the process is called feature extraction. Feature extraction permits detailed patterns to be represented by specific vectors. Deep learning methods are also used to find the boundary range of these vectors. At present time, a data set is used to train the model, and in the end, the model predicts certain objects and labels the new input image into a certain class.

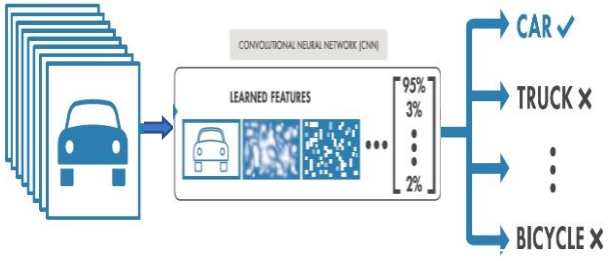


Fig II: Feature Extraction Process

V. Techniques used in Image Processing

Facebook can identify your friend's face with only a few tagged pictures. The efficacy of this technology depends on the ability to classify images. Classification is pattern matching with data. Images are data in the form of 2-dimensional matrices. In fact, image recognition is classifying data into one category out of many. One common and an important example is optical character recognition (OCR). OCR converts images of typed or handwritten text into machine-encoded text.

A. Classification based on the type of training sample used

The image classification techniques can be categorized as supervised and unsupervised, or hard and soft classifiers. Supervised classification technique requires the training data set in order to teach the classifier to define the decision boundary. It recognizes the instance of the necessary information in the image, which are known as training sites. This is then used to expand a statistical description of the reflectance of information for each class, which is known as 'Signature Analysis'. The last step is to classify the image by searching the reflectance for each pixel and evaluating the resemblance to the signatures. [9] The data provided during the signature analysis, also known as training phase, is stored in a file called training data file. The classification phase uses this information for classifying the input images. [10] The advantage of this kind of technique is that the errors can be easily identified and solved. The only disadvantage is the large time required for training phase. [11]

In supervised classification, it is essential for an analyst to have prior knowledge before testing and must be gathered the information after testing. The steps in the supervised classification technique are, i.e.,

i Classifying the training areas for each informational class.
ii. Signatures identifies (variance, covariance, mean etc.)
iii. All pixels are then classified.
iv. Mapping of the informational class.

The main advantage of using supervised classification is an operator can detect errors and correct them. The disadvantages of this technique are that it is time consuming and expensive. Furthermore, the analyst's training data may not highlight all of the conditions seen in the picture, making it vulnerable to human error.

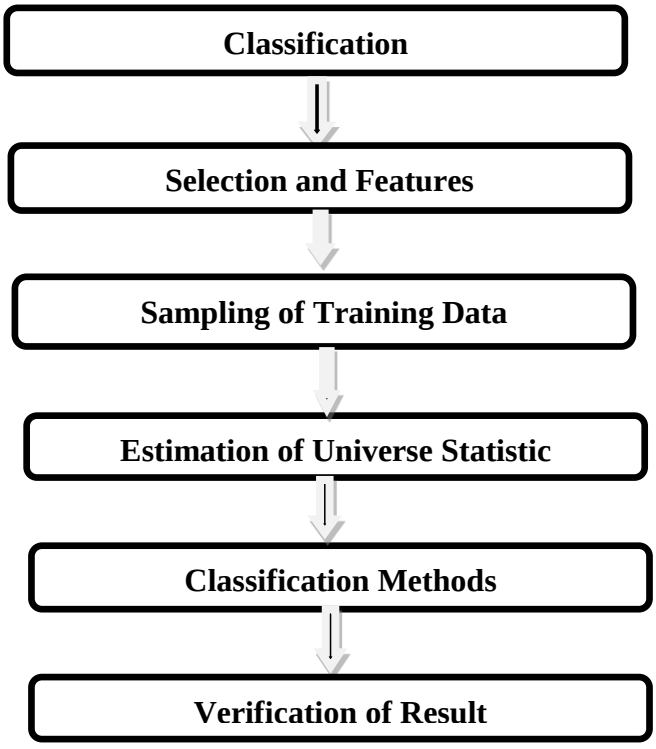
The unsupervised learning explores the fundamental structure of the data and automatically partitions them based on it. It produces a set of centroids which represent the prototypes of the classes. These are used for further classification. [12] divides unsupervised, also known as clustering method, into two groups, namely Hierarchical clustering and Partitioning clustering. The former groups the data with a sequence of partitions, while the later one divides the data into pre-specified number of clusters. The algorithm starts with initialization that is done by executing an initial segmentation rule. Next, the classification is done using

different strategies. The results of these techniques are physically explicable, but the accuracy highly depends on the design of the algorithm. This technique is fast and fully automated. [13]

In unsupervised classification, no prior information is required and doesn't need any form of human intervention. It helps us in identifying clusters in data. The method is divided into steps:

i. Clustering the data.
ii. All pixels are then classified based on clusters.
iii. The Spectral class map.
iv. Cluster labeling done by analyst
v. Map the informational class

The unsupervised technique is faster and free from human errors, the analyst doesn't need any prior knowledge. The advantages of the unsupervised technique are that it is faster, free from human errors and there is no requirement of detailed prior knowledge. The main drawback of this technique is maximally separable clusters.



B. Classification based on the type of various parameter used on data

Parametric and Non-parametric functions are the two types of density-based functions. Classifier based on metrics: Parameters including the covariance matrix and the mean vector are commonly used. When the landscape is complex, parametric classifiers produce noisy results. Training samples are commonly used to capture these parameters. Training samples are commonly used to collect these parameters. The key drawback is that spatial, contextual attributes, ancillary data, and non-statistical information are difficult to incorporate into a classification process. Maximum likelihood, & Linear discriminate analysis is an example.

In the case of an unspecified density function, non-parametric classifiers are used to estimate the probability density function.

So, there are large number of classifiers/methods come to existence that are used for classification task.

Non-parametric classifiers do not use statistical parameters to calculate class separation. Expert systems, support vector machines, artificial neural networks, and decision trees are a few examples.

C. Classification on the basis of nature of pixel information used on data

a) Per-pixel Classification Approaches

Traditional per-pixel classifiers create a signature by combining the spectra of all training-set pixels for a specific feature. The resulting signature incorporates all of the materials present in the training pixels but disregards the impact of the mixed pixels. This classifier aids in the combination of the spectra of all training-set pixels from a given specified feature. The resulting combination will now include the contributions of all spectra present in the training set pixels while ignoring mixed pixel problems. Minimum distance, maximum likelihood, support vector machine, artificial neural network, and decision tree are some examples.

b) Subpixel Classification Approaches

Most classification methods rely on per-pixel information,

in which each pixel is assigned to one of several land-cover classes that are mutually exclusive. Mixed pixels are common in medium and coarse spatial resolution data due to the heterogeneity of landscapes and the spatial resolution limitation of remote-sensing imagery.

Subpixel classification approaches have been designed to provide a more accurate representation and area estimate of land covers than per-pixel approaches, especially when coarse spatial resolution data can be used. The spectral value of each pixel is known to be either a linear or nonlinear combination of pure materials in the case of a sub pixel classifier. It assigns the required membership of each pixel to each end member. Subpixel classifiers can be used on images with medium and coarse spatial resolution. Fuzzy-set classifiers and spectral mixture analysis are two examples.

c) Per-field Classification Approaches

It enhances classification accuracy. In this case, GIS plays an important role in per-field classification. This facilitates the integration of raster and vector data. The vector data are used to divide an image into parcels, and the classification is done based on the parcels. Approaches to classification based on geographic information systems (GIS) are one example.

D. Classification on the basis of number of outputs generated for each spatial data element

a) Hard Classification

In hard classification techniques, we divide an image into different categories based on its pixels.
These algorithms aid in the classification of all pixels within image land cover classes or themes. Hard classification can be used to estimate homogeneous areas (such as croplands and water bodies). Because of the mixed pixel problem, it may result in a large number of errors from coarse spatial resolution data.

Maximum likelihood, support vector machine, ISODATA (Unsupervised classification), parallelepiped, centroid (k means), neural network, and decision tree are a few examples.

b) Soft Classification

Because of its ability to deal with mixed pixels, soft classification has been proposed as an alternative to hard classification. Subpixel scale information is represented in this classification by the output of a soft classification by the strength of membership of a pixel display of each class.

E. Summary of Classification Techniques

While several classification approaches have been introduced, it is not fully understood which approach is appropriate for features of interest in a given study area. Classification algorithms can be per-pixel, subpixel, or field-based. In practice, per-pixel classification is still the most commonly used method. However, due to the impact of the mixed pixel issue, the accuracy may fail miserably of the research requirements. Subpixel algorithms have the ability to solve the mixed pixel problem and achieve higher precision for images with a medium and coarse spatial resolution.
While mixed pixels are reduced in fine spatial resolution data, spectral variation within land classes can reduce classification accuracy. Per-field classification methods are best suited for data with high spatial resolution. Many factors must be considered when selecting a suitable classifier, including classification accuracy, algorithm performance, and computational resources.

CONCLUSION

There are several ways to classify these methods, the most common of which is to divide them into supervised and unsupervised categories. The effectiveness of these techniques is primarily determined by the type of data for which they are used. Our analysis also explored various scenarios for various image classification methods, as well as the benefits and drawbacks of each.

As a result, this paper will assist us in selecting an acceptable classification technique from among those available.

REFERENCES

[1][2] Yunfei Lai, "A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition" Published under licence by IOP Publishing Ltd
Journal of Physics: Conference Series, Volume 1314, 3rd International Conference on Electrical, Mechanical and Computer Engineering 9–11 August 2019: Conf. Ser. 1314 012148
[3] Wischik D, Handley M, Braun M B. The resource pooling principle[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(5):47.

[4] Shan Q, Li Z, Jia J, et al. Fast image/video upsampling[J]. ACM Transactions on Graphics, 2008, 27(5):1.
[5] Schaffer C. Overfitting avoidance as bias[J]. Machine Learning, 1993, 10(2):153-178.
[6] Shubham Gupta, September 28, 2018, Understanding Image Recognition and Its Uses (einfochips.com)
[7] Pulkit Sharma, January 10, 2019 Image Classification | Building Image Classification Model (analyticsvidhya.com)
[8] Lindsay Schardon, October 31, 2020, An Introduction to Image Recognition – Python Machine Learning
[9] Rajesh Sharma R Beaula A, Marikkannu P, Akey Sungheeth, C. Sahana, "Comparative Study of Distinctive Image Classification Techniques", 10th International Conference on Intelligent Systems and Control (ISCO), 2016.
[10] Kalra K., Goswami A.K., Gupta R., "A Comparative Study of Supervised Image Classification Algorithms for Satellite Images", International Journal of Electrical, Electronics and Data Communication, ISSN: 2320-2084, Volume-1, Issue-10, Dec-2013.
[11] Kurian J., Karunakaran V., "A Survey on Image Classification Methods", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 1, Issue 4, October 2012.
[12] Y. Hu and K. Ashenayi, R. Veltri, G. O'Dowd and G. Miller, R. Hurst and R. Bonner, "A Comparison of Neural Network and Fuzzy c-Means Methods in Bladder Cancer Cell Classification", 1994. IEEE World Congress on Computational Intelligence
[13] Yun Yang and Ke Chen, "Unsupervised Learning via Iteratively Constructed Clustering Ensemble", The 2010 International Joint Conference on Neural Networks (IJCNN

A Study Based on Plant Leaf Disease Detection

[1]Ila Sharma, [2]Dr. Varsha Jotwani
[1] Research Scholar CS/IT, Rabindranath Tagore University, Bhopal, India [2] Associate Professor, Department of CS/IT, Rabindranath Tagore University, Bhopal, India
[1] ila.sati23@gmail.com, [2] varsha.jotwani@gmail.com

Abstract— In the last decade due to leaf disease, a huge amount of loss occurred in agriculture as well as horticulture either in plants or crops. To overcome this indomitable problem soft-computing is introduced in this area. With the help of soft computing one can detect the disease in early-stage and adopt medicare to keep the plants. For detection of diseases, different machine learning approaches are being used. This article demonstrate a survey on the use of advanced image processing strategies based on machine learning to distinguish, measure, and group plant illnesses from images. This was accomplished for dual principal explanations: to restrict the length of the venture as well as strategies managing stems, roots, seeds and fruits have a few idiosyncrasies that would permit a particular review. This survey shows a broad and reachable summary of the many approaches used for leaf disease recognition and sorting.

Index Terms— ***Leaf Disease, Soft Computing, Sorting, Bacterial ooze, Water-soaked lesion Neural Networks, Adaptive Neuro Fuzzy and Genetic Algorithm etc.***

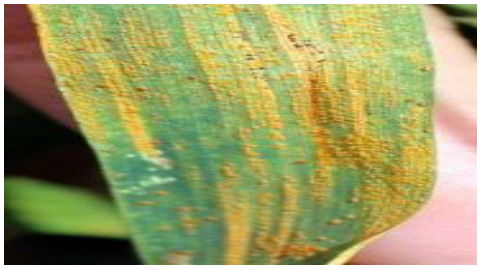
I. INTRODUCTION

India is a predominantly agricultural nation, with agriculture employing roughly 70% of the population. Farmers have a large range of fruit and vegetable crops from which to choose [9]. However, growing these grains for maximum yield as well as product value is as high as possible with the help different disease detection methods. Disease diagnosis is a languidly procedure, and the majority of diseases are difficult to identify [34] [35]. Illnesses are significant elements to limit the development of harvests in horticulture creating, which may diminish yields of yields enormously and nature of items [10]. As of now, the analysis of yields illnesses for the most part relies upon manual acknowledgment, yet a few issues happen: from one perspective, it tends to be erroneously analyzed by ranchers since they as a rule judge the side effect by their encounters; then again, the infection treatment might be dawdled over in light of the fact that the specialist or master can't go to district to analyse eventually [11]. Plant diseases have become a conundrum because they can result in significant reductions in both the quality and quantity of farming crops. Automatic recognition of plant infections is a significant research subject since it could support track vast grounds of crops as a result, identify disease symptoms on plant leaves [13].

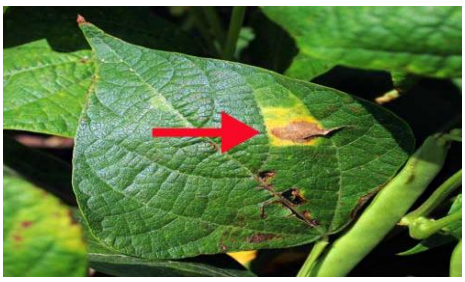
II. LEAF DISEASE

Viruses, fungi, and bacteria cause the majority of leaf diseases. Viruses are small particles made up of protein and genetic material that have no other proteins attached to them [14]. Fungi are classified mainly based on their morphology, with particular attention paid to their reproductive structures. Bacteria are now more complex than fungi and have longer life cycles, and that therefore bacteria originate out from single cells and multiply by splitting into two cells in a process known as binary cell division. Physical evidence of the microorganism is a symptom of plant disease. Fungal fruiting bodies are a sign of infection [15]. Dry mildew on a lilac leaf is the parasitic fungal disease organism itself. Gummosis, a bacterial organizational requirements that emerges from stone fruit cankers, is caused by bacterial canker. Although the canker is made up of plant tissue and is a symptom, the dense, liquid exudates are mostly made up of bacteria and are a sign of the disease. A noticeable impact of disease on the plant is a symptom of plant disease. A visible change in color, form, or function of the plant as it responds to the pathogen can be one of the symptoms. Which is caused by fungal plant pathogens. Brown, necrotic lesions surrounded by a bright yellow halo at the leaf margin or interior are common bacterial blight symptoms on bean plants. You are not seeing the disease pathogen itself, but rather a symptom triggered by the pathogen. Here are some examples of common fungal, bacterial, and viral plant disease signs and

symptoms:



(a)Stripe rust pustules on a winter leaf



(b) Dark red kidney bean leaf bacterial spot

(c) Bacterial Blight Disease

(d) Anthracnose Fungal disease

TABLE -I PROPERTIES OF DIFFERENT LEAF DISEASE

Bacterial disease signs	**Bacterial disease symptoms:**	**Viral disease symptoms**:	**Fungal disease signs**:	**Fungal disease symptoms**:
Bacterial ooze	Leaf spot with yellow halo	Mosaic leaf pattern	Leaf rust (common leaf rust in corn)	Birds-eye spot on berries (anthracnose)
Water-soaked lesions	Fruit spot Canker Crown gall	Crinkled leaves Yellowed leaves	Stem rust (wheat stem rust) Sclerotinia (white meld)	Damping off of seedlings (phytophthora) Leaf spot
Bacterial streaming in water from a cut stem	Sheperd's crook stem ends on woody plants	Plant stunting	Plant stunting	Chlorosis (yellowing of leaves)

III. LITERATURE SURVEY

Singh, U. P., et.al. (2019) - This study aims to improve the production as well as quality of plants. Their products by controlling microbial factors that cause severe losses in crop yield. Pattern detection, classification, and object extraction are only a few of the problems that image recognition coupled with machine learning approaches has surpassed in resolve. As a result, a creative model called Multi-column CNN was introduced in this effort for the arrangement of mango leaves diseased with the fungal disease Anthracnose. The suggested work's superior performance is demonstrated by its accuracy of 97.13 percent as compared to other state-of-the-art approaches [01].

Bhimte, N. R. et.al.(2018) - The suggested technique involves observing different fields to obtain the needed database of various cotton diseases. Support vector machine is used to identify cotton leaf diseases such as Bacterial blight and Magnesium Deficiency using an image processing technique, as well as segmentation and classification techniques. Color segmentation, in which an RGB image is converted to a color space and diseased parts (areas of

interest) are extracted from leaf images with efficient features using k-means clustering. Color, form, and texture are useful features for pattern recognition, classification, and accurate and error-free classification. The development of a more effective, stable machine vision system for early automatic detection of various types of plant diseases will be the focus of future work [02].

Saponaro, P., et.al. (2017) -The authors of this study presented a synthetic fungal hyphal generator that generates 3D image stacks with properties that are similar to those of real data. On their results, they compared segmentation and filtering methods. Although the neural Network approach yielded the better effect, it also took the longest to process. On real microscopy data, they also compares the neural Network method to the Frangi filter, and found that the deep CNN method generated better results overall. Even the best segmentation methods may be affected by objects in real data, causing small gaps to appear. They devised a skeleton gap-closing algorithm based on a minimum spanning tree to deal with this. They used the synthetic fungal generator to measure the gap closing algorithm's output with synthetic gaps, and we got a maximum F1 score of 77.3 percent. They will develop new segmentation algorithms for automated disease resistance analysis of maize in the future, which will run on real microscopy data [03].

Sabrol, H., et.al. (2016) - In this research, the supervised learning technique is used to classify tomato plant leaves into six categories: safe and unhealthy, i.e., five types of diseases caused by fungal, bacterial, and viral infections. Color, form, and texture features were extracted from both types of plant images and combined (diseased affected or normal). The classification results revealed that the classification tree produces accurate results. The approach proposed in this study could also be used to identify and analyze plant disease images. For classification, increasingly advanced techniques such as adaptive neuro-fuzzy, neural networks, and genetic algorithms are available. For image recognition, support vector machines and other methods. These techniques can also be used to identify and analyze plant images [04].

Pujari, J. D., et.al.(2014) -The authors of this article presented an Local binary pattern based framework for addressing and categorizing fungal disease symptoms on both sides of vegetable crop leaves. The analysis indicate that the identification accuracy for the test samples using the Neuro KNN classifier was 91.54 percent. The identification accuracy was 84.11 percent when ANN with BPNN was used for both training and testing. One important observation was that for Local binary pattern, the classifier could have been trained with an ANN classifier and tested with a k-NN classifier. The goal of using k-NN to test the classifier seems to be to enable testing period shorter than training time. Longer training time is possible, but not testing time. To achieve real-time implementations, testing time must be kept to a minimum. The test results demonstrated that the Neuro-kNN technique was superior to the ANN approach in terms of supporting accurate fungal disease analysis with minimal computational determination [06].

Pujari, J. D.et.al.(2013) - The findings of this article presented a new technique that farmers can use to evaluate their crops, look for diseases early enough, then create health decisions, among other things. The use of a machine vision device to identify the symptoms of fungal diseases could help farmers evaluate their crops better effectively. They used commercial crop image samples that displayed visual symptoms of a fungal disease throughout this research. The affected region's features have been extracted and used as inputs to Mahalanobis distance and probabilistic neural network classifiers. The classification accuracy and conceptual speed increase for classifier training demonstrate that the probabilistic neural network classifier performs better in the presented approach [07].

Bauer, S. D., et.al. (2011) - The authors found that on single leaves of sugar beet crops, it's also able to differentiate between healthy leaf areas and those infected with Uromyces betae or Cercospora beticola. Uromyces betae would have an 86 percent recognition accuracy, Cercospora beticola would have a 91% classification performance, and the healthy leaf area would have a 94 percent classification performance. With perhaps the exception of Uromyces betae, they reached the 90% threshold in all categories. Use of neighborhood pixel information to improve the feature representation would have a beneficial effect on these assessment is a technique. Their study demonstrates typical pixel wise classifier defects, such as isolated pixels being misclassified and adjacent pixels being assigned to separate classes while belonging to the same class [08].

Cui, D., Zhang, Q.,et.al. (2010) - Two methods to multi spectral image processing for detecting soybean rust and its magnitude, either with or without manual threshold-setting, were explored in this research study. Two rust indices, RIA

and RCI, were positively associated with rust intensity levels, according to the findings of the manual threshold-setting approach. The investigation of the automated approach, that is, the approach that does not use image segmentation, yielded similar results. The centroid approach is more suitable for practical application in automatic rust detection, according to validation results obtained from a laboratory-scale test on 32 collected leaflets. Both image processing approaches could effectively detect rust severities, with the centroid approach being more suitable for practical application in automatic rust detection. However, more thorough research is required to confirm the accuracy of both proven methods in various environments and to create a reference database for automatic rust severity detection [09].

Rumpf, T., Mahlein, et.al. (2010) **-** The authors provided automated methods for detecting plant diseases early on, which is critical for precision crop safety. The key contribution of this paper is a technique focused on Support Vector Machines and spectral vegetation indices for early detection and differentiation of sugar beet diseases. The aim was to distinguish diseased from non-diseased sugar beet leaves, second is distinguish between Cercospora leaf spot, leaf rust, and powdery mildew, and last third detect diseases even before clear symptoms appeared. For a period of 21 days after inoculation, hyper spectral data were collected from healthy leaves and leaves inoculated with the pathogens Cercospora beticola, Uromyces betae, or Erysiphe betae, which cause Cercospora leaf spot, sugar beet rust, and powdery mildew, respectively. For an automated classification, nine spectral vegetation indices related to physiological parameters were used as features. A Support Vector Machine with a radial base function as kernel can make early distinctions between healthy and inoculated plants, as well as between specific diseases. The ability to distinguish between stable and diseased sugar beet leaves resulted in classification accuracy of up to 97 percent. Furthermore, the ability to detect plant diseases before they became symptomatic was demonstrated. The classification accuracy ranged from 65 percent to 90 percent depending on the form and stage of disease [09]**.**

IV. RESULT PARAMETERS

A. True Positives (TP)

True positive (TP) is the value that predicted correctly in event. There are two main conditions are originated in any event true positive and true negative. If the predicted positive value it mean that assessment of actual class is sure. That is define by TP [24].

B. True Negatives (TN)

True Negative (TN) is the value that predicted vale incorrect in event. There are two main conditions are originated in any event true positive (TP) as well as true negative (TN). If the predicted incorrect negative value it mean that assessment of actual class is not at all. That is denoted by TN [22].

C. False Positives (FP)

The false positive value that actual class is no and predicted class is yes. That is denoted by FP [25].

D. False Negatives (FN)

When actual class is yes but predicted class in no. That is denoted by FN [26].

E. Accuracy (Ac)

The detected part of plant as a disease part, that is known as a true positive (TP). The true negative (TN) is a non-affected leaf of plant detected [27].

$$\mathrm{Ac} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})} \quad (1)$$

F. Precision (Pr)

Precision is the relation of predicated true positive (TP) value in event and remark of the total predicted positive [28].

$$\mathrm{Pr} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \quad (2)$$

G. Recall (Sensitivity)

Recall is the relation of properly predicted positive observations to the all observations in actual class - yes. It is denoted by Re [29].

$$\mathrm{Re} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \quad (3)$$

H. F1 Score

F1 Score is ratio of the weighted average of Precision and Recall [30].

$$\mathrm{F1\_Score} = \frac{2 \times (\mathrm{Re} \times \mathrm{Pr})}{(\mathrm{Re} + \mathrm{Pr})} \quad (4)$$

TABLE – II COMPARISON OF DIFFERENT PREVIOUS METHODS

Ref. No.	Method Adopted	Advantage	Disadvantage	Accuracy
Singh, U.P., et.al[1]	MCNN (Multi-Scale Convolutional Neural Networks) for the classification	Performed in solving a number of plant leaves disease	MCNN, require large time for training data set.	97.13%
Bhimte, N. R., et.al [2]	k-means clustering	Early automatic detection of various type of diseases in plants.	The process of k-means clustering require Cluster section. That is semi auto learning process.	94.63%
Saponaro,P., et.al [3]	CNN (convolutional neural network) method	Frangi filter on real microscopy data, and found that the deep CNN is totally automatic method not require any cluster selection process.	MCNN, require large time for training data set.	77.3 %.
H. Sabrolet.al. [4]	Tomato Plant Illness Classification	otsu's method for segmentation phase computing faster	Preprocessing are required many steps	98.70%
Jagadeesh D. Pujari et al.[6]	Principle component analysis (PCA)	As the level of decomposition increases, it lowers the classification percentage so to manage the classification percentage PCA is used.	Loss of information while compressing the data found to reduce the number of dimensions.	86.48%.
Jagadeesh D. Pujari et al.[7]	Neuro-Knn	Strong method to noisy training data and active if the training data is huge.	Chan vase segmentation is used which was based on an active contour model, working process is slow for large image size and also not capable to segment nearest objects.	91.54%
S D Bauer et al.[8]	Probabilistic neural network	It takes less time to train the system and it has good extension properties.	It requires large memory space and slow execution of the network.	88.59%
Dheeb Al Bashish et al.[18]	Neural network classifier	Efficient in training and good result in test.	Slow in training process also a time consuming	93%
Huang KY et al[16]	Back propagation neural network and GLCM feature extraction	Very easy to implement and able to form difficult nonlinear mapping.	It is difficult to find the required number of neurons and layers, Learning process is also slow.	97.20%
Byadgi AS et al[12]	Segmentation technique-k-means, Classifiers- Artificial neural network and SVM	SVM has a simple geometric definition and it is robust when the training sample has some discrimination.	Training process is slow and difficult to understand the algorithmic structure.	87%

Pujari JD et al[21]	Statistical features like block wise, GLCM GLRM classifier used is nearest neighbor	Co-occurrence matrix is very useful for large and dispersed data sets.	Co-occurrence matrix is sensitive to rotation, which will result a different co-occurence matrix of the same (rotated) image.	91.37%
Anand.H.Kulkarni et al[23]	Gabor filter for feature extraction and Artificial neural network classifier.	Accurate result for texture representation and discrimination, their representations are similar to the human visual system.	Output of Gabor filter is not mutually orthogonal.	91%

V. Conclusion and Future Work

Plants are the main origin for resolving issues such as hunger, carbon dioxide emissions, and climate change [31]. Crop production is a main of revenue for the many of farming community and is important to supply ever-increasing societies. Invasive diseases are present a growing threat to plants. Rapidly evolving, re-emerging, epidemic, and resistant bacteria are causing chaos on crop plants, eventually due to the financial declines[32][33]. Weeds are increasing sharply all across globe, causing serious damage to the plant's normal development becoming one of the major causes of financial decline by decreasing the concentration of agricultural production. Grain farmers and producers require systems that can help us identify early symptoms of pathogenic microorganisms by evaluating digital images of crop collections. By use of an image recognition users to detect the clinical signs of plant diseases as well as provide a disease prevention method may help the farmers in their daily struggle against infectious diseases. It's the first step in developing a smart support for recognizing crop growth as well as effective cure options. In those other words, image pre-processing allows for the collection of characteristic parameters that are not influenced by context, leaf shape and size, light, or camera, as well as the establishment of a good base for successful characteristic parameters for disease diagnosis and pattern recognition systems.

REFERENCES

[1] S. Amar Singh, Uday Pratap, et al. "Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease." IEEE Access 7 (2019).

[2] Bhimte, Namrata R., and V. R. Thool. "Diseases Detection of Cotton Leaf Spot using Image Processing and SVM Classifier." 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2018.

[3] Saponaro, Philip, et al. "Three-dimensional segmentation of vesicular networks of fungal hyphae in macroscopic microscopy image stacks." 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017.

[4] Sabrol, H., and K. Satish. "Tomato plant disease classification in digital images using classification tree." 2016 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2016.

[5] Singh, Vijai, and A. K. Misra. "Detection of unhealthy region of plant leaves using image processing and genetic algorithm." 2015 International Conference on Advances in Computer Engineering and Applications. IEEE, 2015.

[6] Pujari, Jagadeesh D., Rajesh Yakkundimath, and Abdulmunaf S. Byadgi. "Neuro-kNN classification system for detecting fungal disease on vegetable crops using local binary patterns." Agricultural Engineering International: CIGR Journal 16.4 (2014).

[7] Pujari, Jagadeesh D., Rajesh Yakkundimath, and A. S. Byadgi. "Reduced color and texture feature-based identification and classification of affected and normal fruits images." International Journal of Agricultural and Food Science 3.3 (2013): 119-127.

[8] Bauer, Sabine D., Filip Korč, and Wolfgang Förstner. "The potential of automatic methods of classification to identify leaf diseases from multispectral images." Precision Agriculture 12.3 (2011): 361-377.

[9] Cui, Di, et al. "Image processing methods for quantitatively detecting soybean rust from multispectral images." Biosystems engineering 107.3 (2010): 186-193.

[10] Rumpf, T., et al. "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance." Computers and electronics in agriculture 74.1 (2010): 91-99.

[11] Miao, Fengiuan, et al. "Crop Weed Identification System Based on Convolutional Neural Network." IEEE 2nd International Conference on Electronic Information and Communication Technology (2019): 595-599.

[12] Sankaran, Sindhuja, et al. "A review of advanced techniques for detecting plant diseases." Computers and Electronics in Agriculture 72.1 (2010): 1-13.

[13] Mewes, Thorsten, et al. "Derivation of stress severities in wheat from hyperspectral data using support vector regression." 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. IEEE, 2010.

[14] Camargo, A., and J. S. Smith. "Image pattern classification for the identification of disease-causing agents in plants." Computers and Electronics in Agriculture 66.2 (2009): 121-125.

[15] Franke, Jonas, Thorsten Mewes, and Gunter Menz. "Requirements on spectral resolution of remote sensing data for crop stress detection." 2009 IEEE International Geoscience and Remote Sensing Symposium. Vol. 1. IEEE, 2009.
[16] Huang, Kuo-Yi."Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features." Computers and Electronics in agriculture 57.1 (2007): 3-11.
[17] Qin, Zhihao, et al. "Remote sensing analysis of rice disease stresses for farm pest management using wide-band airborne data." IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477). Vol. 4. IEEE, 2003.
[18] Zhang, Minghua, et al. "Detection of stress in tomatoes induced by late blight disease in California, USA, using hyper spectral remote sensing." International Journal of Applied Earth Observation and Geo information 4.4 (2003): 295-310.
[19] Chen, Yud-Ren, Kuanglin Chao, and Moon S. Kim. "Machine vision technology for agricultural applications." Computers and electronics in Agriculture 36.2-3 (2002): 173-191.
[20] Muhammed, Hamed Hamid. "Using hyperspectral reflectance data for discrimination between healthy and diseased plants, and determination of damage-level in diseased plants." Applied Imagery Pattern Recognition Workshop, 2002. Proceedings. IEEE, 2002.
[21] Yang, Yoon Seok, et al. "Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network." IEEE Transactions on Biomedical Engineering 48.6 (2001): 718-730.
[22] Orillo JW, Cruz JD, Agapito L, Satimbre PL, Valenzuela I (2014) Identification of diseases in rice plant (Oryza Sativa) using back propagation artificial neural network. In: IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, pp.1–660.
[23] Phadikar S, Sil J (2008) Rice disease identification using pattern recognition techniques. In: 11th international conference on computer and information technology (ICCIT 2008),
[24] Khulna, 24-27 December. pp. 420–423 Abdullah NE, Rahim AA, Hashim H, Kamal K (2007) Classification of rubber tree leaf diseases using multilayer perceptron neural network. In: Fifth student conference on research and development (SCORed), Selangor, 11-12 December. pp. 1–6.
[25] C. H. Bock, P. E. Parker, A. Z. Cook, T. R. Gottwald, "Visual Rating and the Use of Image Analysis for Assessing Different Symptoms of Citrus Canker on Grapefruit Leaves", 2008.
[26] Noor Ezan Abdullah, Athirah A. Rahim, Hadzli Hashim and Mahanijah Md Kamal, "Classification ofRubber Tree LeafDiseases Using Multilayer Perceptron Neural Network", The 5tn Student Conference on Research and Development -SCOReD 2007, 11-12 December 2007, Malaysia.
[27] Kuo-Yi Huang, "Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features", Computers and Electronics in Agriculture, ELSEVIER, 2007.
[28] G. Corkidi, K. A. Balderas-Ruíz, B. Taboada, L. Serrano-Carreón and E. Galindo , "ackwell Publishing Ltd Assessing mango anthracnose using a new three-dimensional image-analysis technique to quantify lesions on fruit", Plant pathology, 2006.
[29] Maria M. Lo´pez , Edson Bertolini, Antonio Olmos, Paola Caruso, Maria Teresa Gorris, Pablo Llop, Ramo´n Penyalver, Mariano Cambra, "Innovative tools for detection of plant pathogenic viruses and bacteria", Springer-Verlag and SEM 2003.
[30] James W. Olmstead, Gregory A. Lang, Gary G. Grove, "Assessment of Severity of Powdery Mildew Infection of Sweet Cherry Leafsby Digital Image Analysis", HORTSCIENCE, 2001.
[31] Irfan S. Ahmad, John F. Reid, Marvin R. Paulsen, James B. Sinclair, "Color Classifier for Symptomatic Soybean Seeds Using Image Processing", The American Phytopathological Society,1999.
[32] Darrin P. Martin and Edward P. Rybicki, "Microcomputer-Based Quantification of Maize Streak Virus Symptoms in Zea mays", The American Phytopathological Society, 1998.
[33] S.E. Lindow, R.R. Webb, "Quantification of foliar plant disease symptoms by microcomputer-digitized video image analysis", The American Phytopathological Society, 1983.
[34] Rebert M. Haralick, K. Shanmugam, Its'hak Dinstein, "Textual feature of image classification", IEEE, 1973.
[35] Jayme Garcia Arnal Barbedo, "Digital image processing techniques for detecting, quantifying and classifying plant diseases ", Springer, 1970.

AUTHORS PROFILE

First Author Ila Sharma received the Masters of Science and Bachelor of Science degree in Computer Science respectively from Samrat Ashok Technological Institute, Vidisha and Barkatullah University, Bhopal, Madhya Pradesh, India.
Currently she is working toward the Ph.D. degree in discipline of Computer Science and Information Technology at Rabindranath Tagore University, Bhopal, India.

Second Author Dr. Varsha Jotwani is currently working as Associate Professor with Rabindranath Tagore University.

She is PhD in Computer Applications. She has vast teaching and academic developments at leading institution of Bhopal, India. She has published various International and National Research Papers in the high quality journals. She is also well versed in developing curriculum for Undergraduate and Postgraduate students under the field of Information and Technology.

Forecasting UBER demand using SARIMAX Model and ARIMA Model

Jayashree M Kudari
Associate Professor, CS &IT, Jain University

Abstract:
Booking a ride from Uber is continuous. Just open the app, set the pickup location, request a car, get picked up and pay with the tap of a button. But there's more than what catches the eye. There's a lot of data wrangling around to make all of this happen in such a flawless (mostly) process.
Taxi Services today have made it significantly convenient in opting for their services. This is generally done using the enormous volume of data that is collected. With the help of data science and machine learning that data can be analysed and demand can be predicted. Time series forecasting using ARIMA is one of the popular model's used for such predictions. By analysing the data and forecasting trends can be understood and benefit from estimating and advancement strategies. SARIMAX model perform well on detecting the seasonality of the time series, and poorly when it comes to detecting the variations (trends) values, this can be explained as anomaly detection in some problems. For Analysing the peak hour demand, different hours of the day are plotted to the average number of rides for the months. We will perform the well-known statistical method SARIMAX on the aggregated data. Time-series forecasting is one of the sizzling investigation topics in the pitch of data science. Numerous models and methodologies have been confirmed from statistical methods to deep neural networks.[14]

Key Terms: Time series, Machine Learning, ARIMA, Uber, Artificial Intelligence, Data, Prediction, SARIMAX.

Introduction

Traditional cab methods in cities often suffer from inadequacies due to uncoordinated actions as customer demand changes [7]
Enormous database of drivers used by Uber, after a person request for a cab, Uber's algorithm gets to work in a matter of seconds, it matches a person with the driver nearby. [1] It is in the background where Uber is storing all data for every trip taken, even when the driver has no passengers. All of this data is stored and analysed to predict demand and supply, and most prominently pricing. Uber also looks at how transportation is handled across cities and tries to alter for bottlenecks and other common issues. Uber gathers data on its drivers too, in addition to collecting non-identifiable information about their vehicle and their location. [1]
Uber uses the personal data in an anonymous and aggregated way to closely monitor which features of the Service you use the most, to analyse trends and determine where they could possibly focus more on.
While the constant idea of data being misused is biting Uber, there's no denying that the anonymous, aggregated data that they collect insights from is just amazing.
All of this data is collected, analysed and used to predict everything from the
customer's wait time, to recommending where drivers should place themselves through heat maps in order to take benefit of the best charges and maximum passengers. All of these items are applied in present for both drivers and passengers. [2]

Uber's biggest use of data comes in the form of surge pricing, if you're late to an appointment and you need to book a ride in a crowded downtown space, be prepared to pay almost twice as much for it. Using the data these peak hours are what they benefit most out of. Over the course of one night the fares can go from 200INR to 2000INR, while still having riders.

This kind of dynamic pricing scheme is similar to the strategy used by hotels and flights for their weekend or holiday fares and rates. Except Uber makes use of predictive modelling based on real-time traffic

patterns, demand and supply. It has also patented this type of pricing.

Surge pricing considerably, in a long run does affect the rate of demand, while long-term use could be the key to retaining or losing customers. Customer backlash on hiking rates is strong, so Uber has considered using machine-learning and artificial Intelligence algorithms to predict where demand will be strong, so that drivers can prepare to meet that demand, and surge pricing will be significantly reduced.[1] Uber knows that in order to get and maintain a strong customer and driver base, it needs to put data to work for it in new and advanced ways. [1]

Literature Review

Urban liveability is a key concept in the New Urban Agenda (NUA) adopted by the United Nations (UN) in 2016. The UN has documented that effective benchmarks and watching mechanisms are necessary for the successful implementation of the NUA. [4]
For standard modelling is created to forecasts using only time-lagged features. author used the three techniques, with the goal of progressing further with the most promising forecast. [4]. Linear Regression, Seasonal Autoregressive Integrated Moving Average with exogenous repressors (SARIMAX), Facebook Prophet. Models were evaluated using root-mean squared error (RMSE). This way, the error metric would be easily understandable as "number of pickups" the forecast was off by. [5]

Author argued that ARIMA based prediction is not the best solution. [8]

Neighbourhood-level data does help with time series forecasting, but it can also introduce additional noise. Long Short Term Memory (LSTM) Neural Networks perform rather well with Time Series forecasting. [5]

Time series have different time-based consistency. Few algorithms are easy to predict. Given a predictive algorithm such as LSTM or ARIMA (time series), the maximum prediction accuracy that it can reach if it captures all the time-based patterns of that time series. And, given the maximum predictability, which algorithm could approach the upper bound in terms of prediction accuracy? To answer these two question, author use temporal-correlated entropy to measure the time series regularity and obtain the maximum predictability. [6]

Researchers implemented and compare the prediction accuracy of five commonly used and representative predictors and examine their performance under different maximum predictability [9], sequence modelling [10], the auto-regressive integrated moving average (ARIMA) model (time series forecasting) [11], the Neural Network (NN) (machine learning) [13], and the Long Short-Term Memory (LSTM) (deep learning) [12]. Their results indicate that the maximum predictability is an approachable target for the actual prediction accuracy by using LSTM predictors, it provides better accuracy.
LSTM [12] is a Recurrent Neural Network designed specifically for modelling sequential and time series data as it can capture long-term dependency.

Analysing the Data

Six months of UBER data taken from www.kaggle.com Is going to be used. This data consists of pickup date/time along with location and the company code associated with it. Using this data demand can be analysed and predicted, The peak hours and days for the following month.

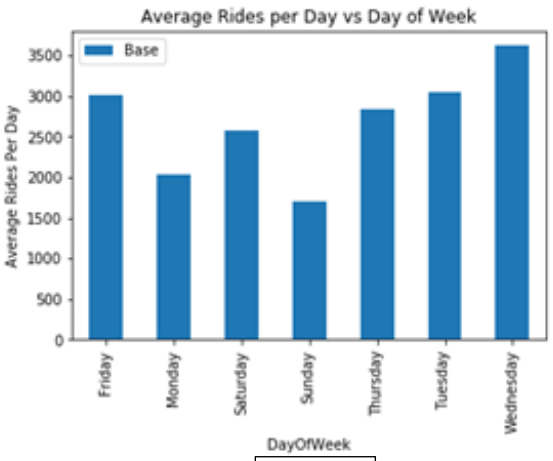


fig 1.1

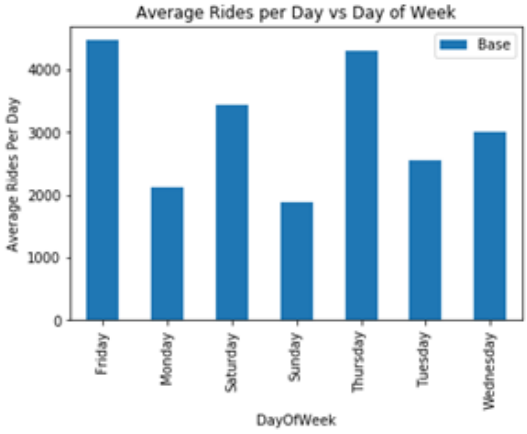


Fig 1.2

Fig 1.1 and Fig 1.2 show demand in rides for april and may respectively,consisting of weekday average of different days in a week and the average number of rides.We can see the variation being not so significant.

For Analyzing the peak hour demand, different hours of the day are plotted to the average number of rides for the months april and may.

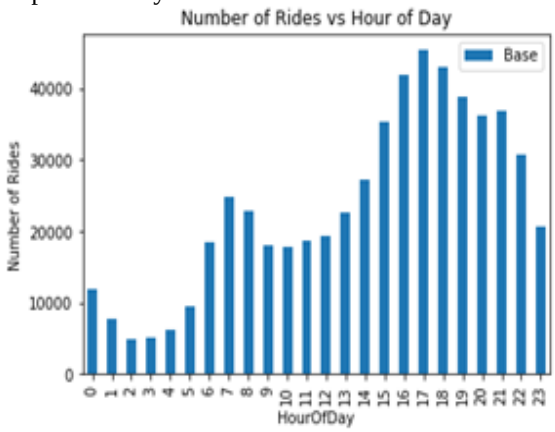


Fig 1.3

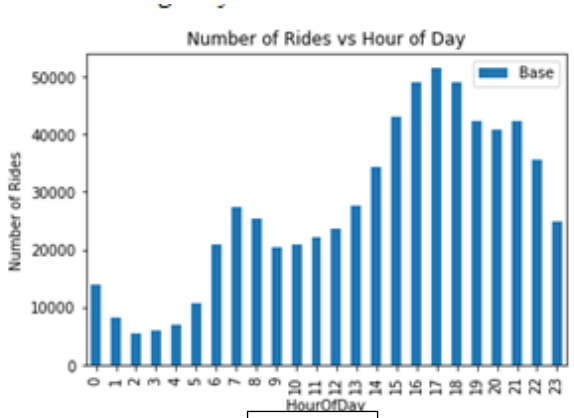


Fig 1.4

By looking at Fig 1.3 and Fig 1.4 high demand can be seen during a particular time of the day.

September will be emphasized upon since prediction will be made for the month of october, highlighting the variation in demand.

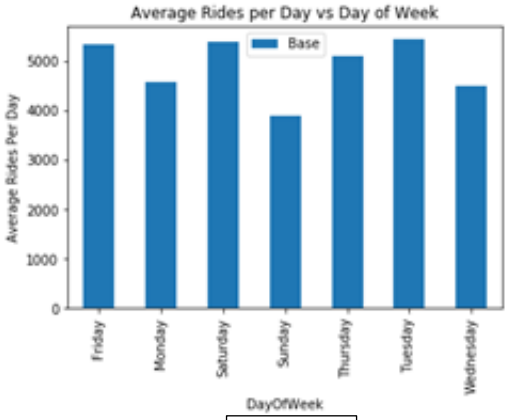


Fig 1.5

Time Series Prediction Using ARIMA

Making predictions about the future is known as extrapolation in the classical statistical handling of time series data. Most focus on the topic and refer to it as time series forecasting. Forecasting often involves taking models fit on historical data and using them to predict future.

An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened (i.e., collected data)

ARIMA stands for Auto Regressive Integrated Moving Average. There are seasonal and Non-seasonal ARIMA models that can be used for forecasting.

ARIMA is going to be used to predict the demand for the month of October using the data collected for six months from august to September. Fig 1.6 shows the actual data plot.

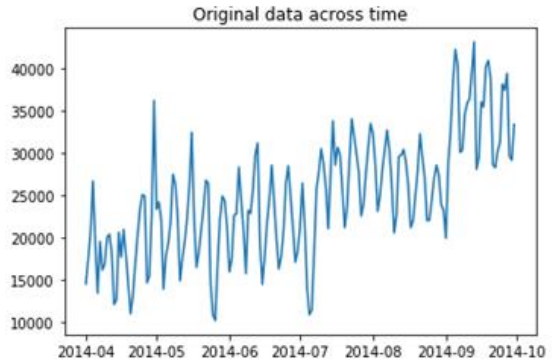


Fig 1.6

Using the ADF test we can tell that the data is not stationary. Hence differentiation is required. We will use a differentiation of 1.

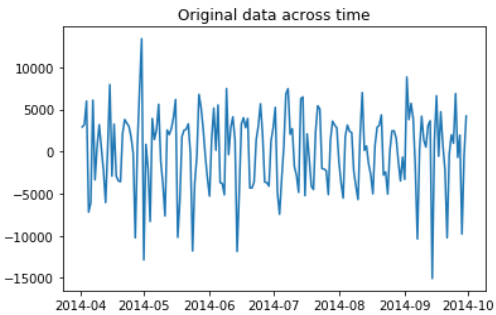


Fig 1.7

The series is now stationary with a confidence level of 95% (Fig 1.7). Below are the autocorrelation and partial autocorrelation plots. From the ACF and PACF plots we can see a clear spike at every 7-day interval. Since this appears clearly in the ACF plot it shows a seasonal MA (moving average) component of 1. [3]

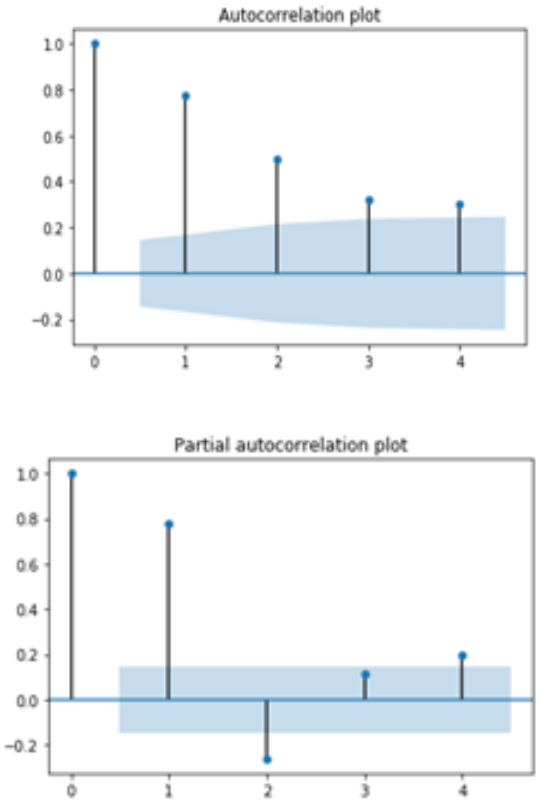


Fig 1.7 shows the predicted data with the actual data. Data for seven days for the month of October has been predicted above. The blue line shows the predicted data. The predicted data does not vary much from the actual data. The RSME Values have been a major deciding factor here forecast has been done with ARIMA (0,1,0) (0,1,2) [7] for best fit and lowest RSME.

Fitting SARIMAX Model and Forecasting

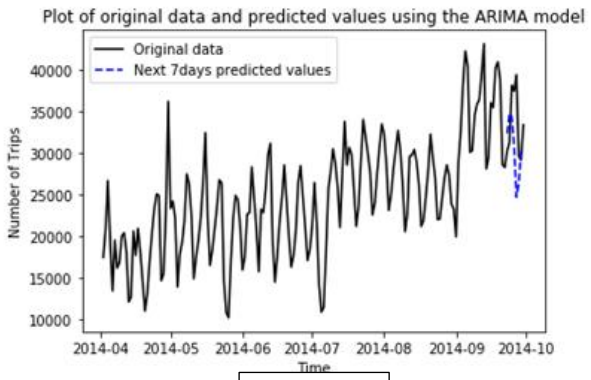


Fig 1.8

In Fig-18 SARIMAX model perform well on detecting the seasonality of the time series, and poorly when it comes to detecting the variations (trends) values, this can be explained as anomaly detection in some problems.

Conclusion

Using this Seasonal ARIMA model we can utilize the forecasted data to plan and alter our marketing strategy. With the generated data peak days and hours can be found. These trends can be used to benefit through surge pricing and promotions.

By analysing the collected data, peak hours and peak days can be found. This can be used to narrow down on the user base. This in turn will help target specific audience and help with promotional techniques. Data plays a very significant role here. The more the data the more accurate the prediction will be.

In future, SARIMAX model can be improved by rearrange the data into smaller time steps in order for the model to better detect the variations, Tune the SARIMAX model on more values and by Clustering the demands into more regions.

References

1. https://neilpatel.com/blog/how-uber-uses-data/
2. https://theonlinemarketingtoday.blogspot.com/2017/04/how-uber-uses-data-to-improve-their.html
3. https://www.kaggle.com/kruthik93/utilizing-arima-to-forecast-uber-s-market-demand
4. A Preliminary Exploration of Uber Data as an Indicator of Urban Liveability June 2019, DOI: 10.1109/CyberSA.2019.8899714, Conference: IEEE Cyber Science 2019At: University of Oxford, Project: Irish Institute of Digital Business (IIDB), Lab: Luiz Affonso Guedes's Lab, Aguinaldo BezerraAguinaldo BezerraGisliany AlvesGisliany, lvesIvanovitch SilvaIvanovitch SilvaShow all 6 authorsTheodore Gerard LynnTheodore Gerard Lynn.
5. Forecasting Uber Demand in NYC, Ankur Vishwakarma, Mar 24, 2018.
6. Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability Kai Zhao, Denis Khryashchev, Huy Vo.
7. Y. Huang and J. W. Powell, "Detecting regions of disequilibrium in taxi services under uncertainty," in SIGSPATIAL'12, Redondo Beach, CA, USA, November 7-9, 2012, 2012, pp. 139–148.
8. L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi-passenger demand using streaming data," IEEE Trans. Intelligent Transportation Systems, vol. 14, no. 3, pp. 1393–1402, 2013.
9. X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility," Scientific. Reports, vol. 3, Oct. 2013.
10. C. T. Cheng, R. Jain, and E. van den Berg, "Mobile wireless systems: Location prediction algorithms," in Encyclopedia of Wireless and Mobile Communications, 2008.
11. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: forecasting and control. John Wiley & Sons, 2015.
12. Hochreiter and J. Schmidhuber, "Long short-term memory,"Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
13. N. Mukai and N. Yoden, Taxi Demand Forecasting Based on Taxi Probe Data by Neural Network. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 589–597.
14. NewYork Taxi demand forecasting with SARIMAX using weather data How to predict future taxi demands in NewYork city using SARIMAX statistical model. By Abdelkader BOUREGAG

Spectral Band Combinations for Land Cover Classification of Satellite Images

[1]Keerti Kulkarni, [2]Dr. P. A. Vijaya

[1]Asst. Professor, Dept of ECE, BNM Institute of Technology, Bangalore

[2] Professor and Head, Dept of ECE, BNM Institute of Technology, Bangalore

Corresponding Author : Keerti Kulkarni. Email : keerti_p_kulkarni@yahoo.com

Abstract: **Land cover classification is one of the important applications of the satellite images. The accuracy of the classification process depends on the feature selection. In multispectral satellite images, the separability of the features depends on the band combinations used. This work demonstrates the change in the accuracy of the classifiers with different band combinations. Landsat-8 images have been classified using the Maximum Likelihood Classifier. In this work, five different band combinations, namely 3-2, 4-3-2, 3-2-1, 7-6-5-4, 7-6-5-4-3, have been considered. The spectral separability of each of the land cover classes is analysed for each of the band combinations using the Jeffries-Matusita Distance measure. It is shown that maximum separability and hence the optimal accuracy of 75.81% is obtained with a three-band combination of 4-3-2.**

***Keywords*: Maximum Likelihood (ML) Classifier, Euclidean Distance, band combination, landsat images.**

I. Introduction:

Machine learning algorithms have been often used for the land cover classification of satellite images. The satellite images are either multispectral (3-10 bands) or hyperspectral (100s of bands). Each land cover class has a unique spectral signature. The spectral signatures of the different land cover types serve as the features for discriminating the various classes. For a learning algorithm to produce accurate results, the features have to be separable. This separability depends on how far apart the spectral signatures are placed in an N-dimensional space. Here, N denotes the number of bands under consideration. The spectral distance of the classes can be evaluated by a variety of distance measures, such as Euclidean distance, Bhattacharya distance and the Jeffries-Matusita distance. It is important to measure the separability of the classes (by using the distance metric) in each on the band combination, before deciding which one is to be used. For the same classifier, keeping all the other parameters constant, a band combination with the greatest distance measure between the classes (greatest separability) will produce the highest classification accuracy. In this work, an urban region of

Bangalore District is classified into 4 different land cover types, namely, water, vegetation, built-up and soil. It is important to understand the implications of the spectral signatures in a heterogenous raw satellite image such as an urban region. Here, the problem of mixed pixels creates overlapping spectral signatures in one band. Hence, when a band combination is used rather than a single band, the spectral signatures tend to be more distinct. Different band combinations give different distance measures between the classes. The band combination which has the highest separability (highest distance) between the classes, is then used as the feature vector for the Maximum Likelihood classifier. The accuracy is maximum when a 3-band combination is used as is depicted in the bar graph.

II. Literature Survey

Generally, a land cover classification problem deals with the sematic segmentation of the raw satellite images. The first step in this kind of sematic segmentation is the preparation of labelled training samples [1, 2]. These classes are labelled depending on their separability index [3, 4]. This index is a discriminating factor, which is helpful in classifying the class of a pixel. Machine learning algorithms for this classification can be unsupervised when there is no spectral information available [5]. When the spectral signature information is available, it can be exploited to discriminate between the classes [6, 7]. Supervised classification also assumes that each of the spectral class can be described by a probability distribution function [8,9]. Each of the bands of the LANDSAT potentially contributes to the information required for the land cover classification. But there will be redundance in the information provided by each band, if the bands themselves are highly correlated [10]. Hence, we need to use only a small subset of all the bands available. The identification of the subsets to be used can be categorized in two basic approaches. The analysis is based either on calculating the eigen values (and eigen vectors) [11] or separability analysis [12]. Separability analysis deals with calculating the statistical distance between the spectral classes. A variety of measures for calculating the distance are available in literature [13, 14]. Each band has its own unique use for the identification of the land cover. For example, Band 1 can identify water, whereas band 4 can identify the urban areas. Near infrared band can be used to identify the vegetation [15, 16]. The performance of the classifier may actually degrade if the number of features (number of bands) are increased [17]. Although quite a few similarity measures have also been discussed in the literature [18], the authors in [19] have shown that the accuracy of the distance measures depend on the domain and the problems statement.

III. Methodology

The methodology adopted for the land cover classification is shown in Figure 1. The raw

satellite images are the LANDSAT-8 images which has 11 bands. Band 1 to band-7 and band-9 have a resolution of 30 meters, band-8, which is a panchromatic band has a resolution of 15 meter and band-10 and band-11 have a resolution of 100 meters.

The study area is the Bangalore Urban district. The land cover is classified into four different categories, namely water, vegetation, built-up and soil.

The first step after downloading the required dataset is correcting the raw images for the atmospheric defects. This constitutes the pre-processing step. Dark Object Subtraction – 1 is a standard algorithm used for atmospheric correction. One more important reason for the atmospheric correction is that the separability between the classes is enhanced after the procedure, because of which histogram equalization need not be done.

Spectral signature of the land cover class is unique and hence acts as the distinguishing factor for the classification. Different land cover classes have different spectral signatures. For example, water reflects only up to 10% of the incident energy back to the sensor, hence it appears as a black form on the satellite image. Vegetation reflects 30% - 45% of the incident energy, soil reflects 20% - 30% of the incident energy and built-up areas reflect 10% - 20% of the incident energy back to the sensors. Even though this is a feature for the classification, the reflected energies and hence the spectral signatures are more pronounced in some bands as compared to some other bands. Hence the need for the spectral signature separability analysis for different band combinations in an N-dimensional space.

Spectral Separability analysis is carried out by calculating the Jeffries-Matusita (JM) distance between all the individual class probability distributions for different band combinations. JM distance calculates the separability between a pair of probability distribution of the training classes. The formula for Jeffries-Matusita Distance is given by

$$\mathbf{JM_{ci,cj}} = \sqrt{\mathbf{2(1 - e^{-B})}} \quad \text{Equation 1}$$

Where B is the Bhattacharya distance given by

$$\mathbf{B} = \frac{1}{8}(\boldsymbol{\mu_i} - \boldsymbol{\mu_j})^{\mathbf{T}} \left(\frac{\mathbf{C_i} + \mathbf{C_j}}{2}\right)^{-1} (\boldsymbol{\mu_i} - \boldsymbol{\mu_j}) + \frac{1}{2}\mathbf{ln}\left(\frac{\left|\mathbf{C_i} + \mathbf{C_j}/2\right|}{\sqrt{|\mathbf{C_i}| + |\mathbf{C_j}|}}\right) \quad \text{Equation 2}$$

Where

i and j = the two signature classes being compared

$\mathbf{C_i}$ = Covariance matrix of signature **i**

$\mathbf{C_j}$ = Covariance matrix of signature **j**

$\boldsymbol{\mu_i}$ = mean of signature **i**

$\boldsymbol{\mu_j}$ = mean of signature **j**

$|C_i|$ = determinant of $\mathbf{C_i}$

The JM distance gives a minimum value of 0 when the signatures are similar and the maximum value of √2, when the signatures are very distinct. The main advantage of using the JM distance is that it tends to suppress the high separability values and the low separability values are emphasized. This gives a fairly better idea of the separability, compared to the Euclidean distance, where the minimum values is 0, but the maximum value is unbounded.

Maximum Likelihood Classifier works on the principle of calculating the probability distribution of each pixel. If the probability that a pixel belongs to **class-i** is greater than the probability that the pixel belongs to **class-j**, then it is classified as belonging to **class-i**.

The discriminant function, is calculated for every pixel as:

$$\mathbf{g_k(x)} = \mathbf{\ln p(C_k)} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{\ln |\sum k|} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{(x - y_k)^t \sum_k^{-1} (x - y_k)} \quad \text{Equation 3}$$

Where,

C_k = land cover class k;

x = spectral signature vector of a image pixel;

$p(C_k)$ = probability that the correct class is Ck;

$|\sum k|$ = determinant of the covariance matrix of the data in class Ck;

$\sum_k^{-1}(x - y_k)$ = inverse of the covariance matrix;

y_k = spectral signature vector of class k.

Therefore:

$$\mathbf{x \in C_k \leftrightarrow g_k(x) > g_j(x) \forall\, k \neq j}$$

Equation 4

The discriminant function for the ML classifier is shown in Figure 2

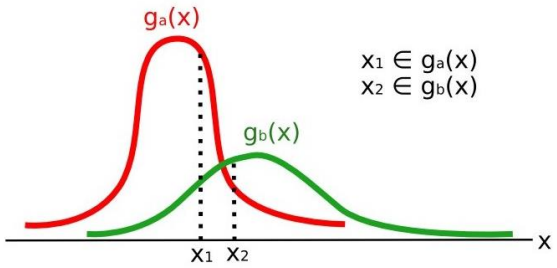


Figure 2: Discriminant Functions

Finally, the classification accuracy of the classifier for all the five band combinations is calculated. Classification accuracy is defined as the ratio of the total number of pixels correctly classified to the total number of pixels.

The visualization of the 2-band and the 3-band combination (2D and 3D respectively) is easier than that for the higher dimension. Nevertheless, the spectral signature of a pixels is a point in the N-

dimensional space, where N depends on the number of bands chosen.

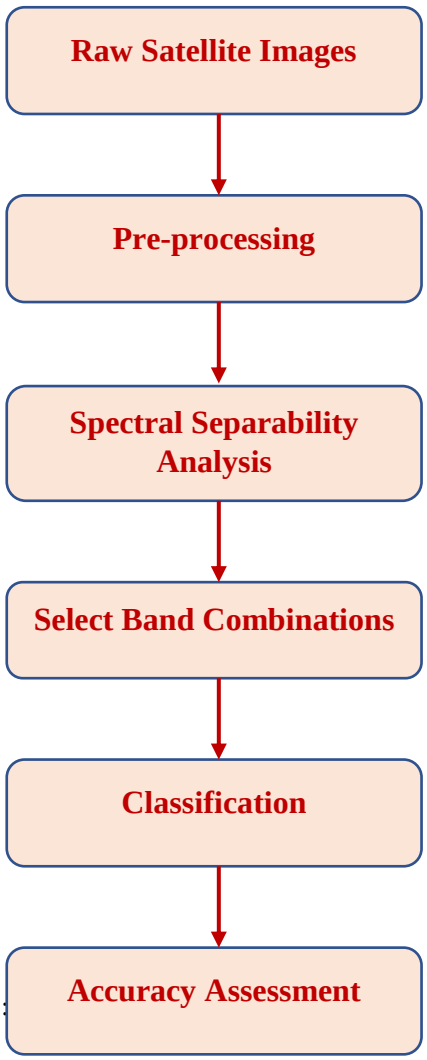


Figure 1:

IV. **Results**

Figure 2 and Figure 3 shows the spectral signatures before and after the atmospheric correction. It is easier to note the separability in the second case. Also, Figure 3 shows the spectral signatures plot for a band combination of 4-3-2. Similar spectral signature plots are obtained for the other band combinations.

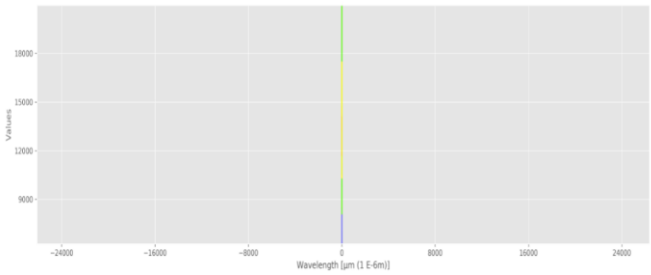


Figure 2: Spectral Signatures before the Atmospheric Correction

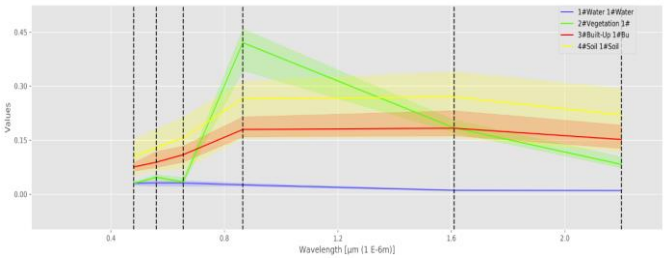


Figure 3: Spectral Signatures after the Atmospheric Correction

The Table 1 shows the Jeffries-Matusita distance or the separability matrix between the classes for a two-band combination of 3-2. The value for the same class should ideally be zero, but a small positive value is obtained because of the mixed pixels.

Table 1: Separability Matrix for a 2-band combination (3-2)

	Water	Vegetation	Built-Up	Soil
Water	0.02	1.1	1.19	1.05
Vegetation	1.1	0.1	1.2	0.96
Built-Up	1.19	1.2	0	0.91
Soil	1.05	0.96	0.91	0.02

Table 1: Separability Matrix for a 3-band combination (4-3-2)

	Water	Vegetation	Built-Up	Soil
Water	0.01	1.23	1.35	1.12
Vegetation	1.23	0.1	1.35	1.01
Built-Up	1.35	1.35	0	1.16
Soil	1.12	1.01	1.16	0.01

Table 1: Separability Matrix for a 3-band combination 3-2-1

	Water	Vegetation	Built-Up	Soil
Water	0.01	1.20	1.29	1.01
Vegetation	1.20	0.1	1.26	0.91
Built-Up	1.29	1.26	0	1.11
Soil	1.01	0.91	1.11	0.01

	Water	Vegetation	Built-Up	Soil
Water	0.1	0.83	1.01	0.97
Vegetation	0.83	0.1	0.93	0.75
Built-Up	1.01	0.93	0.1	0.81
Soil	0.97	0.75	0.81	0.1

Table 1: Separability Matrix for a 4-band combination 7-6-5-4

	Water	Vegetation	Built-Up	Soil
Water	0.01	1.1	1.12	0.92
Vegetation	1.1	0	0.96	0.73
Built-Up	1.12	0.96	0.01	0.75
Soil	0.92	0.73	0.75	0.1

Table 1: Separability Matrix for a 5-band combination 7-6-5-4-3

Figure 4 shows the result of the ML classifier using the band combination of 4-3-2.

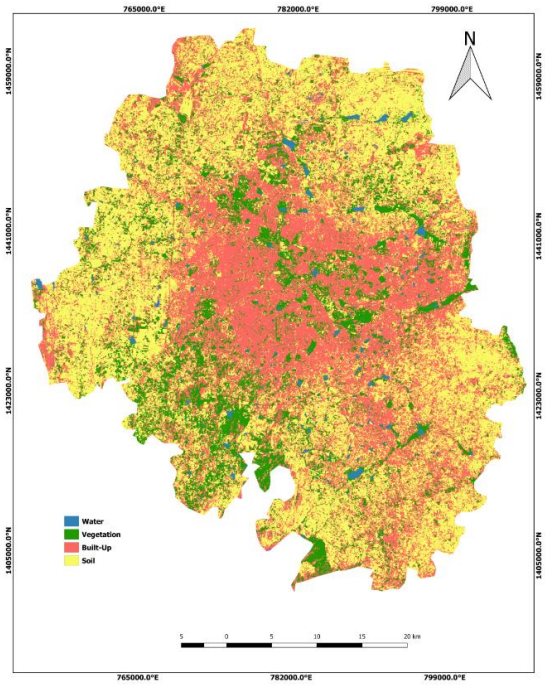
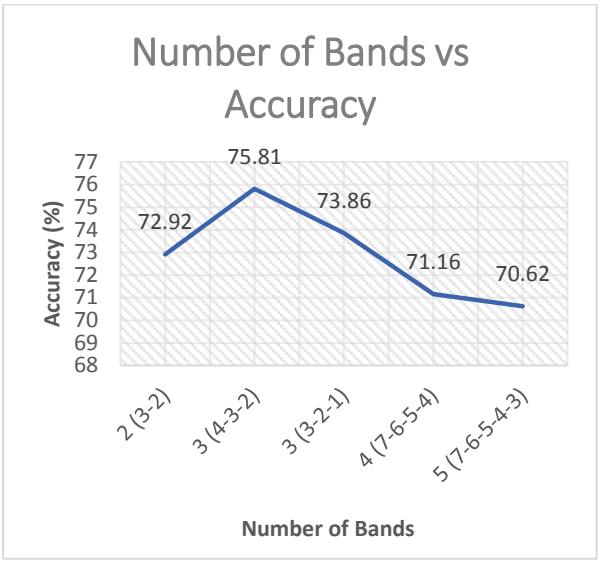


Figure 4: land cover map using the ML classifier with a 4-3-2 band combination.

With the five different band combinations, a change in the accuracy is also plotted as shown in Figure 5. It is seen that when the number of bands are increased from 2 to 3, there is an increase in the accuracy, but any further increase in the number of bands, decreases the accuracy.



V. Discussion and Conclusion

The separability tables show that a maximum separability between the spectral signatures is obtained with a band combination of 4-3-2. This is logical also, because the bands 4, 3 and 2 represent the RGB or the visual bands. Hence the visual interpretation of the separability of the signatures is better in this band combination. Even if the number of bands increase, the separability of the signatures is not improved. Hence, using the band combination of 4-3-2, the accuracy obtained from a ML classifier is 75.81%.

The improvement in the accuracy can be further explored by using a different distance measure. Also, non-parametric approaches to the classification process can improve the classification accuracy.

References

[1] Rogan, J., et al., Land-cover change monitoring with classification trees using Landsat TM and ancillary data. Photogrammetric Engineering & Remote Sensing, 2003. 69(7): p. 793-804.

[2] Elhag, M., A. Psilovikos, and M. Sakellariou, Detection of land cover changes for water recourses management using remote sensing data over the Nile Delta Region. Environment, Development and Sustainability, 2013. 15(5): p. 1189-1204.

[3] Friedl, M.A., et al., MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote sensing of Environment, 2010. 114(1): p. 168-182.

[4] Friedl, M.A., et al., MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote sensing of Environment, 2010. 114(1): p. 168-182.

[5] Congalton, R.G. and K. Green, Assessing the accuracy of remotely sensed data: principles and practices. 2008: CRC press.

[6] Keerti Kulkarni, Dr. P. A. Vijaya, "Parametric Approaches to Multispectral Image Classification using Normalized Difference Vegetation Index", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-2S, December 2019.
DOI: 10.35940/ijitee.B1061.1292S19, pp 611-615.

[7] Keerti Kulkarni, Dr. P. A. Vijaya, "Experiment of Multispectral Images using Spectral Angle Mapper Algorithm for Land Cover Classification", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6S4, pp 96-99, April 2019.

[8] Richards, J.A. and J. Richards, Remote sensing digital image analysis. Vol. 3. 1999: Springer.

[9] Gislason, P.O., J.A. Benediktsson, and J.R. Sveinsson, Random forests for land cover classification. Pattern Recognition Letters, 2006. 27(4): p. 294-300.

[10] Gong, P., et al., Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. International Journal of Remote Sensing, 2013. 34(7): p. 2607-2654.

[11] Li, C., et al., Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. Remote Sensing, 2014. 6(2): p. 964-983.

[12] Lunetta, R.S. and M.E. Balogh, Application of multi-temporal Landsat 5 TM imagery for wetland identification. Photogrammetric Engineering and Remote Sensing, 1999. 65(11): p. 1303-1310.

[13] Davis, S.M., et al., Remote sensing: the quantitative approach. New York, McGraw-Hill International Book Co., 1978. 405 p., 1978. 1.

[14] Mallinis, G., et al., Forest parameters estimation in a European Mediterranean landscape

using remotely sensed data. Forest Science, 2004. 50(4): p. 450-460.

[15] https://www.esri.com/arcgis-blog/products/product/imagery/band-combinations-for-landsat-8/

[16] Coggeshall, M.E. and R.M. Hoffer, Basic forest cover mapping using digitized remote sensor data and automated data processing techniques. 1973.

[17] Duda, R.O., P.E. Hart, and D.G. Stork, Pattern classification. Vol. 2. 1973: Wiley New York.

[18] Sweet, J.N.. (2003). The spectral similarity scale and its application to the classification of hyperspectral remote sensing data. 92 - 99. 10.1109/WARSD.2003.1295179.

[20] Hilda Deborah, Noël Richard, and Jon Yngve Hardeberg, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 8, NO. 6, JUNE 2015, pp 3224- 3234

Authors' Profile:

Smt. Keerti Kulkarni is working as an Assistant Professor, in the Dept of ECE, BNM Institute of Technology. She is currently pursuing her Ph.D under the guidance of Dr. P.A. Vijaya, Professor and Head, Dept of ECE, BNM Institute of Technology, Bangalore

Dr. P. A. Vijaya, Professor and Head, Dept of ECE, BNM Institute of Technology, has a teaching experience of over 35 years. She has an equally rich research experience having guided 4 research scholars for the completion of their Ph.D. Currently she is guiding 6 scholars towards their doctoral degree.

ISSP-Tree: Minimum Item Support Based Improved Single Scan Pattern Tree for Generating Dynamic Frequent and Rare Patterns

[1] Keerti Shrivastava, [2] Dr. Varsha Jotwani
[1] Rabindranath Tagore University, Bhopal ,[2] Rabindranath Tagore University, Bhopal
[1] kirtidev_01@rediffmail.com,[2] varsha.jotwani@gmail.com

***Abstract*— In different human activities, the data mining technique plays a significant role because it discovers hidden meaningful information. This had led to numerous strategies and emphasizes the mining of frequently used patterns to identify patterns that appear frequently, unusually, and rarely. However, conventional techniques for data mining are restricted to static databases. The research has contributed to the optimization by its applications of the efficiency of the technologies. Association rule mining was first introduced to examine patterns among frequent items. The need for rare association rule mining has already been developing very fast. Rare data patterns are a key task for many elevated programs, such as clinical, financial & defense. The main motive to look for such rules was to detect patterns of rare diseases in the medical data by using rare association rules. It is designed to classify correlations of attributes that influence the probability of all other particular disease detection attributes in the existence of a disease. In this paper, we have presented an efficient incremental algorithm for detecting rare patterns. The Improved Single Scan Pattern-Tree (ISSP-tree) algorithm used the minimum support difference of each itemset. If the difference of support counts for each one item and minimum support difference is larger than least support then set MIS (minimum item support). Otherwise, the least support is set as MIS. When an item set contains rare & frequent items, the least MIS of rare items must have complied with it. Regarding runtime & memory use, the ISSP-tree method achieved a tremendous significant advantage.**

Keywords—Data Mining, Association Rule Mining, Frequent Pattern, Rare Patterns, Adverse Diseases, Single Scan Pattern-Tree, MIS, Support Difference.

I. INTRODUCTION

Data mining tackles the unorganized, incorrect, and inadequate data type. The goal of data mining is to look for coherent patterns, data analysis relationships, verify results by using new data sub-sets of the identified pattern then predict new results in new datasets. Data mining can be seen as a challenging task often because the algorithms used can be extremely complicated and the data cannot all be present in one location. It needs integration from a variety of heterogeneous data sources. The key problems are the methodology used for mining & user interactions, different data types & performance problems [1].

Data dependency is growing day by day in the health industry [2]. The most important task of medical sciences is to examine diseases & treatments of patients. Just since recent, manual scribbled notes of experts have also been translated into digital records with a reduction in costs caused by care and improved treatment efficiency[3]. Social welfare data mining applications may be further separated into subsequent classifications:

- Diagnosis and Prediction of Diseases
- Ranking of Various Hospitals
- Improved Treatment Methods
- Appropriate therapies
- Higher quality medical care
- Controlling infection in the Hospitals
- Identifying High-Risk Patients
- Fraud and violence avoidance in insurance
- Proper Hospital Resources Management
- Medical Device Industry

A need to improve and secure the sharing of health data among various parties for efficient use of data mining in healthcare organizations. To resolve security concerns, certain property constraints like contract ties among researchers and health organizations are obligatory. A systematic method is also needed to develop the data warehouse. A large dataset (text & nontext form) is also present on the websites in

coming times due to the upgrade of the Internet facilities. Efficient data mining methods are thus also required to analyze these data, to identify hidden data [4].

The need for rare ARM has been developing very fast. In certain fields, identification of rare patterns is critical, even if subsidized. Rare patterns mining is unavoidable and therefore an evolving research area with many moderate applications, like medical, financial, and safety applications. But it is difficult to uncover rare patterns from databases [5].

The new expansion is a rare pattern mining industry, and some differences still have to be solved. Contrary to frequent patterns, the rare pattern has an occurrence below that of thresholds specified by the consumer. Frequent patterns mining tends to prune such patterns as unnecessary or not of interest. Even so, in most other fields, the researching communities have seen the relevance of rare patterns [6].

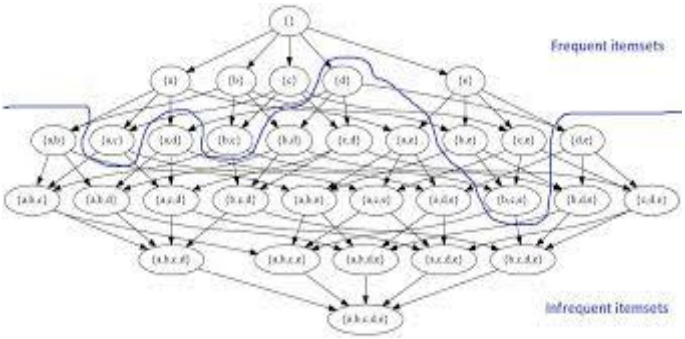


Fig. 1. Rare itemset mining

The remaining paper is organized as following: a comprehensive outline of the advanced methods for mining rare patterns has been discussed in Section II as a literature survey. Section III provided the details about the proposed methodology that describe the concept of MIS and also included the proposed ISSP-tree algorithm. Then experimental results and their evaluation has visualized in Section IV with their descriptions. Finally, we have concluded this paper with future work in Section V.

II. Literature Survey

The purpose of this section is to give a review of state-of-the-art rare patterns mining methods. We shall discuss the issues in the quest for rarity by using conventional association rules mining.

U. Y. Bhatt and P. A. Patel (2015) developed a scheme relying on Maximum Constraints for producing tree-structured rare association rules. Temporary findings show that MCRP-Tree requires less time to create rules than the previous algorithm and found rare items most relevant [7].

D. Zhou et al. (2016) Implemented a bi-level model for rare pattern recognition on temporal data (sequence or segment level). They are focusing on an optimizing system that takes full advantage of the bi-level data structure, — in other words the relationship of irregular sequences with irregular time segments. Also, the sequence-specific secret Markov models are used for obtaining segment labels and for estimating model parameters using the similarity between irregular time segments. They suggested the unsupervised BIRAD algorithm, as well as the semi, supervised BIRAD-K version, which was taken from a single labeled instance, to overcome the optimized model. The performance of the developed algorithms from several different perspectives, superior to state-of-the-art methods with both temporal or even uncommon category analyzes, was shown by explorative results on both synthetically generated and real datasets [8].

Saeed Piri et al. (2017) This study aims to resolve the issue of rare item detection in ARM. For the identification of rare items, a new evaluation metric is developed entitled adjusted support. The proposed metric is tested with a huge dataset with data from around 600,000 patients. Adjusted support is used to detect rare diabetes association rules. The diabetic patient comorbidity index is evaluated in different population groups. At last, every social group of patients analyzed and contrasted the incidence of diabetes-related complications [9].

Brian Delavan et al. (2017) The drug discovery demands for chronic illnesses are enormously unsatisfied. Computation drug repositioning is a viable technique and it has consistently been adapted to the advancement of disease care. But the way this expertise is used and computation drug repositioning methods become successfully performed and applied in rare diseases therapies remains an open question. They concentrated on ways to accelerate and encourage drug repository for rare disease using collected genomic data. They first outlined the current rare diseases genome landscape. Secondly, they suggested some interesting solutions to bioinformatics including computerized drug repositioning pipelines for serious diseases. They eventually addressed emerging incentives for regulators as well as other promoters of the production of rare disease drugs and highlighted the knowledge gaps [10].

Chandrawati Putri Wulandari et al. (2018) This article seeks to find out from stroke clinical datasets rare unusual association rules (RUARs) to provide possible useful information to the domain of users. As data pre-processing steps prior rule generation, a regularization method must be used. This paper used reciprocal details to discrete the data obtained from a health clinic in Taiwan for a stroke study. To simplify the distinctive form & enhance the quality of selected features, the intervals merge approach was presented. As in result, the use of the Apriori-Rare approach has resulted in rare ARs with comparatively low support. Furthermore, a filtering technique was used to detect the predicted RUARs for a physician for the contents of the rule itemsets. The derived RUARs were also evaluated based on the risk factor values of its strokes. The findings suggested that the regularization of shared knowledge was beyond conventional regularization methods to facilitate the removal of RUARs with improved quantities & quality measures for more medical research use. Besides, there were considerably higher RUARs in the developed system. The rare patterns knowledge from rare ARs will provide medical professionals with possible new and different perspectives and raise understanding of the outcomes of the strokes test [11].

Anindita Borah et al. (2018) The successful method has been applied to classify signs & risky factors in 3 adverse diseases: hepatitis, heart disease, and breast cancer with rare ARs. Once the database is changed, the consumer may move to a novel threshold value for desirable rare ARs. The new scheme throughout this study allows you to create new rare ARs in one database scanning from the latest clinical database without re-executing the whole mining procedure. It can manage transaction insertion and removal cases effectively and provides the user with versatility to create a new rare ARs set when the thresholds are changed. Exploratory research reveals the importance of the method provided to the conventional method of the whole revised database constantly being undermined [12].

Q. Pan et al. (2019) based on the Rarity algorithm, this paper presented a Relying on the Rarity methodology, this paper implemented a more successful top-down approach for efficient use of all rare items and their combo rules, using a schematic structure to demonstrate all the existing item combos in the database, defining a pattern matrix for documenting and retaining all items and combining a hash table to speed support value calculation to quickly classify all rarities This study used the existing clinical evidence on diabetes in the test to check this improved approach and to develop some helpful rules for diabetes mellitus issues. Furthermore, this approach reduces the time & space complexity in the ARM in contrast with the 2 above-mentioned methods [13].

Anindita Borah and Bhabesh Nath (2019) Implemented an effective method of efficient rare patterns-based outliers detection (RPOD) to identify outliers from incremental data by mining rare patterns. 1-pass tree-based, the rare patterns-mining method has been implemented to avoid multiple dataset scans & costly candidate generation phases using existing techniques of rare patterns mining & allow exponential mining. The presented scheme for mining the rare patterns is to modify the very famous algorithm of FP-Growth mining. Also, an outliner detection method has been proposed to recognize outliners based on the produced set of rare patterns. Some of the most famous medical datasets illustrated the effectiveness of the implemented RPOD method. The prevalence of the RPOD method over the previous outlier mining method is confirmed by contrasting performance assessment [14].

III. PROPOSED METHODOLOGY

In this section, we provide detailed descriptions of the proposed empirical study. The purpose of this research is to develop and design an effective and efficient model for identifying adverse disease using rare pattern mining. First of all, define the problem definition and then provided the solution for this problem by proposing an ISSP-tree algorithm.

A. Problem Statement

The annual rise in the death rate leads to bad factors has now become a significant global issue. The speed of making decisions including medical diagnostics by examining the rare correlation between various patient characteristics & diseases has been supported by computational intelligence strategies, such as rare ARM. The association rules contain rare items are rare association rules. Rare items are far less frequent items. The solitary minimum support (min_supp) techniques such as the FP-tree technique have a "rare items problem" dilemma for retrieving rare elements.

B. Proposed Methodology

This study implemented an effective methodology for the development of important rare ARs from its following medical datasets: hepatitis, Cleveland heart diseases & breast cancer. This technology has been used for the generation of important rare ARs. Our method to derive association rules on rare has been strengthened. All dataset items are maintained in the form of tree-structures, regardless of occurrences, to avoid any

data loss. The tree-structures only include a single search & retains full database knowledge.

During the scanning of the dataset for the first time, we discard all itemset that relies on minimum support count. After this arranged them in descending order. Assign to each itemset unique support value and overall dataset have least minimum support. We used this approach to define min _supp for items with the definition of the supports difference (SD). The SD refers to the appropriate divergence from (or support) occurrence of an item such that an item set contains this item as a frequent itemset. The min_supp estimation is defined as the min item support (MIS(pq) at each items 'pq'.

1) Multiple Minimum Support Approaches

A multiple support Apriori (MSApriori) method was developed in [15] to enhance the efficiency of retrieving frequent items containing rare itemset. Every item for this method is allocated the min_supp termed as "MIS" value and generated items frequently when an item set matches the least MIS values of the corresponding items. Any item receives the MIS value equal to a support %s. The MIS(pq) is determined in conjunction with the eq. (1) for any item pq to I.

$$MIS(pq) = \beta S(pq)), \text{ when } \beta S(pq) > LSup \quad (1)$$
$$= LSup, \text{ otherwise}$$

wherever,

β= User-defined proportionate value that range [0-1]

S(pq) = Item support equals to O(pq)/N_T,

(O(pq) = Occurrence of pq

N_T = No. of transactions in the transactional dataset

LSup = corresponding to the user-defined least supports value.

This approach allows for frequent items with high MIS values to be extracted depending on the support % & frequent items with comparatively least MIS values to be allocated. For a set of the item that only includes frequent items to be a frequent itemset, it needs to reach comparatively high Min_supp thresholds than the item that comprise either frequent items & rare items or rare items only [16].

C. Proposed ISSP-Tree Algorithm

The steps of SSP-Tree generation on a multiple minimum support basis are carried out adaptively in the following steps:

Input: Cleveland heart, dataset, Breast cancer, Hepatitis dataset, β=0.5, min_sup=15 min_rare=1

Output: Frequent and Rare patterns, No. of rules, Elapsed time, Memory used

Procedure:

Step 1. Initially assign support count to each Itemset and set the least minimum support count
Step 2. Scan all dataset and calculate support count
Step 3. Discard all those items that have minimal support from the specified least support count
Step 4. Arrange all itemset according to their minimum support count in descending order
Step 5. Transactional items insertion into the header table,
Step 6. Restructure the SSP-Tree in arranged header table sequence basis, and
Step 7. The occurrence of transactions has been inserted into the SSP-tree in the descendent sequence.
Step 8. Calculate the minimum support difference of each itemset by eq. (1)
Step 9. If each item's support counts difference & SD is higher than the min support, so set MIS. The least support is set as MIS else.
Step 10. Generation of frequent itemsets has been done, after determining the MIS values by every item to check if the itemset support counts are higher or equal to the itemset's MIS.
Step 11. If an item set comprises all of the frequent items, the frequent items' least MIS must be fulfilled.
Step 12. In the same way, if a set of items includes frequent & rare items, the least MIS of rare items must have complied with it.
Step 13. Obtain frequent and rare patterns and several rules.
Step 14. Measure elapsed time and used memory.

IV. Experimental Results & Evaluation

The experimental findings with a data summary are discussed in this section. Via the MATLAB software design, the proposed project will resolve the above-mentioned issue in part, by defining 3 kinds of adverse diseases such as heart, cancer & hepatitis, based on min_supp counts. This study is necessary to identify a multitude of signs in few seconds, rendering the diagnostic very straightforward & practical. The goal of this project is to identify the various diseases based on the selected features. The MATLAB development framework supports the project. This simulation has performed using MATLAB 2019 tool.

A. Dataset Description

Here, we have provided a brief description of datasets. The dataset included here are the heart dataset, cancer dataset, and hepatitis dataset.

1) Heart dataset

The Dataset for Cleveland Heart Disease comprises 76 features of which about 14 are included in the cardiovascular disease diagnostic testing studies. The occurrence or absence of heart disease is demonstrated by 303 instances & five class naming (0-4). The lack of class 0 and the cardiac condition of class 1 to 4 were indicated. There are 164 cases of class 0, 55 of class1, 36 of class2, 35 of class3, & 13 of class4. The risk of heart disease is seen in classes 1 to 4. Table 1 refers to the features including values of the Cleveland heart dataset.

Table 1. Information of features for the Cleveland heart disease dataset

S. No	Features	Explanation	Data Type	Value
1.	Age	Patient Oldness	Numeric	29 to 77
2.	Sex	Gender	Binary	0= F, 1= M
3.	Chp	Types of chest pain	Nominal	1= Typical angina, 2= Atypical angina, 3= Nonanginal pain, 4= Asymptomatic
4.	Trestbps	Resting blood pressures	Numeric	94 to 200
5.	Ch	Cholesterol	Numeric	126 to 564
6.	Fbs	Fasting blood sugar greater than 120mg/dL	Binary	1= True, 0= False
7.	Restecg	Resting electro-cardiographics result	Nominal	0 = Normal, 1= Abnormal, 2= ST-T wave abnormality, 3= Left ventricular hypertrophy
8.	Thalach	Maximize obtained heart rate	Numeric	71 to 200
9.	Exang	Exercise involved angina	Binary	1 = Y, 0 = N
10.	Oldpeak	Exercise relatively to rest caused ST depression	Numeric	Continuous (from 0 to 6.20)
11.	Slope	Peak workout ST segment slope features	Nominal	1= Upslope, 2= Flat, 3 = Downslope
12.	Ca	No.of fluoroscopy colored vessels	Nominal	0 to 3
13.	Thal	Types of defect	Nominal	3 = Normal, 6 = Fixed defect, 7 = Reversible defect
14.	Class	Healthy or heart disease existence	Binary	0= Safe, 1= Lower Risk, 2 = Medium Risk, 3 = Higher Risk, 4=Severe Risk

Note:- M: Male, F: Female, Y: Yes, N: No

2) Breast Cancer dataset

The data set for Breast Cancer in Wisconsin has 699 instances & 32 features. For diagnostic tests, about ten are included. Two classes of marks classify the types of breast cancer as benign and malignant. This has 458 cases in the benign & 241 cases in the malignant class. The FNA tissue samples for this cancer are studied by physicians in the range 1 to 10 to assign various cell properties. The probability of malignancy rises as the cell characteristics rise in value. Wisconsin Breast Cancer data set feature information is available in table 2.

Table 2. Information of features for the Wisconsin breast cancer dataset

S. No.	Features	Range
I.	Clump thickness	one to ten
II.	Marginal adhesion	one to ten
III.	Cell size uniformity	one to ten
IV.	Bare nucleoli	one to ten
V.	Size of single epithelial cells	one to ten
VI.	Cell shape uniformity	one to ten
VII.	Normal nucleoli	one to ten
VIII.	Bland chromatin	one to ten
IX.	Mitosis	one to ten

X.	Class	Benign or Malignant

3) Hepatitis dataset

There are 155 instances & 20 features in the dataset. The 'Died' & 'Alive' class marks signify whether or not such a patient who has hepatitis is going to alive. There will be 32 cases of 'Died' & 123 'alive.' The data set for hepatitis has several null values. To deal with the null values, the data set that includes over 25% of the null values or substitutes the null values by some of the most frequent values of the respective feature rejects all these tuples. After managing the null values, 147 instances & 18 features are extracted from the Protime feature and the resulting tuple is presented. Table 3 displays the Hepatitis data set features information.

Table 3. Information of features for the hepatitis dataset

S. No	Features	Explanation	Value
1.	Age	Patient Oldness	10 to 80
2.	Sex	Gender	M, F
3.	Steroid	Ingesting of steroid	N, Y
4.	Antivirals	Antiviral handling	N, Y
5.	Fatigue	Sign	N, Y
6.	Malaise	Sign	N, Y
7.	Anorexia	Sign	N, Y
8.	Liver big	Widen liver	N, Y
9.	Liver firms	Solid liver	N, Y
10.	Spleen palpable	Distended spleen	N, Y
11.	Spider	Spider nevus	N, Y
12.	Ascites	Liquid b/w stomach & tissues	N, Y
13.	Varices	Widen vein	N, Y
14.	Bilirubin	Quantity of the bilirubin	0.390 to 4.000
15.	Alk phosphate	Quantity of an alkaline phosphate	33 to 250
16.	Sgot	Liver enzyme	13 to 500
17.	Albumin	Quantity of an albumin	2.10 to 6.00
18.	Protime	Protein for liver	10 to 90
19.	Histology	Histology of liver	N, Y
20.	Class	The patient is dead or alive	Died, Alive

Table 4. No. of Itemsets produced for heart disease dataset

Algorithms	No. of Frequent Itemsets	No. of Rare Itemsets
SSP-Tree (Existing)	486	976
ISSP-Tree (Propose)	195	842

In terms of no. of items sets & produced rules, the efficacy of the proposed ISSP-Tree method regarding previous methods could be assessed. Apart from existing SSP-tree configurations, the ISSP-Tree configuration preserves all relevant data and allows the user to produce desirable suitable patterns set. ISSP-Tree method is also able to extract a significant number of frequent & rare itemsets. Table 4 presented produced no. of item sets by the proposed method and existing SSP-tree method from the Cleveland Heart Disease datasets respectively.

Table 5. Itemsets Generated from breast cancer dataset

Algorithm	#Frequent Itemsets	#Rare Itemsets
SSP-Tree (Existing)	298	2229
ISSP-Tree (Propose)	35	1667

The ISSP-Tree method has also started to achieve a full set of rare ARs & item sets on Wisconsin Breast Cancer datasets. Except for existing SSP-tree configurations, the ISSP-Tree configuration preserves all database items, providing the user with the flexibility to create any number of desirable patterns. From table 5 represents that the produced no. of itemsets for frequent items & rare items for the Wisconsin Breast Cancer dataset.

Table 6. No. of Itemsets produced for Hepatitis dataset

Algorithms	No. of Frequent Itemsets	No. of Rare Itemsets
SSP-Tree (Existing)	436	918
ISSP-Tree (Propose)	56	680

The proposed ISSP-Tree can generate the full set of ARs patterns. From Table 6, it is obvious that the ISSP-Tree method can output a significant no. of item sets & rules for the Hepatitis dataset, as compared to other existing SSP-tree. Thus, the ISSP-Tree algorithm provides important patterns

that required to detect Hepatitis disease based on attributes in terms of no. of itemsets & rare ARs.

Table 7. Produced rare rules from the breast cancer dataset using proposed

Rules	Antecedent	Consequence	Supp (%)
1	Normal Nucleoli=10	Size of Single Epithelial Cells =9	0.0029
2	Bare Nucleoli=1	Size of Single Epithelial Cells =9	0.0029
3	Marginal Adhesion=7	Size of Single Epithelial Cells =9	0.0029
4	Bare Nucleoli= 1 ∧ Marginal Adhesion = 7 ∧ Normal Nuclei=10	Bland Chromatin=6	0.0014
5	Mitoses=2	Bland Chromatin=6	0.0029
6	Mitoses=1	Bland Chromatin=6	0.0100
7	Normal Nucleoli=10	Bland Chromatin=6	0.0029
8	Normal Nucleoli=5	Bland Chromatin=6	0.0029
9	Normal Nucleoli=1	Bland Chromatin=6	0.0029
10	Bare Nucleoli=10	Bland Chromatin=6	0.0086

The rare ARs produced by the proposed ISSP-Tree are shown in Table 7. Rule1 holding rare item "Normal Nucleoli=10", specifies that once the size of the single epithelial cells is assigned 9 as a value at support 0.0029 through the physicians, this indicates a benign or ordinary tumor. In the same way, rule2 taking the rare item "Bare Nucleoli=1", cell size Single Epithelial of value 9 at support 0.0029, which does not signify the tumor. A detailed study has shown that there are many associations between patient features and the existence of tumor disease by complete evaluation of the rare ARs produced by breast cancer results. Since several rules were produced according to the defined support & confidence value, a definition from each rule isn't feasible. Consequently, the findings of a certain rule analysis are tabulated in Table 7.

Table 8. Produced rare rules from the Heart dataset using proposed

Rules	Antecedent	Consequence	Supp %
1	Class=Safe	Class=Severe Risk	0.0264
2	Thal=7	Class=Severe Risk	0.0297
3	Thal=6	Class=Severe Risk	0.0066
4	Thal=3	Class=Severe Risk	0.0198
5	Ca=3	Class=Severe Risk	0.0165
6	Ca=1	Class=Severe Risk	0.0099
7	Ca=0	Class=Severe Risk	0.0066
8	Slope=3	Class=Severe Risk	0.0066
9	Slope=2	Class=Severe Risk	0.0330
10	Oldpeak=[1.56-3.11]	Class=Severe Risk	0.0132

By ISSP-Tree algorithm-generated rules are seen in Table 8. As mentioned in the earlier section, the ISSP-Growth algorithm started to achieve the entire set of rarity items & therefore of all 4 rarity association rules. Rule1 contains the only rare item "Class=Safe". This indicates that Healthy or else heart disease existence has Severe Risk at support 0.0264 from heart diseases. As per rule 2 having rare item "Thal=7", has a reversible defect at support 0.0297and hence does have Severe Risks of heart diseases. Rule 3 taking rare item "Thal=6", have a fixed defect at support 0.0066 and hence does have Severe Risks of heart diseases. According to rule4 taking rare item "Thal=3", have a normal defect and hence does have a Severe Risk of heart diseases. Following a detailed evaluation of the Cleveland dataset's rare ARs, numerous associations have been formed among multiple diagnostic attributes with cardiopathy. Since certain rules are developed based on the user-defined support value & confidence value, a definition of each & every rule cannot be given. Therefore, some sample rules including their interpretations of a rule analysis are summarized in Table 8.

Table 9. Produced rare rules from Hepatitis dataset using proposed

Rule	Antecedent	Consequent	Support
1	Class=Live	Age=[0-20]	0.0194
2	Histology=Yes	Age=[0-20]	0.0194
3	Protime=[0-25]	Age=[0-20]	0.0194
4	Albumin=[3.176-4.25]	Age=[0-20]	0.0129
5	Varices=Yes	Age=[0-20]	0.0194

6	Ascites=Yes	Age=[0-20]	0.0129
7	Spiders=Yes	Age=[0-20]	0.0129
8	Spleen palpable=Yes	Age=[0-20]	0.0129
9	Liver Firm=No	Age=[0-20]	0.0129
10	Liver Big=Yes	Age=[0-20]	0.0194

The rare ARs produced by an ISSP-Tree are given in Table 9. Rule1 taking the first rare item "Class=Live", indicates that live patients have mostly age range (0-20) at support 0.0194 from hepatitis. According to rule 2 taking the rare item "Histology=Yes" people having an age range (0-20) at support 0.0194 have chances of recovery from hepatitis. Similarly for all rules. Here in table 9, we took 10 rules from the complete set of rules.

Table 10. Elapsed time (in seconds) taken for Heart, Breast cancer, and Hepatitis dataset

Algorithm	**Heart dataset**	**Breast cancer dataset**	**Hepatitis dataset**
SSP-Tree (Existing)	58.356289	1367.991366	11.801659
ISSP-Tree (Propose)	52.854660	1353.420129	11.225005

Table 10 represents the elapsed time for both SSP-Tree and ISSP-Tree on Heart, Breast cancer, and Hepatitis dataset. From table 10, we can see that the ISSP-Tree takes time 52.85 seconds, 1353.42 seconds, and 11.23 seconds for the Heart, Breast cancer, and Hepatitis dataset, respectively, that are less than SSP-tree.

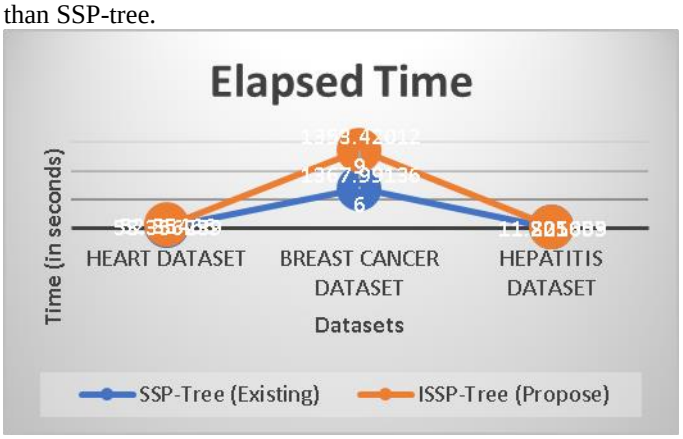


Fig. 2. Elapsed time comparison line graph

Fig. 2 shows the runtime consumed by frequent patterns and rare patterns for different three datasets. Every data set insertion demonstrates the individual impact for frequent & rare pattern mining in fig. 2. Execution time invested by the proposed ISSP-tree algorithm for generating the frequent and rare pattern in several support count values is represented in Fig. 2. The time is calculated in seconds. The proposed ISSP-tree takes less elapsed time in comparison to the existing SSP-tree to all three datasets namely, Heart, Breast cancer, and Hepatitis dataset.

Table 11. Memory Used (in KB) taken for Heart, Breast cancer, and Hepatitis dataset

Algorithm	**Heart dataset**	**Breast cancer dataset**	**Hepatitis dataset**
SSP-Tree (Existing)	3801.243164	5228.993164	2370.102539
ISSP-Tree (Propose)	3559.692383	4533.442383	1977.977539

Table 11 represents the memory consumption for both SSP-Tree and ISSP-Tree on the Heart, Breast cancer, and Hepatitis dataset. From table 11, we can see that the ISSP-Tree used memory 3559.69 KB, 4533.44 KB, and 1977.98 KB for Heart, Breast cancer, and Hepatitis dataset, respectively, that are much less than SSP-tree.

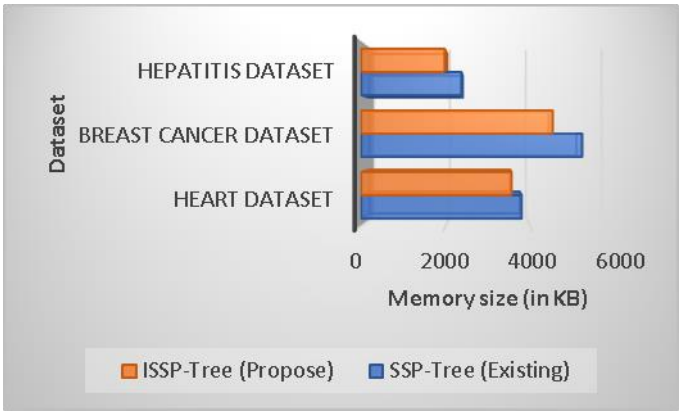


Fig. 3. Used memory comparison bar graph

The memory usage of the existing SSP-tree and proposed ISSP-tree for generating a frequent and rare pattern on the Heart, Breast cancer, and Hepatitis dataset has illustrated in fig. 3. The memory size is taken in KiloByte (KB). Fig. 3 indicates memory usage in various upgrade sizes & support thresholds for frequent & rare pattern generation. In terms of efficiency, the proposed ISSP-tree algorithm for all data sets

still outperform. The use of memory of the algorithm continues to reduce as the thresholds remain stable with increasing.

V. CONCLUSION

Mining of Frequent item sets is a big challenge to explore the unknown, meaningful pattern from datasets. The rare items are the items rarely contained in the database. Rare item sets are much more interesting often because they provide interesting knowledge which sometimes does not include infrequent patterns. Only if the threshold is quite the least will a rare itemset occur. To find correlations between rarely bought (i.e. costly or high-profits) retail products, analyzing clinical data as rare patterns assist physicians to find that diseases through rare symptoms, too, rare items are important. The mining of rare items is a major undertaking. In this paper, we have an effective approach to the mining of rare items for dynamic data sets of time-variant data such as the Cleveland Cardiovascular Disease dataset, Hepatitis & Wisconsin Breast Cancer. Here, we implemented a minimal support difference-based algorithm termed as ISSP-Tree.

In terms of runtime & memory utilization, the ISSP tree algorithm provides impressive efficiency gains. Decreasing the no. of database scanning is a vital requirement for the design of a method as it greatly decreases I/O cost. The suggested solution also allows it effective to produce complete patterns set & rare association rules with lower runtime & memory than that of the existing SSP-tree algorithm. Comparative outcome theoretical analysis shows the significance and effectiveness of the proposed method in producing major and most significant medical diagnostic rare association rules over existing SSP-tree algorithms.

In future work, we will attempt to use other MIS support different functions to the dynamic rare ARMs.

References

[1] B V Chowdary and Dr. Y. Radhika, "A Survey on Applications of Data Mining Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 5384-5392.

[2] Nada Lavrac, BlažZupan, "Data Mining in Medicine" in Data Mining and Knowledge Discovery Handbook, 2005.

[3] Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.

[4] Divya Tomar and Sonali Agarwal, "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25

[5] Hana Alharthi, Healthcare predictive analytics: An overview with a focus on Saudi Arabia, Journal of Infection and Public Health, Volume 11, Issue 6, 2018, Pages 749-756, ISSN 1876-0341, https://doi.org/10.1016/j.jiph.2018.02.005.

[6] Koh, Yun Sing & Ravana, Sri Devi. (2016). Unsupervised Rare Pattern Mining: A Survey. ACM Transactions on Knowledge Discovery from Data. 10. 10.1145/2898359.

[7] U. Y. Bhatt and P. A. Patel, "An effective approach to mine rare items using Maximum Constraint," 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2015, pp. 1-6, DOI: 10.1109/ISCO.2015.7282234.

[8] D. Zhou, J. He, Y. Cao, and J. Seo, "Bi-Level Rare Temporal Pattern Detection," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016, pp. 719-728, DOI: 10.1109/ICDM.2016.0083.

[9] Saeed Piri, Dursun Delen, Tieming Liu, William Paiva, Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications, Expert Systems with Applications, Volume 94, 2018, Pages 112-125, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2017.09.061.

[10] Brian Delavan, Ruth Roberts, Ruili Huang, Wenjun Bao, Weida Tong, Zhichao Liu, Computational drug repositioning for rare diseases in the era of precision medicine, Drug Discovery Today, Volume 23, Issue 2, 2018, Pages 382-394, ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2017.10.009.

[11] Chandrawati Putri Wulandari, Chao Ou-Yang, Han-Cheng Wang, Applying mutual information for the discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset, Expert Systems with Applications, Volume 118, 2019, Pages 52-64, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.09.044.

[12] Anindita Borah, Bhabesh Nath, Identifying risk factors for adverse diseases using dynamic rare association rule mining, Expert Systems with Applications, Volume 113, 2018, Pages 233-263, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.07.010.

[13] Q. Pan, L. Xiang, and Y. Jin, "Rare Association Rules Mining of Diabetic Complications Based on Improved Rarity Algorithm," 2019 IEEE 7th International Conference on Bioinformatics and Computational

Biology (ICBCB), Hangzhou, China, 2019, pp. 115-119, DOI: 10.1109/ICBCB.2019.8854639.

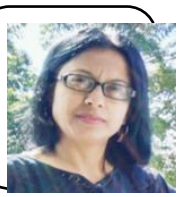
[14] Anindita Borah, Bhabesh Nath, Incremental rare pattern-based approach for identifying outliers in medical data, Applied Soft Computing, Volume 85, 2019, 105824, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.105824.

[15] Liu, B., Hsu, W., and Ma, Y. "Mining Association Rules with Multiple Minimum Supports." SIGKDD Explorations, 1999.

[16] R. Uday Kiran and P Krishna Reddy, "An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules", 2009 IEEE Symposium on Computational Intelligence and Data Mining (IEEE CIDM 2009), 2009, pp. 1-9.

AUTHORS PROFILE

Keerti Shrivastava is a research scholar of Rabindranath Tagore University ,Bhopal(M.P).
She has completed M.Tech in computer science.
Her specialization area is data mining.
She has published International and National Research Papers in the high quality journals



Dr. Varsha Jotwani is currently working as Associate Professor with Rabindranath Tagore University .
She is PhD in Computer Applications She has vast teaching and academic developments at leading institution of Bhopal, India. She has published various International and National Research Papers in the high quality journals. She is also well versed in developing curriculum for Undergraduate and Postgraduate Students under the field of Information and Technology.

Heart Disease Prognosis System with Nearest Clinic Recommendation

Chitra Bhole [1], Ulkesh Chendwankar [2], Jainam Jatakia[3], Mayuresh Pujari [4]
[1] Professor, Department of Computer Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India cbhole@somaiya.edu
[2][3][4] Undergraduate Students, Department of Computer Engineering,
K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India
ulkesh.c@somaiya.edu, jainam.jatakia@somaiya.edu, mayuresh.pujari@somaiya.edu

***Abstract*- Now-a-days, any age group of people can suffer through heart attack or any other heart disease so there's need for people to regularly do check-ups. Those check-ups can be costly if they are done directly at the cardiologist or physiologist for their particular test reports generated. So, what we intended to build a system which general clinic doctors can use for their patients for checking chances of getting heart disease by either making them undergo the test in their clinics or from already generated test reports for testing labs. So, we thought to use machine learning algorithms KNN and Logistic regression for computing the entries filled by general doctors of user to be tested. This system finds the final output with help of Cleveland Heart Disease Dataset from UCI machine learning repository and current patient data entries. After the generating final output, it suggests the location of the nearest Clinic or Hospitals in patient's vicinity. This suggestion feature is mostly for proper educated users who are knowledgeable about their reports parameters and who track their health regularly.**

Keywords— Logistic Regression, K-Nearest Neighbour. Cleveland Heart Disease, UCI machine learning repository

I. INTRODUCTION

According to Centres of Disease Control Prevention (CDC), Heart Disease is the leading cause of death in the United States and among 4 deaths every 1 is due to the result of heart disease. That sum the amount of almost 610,000 deaths from heath disease conditions each year [1] and the statics are predicting that 45% of the people in America will have at least one issues related to the disease by 2035 from American Heart Association (AHA). AHA also predicts that costs related to disease might get doubled till 2035 which could bankrupt the nation's economy and health care system; since the complication from heart disease are spreading faster than their original thoughts [2]. Even thought other diseases like cancer, diabetes is gaining more attention, but still cardiovascular disease remains the world most costly killer and most of time, people may not realize the risk or audacity of heart disease until they know someone who suffer from it or they themselves are suffering from it.

Previous decade had witnessed huge advances in the quantity of data that is usually generated and collected. Various Private and public sector industries generally generate tons of data that can be stored and analysed for the improvement of their services. Rather that increasing the profits and cutting down overheads; analysing the past data from Health care industry to predict epidemic, pandemic, curing diseases, increasing quality of life. These decisions can be made due to the changes driven by the events happening in the current or past to predict what can happen in the mere future. So, the focus can be now solely depending on patients records for detecting any signs of diseases early and making the treatment quicker and less painful. Nearly millions are affected by heart disease due to changes in their lifestyle. Mainly it is caused by blockage of cholesterol in arteries and which can lead to heart stroke and can be caused by hypertension.

In this system, we have implemented a heart disease prediction application which works on K Nearest Neighbours and Logistic regression algorithm. Initially, we input the values in the application as per the fields provided in system which is then analysed with respected

to the dataset that we provided to the system containing 14 fields including a target field which helps in predicting when given processed inputs what output can be predicted from current inputs. This processing and analysing is done by K Nearest Neighbours and logistic algorithm. After that our system will predict the output and display it in window whether there is presence of heart disease and if there is presence of heart disease it will suggest the clinics with certain distances.

II. Literature Survey

Before building the actual system, we did some research on the various ML based diagnosis techniques that have been already proposed in research papers. This analysis study presents a number of the usual machine learning based mostly identification techniques to clarify the vitality of the proposed system.

- Detrano et al. [3] developed they developed and Heart disease classification which gave accuracy of 77% in terms of accuracy. They used Cleveland dataset in this paper.
- Mohammed Jaeead Ali Junaid et al. [6] They proposed Naïve Bayes, ANN, SVM which gave accuracy of 82.97%, 85.30%, 86.12% respectively.
- Palaniappan et al. [5] proposed an Heart disease System using Decision Tree and ANN having accuracy of 80.41%
- Pahulpreet Singh Kohli, Shriya Arora, [4] used LR, DT, Adaboost with accuracies of 87.1%, 70.97%, 83.87% respectively.
- Khairina, D. M. (2017), Haversine formula is used to find the courier location of the nearest web-based JNE, the result of this research is a web that can be accessed by people in Samarinda, Indonesia [8].
- Dhanashri Gujar, Rashmi Biyani et al. [7] had recommended best 5 specialist by filtering the data and using Core NLP.

Logistic Regression and KNN has the classification accuracy better than above discussed methodsin predicting the results and the Suggestion of the nearest clinic will be done by obtain users gps coordinates and then using the Haversine Formula within location coordinates of Clinic in the database.

III. Existing Systems

Prior to any system, the patient had to visit the cardiologist with their report wherein cardiologist go through their report and tell the patient about their anomalies in health parameter and if there are serious report output, they would had got admitted to hospital and further so. So, this process is still used as priority and most trustful method. But still there are some efforts made to build system to automatize this analyses process. This system used one of the various algorithms of machine learning like Support vector method, Linear Regression, K Nearest Neighbour but this method did not provided sufficient accuracy.

IV. Proposed System

In this system we are implementing effective heart attack prediction system using KNN algorithm and Logistic Regression. We have to input manually in the input field according to the label given aside. After taking inputs we perform analysis on those inputs with K Nearest Neighbour and logistic regression algorithms. After accessing data set which is in Comma Separated Values format the operation is performed and prediction result is produced. The heart attack prediction system designed to help the identify presence of disease and with suggestion of the nearest clinic for further treatments.

V. System Architecture

Proposed System act as a whole providing the user with analysis on their inputs. This system gets the input from the user which with the help of dataset provided to it and KNN and logistic algorithm is processed on the Dataset as well as user inputs through which a final decision function generate the output which is activated if either of the algorithm provides the output as disease present as shown Figure 1.

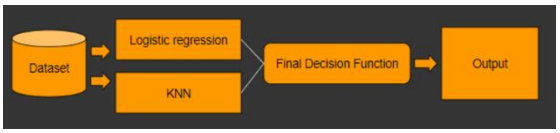


Figure 1: System Architecture of Prediction Module.

After generation of Output, if user has presences of heart disease, then there will be an suggest button which will take the user over the browser and ask for the access of location as you can see in figure 2, user's current GPS location is sent by using POST request. POST method is one of the formats for transmission of data in PHP for sending sensitive data. Furthermore, this location is processed with locations of the various other clinics

present in the MySQL database using Haversine Formula and then generated response of all the clinics data within a particular range is sent in form of JSON (JavaScript Object Notation) to the User interface. We have used postman application to create an API for saving complex http request and for the creation of map we have used leaflet.js one of the JavaScript libraries used for creating interactive maps.

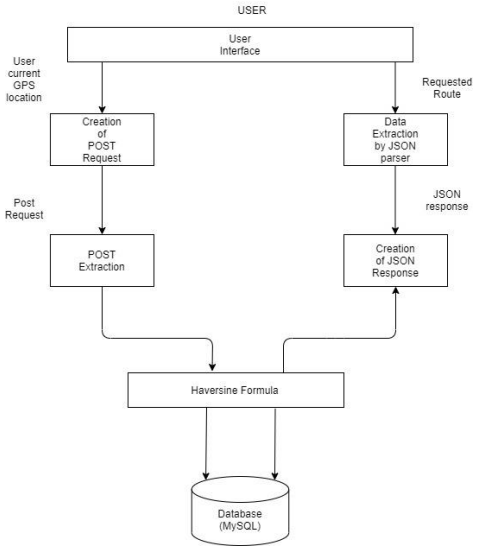


Figure 2: System Architecture of suggestion Module.

B. *Dataset Collection*

The dataset we used is Cleveland database and it contain 14 attributes as given below in Table 1.

S.N.	Attributes	Description
1	Age	Age (years)
2	Sex	Male or Female
3	Cp	Chest pain type
4	Thestbps	Resting Blood Pressure
5	Chol	Serum Cholesterol
6	Restecg	Resting electrographic results.
7	Fbs	Fasting Blood Sugar.
8	Thalach	Maximum Heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest.
11	Slope	Slope op the peak exercise ST segment
12	Ca	No. of major vessels colored
13	Thal	Defect type
14	Target	Target value

Table I. Dataset Parameters Information

VI. IMPLEMENTATION

The implementation included two modules namely prediction module and suggestion module. These application within modules are built for optimized of easy usage of use for the users and to get approximate location of nearest cardiologists clinic for user within the database.

A. Prediction Module

Prediction module consist of KNN and Logistic regression which are built using Python libraries like sklearn with an User Interface made of Tkinter which is also Python Library.

1. KNN (K-Nearest Neighbour Algorithm)

K- Nearest Neighbour Algorithm is used for classification of given input whether they belong any of the given class and in which class. There can be k classes. For, our system we used k=3 so the accuracy of our system is 87% as shown in figure 3 below.

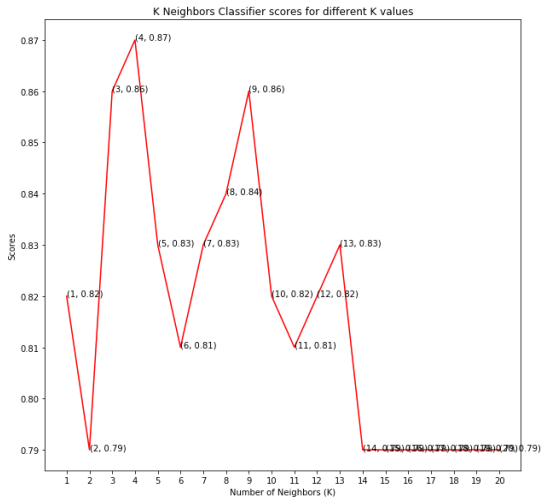


Figure.3: KNN (K-Nearest Neighbour Algorithm) Accuracy

2. Logistic Regression

Logistic Regression is an algorithm that gives the output in the form of binary values i.e., like 0's and 1's. Threshold is used for distinguishing this values. When we did accuracy check it gave us accuracy of 88.14% as shown in figure 4.

```
In [30]: from sklearn.linear_model import LogisticRegression
         logreg=LogisticRegression()
         logreg.fit(x_train,y_train)
         y_pred=logreg.predict(x_test)

In [31]: sklearn.metrics.accuracy_score(y_test,y_pred)
Out[31]: 0.881491344873502

In [35]: Accuracy of the model is 0.88
```

Figure 4: Logistic Regression Accuracy.

B. Suggestion Module

In Suggestion module, we take users current location using gaps based on user's current location nearby clinics are shown having distance of 5km. We have used Haversine formula for the calculation of distance of user and clinics. We have used postman application to create an API for saving complex http request. For the creation of map, we have used leaflet.js one of the JavaScript libraries used for creating interactive maps.

1. Haversine Formula:

Haversine Formula is the calculation of the distance from a point to another point on the surface of the earth. This calculation is affected due to certain degree of curvature [10]. This formula calculates the considered distance accurately unlike other methods like choice of distance calculation. So, Haversine formula is a method that exactly is calculation of the distance between the distance with two places longitude and latitude data.

Haversine formula: d =

$$2r\arcsin\sqrt{sin^2\left(\frac{\phi_2-\phi_1}{2}\right)+\cos(\phi_1).\cos(\phi_2).sin^2\left(\frac{\psi_2-\psi_1}{2}\right)}$$

Where d is the distances in Km, r is radius of earth (637 km), Φ is the latitude and Ψ is longitude. 1 degree is equal to 0.0174532925 radians.[9]

VII. RESULTS

Each Module of our System i.e., Prediction Module and Suggestion Module is depicted with following figures.

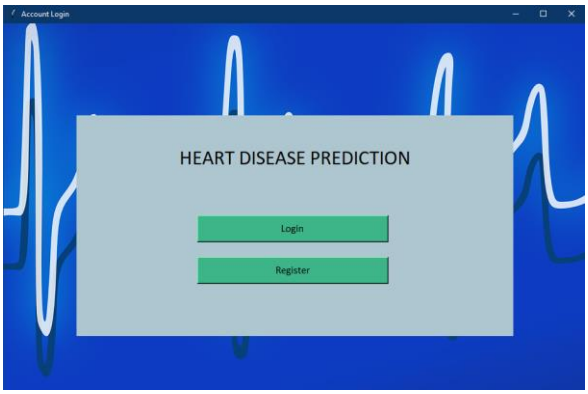


Figure 5: Login/Register Window of the System.

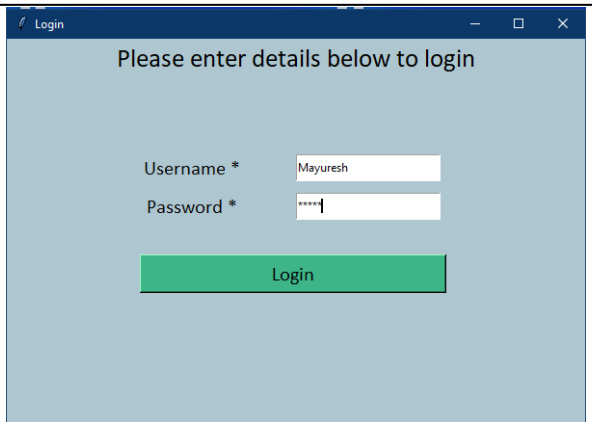


Figure 6: Login Window of the System.

Figure 7: Home Window of the System.

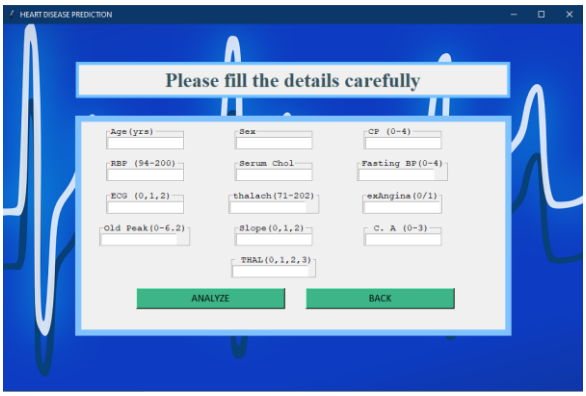


Figure 8: Analysis Window of the System

Figure 9: Result Page (Absence of Disease)

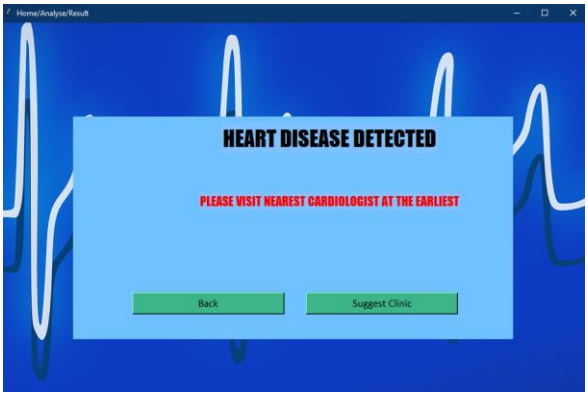


Figure 10: Result Page (Presence of Disease)

Figure 5 show the Window wherein user can login or register into the system.

Figure 6 shows the Login window where user has to enter his credentials.

Figure 7 shows the Home window wherein there are 3 option namely Analysis, about us and Log out.

Figure 8 shows Analysis window where the user has to enter all the input fields and press Analyse Button to start prediction on their inputs.

Figures 9 shows Window which is produced by the input which weren't leading to presences of heart disease.

Figure 10 shows Window which appears when user enters the input are leading to presence of heart disease. Here user can press the suggest button for suggesting a

clinic nearest to their current location, this button will take user to browser opening the webpage.

Figure 11: Suggestion module Home Page

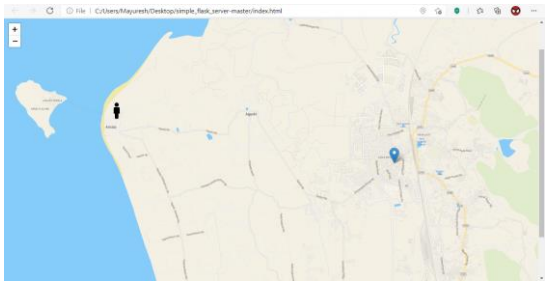


Figure 12: Suggesting a Clinic Page.

Figures 11 shows the webpage when user press the Suggest button so here when user press Nearby Clinics, they may be asked of get access to their current location as a pop-up where they have to enter Ok button.

Figure 12 shows the webpage where they are been suggested nearest clinic around their current location.

VIII. Conclusion

The Proposed system will be used by the general practitioner doctor or user with knowledge about the input parameters which will be enter from referring to the report of their tests and this it can be used by the medical students as a simulator. This system has improved accuracy since we are using both KNN and Logistic Regression, so this system will basically find if either of the algorithms produce output as presence of disease then it gives it has final output. This system can be help in reducing check-up cost and treatment costs by providing initial diagnostics in time. The User then will be suggested the clinic nearby with its name, the city in which it is located, the rating of the clinic on to the map with also their current location.

IX. Future Scope

In the future, our system can be improvised by using more accurate algorithm then current system uses and also the system can have single image input which will extract inputs from the image of the report and do analysis on those input so by this way user may not need doctor to input their values. Systems Suggestion system can be improvised by adding more features like tracking user's current location and showing the path towards the destinated nearest clinic. Mainly this system can be switched from desktop to mobile version so more users can use it.

References

[1] https://www.healthline.com/health/heart-disease
[2] https://www.healthline.com/health-news/why-is-heart-disease-on-the-rise
[3] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ''International application of a new probability algorithm for the diagnosis of coronary artery disease,'' Amer.J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989.
[4] Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Dease Prediction", International Conference on Computing Communication and Automation (ICCCA),2018.
[5] S. Palaniappan and R. Awang, ''Intelligent heart disease prediction system using data mining techniques,'' in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 108–115
[6] Mohammed Jawwad Ali Junaid, Dr, Rajeev Kumar, "Data Science And Its Application In Heart Disease Prediction", International Conference on Intelligent Engineering and Management (ICIEM),2020.
[7] Dhanashri Gujar, Rashmi Biyani, Tejaswini Bramhane, Snehal Bhosale, Tejaswita Vaidya, "Disease Prediction and Doctor Recommendation System", International Research Journal Of Engineering and Technology (IRJET), Pune, India, Maar-2018
[8] Khairina, D. M., Ramadhinata, F. W., & Hatta, H. R. (2017). Determinating of the Nearest Jalur Nugraha Ekakurir (JNE) Office Using Haversine Formula (Case Study in Samarinda). Prosiding SENIATI, 3(1), 10- 1.
[9] M.Basyir, M.Nasir, Suryati, Widdha Mellyssa, "Determination of Nearest Emergency Service Office using Haversine Formula Based on Android Platform", EMITTER International Journal of Engineering Technology, Vol. 5, No. 2, December 2017, ISSN: 2443-1168.
[10] Putra, D., & Herwan, R. (2015). Implementation of Haversine method formula in the information system of geographical landscaping. Jurnal Sistem dan Teknologi Informasi (JustIN), 1(1).

Analysis of Opportunities & Challenges for growth of E-commerce in India

Monojit Kumar
APJ Abdul Kalam University Indore

Abstract- This paper is about the analysis of opportunities & challenges for growth of E-commerce in India. E-commerce is certainly one in all the business choices that one can need to explore within the future. Ecommerce is said to bring forth paradigm shift within the world for commercialism. Prediction e-commerce is showing tremendous business growth in India. Backed by redoubled on-line user base and mobile phone presentation, Indian e-commerce has seen spectacular growth within the previous couple of years. Considering India's demographic dividend & rising net accessibility, the world is slated to scale larger heights. The present study has been undertaken to explain the current standing & future growth of e-commerce in India. Everybody in the world is getting mad for online shopping and online work. In the fast changing world the India is also now fast moving country as we start to discuss about development in the information communication technology era. Today India is the main market for online trades or shopping. The Indian government is also now going for online selling and buying the products, cars etc. The education in India is changing to online educations. The seminars are changing to webinars. So it is necessary to make changes in you with the flow of the market. Thus, India is now become a biggest marketplace for many big countries and big products. The paper is showing how will be the future of Indian commerce and condition after the maximize use of E-Commerce.

Keywords- E-commerce, Customer satisfaction, Trust, Loyalty, Customer loyalty, India

Introduction-

E-commerce is a trading or facilitation of trading in products or services using computer networks, such as Internet. E-Commerce is one of the biggest forms of doing E-business, that has happened to the Indian cashless economy in recent years. This has created a new flavor of doing business, which has a huge potential and is fundamentally changing the way businesses are done. This provides advantage for both buyers as well as sellers at the core of its phenomenal rise. The economic reforms of India that were amended in 1991, has resulted in opening of the economy with a view to integrate itself with the worldwide economy. As a result, in last few years we have witnessed a technological revolution accompanied by the widespread use of the Internet, web technologies and their applications. As a symbol of globalization, E-commerce represents the cutting edge of success in this digital age and it has changed and is still changing the way business is conducted around the world.

Literature Review-

India has an online user base of one hundred fifty million as of June 2014. The access of E-commerce is low as compared to markets just like the USA and UK. However, this is growing at a way quicker rate with several new entrants. E-commerce in India continues to be in mushrooming stage; however, it offers intensive chance in developing countries like India which has ntense urban areas with terribly high skill rates, vast rural population with quick increasing skill rate, a speedily growing web user base,

technology advancement and adoption. Cash on delivery in an unique and distinctive payment mode introduced for Indian customers and may be a most popular payment technique. Moreover, reduction in price of personal computers, reduction in the internet prices, easy access to internet and lot of more competitive Internet Service supplier (ISP) has added fuel to the hearth in augmenting E-commerce growth in Asia's second most densely settled nation. India's E-business trade is on the expansion curve and experiencing a surge in growth. the web Travel trade is that the biggest section in e business and is flourishing mostly thanks to the Internet-savvy urban population. An outline by the web and Mobile Association of India has exposed that India's E-commerce market is mounting at a median rate of 70% annually and has big over five hundred percent since 2007. This estimate of US $ 1.79 billion for year 2010 is much prior to the market size within the year 2007 at $1.75 billion. Apparently, additional on-line users in India area unit willing to form purchases through the web. Overall e-commerce trade is on the sting to expertise a high growth within the next few years. The E-commerce market in India was mostly dominated by the web travel trade with eightieth market share whereas electronic retail (E-Tailing) command second place with 1.48% market share. E-Tailing and digital downloads area unit expected to grow at a quicker rate, whereas on-line travel can still rule the main proportion of market share. Due to increased ecommerce initiatives and awareness by brands, e-Tailing has practiced tight growth. in keeping with the Indian Ecommerce Report discharged by web and Mobile Association of India (IAMAI) and IMRB International, "The total on-line transactions in India was Rs. 7000 crores (approx. $1.55 billion) within the year 2007-2008 and it absolutely was big by thirty third to the touch Rs. 9200 crores (approx. $2.10 billion) by the year 2007-2008. Overall E-commerce market is anticipated to succeed in Rs 17,800 crores (US$ 24 billion) by the year 2015 with each on-line travel and e-tailing platform equally.

Advantages of E-Commerce to Indian market-

E-commerce is one amongst the main forms within the electronic ways of doing business. Awareness among the companies in India regarding the opportunities offered by e-commerce has seen gradual increase year on year. Ease of doing business and web access square measure few of the numerous factors that has resulted in speedy adoption of E-commerce. Safe and secure payment modes have also contributed to create and popularize innovations like Mobile Commerce. E-commerce conjointly provides an alternative platform for connecting with customers and conducting transactions. Virtual stores operate throughout the year, all days in a very week, twenty-four hours giving customers to buy at their comfort successively providing ease of doing business for E-commerce merchants. Indian E-commerce has big at a combined annual rate of 35% since FY10 and is predicted to be $20 billion chance by FY18. The web brings low search prices and high worth flexibility. E-commerce has proved to be extremely value effective for e-business considerations because it cuts down the value of promoting, processing, inventory management, client care etc. It conjointly reduces the load of infrastructure needed for conducting business. Customers are empowered to do transactions for the merchandise or enquiry regarding any product/services provided by a online website anytime, anyplace from any location.

GreaterEconomic Potency -
We have achieved larger economic potency (lower cost) and additional speedy exchange (high speed, accelerated, or time period interaction) with the assistance of electronic business. Key drivers in Indian e-commerce are:

1. Increasing broadband web growing at 20% and 3G penetration.
2. Rising living standards and a growing, up mobile socio-economic class with high incomes.
3. Handiness of a lot of wider product compared to what's obtainable at conventional retailers.
4. Busy lifestyles, urban tie up and lack of your time for offline looking.
5. Lower costs compared to conventional retail driven by reduced inventory prices.
6. Accumulated usage of on-line classified sites, with additional shopper shopping for and merchandising product.
7. Evolution of the net marketplace model with sites like eBay, Flipkart, Snapdeal.

The evolution of E-commerce has come back a full circle with marketplace models taking centre stage once more.

Opportunities-
India has prospect of market potential with E-Commerce business growth doubling each year. Morgan Stanley noted that market of E-Commerce can rise to $137 million by 2020. Asian country's annual house financial gain has accumulated in year 2015 of total 246 million in India. Statistics indicate that house financial gain is predicted to achieve $3823 in 2015 & become $6790 in 2020 which can conjointly contribute in growth of E-Commerce.

Challenges-
E-commerce companies got to address problems like-
Strengthen supplying infrastructure & service levels as storage demand can increase with increase in E-commerce activity in returning years.
Security, privacy breaches & fictitious dealing problems.
Rules & rules for taxation & evaluation of product for international & native corporations.
Customers square measure involved on security once it involves use of credit & debit cards, thus money on delivery is most well-liked mode of payment ought to be created obtainable.

Conclusion-
The e-commerce market in the country has see growth by thirty four percent in the last seven years, was regarding USD 600 million in 2011-12 and is predicted to the touch USD nine billion by 2016 and USD seventy billion by 2020.
Companies need to address all above among many or challenges to mitigate any prospective slowdown in growth.

RELATIONSHIP BETWEEN CLOUD COMPUTING AND BIG DATA

[1]Mr Bharat Batham, [2]Dr. Shailja Sharma
[1]Research Scholar, [2]Associate Professor
[1] batham_bharat@yahoo.in,[2] shailja.sharma@aisectuniversity.ac.in,

Abstract
Communicating by using information technology in various ways produces big amounts of data. Such data requires processing and storage. The cloud is an online storage model where data is stored on multiple virtual servers. Big data processing represents a new challenge in computing, especially in cloud computing. Data processing involves data acquisition, storage and analysis. In this respect, there are many questions including, what is the relationship between big data and cloud computing? And how is big data processed in cloud computing? The answer to these questions will be discussed in this paper, where the big data and cloud computing will be studied, in addition to getting acquainted with the relationship between them in terms of safety and challenges. We have suggested a term for big data, and a model that illustrates the relationship between big data and cloud computing.

Keywords: big data, Hadoop, Cloud, MapReduce, resources, Five (Vs).

I. INTRODUCTION

Data is the uncooked material for Data before sorting, arranging and processing. It cannot be used in its primary shape prior to processing. Information represents records after processing and analysis [1]. The technology has been evolved and used in all factors of lifestyles, increasing the call for storing and processing greater information. As an end result, several systems were developed consisting of cloud computing that assist huge records. at the same time as large information is liable for records storage and processing, the cloud gives a reliable, handy, and scalable surroundings for huge records systems to characteristic [2]. large records is defined as the amount of digital data made out of different resources of era for instance, sensors, digitizers, scanners, numerical modeling, mobile phones, net, movies, e-mails and social networks. The facts types include texts, geometries, photos, motion pictures, sounds and combos of each. Such facts may be without delay or in a roundabout way related to geospatial facts [3].

II. BIG DATA

Big Data comes and is composed through electronics operations from a couple of assets. It calls for proper processing strength and high talents for evaluation [9]. The importance of big Data lies inside the analytical use which can help generate a knowledgeable selection to offer better and quicker services [4].

The main traits of huge data, known as V's five As in parent 1 , may be summed up inside the truth that the issue isn't always simplest approximately the volume of Data, other dimensions of large records, referred to as 'five Vs', are as follows:

1. Volume: It represents the quantity of records made out of multiple sources which

show the huge records in numbers by means of zeta bytes.

2. Variety: It represents Data sorts, with, growing the wide variety of net customers everywhere, clever telephones and social networks customers, the familiar form of facts has modified from established facts in databases to unstructured data that consists of a big range of formats consisting of photos, audio and videos, SMS, and GPS data [5].

3. Pace: It represents the rate of facts frequency from one of kind assets, that is, the velocity of data manufacturing which include Twitter and Facebook.

4. Veracity: The exceptional of the information captured can range substantially, which affects the accuracy of evaluation. Despite the fact that there may be wide agreement on the ability cost of massive information, the records is almost worthless if it isn't correct [6].

5. Fee: It represents the fee of massive data, i.e. it suggests the importance of information after analysis. That is because of the fact that the information on its own is sort of worthless.

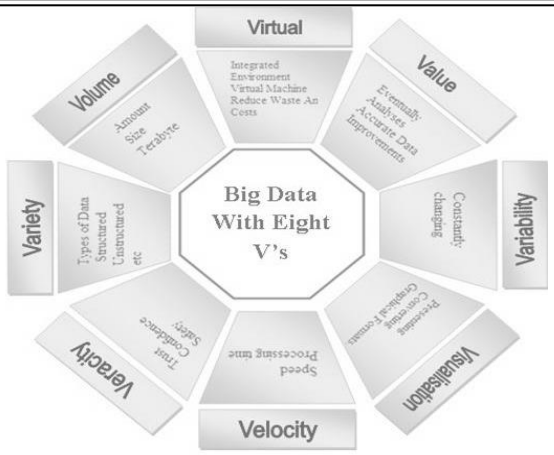


Figure 1. Characteristics of Big Data

There had been several revisions to the big information till they reached (7 v) . on this paper, based on the relationship between cloud computing and large data, will endorse a brand new time period, virtualization, which simply represents The statistics structure is through default. The virtualization of huge information is a technique that focuses on creating digital systems for massive statistics structures. Virtualization technology is the key generation used to help cloud computing handle large amounts of information flexibly and facilitate the procedure of coping with massive statistics.

Difference between traditional data and big data

In general, the data in the world of technology is a set of letters, words, numbers, symbols or images, but with the evolution of multitasking technology tools the data has become different in content and source[7]. In light of this, big data emerged which differs from traditional data. Differences between traditional data and big data are shown in Table1:

Table 1 Comparison between traditional and big data[18]

	Traditional Data	**Big Data**

Volume	MB and GB	PBs And ZBs
Data Generation Rate	Long periods Of time	More rapid
Data Type	S	Sim-Structure , Unstructured
Data sources	C	multiple sources, and distributed
Data Store	R	HDFS, No SQL

Cloud Storage

The concept of cloud storage is the same as that of storing files on a remote server to retrieve them from multiple devices at any time we need. Cloud storage is basically a system that allows storing data on the internet. Examples of this system are Google Drive, Dropbox, etc. [8]. Cloud storage , it is stored data while cloud computing is used to complete the specified digital tasks. In most cloud computing applications, data is sent to remote processors over the internet for complete operation, and the resulting data is sent back where you can use the program interface but the bulk of the program activity is remote instead of the computer. Cloud computing is usually more useful for companies than individuals in most cloud computing applications. It is a set of technologies hosting a cloud, and giving resources to hire and consume on demand over the internet on the basis of pay-per-user. Among the best known cloud computing providers are Amazon, Google, and Microsoft.

The increasing amount of data requires equipment to store them. The cloud provides storage units, making it easier to navigate without having to carry physical storage equipment while on the move. Limited storage space is a real concern for both consumers and businesses. The storage of data in the cloud is done through a cloud service provider (CSP) in a set of cloud servers where the user interacts with the cloud servers via CSP to access or retrieve its data. Since they no longer have their data locally, it is important to assure users that their data is properly stored and maintained. This means that users should be provided with security means so that they can ensure that their stored data is consistently maintained even without local copies.

III. DATABASE MANAGEMENT SYSTEM.

Data is collected in the form of an organized structure called the database which is the food of any information system. Data huge amount is the major component of the cloud infrastructure. Data can be shared among many tenants. As a result, data management in particular is a key aspect of storage in the cloud [9]. Data in the cloud is distributed across multiple sites and may contain certain privileges and authentic information. It is therefore very important to ensure that data consistency, scalability and security are maintained. In order to address these issues and many other important data issues, there is a need for a database management system for cloud data. The database management system shows the mechanism of storage and retrieval of user data with maximum efficiency, taking into consideration the appropriate security policies. The database management system always provides data independence. No change is made to the storage mechanism and shapes without modifying the entire application. There are several types of database organization, relational database, flat database, object oriented database, hierarchical database.

IV. THE RELATIONSHIP BETWEEN THE CLOUD AND BIG DATA

Cloud computing is a trend in the development of technology, as the development of technology has led to the rapid development of electronic information society. This leads to the phenomenon of big data and the rapid increase in big data is a

problem that may face the development of electronic information society [10]. Cloud computing and big data go together, as big data is concerned with storage capacity in the cloud system, cloud computing uses huge computing and storage resources. Thus, by providing big data application with computing capability, big data stimulate and accelerate the development of cloud computing.

Clouds are evolving and providing solutions for the appropriate environment of big data while traditional storage cannot meet the requirements for dealing with big data, in addition to the need for data exchange between various distributed storage locations. Cloud computing provides solutions and addresses problems with big data. Cloud computing environments are built for general purpose workloads and resource pooling is used to provide flexibility on demand. Therefore, the cloud computing environment seems to be well suited for big data.

Big data processing and storage require expansion as the cloud provides expansion through virtual machines and helps big data evolve and become accessible. This is a consistent relationship between them. Google, IBM, Amazon and Microsoft are examples of the success in using big data in the cloud environment. In order for the cloud environment to fit with big data the cloud computing environment must be modified to suit data and cloud work together.

Virtual Machine (VM) between the cloud and big data

Virtual Machine (VM) is a software application that simulates a virtual computing environment that can run the operating system (OS) and its associated applications with multiple virtual machines installed on a single machine. Distributed systems, network computing and parallel programming are not new as one of the key enabling factors of the cloud is virtual technology. By using virtualization technology, one virtual machine can often host multiple virtual machines [11]. Virtualization technology is the best platform for big data as well as traditional applications. Assuming big data applications simplifies managing your big data infrastructure, providing faster results and is more cost-effective. The role of infrastructure, whether real or virtual, is to support applications. This includes important traditional business applications, modern cloud, and mobile and big data applications.

Big data and cloud computing point to the convergence of technologies and trends that make IT infrastructure and their applications more dynamic, more modular and more expendable. Currently, the virtual platform building technology is only in the primary stage, which is mainly based on cloud data center integration technology. Cloud computing and big data projects rely heavily on virtualization. Virtual data is the only way to access and improve heterogeneous environments, such as environments used in big data projects. The cloud computing model allows users to have a default data center that can access data sets that were not previously available by using a shared (API) for disparate data sets.

Big data Security in cloud computing

Big data and cloud are among the most important stages of IT development. Information privacy and security are one of the most important issues for the cloud because of its open environment with very limited user control. Security and privacy affect big data storage and processing because there is a huge use of third party services and the infrastructure used to host important data or to perform operations as growing data and application growth bring challenges.

A solution is provided for the security services and the level of confidence needed through the third party services within the cloud. The data is stored in a central location known as the cloud storage server, where the data is processed somewhere on the servers, so the client has

confidence in the service provider as well as data security. The service level agreement must be standardized to gain trust between service providers and customer . The security of cloud client data varies in protection requirements. Customers require protection of their data only through basic logical access controls, while intellectual property, structured or classified data are confidential and require advanced security controls including encryption, data hiding, login, logging, etc..

The Service Level Agreement (SLA) reflects a service level contract between the user and the service provider. There are rules with service level agreements to protect the data, capacity, scalability, security, privacy, and availability of issues such as data storage and data growth. The technologies available to secure big data, such as registry entry, encryption, and trap detection are essential. In many organizations, big data analytics can be used to detect and prevent malicious hackers and advanced threats. The security of big data in cloud computing is necessary because of the following issues:

1. Protection of big data from malicious intruders and advanced threats.
2. Knowledge about how cloud service providers securely maintain huge disk space and erase existing big data.
3. Lack of standards for checking and reporting big data in the public cloud.

V. CHALLENGES IN BIG DATA AND CLOUD COMPUTING

The security challenges in cloud computing environments fall under several levels: Cloud computing follows the policy of shared resources, where the privacy of data is very important because it faces some challenges like integrity, authorized access, and availability of (backup / replication). Data integrity ensures that data is not corrupted or tampered with during communication. Authorized access prevents data from infiltration attacks while backups and replicas allow access to data efficiently even in case of technical error or disaster in some cloud location.

Big data face some challenges as they can be classified into groups: data sets, processing and management challenges. When dealing with big amounts of data we face challenges such as volume, variety, velocity and verification which are also known as 5V of big data[12]. Among the factors and challenges that affect the processing of big data in a timely manner is the bandwidth and latency. Where several challenges can be summarized in the relationship between big data and cloud computing.

Data Storage: The storage of big data through traditional storage is problematic because hard drives often fail, data protection mechanisms are not effective, and the speed of big data requires storage systems in order to expand rapidly, which is difficult to achieve with conventional storage systems. Cloud storage services offer almost unlimited storage with a great deal of error tolerance, which offers potential solutions to address the challenges of big data storage.

Variety of data: Big data naturally grow, increase and vary, which is the result of the growth of almost unlimited sources of data. This growth leads to the heterogeneous nature of big data. Generally speaking, data from multiple sources of different types and representations are highly interrelated. They have incompatible shapes and are inconsistent. A user can store data in structured, semi- structured or unstructured formats. Structured data format is suitable for today's database systems, while semi- structured data formats are only fairly suitable. Unstructured data is inappropriate because it contains a complex format that is difficult to represent in rows and columns.

Data transfer: The data goes through several stages: data collection, input, processing, and output. Big data transfer is a challenge, so data compression techniques need to be reduced to

reduce the volume, where data volume is a hindrance to transfer speed. It also affects the cost, while cloud computing provides distributed storage resources and data transfer on high-speed lines, reducing costs through virtual resources and resource use at user's request.

Privacy and data ownership: According to (IDC) estimates, by 2020, around 40% of global data will be accessed by cloud computing. Cloud computing provides strong storage, calculation and distribution capability to support big data processing. As such, there is a strong demand to investigate the privacy of information and security challenges in both cloud computing and big data.

What Is Big Data's Relationship To The Cloud?

How does the cloud computing environment correspond to big data? The answer to this question reflects the relationship between them. This is done through the cloud computing features to handle big data, the resources provided by cloud computing, the resource service to provide service to many users where the various physical and virtual resources are automatically set and reset upon request. Cloud computing has access from anywhere to data resources that are spread all over the world by using a (public) cloud to allow those sources faster access to storage. The nature of big data is generated by technologies and locations worldwide, so the cloud resource service provides and helps in the collection and storage of big amounts of data resulting from the use of technologies.

Any system in technology must pass through several main stages. The computer system follows the input, processing and output model. Input is done through devices and then processed through the CPU. Thus, the results of the information are produced. In the relationship between the data and cloud computing, the data is stored on external and remote storage units. On the other hand, in the computer system, the data is stored internally or locally.

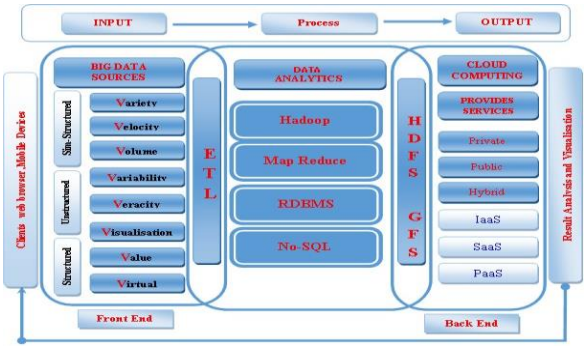


Figure 3 . A Model Showing The Relationship

Common factor between cloud computing and big data

The internet of things represents the new concept of the Internet network, which enables communication between several parties to communicate together, and these parties include smart devices, mobile devices, sensors and other [13] where it is considered effective communication between all elements of architecture so that it can Rapidly deploy applications, process and analyze data quickly to make decisions as quickly as possible.

The architecture represents several systems: objects, gates, network infrastructure, cloud infrastructure. [14] Internet objects can benefit from the scalability and performance of cloud computing infrastructure.

The Internet of Things (IOT) is going to generate a massive amount of data and this in turn puts a huge strain on Internet Infrastructure. As a result, this forces companies to find solutions to minimize the pressure and solve their problem of transferring large amounts of data. Many cloud providers can allow your data to either be transmitted over your traditional Internet connection or via a dedicated direct link. The Internet of things generates huge amounts of data, and cloud computing provides a pathway for these data to navigate. By storing data in the cloud, most companies find that it is possible to access large amounts of big data through the

cloud. And internet of things is all parts of a continuum. Difficult to think of Internet things without thinking about the cloud, it is difficult to think of the cloud without thinking about the big data analyzes. Which generates a lot of data, this data is stored in the cloud computing, cloud computing is the only technology suitable for filtering, analysis, storage and access to IoT and other data in ways that are useful, as these data constitute large quantities must be analyzed, Objects is a common factor between the erased cloud and big data.

Common points between big data and the cloud

The cloud computing environment consists of several user terminals and service provider. The service provider must ensure that users have on-demand resources or otherwise access their data, systems and applications on a regular basis and is available throughout the service without any interruption.

Data, whether small or big, require storage, processing and security, but the volume and capacity of data requirements differ in accordance with the volume of the data, so cloud computing must provide storage, processing and security requirements for big data in its environment.

Cloud computing provides security, depending not on data volume but the availability of security and protection for small and big data. The service provider guarantees complete confidentiality of user data of all kinds and only allows access to authorized users. The user can connect to the network in these resources through a simple software interface that simplifies and ignores many internal details and processes.

Cloud computing saves the cost of storing and processing data to the user through the availability of geographically dispersed servers and the availability of virtual server technology. The service provider must ensure that the devices and equipment are sufficiently available, and restricted by an integrated and documented entry system for reference when needed.

VI. CONCLUSION

Big data and cloud computing have been studied from several important aspects, and we have concluded that the relationship between them is complementary. Big data and cloud computing constitute an integrated model in the world of distributed network technology. The development of big data and their requirements is a factor that motivates service providers in the cloud for continuous development, because the relationship between them is based on the product, the storage and cloud for continuous development, because the relationship between them is based on the product, the storage and processing as a common factor. Big data represents the product and the cloud represents the container. The big data is concerned with the capacities of cloud computing. On the other hand, cloud computing is interested in the type and source of big data. Compatibility between them is summarized in Table 2. Cloud computing represents an environment of flexible cloud for continuous development, because the relationship between them is based on the product, the storage and concerned with the capacities of cloud computing. On the other hand, cloud computing is interested in the type and source of concerned with the capacities of cloud computing. On the other hand, cloud computing is interested in the type and source of big data. Depending on the relationship between them, a model was prepared to show the relationship between them as in Figure 3. Compatibility between them is summarized in Table 2. Cloud computing represents an environment of flexible distributed resources that uses high techniques in the processing and management of data and yet reduces the cost. All these cloud for continuous development, because the relationship between them is based on the product, the storage and processing as a common factor.

Characteristics show that cloud computing has an integrated relationship with big data. Both are moving towards rapid progress to keep pace with progress in technology requirements and users.

REFERENCES:

1. Kshetri, Nir. "Cloud computing in developing economies." Computer 43, no. 10 (2010): 47-55.
2. https://en.wikipedia.org/wiki/Cloudcomputing
3. Klous, Sander, and Nart Wielaard. We are Big Data: The Future of the Information Society. Springer, 2016.
4. https://www.internetworldstats.com/stats.htm
5. https://www.ibm.com/big-data/us/en/ Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Information Fusion.2016 Mar 31;28:45-59.
6. Boyd, D., & Crawford, K. (2011, September). Six provocations for big data. In A decade in internet time:.
7. SHAN, Y. C., Chao, L. V., ZHANG, Q. Y., & TIAN, X. Y. (2017). Research on Mechanism of Early Warning of Health Management Based on Cloud Computing and Big Data. In Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management 2016 (pp. 291-294). Atlantis Press, Paris.
8. Parvin Ahmadi Doval Amiri and Mina Rahbari Gavgani, 2016. A Review on Relationship and Challenges of Cloud Computing and Big Data: Methods of Analysis and Data Transfer. *Asian Journal of Information Technology, 15: 2516-2525*
9. Chen, Min, et al. Big data: related technologies, challenges and future prospects. Heidelberg: Springer, 2014.
10. Demchenko, Yuri, et al. "Big security for big data: Addressing security challenges for the big data infrastructure." Workshop on Secure Data Management. Springer, Cham, 2013. Environments and evaluation of resource provisioning algorithms." Software: Practice and experience 41.1 (2011): 23-50.
11. R.Subhulakshmi, S.Suryagandhi, R. Matubala, P.Sumathi, An evaluation on Cloud Computing Research Challenges and Its Novel Tools, International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST) Volume 2, Special Issue 19, October 2016.
12. Fonseca, N., & Boutaba, R. (2015). Cloud services, networking, and management. John Wiley & Sons.
13. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 1(1), 7-18.
14. Ahmed, F. F. (2015). Comparative Analysis for Cloud Based e-learning. Procedia Computer Science, 65, 368-376.

AUTHORS PROFILE

First Author:

Bharat Batham is currently working as a Asst. Professor in Computer Science and Application department of Atal Bihari vajpayee Hindi University Bhopal. His research area is Classification of cloud computing and security in Big Data environment. He has various research papers are published in National (4) and International Journals (3) of repute. He has vast teaching and academic development at leading institutions of Bhopal.

Second Author:

Dr. Shailja Sharma received her Ph.D.in Computer Science from Jiwaji University, Gwalior INDIA. Presently She is Associate Professor of computer Science & Engineering, Rabindranath Tagore University ,Bhopal, INDIA. She served as

Associate Professor in Computer Science in Career College, Bhopal also Guest Faculty in Barkatullah University, Bhopal. She has various research papers are published in National and International Journals of repute. She is reviewer of International Journal of Network Security (IJNS).Her research interest includes Computer Networks, Network Security, Internet & Web Technology, Client-server Computing ,Image processing and IT based education. She has 20 years of teaching experience.

Innovative Technique in Combating With Stress in Employees at Workplace

Pushpa Tiwari[1] S.Veenadhari[2]
[1,2] Computer Science and Engineering, RNTU, Bhopal, India

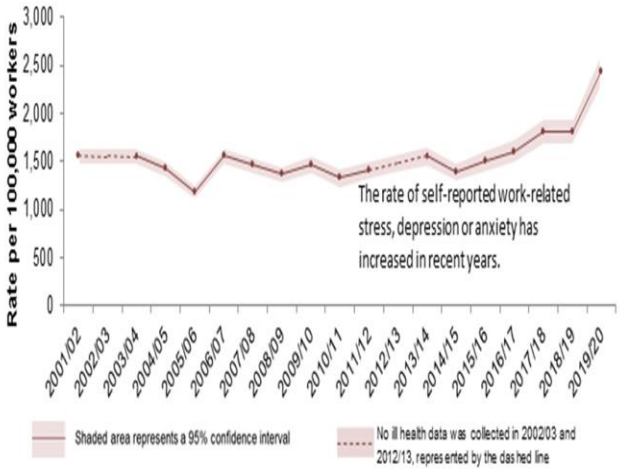
ABSTRACT
Stress at workplace has become an increasingly growing phenomenon. The reasons can be meeting work deadlines, complex tasks, challenge to cope up with changing technology, role ambiguity, job insecurity, etc. at the workplace. Such situations may lead to anxiety, depression or even heart ailments in the stressed employees. The risk to organization with such employees will result in increased absenteeism, poor communication, increase in conflict with colleagues and reduction in quantity and quality of work, thereby impacting the productivity of the organization. This covid-19 times have added to the stress of employees in terms of cost cut, job loss and productivity. It is a good employment practice to assess the risk of stress amongst employees. It becomes a necessity for organizations to look for some modern ways, which help in reducing stress as the traditional ways are time consuming, expensive and not so effective. It becomes the need of time to come up with ways such that employees are able to cope up with stress on their own. With the advancement in modern information technology, virtual reality (VR) is being used for the diagnosis, assessment, and treatment of various disorders and phobias. Virtual Reality is a technology which presents a very realistic and immersive environment. This paper is a study to investigate whether virtual reality can help in reducing the stress of employees at workplace. It also checks for the availability of infrastructure in implementing virtual reality at workplace. Thus understanding of new technology such as virtual reality could open new doors in combating stress in employees at workplace.
Keywords : Stress, employees, virtual reality, workplace.

I. INTRODUCTION

Mental stress can be regarded as the disturbance in the mental balance of a person. According to Labour Force Survey (LFS) for the year 2019-20, 828,000 workers have

suffered from work-related stress, depression or anxiety and 17.9 million working days were lost due to work-related stress, depression or anxiety. In 2019/20 stress, depression or anxiety accounted for 51% of all work-related ill health cases [1]. The employees stated that the main factor of work related stress was huge work load pressure to be completed in limited time. The respondents were given lots of responsibilities and different roles with little or no managerial support.

Stress is the process in which an individual react when opened to external or internal problems and challenges[2]. According to the National Institute of Safety and Health (NIOSH) Job stress can be defined as the harmful physical and emotional responses that occur when the requirements of job, do not match the capabilities, resources or needs of the worker[3]. The impact of stressful working environment may lead to increased absenteeism or the employee may become casual or unpunctual. The employee may plan to quit the job or seek for another. It is stated that work related stress , anxiety or depression has increased significantly higher in 2019/20 than the previous years [1].



Source: LFS annual estimate, from 2001/02 to 2019/20

II. CAUSES OF STRESS AMONG EMPLOYEES

Productivity and performance are the two important decisive factors in the growth and success of an organization. These two factors are dependent on the well-being of the employees. Organizations are facing great challenge due to rapid changing in global scenario which employees are not able to handle. Various stressors are discussed below:

A. The Design of Task

Heavy workload, infrequent rest breaks, long work hours and in shifts, hectic and routine tasks with little inherent meaning and do not utilize workers' full and actual capability. Employees work to the point of exhaustion.

B. Management Styles

Generally work is entrusted on the employees irrespective of their capability and choice. The employees are not made part of decision making nor are they asked. Employees complain that important information do not reach them on time and not through proper channel because of poor communication in the organization. Many organizations do not have proper family friendly policies.

C. Interpersonal Relationships

Employees complain about poor social environment and lack of support or help from coworkers and supervisors.

D. Work Roles

It is proved by the scientists that employees with uncertainty in job are more stressed and have more health issues as compared to employees who actually lose their jobs. Employees in job have multiple roles and responsibilities to perform.

E. Career Concerns

With the onset of technological growth in IT industry, the working pattern has changed, for which the workers may not be prepared. Employees resist the change as they feel that the new change will make their work more difficult or they will be forced to learn new technologies. They may not be interested to come out of their comfort zone. They also resist because of bad execution of change in the organization. These career concerns make them feel stressed. They fear losing their jobs. In existing job also, employees do not get better opportunities to grow. Many times promotion to deserving employees is with held which makes the employees dissatisfied. With rapid changes, the workers might feel like quitting their existing job. But they may not get equivalent opportunity which in turn makes them stressed. Thus career concerns are responsible for hyper arousal response of employees.

F. Environmental Conditions

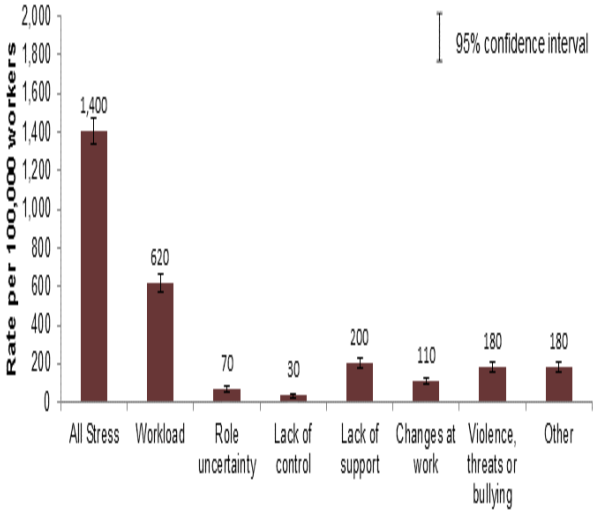
Unpleasant or dangerous physical conditions such as crowding, noise, air pollution, or ergonomic problems are reasons for stress in employees.

G. Economic Uncertainties

Employees are also affected by the economic uncertainties such as recession of jobs and downsizings. Changes in political situation or economic disability affects the employees. The world lost nearly 400 million full-time jobs in the year's second quarter (April-June

2020) due to covid -19 pandemic, said the International Labour Organization (ILO) [1]

The bar graph depicts various causes resulting in stress in Great Britain between 2009/10 – 2011/12 reasons of stress. Estimated prevalence rates of self-reported stress, depression or anxiety in Great Britain, by how caused or made worse by work, averaged 2009/10 - 2011/12. [1]

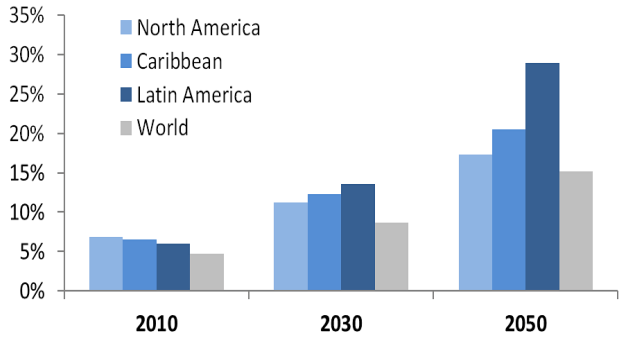


Source: LFS, estimated annual average 2009/10 - 2011/12

III. IMPACT OF JOB STRESS

Dr. Hans Selye used the term eustress for positive stress and distress for negative stress
Moderate as well as appropriate level of stress proves to be significant in triggering the passion for work, igniting inspiration and identifying hidden talents. This type of stress is known as Eustress, while distress slow down the work performance and affects adversely[2].

Early signs of Job Stress may be headache, sleep disturbances, difficulty in concentrating, short temper, upset stomach, job dissatisfaction or low morale. These symptoms if not controlled may end up in chronic diseases like heart problems, musculoskeletal Disorders, Psychological Disorders, suicides, cancer, ulcers, and impaired immune function[3]



Source: Pan American Health Organization

The above bar graph depicts the increasing number of stressed employees and ailments related to it, in various countries from 2010 to 2050. This indicates there should be some coping mechanism. Traditional ways of coping with workplace stress[4,5] are as follows:

- Encourage workplace wellness.
- Revamp the habitat
- Allow for flexible hours and remote working.
- Encourage social activity
- Provide onsite or distance counseling.
- Recognize your employees
- Guided Imagery's Effects

Conventional methods are quite cumbersome. Traditional anxiety treatments are expensive, inaccessible, and underutilized. These forms of therapies expose the patients to life-threatening hazards [6,7] Job stress needs to be minimized to the extent that the productivity and

health of the employee is maintained which could lead to a productive organization.

To help people reduce anxiety, we need to start thinking outside of the therapist's office. A review of virtual reality research shows that this modality has been successfully used in the treatment of a variety of anxiety disorders. VR can generate strong feelings of "presence". "Presence" is subjective feeling of being in another place. Virtual environment provide a great way to quickly transfer the employee out of a stressful situation into a personalized experienced tailored to induce positive emotions and thoughts.

IV. ABOUT VIRTUAL REALITY

Virtual Reality (VR) is a technology which allows user to interact with computer-generated 3D environment. The impact is so immersive that the user feels like being transported to new world and starts living and feeling it. VR is a computer interface, which tries to mimic real world beyond the flat monitor to give an immersive three dimensional visual experiences. The immersive effect of VR gives feeling of enjoyment or engagement or excitement either through place illusion, plausibility illusion or body ownership illusion.

Virtual Reality Therapy is commonly being seen used in the domains of stroke rehabilitation, therapeutic application in Post-Traumatic Stress Disorders seen in war veterans, pain alleviation for burn victims and many more[8,9,10]

Virtual Reality has been successful in curing various ailments and phobias. It is used in various industries like gaming, entertainment and education. Listed in the table are few ailments or phobia's which have been successfully treated using virtual reality.

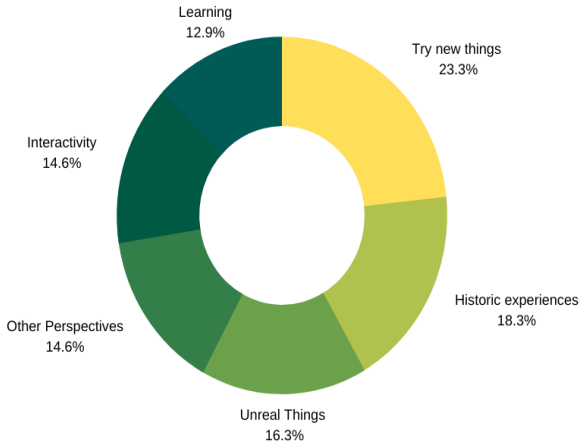
Sr. No.	Ailment or Phobia
1	Mental Illness
2.	Flight Phobia
3.	Acrophobia and Agoraphobia
4.	Visual auditory impairment
5.	Relief from burning pain
6.	Cerebral Palsy
7.	Controlling addictive disorders
8.	Developing positive social behavior
9.	Spider phobia
10.	Eating disorder

Table No. 1

Today's VR systems are more affordable, lightweight, smaller and comfortable. 84.4% people using VR have confessed of being comfortable in virtual reality environment. The advent and arrival of readily available virtual reality equipments has the technology easily available to users. The emergence of commercially accessible virtual reality equipment has made the technology more available to everyday users.[11]

Around twelve people were asked to use VR systems one after the other. The participants were made to see some old historic monuments or were engaged in various exciting sports like sky diving, scuba diving and bungee jumping. Many participants were exposed to these activities for the first time many feared participating in these activities in reality. But with VR exposure their interest level increased most of them wanted to experience it again and again. This was possible only because of the interactive and immersive feature of VR systems.

The pie chart shows the increased interest levels in various activities emphasizing that VR is an exciting new technology that will redefine the way we socialize, learn or entertain ourselves. Various categories taken into consideration are try new things, experience historic content, experience unreal things, develop other perspectives, have an interactive experience and using VR for learning, there is a positive increase in each category [11].



VR usage enhances interest levels in various categories

V. AIM OF RESEARCH

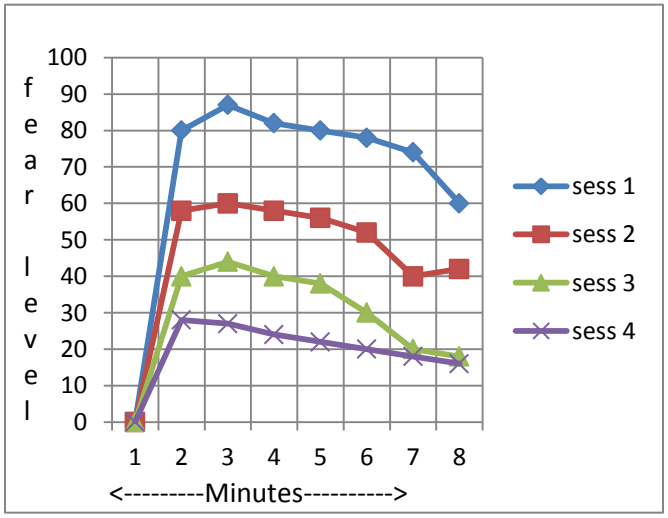
Aim of the study is exploring innovative way in combating with stress in employees at workplace using virtual reality.

VI. DISCUSSION

Cognitive behaviour therapy (CBT) is a common treatment for anxiety disorders. [12]. However, many people avoid seeking mental health treatment for reasons including the stigma of seeking therapy, uncertainty whether the symptoms are considered severe enough and financial and time constraint.[13]. Despite the availability of treatments for anxiety disorder, only 40% of patients seek out treatment, and of those who do, less than half actually benefit from the treatment [14].

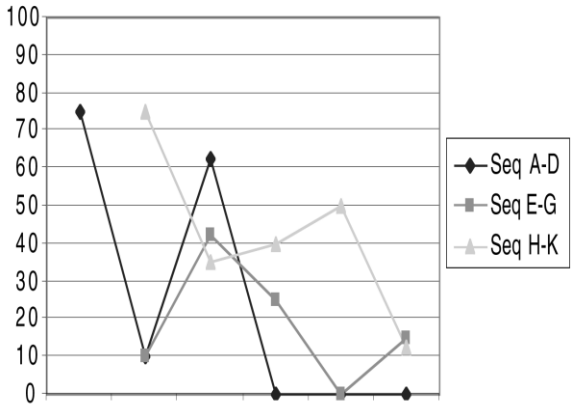
VR has brought a sharp change in the technological development with it immersive and interactive characteristics.In the conventional stress therapies the patients are made to visualize the calming images on their own. Many struggle in doing so, possibly because of anxiet. But VR based stress therapy generates such environment for the patients[15]. The release of modern consumer-grade virtual reality devices, such as Oculus Rift CV1 (Oculus, 2016) and Gear VR (Oculus, 2015) that are compatible with modern smart phones, laptops or desktop computers has provided an opportunity for this technology to be used in various fields. The currently available VR devices provide more realistic and immersive experiences, are portable, have better graphics and are more user friendly [16]. The sense of being present in the virtual environment is facilitated through the use of high-resolution HDMs and motion tracking systems that are in place to track and capture the position of the user's head in order to produce a dynamic 3D scene that changes with the position of the head. VR technology has the potential to help these people, overcome mental problems in a relatively low-cost environment[17]. The first documented use of VR in exposure therapy was in 1995 and it focused on the fear of heights [18].

One of the current clinical uses of this technology is through VR exposure therapy (VRET), where the patient is gradually exposed to a negative stimulus in a controlled and safe environment to reduce anxiety and post traumatic stress disorder [19]. The first documented use of VR in exposure therapy was in 1995 and it focused on the fear of heights [20]. The results of this study showed VR to be an effective tool in reducing fear of heights. The graph indicates the fear of heights decreasing after each exposure therapy.



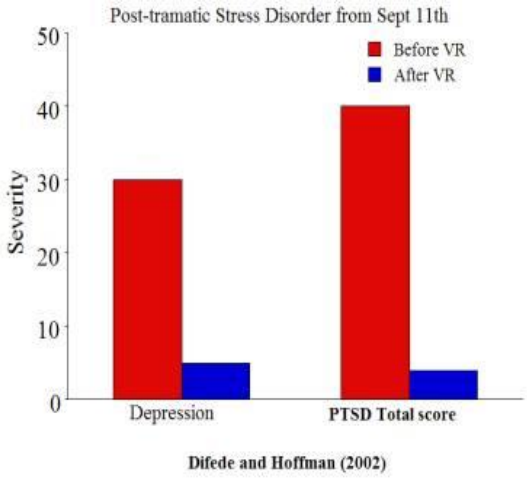
VRET Exposure

A very interesting case of WTC attack survivor, who developed acute post trauma stress disorder (PTSD) after the attack, is an excellent success example of VRET. The traditional therapies had minimal impact on her ailment. But when she was given a course of six – one hourly VRET exposure, there was 83% reduction in her depression and 90% reduction in her PTSD symptoms [21].



Source : Cyber Psychology and behavior, Volume 5

The above graph depicts the sequence of exposures of Virtual Reality Exposure Therapy on the survivor of World Trade Center(WTC) attack. During the therapy the survivor was exposed to virtual environment which had planes flying at a very high speed, planes crashing with WTC building with loud sound effect and explosion. The content contained virtual people jumping to death from burning buildings, tower collapsing and dust clouds [21]



The psychology of reducing pain through virtual reality is diverting attention and engaging in an activity liked by the patient. The incoming pain signals can be interpreted as painful or not, depending on what the patient is thinking. But if the patient having severe pain is taken to another virtual world, his entire concentration is involved there and hence the patient gives less attention to the pain signals. The pain signals are suppressed by the computer-generated environment [22].

The upper edge of using VRET is that the patients are aware that they are in artificial environment, but still they give response as in real world. This makes the patient confidant enough to face new challenges in VR. He is ready for different treatment strategies. The content in the virtual environment stimulate the sense organs of the user. The output devices present the VR content or environment to the users which generate an immersive feeling. Audio is an equally important component to stimulate a user's senses and achieve immersion. The audio system provides three-dimensional sound effect. Most virtual reality headsets provide users with the option to use their own headphones in conjunction with a headset.

VII. INFRA STRUCTURE SUPPORT AND AVAILABILITY

The advent of commercially available VR headsets, across a range of price points, with varying complexity, alongside promises of VR for empathy, engagement and the chance to see the world is an appealing prospect in diverse fields.
Many firms, these days are trying to bring more immersive effect in the VR systems by developing near to real content to be displayed on the VR's output devices.

Relax VR virtually transports the user to a quiet and peaceful location. The app is compatible with Google's Cardboard and Daydream, as well as Samsung's Gear VR headsets. The virtual locations can be beaches or oceans combined with soothing voice of therapist and therapeutic music. The app helps the user to self manage his stress by diverting attention from the current scenario [23].

Starflight VR is another app that gives user a 3D flight in space. Objective of this app is also to bring peace and tranquility in the mind of the user. The virtual flight is accompanied with mesmerizing music[24] .

Lumen a VR experience that allows the user to take breathing exercises. It is launched through LIFE VR's mobile app for iOS and Android devices. Here also the user can interact with the sights and sounds of the environment. [24]

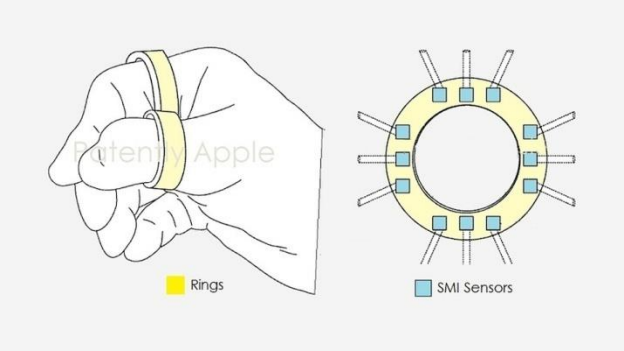
Build VR by Victorian-startup sets off positive feeling in barely responsive dementia patient. The experiences offer distraction when dementia patients are experiencing boredom or displaying repetitive behavior. [24]

Various governments like Australia have released its strategy to support the timely rollout of 5G in Australia [25]. New ministers for regional communications, cyber security, intelligence, innovation, and digital transformation have been announced by government.

Many countries provide free wifi for internet access. Indian government's move to free up spectrum in 5Ghz frequency band has increased the availability of radio waves for WiFi services by about 12-folds. PM WANI WiFi has been introduced with an intention to elevate wireless internet connectivity.

The mobfish VR STUDIO uses cloud transcoding to ensure perfect interaction between hardware and media. With its offline function, the system becomes completely independent of Internet connection.

Facebook has developed neural wristband to support augumented reality glasses which will be capable of detecting nerve signals to interpret complex hand gestures.[26]. Apple is also working on smart ring which will act as controller for VR and will also give haptic feedback[27]



Source: Patently Apple

Apple's patent shows the use of two rings as a controller along with the SMI sensor layout. With all these exceptional constructive developments in the field of VR, will lead to, making virtual reality a house hold commodity in near future.

VIII. CORPORATE WELLNESS PROGRAMMES

An employee spends 40 hours in a week at his workplace. Hence the workplace should promote mental and physical health. Employers like Google, Microsoft, earth friendly products to name a few understand this and have given rise to many corporate wellness programs are the result of these employers. They feel wellness should be integrated

into company structure [28]. Companies provide wellness programs because they understand that healthier teammates enjoy their work more and can do more for their customers and their clients.
Employers constantly look for new ways to mitigate stress at work. Some have created perks and policies that enable a better work-life balance. Organizations are converting extra space into wellness spaces where employees can take a nap, relax, or even do some yoga [28].

In this corporate wellness structure, employees may utilize VR facility by visiting a wellness area where the technology is housed. Employees can maintain their own headset. Personal headsets might be more practical in big companies with many employees, especially if individuals buy them on their own. In this era of social distancing personal headset is more advisable. Since this will enable workers to pop into relaxing VR environments whenever need be. It would also prevent potential over bookings of wellness rooms or area. It offers a seamless and immediate transition to the de-stressing environment. Individual VR equipment especially during this covid-19 times, also helps with sanitation.

IX. CONCLUSION

The nature of work is changing at whirlwind speed. Perhaps now more than ever before, job stress poses a threat to the health of workers and, in turn, to the health of organizations. Virtual reality (VR) may be one such management technique for individuals suffering from anxiety and wanting to be free from it. VR devices are available that can help alleviate anxiety pressures, by immersing the user in an interactive yet calming synthetic environment. Virtual reality provides a refreshing option to include in a workplace wellness program. The technology offers employees an ability to transport, at least visually, to different environments that are conducive to relaxation and enjoyable experiences. With displays that continue to improve, VR presents an increasingly realistic simulation of environments. This couples with enhancements to motion tracking quality and graphics that further round out the effectiveness of creating a convincing presence in spaces created by VR. Now we are seeing the development of VR hardware in the form of wearable devices and smart phones with decrease in price of VR devices. Thus employees can use virtual reality therapy at their workplace for a therapeutic experience. It will be an innovative approach to cope with stress at workplace with no logistic effort.

REFERENCES

[1]https://www.hse.gov.uk/statistics/causdis/stress.pdf. "Work-related stress, anxiety or depression statistics in Great Britain, 2020"

[2] George Essel and Patrick Owusu Causes of students' stress, its effects on their academic success, and stress management by students

[3] https//www.cdc.gov/niosh/docs/99-101/pdfs/99-101.pdf

[4] Kristin Ryba ,"7 ways to reduce stress in workplace"

[5] "Using guided imagery for stress management", https://www.verywellmind.com/using-guided-imagery-for-stress-management-3144610

[6] Schumacher S, Miller R, Fehm L, Kirschbaum C, Fydrich T, Ströhle A. " Therapists'and patients' stress Therapists' and patients' stress responses during graduated versus flooding in vivo exposure in the treatment of specific phobia: A preliminary observational study", Elsevier publication, volume 230, Issue 2, https://doi.org/10.1016/j.psychres.2015.10.020

[7] Turner R, Susan M, Napolitano S., " Cognitive behavioral therapy. Encyclopedia of Cross-Cultural School Psychology", p. 226-229. Springer, 2010.

[8] Bohil CJ, Alicea B, Biocca FA. Virtual Reality in neuroscience research and therapy. Nature Reviews Neuroscience. 2011;12:753-62.

[9] Rizzo A, Hartholt A, Grimani M, Leeds A, Liewer M. Virtual reality exposure therapy for combat-related posttraumatic stress disorder. Computer. 2014;47(7):31-37.

[10] Rizzo A, Schultheis M, Kerns KA, Mateer C. Analysis of assets for virtual reality applications in neurophysiology. Neurophysiological Rehabilitation: An International Journal. 2004;14(1-2):207-39.

[11]https://uocsweb03.uocslive.com/ISTE/ISTE2018/PROGRAM_SESSION_MODEL/HANDOUTS/110768101/ActualDatafromVRUsageWhatHappensWhenVRisinSchools_RP.pdf ,

[12] Hans, Eva, & Hiller, Wolfgang. (2013). A meta-analysis of nonrandomized effectiveness studies on outpatient cognitive behavioral therapy for adult anxiety disorders. *Clinical Psychology Review*, *33*(8), 954–964. https://doi.org/10.1016/j.cpr.2013.07.003

[13] Daniel Eisenberg, Marilyn F. Downs, Ezra Golberstein, Kara Zivin, "Stigma and Help Seeking for Mental Health Among College Students", https://doi.org/10.1177/1077558709335173

[14] John W G Tiller, 'Depression and anxiety", https://doi.org/10.5694/mja12.10628

[15] Farhah Amaliya Zaharuddin, Nazrita Ibrahim, Eze Manzura Mohd Mahidin, Azmi Mohd Yusof, Mohd Ezanee Rusli), " Virtual Reality Application for Stress Therapy: Issues and Challenges "

[16] Philip Lindner, Alexander Miloff, William Hamilton, Lena (Lotta) Reuterskiöld, Gerhard Andersson,Mark B Powers, Per Carlbring, "Creating state of the art, next-generation Virtual Reality exposure therapies for anxiety disorders using consumer hardware platforms: design considerations and future directions", DOI: 10.1080/16506073.2017.1280843

[17] Jessica L Maples-Keller 1, Brian E Bunnell, Sae-Jin Kim, Barbara O Rothbaum, "The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders", DOI: 10.1097/HRP.0000000000000138

[18] Rothbaum, Barbara Olasov Hodges, Larry F. Kooper, Rob Opdyke, Dan Williford, James S. North, Max**,** " Effectiveness of computer-generated (virtual reality) graded exposure in the treatment of acrophobia"

[19] Dayne Camara, Richard Edward Hicks, "Using virtual reality to reduce state anxiety and stress in university students : an experiment."

[20] Barbara Olasov Rothbaum, Larry F.Hodges, Rob Kooper, Dan Opdyke, James S.Williford, MaxNorth, "Virtual reality graded exposure in the treatment of acrophobia", https://doi.org/10.1016/S0005-7894(05)80100-5

[21] Joann Difede, Hunter G. Hoffman, "Virtual Reality Exposure Therapy for World Trade Center Post-traumatic Stress Disorder", CYBERPSYCHOLOGY & BEHAVIOR, Volume 5, Number 6, 2002

[22] Hunter G Hoffman 1, Gloria T Chambers, Walter J Meyer 3rd, Lisa L Arceneaux, William J Russell, Eric J Seibel, Todd L Richards, Sam R Sharar, David R Patterson, "Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures", DOI: 10.1007/s12160-010-9248-7

[23] https://www.zdnet.com/article/virtual-reality-startup-relax-vr-wants-to-relieve-stress-in-corporate-environments/

[24] https://techhq.com/2018/05/feeling-stressed-at-work-let-vr-help-you/"

[25]https://www.communications.gov.au/departmental-news/5g-enabling-future-economy

[26] https://www.reuters.com/article/us-facebook-technology-idUKKBN2681OS

[27] https://gadgets.ndtv.com/wearables/news/apple-ring-smart-patent-vr-headset-glass-ar-mr-uspto-virtual-reality-2391826

[28] https://snacknation.com/blog/successful-corporate-wellness-programs/

[29] Suji Kim, Eunjoo Kim, "The Use of Virtual Reality in Psychiatry: A Review**",** doi: 10.5765/jkacap.190037

INTRUSION DETECTION AND LARGE MIXED DATA

[1]Neelu Singh, [2]Dr. Varsha Jotwani
Research Scholar (RNTU)
(Rabindranath Tagore University)

ABSTRACT: In recent days due to technological growth, there is a tremendous increase in data generation and data distribution. An intrusion detection system (IDS) can be a hardware device or a software application that monitors the network or a host for malicious activity or policy defilement. An IDS is categorized into two types based on the audit source location as network IDS (NIDS) and host IDS (HIDS). An intrusion detection system has become an important mechanism to detect a wide variety of malicious activities in the cyber domain. However, this system still faces an important limitation when it comes to detecting zero-day attacks, concerning the reduction of relatively high false alarm rates. It's no longer consider the tasks of monitoring and analyzing network data in isolation, but instead optimize their integration with decision-making methods for identifying anomalous events. Represents a scalable framework for building an effective and lightweight anomaly detection system. This framework includes three modules of capturing and logging, pre-processing and a new statistical decision engine, called the Dirichlet mixture model based anomaly detection technique. The first module sniffs and collects network data while the second module analyses and filters these data to improve the performance of the decision engine. The empirical results showed that the statistical analysis of network data helps in choosing the best model which correctly fits the network data. Additionally, the Dirichlet mixture model based anomaly detection technique provides a higher detection rate and lower false alarm rate than other three compelling techniques. These techniques were built based on correlation and distance measures that cannot detect modern attacks which mimic normal activities, whereas the proposed technique was established using the Dirichlet mixture model and precise boundaries of interquartile range for finding small differences between legitimate and attack vectors, efficiently identifying these attacks.

Keywords: Scan Detection; Virus Detection; Anomaly Detection; Security.

1.INTRODUCTION

Data mining is a famous technological innovation that converts portions of data into useful understanding which can help the data owners/users make knowledgeable choices and take clever actions for his or her personal gain. In specific terms, information mining seems for hidden patterns amongst huge sets of information which could help to understand, expect, and manual destiny behavior. A extra technical explanation: information Mining is the set of methodologies used in reading records from numerous dimensions and perspectives, finding previously unknown hidden styles, classifying and grouping the statistics and summarizing the recognized relationships. Facts mining is, at its core, sample locating. Records miners are specialists at the usage of specialized software program to locate regularities (and irregularities) in massive statistics units. [1]

right here are a few specific matters that information mining may make contributions to an intrusion detection project: eliminate regular activity from alarm data to allow analysts to focus on actual attacks discover false alarm generators and "bad" sensor signatures discover anomalous activity that uncovers a real attack, discover long, ongoing patterns (exclusive IP address, equal activity) to perform those responsibilities, data miners use one or more of the following strategies: data summarization: with records, which includes finding outliers Visualization: imparting a graphical summary of the data clustering of the facts into natural categories.

Association rule discovery: defining normal activity and permitting the discovery of anomalies category: predicting the category to which a specific record belongs data mining has many programs in security along with in country wide protection as well as in cyber safety (e.g., virus detection).

The threats to national security encompass attacking buildings and destroying important infrastructures which includes electricity grids and telecommunication structures. Statistics mining techniques are getting used to discover suspicious individuals and organizations, and to discover which people and corporations are able to carrying out terrorist sports. Cyber security is concerned with shielding laptop and community structures from corruption due to malicious software program Consisting of Trojan horses and viruses. Information mining is also being carried out to provide solutions consisting of intrusion detection and auditing.

On this paper we are able to focus specially on data mining for cyber protection programs. Records mining for cyber security packages for instance, anomaly detection techniques might be used to discover unusual styles and behaviors. Link analysis can be used to trace the viruses to the perpetrators. Category may be used to institution numerous cyber-attacks and then use the profiles to discover an attack while it occurs. Prediction may be used to determine capacity future assaults depending in a manner on statistics learnt approximately terrorists through e-mail and make contact with conversations. Data mining is likewise being carried out for intrusion detection and auditing the conventional method to securing pc structures

towards cyber threats is to layout mechanisms including firewalls, authentication gear, and digital private networks that create a protecting guard. However, these mechanisms nearly continually have vulnerabilities. They cannot ward assaults which are usually being adapted to take advantage of system weaknesses that are frequently due to careless design and implementation flaws. [2] This has created the need for intrusion detection, security era that enhances conventional protection procedures by way of tracking structures and figuring out pc assaults. Conventional intrusion detection strategies are primarily based on human specialists full-size knowledge of assault signatures which can be individual strings in a messages payload that imply malicious content. Signatures have several obstacles. They cannot stumble on novel assaults, due to the fact someone have to manually revise the signature database ahead for each new sort of intrusion found. once someone discovers a brand new assault and develops its signature, deploying that signature is frequently delayed. Those barriers have brought about an growing hobby in intrusion detection strategies primarily based on information mining.

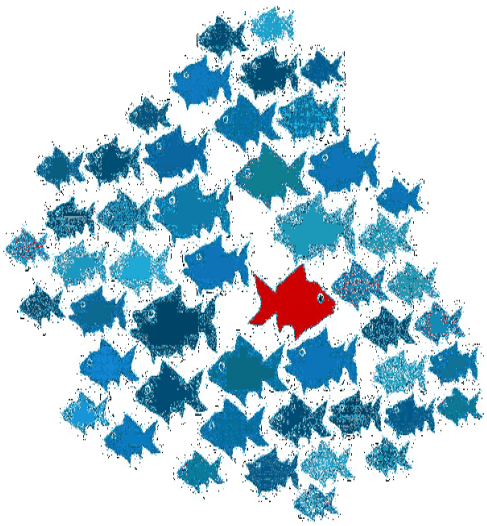
2. DATA MINING FOR NETWORK SECURITY

Source: Internet Based Image 1

2.1 Overview

This section discusses information related terrorism. By information related terrorism we mean cyber terrorism as well as security violations through access control and other means. Malicious software such as Trojan horses and viruses are also information related security violations, which we group into information related terrorism activities. In the next few subsections we discuss various information related terrorist attacks. [3]

2.2 Anomaly Detection



Source: Internet Based Image 2

Anomaly detection processes construct models of everyday records and locate deviations from the regular model in discovered information. Anomaly

detection implemented to intrusion detection and computer security has been an lively region of studies because it was initially proposed by Denning. Anomaly detection algorithms have the benefit that they could discover rising threats and attacks (which do not have signatures or classified information similar to them) as deviations from everyday usage. furthermore, not like misuse detection schemes (which build category models using classified records and then classify an observation as normal or attack), anomaly detection algorithms do no longer require an explicitly classified training data set, which is very suitable, as classified records is tough to achieve in a actual network setting.

2.3 Profiling Network Traffic Using Clustering

Clustering is broadly used information mining technique which corporations comparable items, to gain meaningful agencies/clusters of statistics items in a records set. These clusters constitute the dominant modes of conduct of the records objects determined using a similarity measure. A facts analyst can get high degree information of the characteristics of the records set with the aid of studying the clusters. Clustering offers an powerful approach to find out the expected and surprising modes of behavior and to reap a excessive level information of the community visitors.[4]

2.4 Scan Detection

A precursor too many attacks on networks is usually a reconnaissance operation, extra normally referred to as a scan. Figuring out what attackers are scanning for can alert a system administrator or security analyst to what services or kinds of computer systems are being focused. Knowing what services are being targeted before an attack permits an administrator to take preventative measures to shield the assets e.g. installing patches, firewalling services from the outside, or eliminating services on machines which do not need to be running them.

2.5 Cyber-terrorism

Source: Internet Based Image 3Insider Threats, and outside assaults Cyber- terrorism is one of the fundamental terrorist threats posed to our state today. As we've mentioned in advance, this threat is exacerbated by the massive quantities of information now available electronically and on the internet.[5] assaults on our computer systems, networks, databases and the internet infra-structure may be devastating to agencies. Its miles anticipated that cyber-terrorism should cause billions of greenbacks to businesses. A classic instance is that of a banking statistics device. If terrorists assault such a gadget and burn up money owed of price range, then the bank ought to lose millions and possibly billions of dollars. By crippling the computer gadget thousands and thousands of hours of productivity will be misplaced, that is ultimately equal to direct monetary loss. Even a easy energy outage at work via a few twist of fate ought to purpose numerous hours of productiveness loss and as an end result a major financial loss. Therefore its miles essential that our records systems be secure. [6]

We talk various styles of cyber-terrorist attacks. One is the propagation of malicious cellular code that may harm or leak touchy files or different statistics; any other is intrusions upon laptop networks. Threats can arise from outdoor or from the internal of an organization. Out of doors attacks are assaults on computers from someone out of doors the agency. We hear of hackers breaking into laptop structures and inflicting havoc within an organization. Some hackers spread viruses that damage files in various pc systems. however a more sinister hassle is that of the insider threat. Insider threats are noticeably well understood within the context of non-information associated assaults, but records related insider threats are frequently disregarded or underestimated. human beings interior an agency who have studied the business' practices and approaches have an substantial advantage while growing schemes to cripple the organization's information assets. These people may be regular personnel or even those running at computer centers. The trouble is pretty critical as a person can be masquerading as a person else and causing all styles of harm. in the following couple of sections we can look at how facts mining may be leveraged to come across and perhaps prevent such attacks.

2.6 Credit Card Fraud and Identity Theft

We're hearing lots nowadays about credit score card fraud and identification theft. In the case of credit card fraud, an attacker obtains someone's credit score card and uses it to make unauthorized purchases. By the time the proprietor of the cardboard turns into aware of the fraud, it might also be too past due to reverse the damage or recognize the culprit. A comparable trouble takes place with telephone calling playing cards. In fact this type of assault has come about to me personally. Perhaps at the same time as i was making smartphone calls the usage of my calling card at airports a person noticed the dial tones and reproduced them to make free calls. This becomes my organization calling card. Fortunately our telephone organization detected the trouble and informed my organization. The problem changed into dealt with right now. An extra serious theft is identity robbery. Right here one assumes the identification of any other person by obtaining key private facts including social protection quantity, and uses that statistics to perform transactions underneath the other man or woman's

name. Even a single such transaction, together with selling a house and depositing the earnings in a fraudulent bank account, will have devastating results for the victim. By the point the proprietor unearths out it is going to be a long way too late. It's far very possibly that the proprietor may have lost thousands and thousands of greenbacks because of the identification theft. We want to explore the use of data mining each for credit score card fraud detection as well as for identification theft.[7]

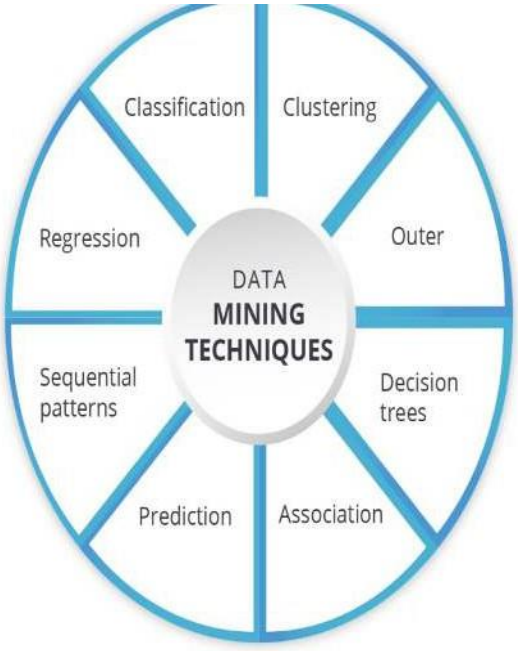
There were a few efforts on detecting credit card fraud. We need to start operating actively on detecting and preventing identification thefts.

2.7 Attacks on Critical Infrastructures

Attacks on crucial infrastructures may want to cripple a country and its financial system. Infrastructure assaults consist of attacking the telecommunication lines, the electrical, strength, gas, reservoirs and water supplies, food materials and different simple entities which might be critical for the operation of a country. Attacks on essential infrastructures ought to arise for the duration of any kind of attack whether they may be non- information associated, facts associated or bio- terrorism assaults. As an example, one could attack the software that runs the telecommunications enterprise and close down all of the telecommunication strains. In addition, software That runs the strength and fuel elements can be attacked. Attacks can also arise thru bombs and explosives. This is, the telecommunication strains may be physically attacked. Attacking transportation strains which includes highways and railway tracks also are attacks on infrastructures. Infrastructures can also be attacked via natural catastrophe such as hurricanes and earth quakes. Our main interest right here is the attacks on infrastructures via malicious attacks, each records associated and non-records related. Our aim is to have a look at facts mining and related statistics control technology to hit upon and save you such infrastructure assaults.

3. DATA MINING TECHNIQUES

The art of data mining has been continuously evolving. There are a number of progressive and intuitive techniques which have emerged that great- track data mining ideas in a bid to give organizations extra complete perception into their personal data with useful future developments. Many techniques are employed by means of the facts mining professionals



Source: Internet Based Image 4

3.1 Seeking Out Incomplete Data:
Information mining is based at the actual data aptitude, subsequently if records is incomplete, the effects could be absolutely off-mark. Subsequently, it's far imperative to have the intelligence to sniff out incomplete statistics if viable. Techniques including Self-Organizing-Maps (SOM's), help to map lacking data based totally via visualizing the version of multi-dimensional complicated information. Multi-mission learning for lacking inputs, wherein one current and legitimate facts set in conjunction with its procedures is in comparison with any other likeminded but incomplete information set is one manner to are searching for out such data. Multi-dimensional preceptors the use of shrewd algorithms to construct imputation strategies can address incomplete attributes of information. [8]

3.2 Dynamic Data Consoles:
This is a scoreboard, on a manager or supervisor's computer, fed with real-time from data as it flows in and out of various databases within the company's environment. Data mining techniques are applied to give live insight and monitoring of data to the stakeholders.

Record Analysis:
Databases keep key statistics in an established layout, so algorithms constructed the use of their personal language (which includes sq. Macros) to locate hidden patterns inside prepared facts is maximum beneficial. These algorithms are occasionally inbuilt into the statistics flows, e.g. tightly coupled with person-defined capabilities, and the findings offered in a geared up-to-refer-to file with significant analysis. a very good technique is to have the snapshot dump of information from a massive database in a cache file at any time after which analyze it further. In addition, information mining algorithms must have the ability to tug out records from more than one, heterogeneous databases and are expecting changing traits. [9]

Print Analysis:
This concept may be very beneficial to automatically find patterns inside the textual content embedded in hordes of text files, word- processed files, PDFs, and presentation files.

The textual content- processing algorithms can as an instance, discover repeated extracts of information, that's quite beneficial within the publishing enterprise or universities for tracing plagiarism.**Efficient Handling of Complex and Relational**

Data:

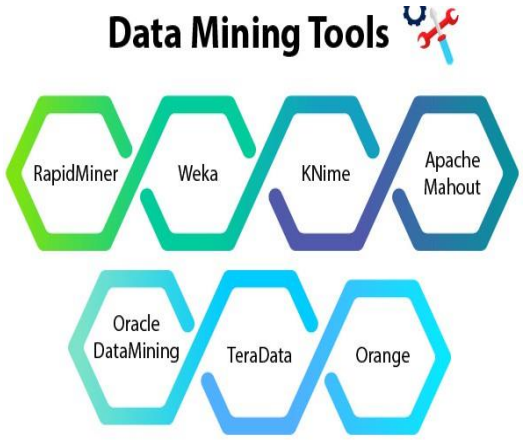
A records warehouse or big facts stores ought to be supported with interactive and question-based information mining for all sorts of records mining features inclusive of classification, clustering, association, prediction. OLAP (online Analytical Processing) is one such beneficial methodology. Other standards that facilitate interactive facts mining are reading graphs, mixture querying, photograph class, meta-rule guided mining, change randomization, and multidimensional statistical evaluation.

Data Mining Relevance and Scalability Algorithms:

While selecting or choosing facts mining algorithms, it's far imperative that corporations preserve in mind the commercial enterprise relevance of the predictions and the scalability to reduce charges in future. More than one algorithms must be capable of be finished in parallel for time performance, independently and without interfering with the transnational enterprise programs, specifically time-important ones. There need to be help to consist of SVMs on large scale. [10]

Popular Tools for Data Mining:

Famous equipment for data mining: there are many
ready-made tools available for data mining inside the market nowadays. a number of these have common functionalities packaged inside, with provisions to add-on functionality by using assisting building of business-precise analysis and intelligence.



Source: Internet Based Image 5

SOME OF THE POPULAR MULTI-PURPOSE DATA MINING TOOLS THAT ARE LEADING THE TRENDS LISTED BELOW:

Rapid Miner (erstwhile YALE):

That is very common in view that it is a prepared-made, open source, no- coding required software program, which offers superior analytic s. Written in Java, it includes multifaceted data mining capabilities which includes data preprocessing, visualization, predictive analysis, and can be easily integrated with WEKA and R-tool to directly provide models from scripts written within the former two.

WEKA:

This is a JAVA based customization tool, which is free to use. It includes visualization and predictive analysis and modeling techniques, clustering, association, regression and classification.

R-Programming Tool:

This is written in C and FORTRAN, and allows the data miners to write scripts just like a programming language/platform. Hence, it is used to make statistical and analytical software for data mining. It supports graphical analysis, both linear and non-linear modeling, classification, clustering and time-based data analysis.

Python based Orange and NTLK:

Python may be very popular due to ease of use and its powerful capabilities. Orange is an open source tool that is written in Python with useful facts analytic s, textual content analysis, and system- studying capabilities embedded in a visible programming interface. NTLK, additionally composed in Python, is a effective language processing information mining tool, which consists of records mining, device studying, and statistics scraping capabilities that may without problems be built up for custom designed desires.

Knime:

Frequently used for information preprocessing – i.e. information extraction, transformation and loading, Knime is a powerful tool with GUI that indicates the community of data nodes. Popular amongst monetary records analysts, it has modular statistics pipe lining, leveraging system getting to know, and facts mining ideas liberally for building enterprise intelligence reports. statistics mining gear and techniques are now extra important than ever for all companies, big or small, in the event that they would love to leverage their existing facts stores to make commercial enterprise decisions on the way to provide them an aggressive part. Such Actions based on data evidence and superior analytics have higher chances of increasing income and facilitating increase. Adopting nicely- set up strategies and tools and availing the help of data mining professionals shall help agencies to make use of applicable and powerful records mining principles to their fullest capability.[11]

CONCLUSION:

As per the above study the paper has discussed data mining for security purpose. We first started with a discussion of data mining for cyber security applications and then provided a brief overview of the tools we are developing. Data mining for national security as well as for cyber security is a very active research area.

Various data mining techniques including link analysis and association rule mining are being explored to detect abnormal patterns. Because of data mining, this also raises privacy concerns. One of the areas we are exploring for future research is active defense. Here we are investigating ways to monitor the adversaries. For such monitoring to be effective, the monitor must avoid detection by the static and dynamic analyses employed by standard anti-malware packages.

Consequently Over all the Data Mining is very essential for Information Security.

REFERENCE:

[1]. Data Mining for Security Applications: Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen

[2]. Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD

international conference on Management of data,

[3]. Daniel Barbara and Sushil Jajodia, editors. Applications of Data Mining in Computer Security. Kluwer Academic Publishers

[4]. Markus M. Breunig, Hans-Peter Kriegel,
Raymond T. Ng, and J Sander. Lof: identifying density-based local outliers. In Proceedings of the 2000 ACM SIG-MOD international conference on Management of data, pages

[5]. Varun Chandola and Vipin Kumar.
Summarization {compressing data into an Informative representation. In Fifth IEEE International Conference on Data Mining, pages.

[6]. Thuraisingham, B., "Web Data Mining Technologies and Their Applications in Business Intelligence and Counter- terrorism", CRC Press, FL, 2003.

[7]. Chan, P, et al, "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems.

[8]. Lazarevic, A., et al., "Data Mining for Computer Security Applications", Tutorial Proc. IEEE Data Mining Conference, 2011.

[9]. Thuraisingham, B., "Managing Threats to Web Databases and Cyber Systems, Issues, Solutions and Challenges", Kluwer, MA 2004 (Editors: V. Kumar et al).

[10]. Thuraisingham B., "Database and
Applications Security", CRC Press, 2005.

[11]. Thuraisingham B., "Data
Mining, Privacy, Civil Liberties and National Security", SIGKDD Explorations, 2012.

Current Trends of Green Cloud Computing –A Comparative Study

[1]Neeta Verma, [2]Dr.Varsha Jotwani
[1] Research Scholar, Rabindranath Tagore University,[2]Associate Professor, Rabindranath Tagore University
[1] neeta.dhani@gmail.com,[2] varsha.jotwani@gmail.com

Abstract— ***'Cloud Computing' is that technology which has seen tremendous growth in the recent years with goals to deliver green computing as a utility. Fulfilling this technological demand is not an easy task. Enormous data centers will be having vast amount of various kind of data, enormous number of servers and various other additional hardware comprises of various infrastructure required to support Green Cloud Computing. The data centers are set up on a large-scale consisting of many computing nodes and use a tremendous quantity of electrical power which having higher executional working outlays. Hence therefore increases the energy consumption, demands to rethink the energy efficiency of cloud infrastructures. Nevertheless, there is a need to review the present status prior to changing anything to the infrastructure. One of the important research point in the cloud computing researchers have been trying to resolve is to lessen the carbon emission by the data centre. In this paper in which the modern systems, are developed for the reformation of energy efficiency in case of large-scale cloud computing infrastructures while not compromising the quality of services and performance, by determining all possible factors in the cloud infrastructure supervise for the significant amount of quantity of energy consumption. A relative analysis of various survey papers published in recent times is carried out and their main contributions are also recorded. It discusses the different tools that are worked for the designing of energy efficient schemes. Based on the current trends of energy-efficient algorithms, pioneering scope in the domain is also provided for the cloud computing research community.***

.

Index Terms— Data Centre, cloud Computing, Energy Efficiency

INTRODUCTION

Due to number of reasons cloud computing is an area of interest, which is basically focused to provide dynamic and customizable infrastructure of computing infrastructure to the end users, with a top level of reliability, and quality of services[1]. Sharing of resources, on requisition and services within parameter these are some characteristic which clouds provide in highly customizable manner. Cloud computing provides a cost-effective virtual environment that can replace the high-cost computing infrastructure for various Information technology solutions. This special feature of scalability makes it flexible choice that can reduce or add on the services based on demand by the end users. The consumers who would be availing services of cloud varies depend on what kind of service they want and according to individual service they have to charge the bill. These various kinds of resources could be a consist of variety of virtual and physical resources that will be availed a anywhere with not any particular time. Cloud service providers deliver their facility in three formats. **Software as a Service (SaaS):** This makes provide the application as a service to clients. **Infrastructure as Service (IaaS):** It promotes a computing platform for the client who and made the service accessible from the cloud. It permits the client for the better storage of data in the cloud. **Platform as a Service (PaaS):** It present a platform for the developers to filter their application and helps to access the tools provided by the provider. An implementation of cloud can be divided into public or private. **The public cloud has** limitless users and shared access whereas **private cloud** is non sharable and accessible by the single user.

With the number of mobile devices growth and lots of storage is required to store the data there is a need of space in cloud beside that there are some issues are regarding the data centre, the recent technology in computing and their related problem like global warming, increase in fuel consumption and expenses of energy is to be determined. The quantity of energy is absorbed by the data centre of the service provider. An implementation can be divided as public and private. The public cloud has unlimited users and shared access whereas

`

private cloud is non sharable and access by the single users. The quantity of energy absorbed by the ICT devices need to be lessen. It has been estimated that the quantity of energy consumed by the data centres of the service providers is equal to 1.5% of the power supplied to an entire city [3]. When the cloud applications are hosted in a data centre the number of resources utilized consumes a huge percentage of electrical energy which increases the heat in the environment.

There are various procedures to lessen energy consumption in cloud systems. They are: [8]**(i) Energy Saving Hardware:** Intel and Power Now proposed the 'Speed Step' technique by AMD, lessen the power and heat emission. As the absorption of energy is directly proportional to the utilize any of the resource, and still the problem remains unresolved. So, energy saving policies-based software were developed.
(ii) Energy-Aware Scheduling: Various energy-aware procedures like Dynamic Voltage Scaling (DVS), Dynamic Voltage/Frequency Scaling (DVFS), Request Batching were developed. But all these procedures don't attain energy optimization significantly. **(iii) Alliance:** The alliance technique includes server alliance in which different kind of server consolidate their working, task alliance procedure includes alliance of various tasks, Energy Task Consolidation (ETC) is an amalgamation of various kinds of energy, Task-Based Energy Consumption and Energy Conscious Task Consolidation (ECTC).

All these procedures have their own advantages and flaws. **(iv)Energy absorption in Conglomeration of Servers:** Basically, it works by accumulating load on the system, after that getting the lesser number of servers which will be enough for the task. Some of the techniques only work with homogeneous servers and can't allot the load potentially. Wherease in few techniques Virtual Machine mapping problem is not considered fully. So, new methodology is created that main purpose is to lessen this power disperse in a group at clusters of servers by considering the system's throughput and latency. Therefore, reducing energy usage in a data center is now a hot issue in the IT industry. So, in this work, a concise review unfolding the current trends in energy efficiency and following research scope has been focused.

An existing calculation figure out is 60kWh to 160kWh to 400kWH consumption for the laptop and desktop respectively. Similarly to lower down a temperature of server 200 watts power of server is absorbed thinking 50 employees are using one server, PUE (Power usage effectiveness) as 2.0 and 8.766 operating hours per year)[5],by 2020 the quantity of carbon dioxide a data centre emits will become all most four times than the 2008.In cloud computing saving of energy is quite important because of any concern for the environment and the necessity to lessen down the green house emissions. To lessen the power requirements few data centre have been created at the high cliff point. Several implementation complications generated on developing a data centre. Hence, lessening the requirement of energy in cloud computing has always grab the attention of researchers.
In cloud computing saving of energy is very significant because of the concern for the surrounding and the requirement to lessen the greenhouse emissions. To decrease the power requirements some data centers have been created in high altitudes. On the other hand, building a data center at a high altitude also gives rise to various issues of implementation. Researchers always get attracted and focus on the issue which lessens the energy requirement.

II. Progress of Energy Efficiency In The Last Decade

Over the most recent couple of years, different planning methodology have been made to resolve various issues in distributed computing. All things considered, barely any examinations have been completed to zero in on the energy productive procedures in cloud server farms. Andreas et al.[9] examined a few methodologies for executing explicit modules and energy control focuses to diminish the equipment and programming costs alongside improved burden adjusting. Amandeep et al. [10] thought about and broke down the current energy proficiency methods like ECTC, MaxUtil, and BTC calculations. Nonetheless, the BTC calculation gave the well energy-effective arrangements. Elizabeth et al.[11] examined the errand aggregation strategy for energy advancement. Backialakshmi et al. [5] explored the energy saving strategies for distributed computing organizations and PC frameworks.
Abdul et al. [12] examined the current asset portion procedures that utilization energy effectively. Nazmul et al. [13] zeroed in on energy proficiency of various cloud frameworks alongside the current energy productivity conventions. Altaf et al. [14] talked about and looked at changed programming based energy proficiency procedures like responsibility solidification, asset the board, DVFS, and equal programming. Krishnaveni et al. [15] recognized the prerequisites of green registering, issues, and to save energy alongside the boundaries by applying different strategies.

To expand the assimilation of energy, Mahendra et al. [16] have talked about the designation of asset distribution and presumed that the usage limit of CPU is 70% to utilize the energy in an effective manner. Sara et al. [17] depicted the energy effectiveness strategies by controlling the actual machines that have the virtual ones. Rajat et al. [18] analyzed the current planning calculation for task solidification dependent on different boundaries. Mohammed et al. [19] examined Virtual Machine movement, union and asset arrangement strategies for energy proficiency. A portion of the milestone study papers and their significant contribution are accounted for in Table-I. This table is sort out the center thought of a portion of the all around introduced studies that were completed by different specialists somewhat recently thus.

Year of Publication	Author Name	Key Involvement
2010	Andreas Berl et al.	Determined few approaches for large scale networked software and hardwares that can adopt energy control centers, and specific plug-ins and has a major effect on lessening software and hardware related costs, reforms load balancing, safeguard greenhouse gas, and emission of CO_2 resulting from data centers.
2013	Amandeep Kaur et al.	Current energy saving techniques such as ECTC (Energy Conscious Task Compilation),MaxUtil(Maximum Rate Utilization) and Bi-objective task compilation(BTC) algorithms were reviewed and concluded that the BTC algorithm provided the best energy-efficient outcome, however it is the gentle among all the techniques.
2014	Elizabeth Sylvester Mkoba et al.	Variety of maximization of energy techniques with task consolidation is adopted.
2015	Backialakshmi M et al.	In depth review on the potential influence of strategies that can reserve energy for computer systems and web of networks. Cloud computing with virtualization helps in recognizing the key aspects responsible for the absorption of energy. Major point of concern is between the energy efficiency, Quality of Service, and performance.
2016	Abdul Hameed et al.	For in depth analysis the objective function is placed to meet the state-of-the-art software and hardware-based techniques that associated with the energy-efficient resource allocation, strategy for adaption of resources, procedures and execution for allotment and interoperability.
2016	Sobinder singh et al.	Various energy efficiency techniques such as hardware and scheduling procedures for energy saving and servers to support network and clusters and compilation steps are evaluated
2017	Krishnaveni. S et al.	Various kinds of precautions like hardware devices which absorb less power, virtualization, electricity enabled computer, dynamically voltage meter, green cloud architecture to save energy and lessen down the CO2 emission can be taken into consideration.
2018	Mahendra Kumar et al.	In order to attain the best energy efficiency, the problem of resource allocation is considered and recognizes the threshold for CPU
2019	Mohammed Deiab et al.	Current methods and tricks in cloud computing such as migration of virtual machine, compilation and resource management is discussed.

III. **Development of Energy Techniques in the Past years**

The challenging research issue in cloud computing is saving energy. At the data centre level energy efficiency can be divided into hardware based, infrastructure based, location based. There are five groups which are classified into software-based structure. Many researchers have been proposed a various technique for energy efficiency.

An algorithm for live migration of virtual machines was proposed by Beloglazov et al.[20] which shows considerable improvement in energy consumption according to resource requirements . However, at run time the procedure doesn't set the threshold. Ching et al. [21] presented an energy-aware task consolidation (ETC) system to reduce the consumption of energy by controlling the CPU usage upto definite limits. The proposed algorithm attains a 17% reformation in energy absorption over the MaxUtil algorithm. Hence the breakthrough will not work between the data centre and fixed to network bandwidth. Three existing energy-conscious task consolidation algorithms Young et al. [22] implemented the variants of that identified better energy consumption by

reducing the operational cost. But, the performance of the proposed algorithm is difficult to find out in a dynamic environment. An energy saving task compilation (ESTC) algorithm has been proposed by Sanjay et al. [23], which lessen absorption of energy by using the time duration for which the resource will remain idle in cloud-based environment. The outcome of 10 % optimum result shows the improvement then the existing ETC algorithm. Inactive period is completely taken if taken by arrived task. Nathan et al.[24] For the energy efficiency IAAS type clouds designed a simulation platform so it methodology get improved. For CPU usage on server consolidation cloud Sim environment is implemented.

The challenging research issue in cloud computing is saving energy. At the data centre level energy efficiency can be divided into hardware based, infrastructure based, location based. There are five groups which are classified into software-based structure. Many researchers have been proposed a various technique for energy efficiency.
In algorithm for live migration of virtual machines was proposed by Beloglazov et al.[20] which shows considerable improvement in energy consumption according to resource requirements . However, at run time the procedure doesn't set the threshold. Ching et al. [21] presented an energy-aware task consolidation (ETC) system to reduce the consumption of energy by controlling the CPU usage upto definite limits. The proposed algorithm attains a 17% reformation in energy absorption over the MaxUtil algorithm. Hence the breakthrough will not work between the data centre and fixed to network bandwidth. Three existing energy-conscious task consolidation algorithms Young et al. [22] implemented the variants of that identified better energy consumption by reducing the operational cost. But, the performance of the proposed algorithm is difficult to find out in a dynamic environment. An energy saving task compilation (ESTC) algorithm has been proposed by Sanjay et al. [23], which lessen absorption of energy by using the time duration for which the resource will remain idle in cloud-based environment. The outcome of 10 % optimum result shows the improvement then the existing ETC algorithm. Inactive period is completely taken if taken by arrived task. Nathan et al.[24] For the energy efficiency IAAS type clouds designed a simulation platform so it methodology get improved. For CPU usage on server consolidation cloud Sim environment is implemented.
In table II there are few papers which show the efficiency and their involvement with respect to its advantage and disadvantages.

Year of Publication	Author Name	Proposal	Advantage	Disadvantage
2010	Anton Beloglazov et al.	The author suggests heuristics for live migration and dynamic reallocation of virtual machines as per existing resource needs.	The methodology of reallocating the Virtual Machines dynamically and turning off the servers that are idle save lot of energy that can also be implement to cloud data centre.	Dynamically Utilization Threshold is not fixed .
2011	Ching-Hsien Hsu et al.	An energy-aware task consolidation(ETC) scheme is depicted to lessen down the utilization of energy by adding constraints utilization of CPU to certain extent and task compilation amid the virtual clusters.	The presented ETC algorithm more importantly will lessen down the power consumption (with 17% reformation) than the current MaxUtil algorithm.	Without the proper usage of idle period (i.e. 1 to 20%), the algorithm has snatched the power absorb by the resources into account even if they are in not active conditions .Between the data centers solution won't work.It is fixed to constant network bandwidth.
2012	Young Choon Lee et al.	The variety of three current Energy conscious task compilation algorithms are effectively adopted that contemplate the vital as well as inactive energy absorption.	Two energy-conscious task compilation heuristics are developed which depicts better energy absorption than the current heuristics by effectively using the resources. The proposed heuristic also reduces operational costs.	Under a changing surrounding, the supremacy of the working of the proposed heuristics is difficult to find out.
2014	Sanjay Panda et al.	An energy-saving task consolidation (ESTC) algorithm is proposed that lessen the absorption of energy by the utilization of the time period for which a resource remain inactive in a cloud.	The proposed ESTC algorithm gives a 10 % better result than the existing ETC algorithm in terms of energy consumption and total no of task completion.	The tasks won't be set up according to the proposed strategy if the inactive time-period is fully alloted by the arrived tasks. The task characteristics may not be content due to certain constraints of resources.
2015	Nathan Whittington et al.	To refine the energy efficiency for IAAS type clouds by creating a simulation platform that can recognize the policy which lower down the energy and maintains the service Level Agreement (SLA) at its best.	With the help of virtual machines promoting to improvisations were made on a basic cloud scheduler and real-time consolidation is possible through the usage of virtual machine	The simulation result is fixed to CloudSim software. Only CPU usage has been point it on server consolidation.
2016	Madhu B.R. et al.	Allocate efficiently the virtual machine to the available servers and execute the user requests or tasks to the appropriate virtual machine that can prevent the energy inefficiency of the data center.	The proposed task consolidation mechanism outperforms the current Round Robin scheduling. The modified Local Linear Regression (LRR) mechanism also show the refinement than the current Linear Regression mechanism in form of energy efficiency	The breakthrough is limited to the CloudSim environment.

2016	Thusoyaone Joseph Moemi et al.	Balancing of load on Virtual Machine Aware model and Efficient Energy Usage (EEU) metric has been introduced	Lessen down the energy absorption and refine Quality of Service.	Data Center with lesser machines doesn't get optimum response time and greater energy is absorbed for memory configurations with greater frequencies.
2018	Parthasarathi et al.	Task compilation using minimization of IDLE VM algorithm	The proposed algorithm was shown to get optimum CPU utilization and lower down the energy absorption than the existing ECTC (Energy Conscious Task Consolidation) and MaxUtil (Maximum Rate Utilization) algorithm.	It won't work in a alliance cloud environment

IV. Energy Coherent Algorithm Tools

T. Guérout et.al. [28] considered lots of simulators which provide operations to simulate and execute the energy aspect in variant computing situations. SIMGRID is the one of the authors first discussed. It assists simulation environment which executes on distributed kind of atmosphere. In that situation the system comprises of various configurations.

GroudSim is another simulation tool. And it supports various functionalities for simulating events in Cloud as well as Grid computing. It has a feature to consider the probability of failure of hardware in those scenarios.

Other simulation tool GSSIM is used to review the policies involved in scheduling. Standard formats are adopted to figure out about the workflows while simulating the polices of schedule. Various kinds of models for reviewing energy absorption are also exist in GSSIM. By extending NS2, Green Cloud simulator is framed. It describes various sources of energy absorption. Calculation at CPU is one of them, communication and the cooling systems applied in data centre are also sources of energy absorption. Using this simulator an in depth study about consumption of energy can be attained.
One of the renowned Simulator is CloudSim. It assists to model IaaS (Infrastructure as a Service) layer. In CloudSim one can add data centres in CloudSim. One can create different models to describe energy consumption in cloud scenarios using CloudSim. Migrating virtual machines, terminating host machines etc. can be simulated while creating energy models. iCanCloud can be applied to simulate and check the cost versus performance measure for various applications. It provides working to simulate new frameworks along with the current models.

Parameter	CloudSim	Greencloud	iCanCloud
Framework	-	NS2	OMNET,MPI
Programming Language	Java	C++/OTCL	C++
Availability	Open Source	Open Source	Open Source
Communication Model	Fixed	Full	Full
Support to model consumption of power	Fixed	Yes	Fixed

V.Future Scope

As future progress, will be proposing a dynamically configured resource management step by the right allotment of Virtual Machines in the cloud data center that takes into account most of the major energy parameters and most feasible PM allocation constraints.

VI. CONCLUSION

For cloud computing today, the field of resource handling and energy absorption is an pivot and inspirational one. In reality, data centers absorb a huge amount of electrical energy that causes a decrease in efficiency and a significant amount of carbon dioxide to be released. Many procedures, such as virtualization of server, duplicacy, and compilation, are implemented to increase resource usage and lessen down the usage of energy. This paper is an effort to provide a comparative analysis of the existing state-of the-art energy-efficient techniques that are exclusively devised for cloud computing applications. It also focused some of the most appropriate review works and discussed the tools and platforms compatible with such kind of applications.

REFERENCES

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology,vol. 53, no. 6, 2009.

[2] J.Srinivas,K.VenkataSubba Reddy, Dr.A.MoizQyser, "Cloud Computing Basics," International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2012.

[3] A. Beloglazov and R. Buyya," Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience, vol. 24, no. 13,pp. 1397-1420, 2012

[4] A. Beloglazov, R. Buyya, Y.C.Lee, and A.Zomaya,"A taxonomy and survey of energy efficient data centers and cloud computing systems," Advances in Computers,Elsevier, 2011, vol. 82, pp. 47-111.

[5] Backialakshmi M and Hemavathi N," Survey on Energy Efficiency in Cloud Computing," Journal of Information Technology & Software Engineering, vol. 06, issue 1, 2015 .

[6] Kaplan, J.M., Forrrest W. and Kindler, N. 2008. Revolutionizing Data Center Energy Efficiency. McKinsey & Company.

[7] Beloglazov, A., Abawajy, J. and Buyya, R. 2012."Energy- Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing," Future Generation Computer Systems, Elsevier. 28, 755-768.

[8] Sobinder Singh, Ajay Kumar, AbhishekSwaroop, Anamika, "A Survey on techniques to achieve energy efficiency in cloud computing," International Conference on Computing, Communication and Automation (ICCCA 2016).

[9] Andreas Berl, ErolGelenbe, Marco Di Girolamo,Giovanni Giuliani, Hermann De Meer, Minh Quan Dang,Kostas Pentikousis,"Energy Efficient Cloud Computing,"The computer Journal, Vol.53,No. 7, 2010.

[10] Amandeep Kaur, Rupinder Kaur, Prince Jain,"Algorithms for Task Consolidation Problem in a Cloud Computing Environment,"International Journal of Computer Applications, Vol. 75, No. 4, August 2013.

[11] Elizabeth Sylvester Mkoba, Mokhtar Abdullah AbdoSaif, " A Survey on Energy Efficient With Task Consolidation in the Virtualized Cloud Computing Environment ," International Journal of Research in Engineering and Technology, Vol. 3, Issue. 3, March, 2014.

[12] Abdul Hameed,Alireza Khoshkbarforoushhsa, Rajiv Ranjan et al., "A Survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems," Springer Computing Journal, 2016.

[13] Nazmul Hossain, Md. AlamHossain, A.K.M. Fayezul Islam, PriyankaBanarjee, TahiraYasmin, "Research on Energy Efficiency in Cloud Computing," International Journal of Scientific & Engineering Research, Vol. 7, Issue. 8, August 2016.

[14] Altaf Ur Rahman, FiazGul Khan, WaqasJadoon, " Energy Efficiency Techniques in Cloud Computing," International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 6, June 2016.

[15] Krishnaveni. S, Baddam Indira, "Issues on Green Cloud Computing Towards Energy Saving," International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT) Vol. 3, Special Issue- 1, March, 2017.

[16] Mahendra Kumar Gourisaria, S.S. Patra, P.M. Khilar, "Energy Saving Task Consolidation Technique in Cloud Centers With Resource Utilization Threshold," Advances in Intelligent Systems and Computing 563, Springer Nature Singapore Pte Ltd. 2018.

[17] Sara Diouani, HichamMedromi, " Survey: An Optimized Energy Consumption of Resources in Cloud Data Centers," International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No.2, February,2018.

[18] RajatPugaliya, Madhu B. R, "Algorithm for Task Consolidation in Cloud Computing: A Comparative Survey," International Journal of Research Granthaalayah, Vol. 6, Issue. 5, May, 2018.

[19] Mohamed Deiab, Deena EI-Menshawy, Salma EI-Abd,Ahmad Mostafa, M. Samir Abou EI-Seoud, "Energy Efficiency in Cloud Computing," International Journal of Machine Learning and Computing, Vol. 9, No. 1, February, 2019.

[20] Anton Beloglazov, RajkumarBuyya, " Energy Efficient Allocation of Virtual Machines in Cloud Data Centers," 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[21] Ching- Hsien Hsu, Shih-Chang Chen, Chih-Chun Lee et al., "Energy –Aware Task Consolidation Technique for Cloud Computing," Third IEEE International Conference on Cloud Computing Technology and Science, 2011.

[22] Young Choon Lee, Albert Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," Journal of Supercomputing, 2012.

[23] Sanjay K. Panda, Prasanta K. Jana, "An Efficient Energy Saving Task Consolidation Algorithm for Cloud Computing Systems," IEEE International Conference on Parallel, Distributed and Grid Computing, 2014.

[24]Nathan Whittington, Lu Liu, Bo Yuan, Marcello Trovati, "Investigation of Energy Efficiency on Cloud Computing," 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing.

[25] Madhu B.R., A.S. Manjunatha, Prakash Chandra, Chidananda Murthy P, "Minimizing Energy Consumption in Cloud Datacenters using Task Consolidation," International Journal of Engineering and Technology (IJET), Vol. 8, No. 5, Nov, 2016.

[26] Thusoyaone Joseph Moemi, Obeten Obi Ekabua, "Energy Efficiency Models Implemented in a Cloud Computing Environment," International Conference on Computing, Communication and Automation (ICCCA), 2016.

[27] Parthasarathi Pattnayak, Pamela Pal,"Energy-Efficient Cloud Computing With Task Consolidation," International Journal of Latest Trends in Engineering and Technology,2018.

[28]T.Guérout.et.al.,Energy-aware simulation with DVFS,Simulat.Modell.Pract.Theory(2013), http://dx.doi.org/10.1016/j.simpat.2013.04.007

AUTHORS PROFILE

Neeta Verma have completed M.Sc.in Computer Science..she have 16 years experience of teaching in computer science. . I is pursuing as Research Scholar with Rabindranath Tagore University.

Dr.Varsha Jotwani is currently working as Associate Professor with Rabindranath Tagore University. She is Ph.D. in computer Application. She has vast teaching and academic developments at leading institution of Bhopal, India. She has published various international and national research papers in the highly quality journals. She is also well versed in developing curriculum for undergraduate and postgraduate students under the field of Information technology

A Study on Challenges Associated With Antenna Design and Future Antenna Models

Nikunj Goyal

Galgotias College of Engineering and Technology, Greater Noida

Abstract

The antenna is implemented to transfer signals when no other way is possible for appropriate communication over remote locations, rugged regions, and so on. Design involves maintaining, regulating the electric current to generate the coveted radiation exemplar called a pattern. An antenna design was initiated with an ideology for the configuration and fundamental determinants based on the Maxwell equation for the sizing of antenna components. The motive of the research is to comprehensively analyze the advanced challenge associated with the antenna design and the future of antenna design, for exploring and enhancing the understanding and fulfilling the objective paper, implementing secondary methodology and obtaining appropriate conclusions. The paper exclusively addresses loopholes that must be overcome to identify the functionality of this bandwidth and provide a peculiar perspective for new generation networks, novel multiple access techniques, antenna array, and real-time signal processing.

Keyword- Challenges in antenna design, future antenna design,

Introduction

Communication is one of the indispensable tools to establish an association, whether it is human or equipment. To effectively communicate between two discrete points is invariably a trial from conventional to the most modern wireless communication technology based on electromagnetic signals. Innovative technology executed voice-over data, digitalization, and so on utilizes wireless hand-held automated equipment for communication. In this respect, the antenna society usually performs a crucial role emphasizing the conception of low-profile miniature and multiband antennas synchronically with repetitious antenna operations proficient in fulfilling the stringent requirement of emanating multifunctionality wireless tools. The antenna is implemented to transfer signals

when no other way is possible for appropriate communication over remote locations, rugged regions, and so on.

1.1 Antenna Design

An antenna turns confined circuit domains into disseminating electromagnetic waves and, by retaliating, gathers power from progressing electromagnetic waves. It is prominent to recognize that the antenna transmits from an electric flow. Design involves maintaining, regulating the electric current to generate the coveted radiation exemplar called a pattern. An antenna design was initiated with an ideology for the configuration and fundamental determinants based on the Maxwell equation for the sizing of antenna components. Antenna radiated stellar waves that disseminated in the spiral path for a correspondent operation focused on the antenna. It is an antenna that transforms electrons into photons and vice versa. The allowable beamwidth antenna means at half of the most power obtained by an antenna. Energy transmitted from an antenna per unit solid angle is called radiation intensity.

1.2 Type of Antenna

There are various type of antenna and can be permanently classified into modern and contemporary antennas-

Wire antenna

Reflector antenna

Lens antenna

Array antenna

Aperture antenna

Patch antenna

1.3 Modern Antenna Design-

a. Microstrip Antenna

Microstrip patch antennas comprising transmitting patches on one front of a dielectric substrate hold a terrain plane on another show. Identify the expansion and perimeter of a rectangular patch antenna. Microstrip antenna planar reverberating basin that departs from their edges and radiates. The antenna is composed of submissive substrates to endure suggestions of collapse and oscillation surrounding.

Envisioned for wireless portable communication base location of a fabricated antenna. This variety of antenna also eradicates the hurdle of transmission from covering waves induced in a sheer dielectric substrate domain utilized to enhance the bandwidth. Microstrip patch antenna incorporating metal patches extended concerning conventional broadcast line width. A patch transmitted from a fringing range encompassing its edge.

b. Horn Antenna Design

Horn has a widespread dimension of utility from a miniature hole antenna to serve reverberator to a comprehensive specialty antenna employed by themselves as an ordinary accumulation antenna. Horn can be stimulated in any polarization or sequence of polarization. Horn also strictly accompanies the quality prophesied by uncomplicated hypotheses and can be investigated as a sectoral horn by determining the terminal equivalent query in the drive.

c. Lens Antenna

Lenses are outfitted with a parabolic reverberator which achieves loose range as a meadow arrangement to attract a considerable aperture. Lenses produce only half the endurance condition of a reverberator because the wave crosses by the irregularity only formally. At below microwave frequency, the lens is prohibitively troublesome, but administration and the service of synthetic dielectric subdue this issue. A single lens with an identical dielectric has two coverings and is comparable to a binary reverberator because its coverage measures autonomy. Lenses have no intrinsic frequency bandwidth constraint. The implementations are restricted by the enclosure and construction obstacles to a considerable extent.

1.4 Technology Trends and Challenges of Antennas for Satellite Communication Systems.

The trend of the technology evolution of antennas for an extensive spectrum of marketing satellite broadcasting assistance

and the hurdles encountered in their improvement are chronologically renewed. The geometric optics procedure is one of the most precise means to evaluate reverberator practices enlarged by the geometric approach to address edge diffraction. The physical principles of diffraction performed the physical optics classification. The demand for commercialization and revenue extensively inspires the trending objection of antenna advancement for satellite communication. The principal drift in both spaces' in-ground antenna pattern is to operate with multifunctional antennas. The potentiality of autonomous gathering in the span of a very comprehensive satellite antenna after possibly more than an individual launching vehicle has performed the entire engagement of equipment components may implement such an abundant cost-effective approach in the reach of private stakeholders. Innovation reinforcements are suspected of accommodating the synergy of satellites with wireless co-operations.

1.5 Advances in Antenna Technology for Wireless Handheld Devices

The steady progression of wireless hand-held gadgets and the phantom of versatile wireless transmission practices stimulate the antenna association to compose innovative radiating and computation practice ability providing the business interest. The antenna blended with smart wireless devices functions in extraordinary surroundings. These specific circumstances require the antenna society to distinguish the unified antenna in smart broadcast equipment to achieve an adequate antenna regularity. The noticeable aspect of antenna expedited explaining its performance modern extensive from execution server to compose robust mechanism with the productive utilization of electromagnetic review to examine how antenna radiation harms the human body. The multipath atmosphere incites a new estimation practice, such as a repercussion antechamber, which can imitate a stable dispersion situation.

Literature Survey

The study focuses on the trend of the technology evolution of antennas for an extensive spectrum of marketing satellite broadcasting assistance, and the hurdles encountered in their improvement are chronologically renewed (Rowell, C., & Lam, E. Y., 2012). The geometric optics procedure is one of the most precise means to evaluate reverberator practices enlarged by the geometric approach to address edge diffraction. The physical principles of diffraction performed the physical optics classification. The demand for commercialization and revenue extensively inspires the trending objection of antenna advancement for satellite communication (Rahmat-Samii, Y., & Densmore, A. C., 2015). The principal drift in both spaces' in-ground antenna pattern is to operate with multifunctional antennas. The potentiality of autonomous gathering in the span of a very comprehensive satellite antenna after possibly more than an individual launching vehicle has performed the entire engagement of equipment components may implement such an abundant cost-effective approach in the reach of private stakeholders (Saad, W., 2021). Innovation reinforcements are suspected of accommodating the synergy of satellites with wireless co-operations (Stutzman, W. L., & Thiele, G. A., 2012). Another research focuses on modern technology in wireless smart communication devices. The steady progression of wireless hand-held gadgets and the phantom of versatile wireless transmission practices stimulate the antenna association to compose innovative radiating and computation practice ability providing the business interest. The antenna blended with smart wireless devices functions in extraordinary surroundings (Hornby, G.,S et., al., 2006). These specific circumstances require the antenna society to distinguish the unified antenna in smart broadcast equipment to achieve an adequate antenna regularity. The noticeable aspect of antenna expedited explaining its modern extensive performance from execution server to compose robust mechanism with the

productive utilization of electromagnetic review to examine how antenna radiation harms the human body (Anguera, J., et al., 2013). The multipath atmosphere incites a new estimation practice, such as a repercussion antechamber, which can imitate a stable dispersion situation. The research (Hong, W., 2017) emphasizes the miniature antenna design with high bandwidth, gain and numerous antennas at transmission and retrieving to improve the channel capacity. In the rapidly advancing technology, many improvisational approaches enhance the antenna's performance; one of the novel and impressive methodologies was metamaterials (MTM) utilized in antenna design. It is equipped with artificial electromagnetic properties, which stimulate the designing advancement encompassing high-performance antennas and microwave equipment, making them novel and unique compared to contemporary antennas (Kumar, P.,et al., 2021).

Research Objective

The motive of the research is to comprehensively analyze the advanced challenge associated with antenna design and the future of antenna design.

Research Question

What are the challenges associated with antenna design?

What will be the future model of antenna design?

Methodology

The study opts for a secondary approach that involves already existing resources to explore, understand and enhance effectiveness in the research. It utilizes online and offline resources like research papers, literature, public reports, international organization reports and other internet website data relevant to the research. In this research, more than 50 research papers were initially selected to understand the objective of the research thoroughly. Subsequently, for research writing, the study finally selected 19

research papers and relevant internet resources for obtaining appropriate finding and conclusion.

Q. What will be the future model of antenna design?

With the emerging technology in communication, system volume challenges are particularly facing in the antenna field. The considerable function of design and advancement of antenna must be well-advised with concern and exclusively because it plays a vital ingredient in wireless communication. With the motive to utilize in hands and peculiar antenna designing resolving the challenges of language and another frequency issue with handheld devices being the most concern research (Kumar, P. et al., 2021). One of the most innovative designing techniques of material implemented to resolve the challenges facing the antenna community is MTM (metamaterial). The prominent advantage of this material for appropriately used in the implementation of antenna designing is its outstanding miniature quality which enhances the requisite for automated equipment and unwire communication and devices appropriately associated with the internet of things urge for miniature size (Alibakhshikenari, M., et al., 2021). Presently various MTM enables methodology to minimize the size of distance type of integrated antenna patch antenna, loops and slot antenna, and so on. It has a significant implementation over MIMO (massive multiple input and multiple outputs) antenna systems, enhancing its parameter and its performance with MTM materials.

In this emerging technology world where the fifth generation acts as a revolutionary innovation that provides outstanding functionalities like high-frequency network function virtualization with revolutionary technology MIMO, which has a capacity. Although with this fifth-generation cellular network approach, significant challenges came in front of a world to deal with, like holographic communication message deployment of innovative material edge

devices, the bulk of fiber transportation has grown considerably, and the prevailing portable interface is inefficient to meet this enormous load (Pedram, K. et al., 2019). Thus it is envisioned that the 6th generation cellular network will stimulate and play a significant role in avoiding these hurdles by rendering further transmission amenities, networkability, and ultra-low latency experiences (Abed, A. T., et al.,2021). With the enclosure of advanced technology, certain challenges encompassed as this technology utilization of Terahertz communication innovative antenna design is required, emphasizing the miniature size of the antenna and current flow with the high-frequency transmission. Another issue was the large intelligence office used for the 6G network for effective communication applying artificial electromagnetic metasurface as a large antenna with the motive to enhance the network's capability (Saad, W., 2021). Another challenge was holographic MIMO surface as in 5th generation large antenna array is employed with enormous MIMO architecture which was incapable of handling immense traffic load.

Improved technology in communication systems requires enhancing marine communication methodology, which has a crucial role in the safety of the nation's geographical boundary and the fishermen community, which vogue in the deep sea for their livelihood. The prospective smart antenna can be implemented either as a terminal service antenna at the nearshore zone or as a consumer premises equipment (CPE) antenna and different transportable device. Various types of antenna components and distance among the elements are the factors that compose the antenna pole as best linear (Jayakrishnan, V. M., & Vijayan, D. M., 2020). The complete interpretation of adjusted modulation achievement is decided based on intelligent antenna technique, and the consequences show that the approach has significant characteristics like it is more specific in the duration of half-wavelength displacement among the space antenna ingredient and unusual signal to noise ratio, and reveals the

potential to be implemented for the base station antenna.

Q. What are the challenges associated with antenna design?

Wireless movable communication devices were based on the technology. Various methodologies were employed in the present application, such as WLAN, WiMAX, LTE, ISM, and 5G, which are implemented on more than three different bands to illustrate the impact multipath issues are utilized in urban regions (Abed, A. T., et al., 2021). The paper comprehensively exhibits two types of microstrip antenna with prominent ingredients, such as operating bandwidth gain efficiency and size. The loophole in the antenna work design was fully identified and manifested various significant challenges in designing this antenna (Saad, W., 2021).

- One of the significant challenges identified is the technique utilized by Wi-Fi, ISM, LTE, WiMAX, and 5G triples-based bands, and it is critical to design an antenna suitable for all these bands.
- Another issue with antenna designing radiation properties is high-efficiency, gain, bandwidth that did not specifically interest Wi-Fi application.
- Moreover, the usual concern regarding antennas was their size, weight, and cost. Because in modern communication devices, portable antennas have characteristics of compact size, lightweight, low profile, and cost-effectiveness being the most crucial requirement.
- Most antennas are geometrically polarized to diminish their extent or challenges of CP production and accordingly have severe obstacles such as unsuitable polarization and beamforming intervention.

Result

The steady progression of wireless handheld gadgets and the phantom of versatile wireless transmission practices stimulate the antenna association to compose creative radiating and computation practice ability

providing the business interest. The antenna blended with smart wireless devices functions in extraordinary surroundings. These specific circumstances require the antenna society to distinguish the unified antenna in innovative broadcast equipment to achieve an adequate antenna regularity. The noticeable aspect of antenna expedited explaining its performance modern extensive from execution server to compose robust mechanism.

Although, with this fifth-generation cellular network approach, significant challenges came in front of a world to deal with, like holographic communication message mobilization of innovative equipment, the bulk of network congestion has risen notably, and the contemporary portable network is unable to meet this enormous load. Thus it is envisioned that the 6th generation cellular network will stimulate and play a significant role in avoiding these hurdles by rendering further transmission amenities, networkability, and ultra-low latency experiences.

With the enclosure of advanced technology, particular challenges encompassed as this technology utilization of Terahertz communication innovative antenna design is required, emphasizing the miniature size of the antenna and current flow with the high-frequency transmission Moreover, the usual concern regarding antennas was their size, weight, and cost. Because in modern communication devices, portable antennas have characteristics of compact size, lightweight, low profile, and cost-effectiveness being the most crucial requirement. Most antennas are geometrically polarized to diminish their extent or challenges of CP production and accordingly have severe obstacles such as unsuitable polarization and beamforming intervention.

Future antenna or desired antenna must possess specific characteristics, such as it must have a multi-operating band that covers all required spectrum. The radiating property necessity is constant in the running

bands, begetting tremendous performance and an adequate assessment of gain. For an appropriate portable interface, the transmission design of the antenna must be multidirectional. The polarization of the antenna in mobile equipment must be circular because of its immense utility.

Conclusion

With the emerging technology in communication, system volume challenges are particularly facing in the antenna field. The substantial role of the design and development of antenna must be considered with concern and exclusively because it plays a vital ingredient in wireless communication. Innovative technology executed voice-over data, digitalization, and so on utilizes wireless hand-held automated equipment for communication. In this respect, the antenna society usually performs a crucial role emphasizing the conception of low-profile miniature and multiband antennas synchronically with repetitious antenna operations proficient in fulfilling the stringent requirement of emanating multifunctionality wireless tools.

With the motive to utilize in hands and peculiar antenna designing resolving the challenges of language and another frequency issue with handheld devices being the most concern research. Moreover, the usual concern regarding antennas was their size, weight, and cost. Because in modern communication devices, portable antennas have characteristics of compact size, lightweight, low profile, and cost-effectiveness being the most crucial requirement of antennas are geometrically polarized to diminish their extent or challenges of CP production and accordingly have severe obstacles such as unsuitable polarization and beamforming intervention.

The paper exclusively addresses loopholes that must be overcome to identify the functionality of this bandwidth and provide a peculiar perspective for new generation networks, novel multiple access techniques, antenna array, and real-time signal processing. The research emphasizes the

miniature antenna design with high bandwidth, gain, and numerous antennas at transmission and retrieving to improve the channel capacity.

References

1. Rowell, C., & Lam, E. Y. (2012). Mobile-Phone Antenna Design. IEEE Antennas and Propagation Magazine, 54(4), 14–34. doi:10.1109/map.2012.6309152
2. Rahmat-Samii, Y., & Densmore, A. C. (2015). Technology Trends and Challenges of Antennas for Satellite Communication Systems. IEEE Transactions on Antennas and Propagation, 63(4), 1191–1204. doi:10.1109/tap.2014.2366784
3. Stutzman, W. L., & Thiele, G. A. (2012). *Antenna theory and design*. John Wiley & Sons.
4. Hornby, G., Globus, A., Linden, D., & Lohn, J. (2006). Automated antenna design with evolutionary algorithms. In *Space 2006* (p. 7242).
5. Anguera, J., Andújar, A., Huynh, M.-C., Orlenius, C., Picher, C., & Puente, C. (2013). Advances in Antenna Technology for Wireless Handheld Devices. International Journal of Antennas and Propagation, 2013, 1–25.
6. Muirhead, D., Imran, M. A., & Arshad, K. (2016). A survey of the challenges, opportunities and use of multiple antennas in current and future 5G small cell base stations. *IEEE access*, *4*, 2952-2964.
7. Milligan, T. A. (2005). *Modern antenna design*. John Wiley & Sons.
8. Shaker, G., Safavi-Naeini, S., & Sangary, N. (2015). Modern antenna design using mode analysis techniques. *Progress In Electromagnetics Research*, *62*, 153-165.
9. Chandrasekharan, S., Gomez, K., Al-Hourani, A., Kandeepan, S., Rasheed, T., Goratti, L., … Allsopp, S. (2016). Designing and implementing future aerial

communication networks. IEEE Communications Magazine, 54(5), 26–34. doi:10.1109/mcom.2016.7470932

10. Tsunekawa, K. (2005). Recent antenna system technologies for next-generation wireless communications. *NTT Technical Review*, *3*(9).
11. Hong, W. (2017). *Solving the 5G Mobile Antenna Puzzle: Assessing Future Directions for the 5G Mobile Antenna Paradigm Shift. IEEE Microwave Magazine, 18(7), 86–102.* doi:10.1109/mmm.2017.2740538
12. Pedram, K., Karamirad, M., & Pouyanfar, N. (2019). *Evolution and Move toward Fifth-Generation Antenna. The Fifth Generation (5G) of Wireless Communication.* doi:10.5772/intechopen.74554
13. Zhang, S., Liu, Y., Gao, F., Xing, C., An, J., & Dobre, O. A. (2021). Deep learning based channel extrapolation for large-scale antenna systems: Opportunities, challenges and solutions. *IEEE Wireless Communications*.
14. Shahraki, A., Abbasi, M., Piran, M., Chen, M., & Cui, S. (2021). A comprehensive survey on 6g networks: Applications, core services, enabling technologies, and future challenges. *arXiv preprint arXiv:2101.12475*.
15. Saad, W. (2021). 6G Wireless Systems: Challenges and Opportunities. *5G and Beyond: Fundamentals and Standards*, 201.
16. Kumar, P., Ali, T., & Pai, M. M. (2021). Electromagnetic Metamaterials: A New Paradigm of Antenna Design. *IEEE Access*, *9*, 18722-18751.
17. Alibakhshikenari, M., Virdee, B. S., Althuwayb, A. A., Aïssa, S., See, C. H., Abd-Alhameed, R. A., ... & Limiti, E. (2021). Study on on-chip antenna design based on metamaterial-inspired and substrate-integrated waveguide properties for

millimetre-wave and THz integrated-circuit applications. *Journal of Infrared, Millimeter, and Terahertz Waves*, *42*(1), 17-28.

18. Abed, A. T., Singh, M. S. J., Thiruchelvam, V., Duraikannan, S., Tawfeeq, O. A., Tawfeeq, B. A., & Islam, M. T. (2021). Challenges and limits of fractal and slot antennas for WLAN, LTE, ISM, and 5G communication: a review paper. *Annals of Telecommunications*, 1-11.
19. Jayakrishnan, V. M., & Vijayan, D. M. (2020, March). Performance Analysis of Smart Antenna for Marine Communication. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 88-91). IEEE.

Iot base Transformer Monitoring System

[1]Prof. Mohammad Hassan, [2]Nishant Gadhawe, [3]Bhushan Kohade , [4]Sameer Dongre, [5]Rahul Urkude, [6]RahulTayde, [7]Parish Swami

[1]Assistant Professor, Department of Electronics and Telecommunication, J D College of Engineering & Management, Nagpur, India

[2,3,4,5,6,7] Student, Department of Electronics and Telecommunication, J D College of Engineering & Management, Nagpur, India

Abstract – At present the observation during night clad to be exceptionally testing task. There are some spots where people cannot be engaged with watching [1]. A fundamental prerequisite of this circumstance could be a robot which consequently identifies trespassers within the territory like workplaces, home, building so forth and report handy board security control unit. In the current work, A late evening guarding robot is formed with upgraded capacity to recognize and alarm if there's any human movement within the territory to present exact observing framework [2]. The Night Patrolling Robotic vehicle moves during a random path while watching. The framework utilizes IR based way following framework for watching allocated zone. The development of a robot is additionally controlled consequently through deterrent recognizing sensors to remain far from the crash. It screens every zone to acknowledge any

Interruption utilizing camera which is mounted on the highest of the robot to catch the images, record and sends them to the client. It can likewise impart the continued video signs to the client [3]. The principle goal of this undertaking is to acknowledge the dubious exercises within the regions where human presence cannot be seen.

Key Words: Arduino, Surveillance, ESP32, IoT, Robot, Security, Microcontroller, Embedded frameworks.

1. INTRODUCTION

Technology has presented to us a dynamic and colossal change in apply autonomy and Robotics field which runs in a wide range of regions. Reconnaissance is the procedure of close precise perception or oversight kept up over an individual, gathering, and so forth particularly one in authority or under doubt. In this way reconnaissance is primarily required in the zones, for example, outskirt zones, open spots, workplaces and in enterprises. It is for the most part utilized for checking exercises. The demonstration of observation can be performed both indoor just as in open air territories by people or with the assistance of installed frameworks, for example, robots and other mechanization gadgets. A robot is only a programmed electronic machine that is equipped for performing customized exercises in this way supplanting human work, giving profoundly precise outcomes and beating the constraints of people. In this way supplanting people in the reconnaissance fields is one of the extraordinary progressions in mechanical autonomy.

Patrolling is nothing but to keep monitoring over an area by regularly moving or travelling a route of the corresponding area. The robot captures the images with the help of camera. These images are then sent to the user in a real time, user will analyses it and if there is any problem observed then alarm is triggered manually. Robot patrolling

is mostly used in Military area, Hospitals, Shopping mall, Restricted Zones, Industrial area, Agricultural area etc. The robot uses ESP32 based camera sensor which cuts down the price of using a raspberry pi. This also reduces the instructions and enables programming the robot with a least programmable skills.

2. Methodology

A] Vehicle Assembly:

- The vehicle consists of 4 dc motor connected to a motor driver to perform linear motion.
- The vehicle is powered by 12volts Lithium ion battery.
- The vehicle is operated by user through Blink server.
- The vehicle can also be controlled by wired transmitter.

B] System Assembly and Working:

- Our system consists of following sensors such as, Temperature sensor, Gas sensor, Camera module, Sound sensor and ultrasonic sensor, etc.
- Above mentioned sensors are interfaced with ESP 8266(Node MCU) module.
- The above mentioned sensors coverts stimuli such as heat, light, sound and motion into electrical signals.
- These signals are passed through ESP 8266 module that converts them into a binary code and passes to Blynk server to be processed.
- The temperature sensor is connected to motor driver, which gives an alert (notification) above critical temperature .
- If the heat is detected by a gas sensor then the notification will be displayed.
- Camera module will survey the surroundings environment and capture images and videos which later been processed by a server.
- Sound sensor is used to capture the audio from nearby.
- Ultrasonic sensor is used to detect the object detection.
- The data which has been fetched from the sensors will later has been processed simultaneously and will give corresponding output.
- The real-time data is being graphically visualized on Blynk server.
- All the code and algorithm of the system is executed on Arduino IDE.

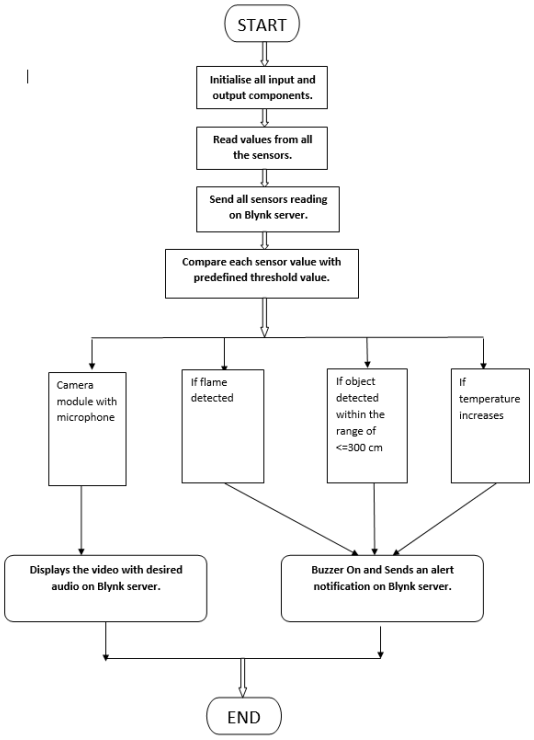


Fig.1 Flow Chart

3. SYSTEM ARCHITECTURE

In proposed paper we are collecting all parameters with the help of sensors. The system comprises all components in transformer monitoring system as shown in Fig. 2.

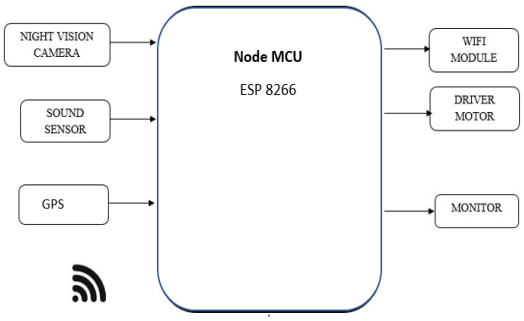


Fig.2 Block Diagram Project

3.1 Hardware

- ESP8266 Node MCU
- GPS Module
- Camera Relay Module
- Motor Driver
- Sound Sensor
- Ultrasonic Sensor
- LM-35 Temperature Sensor

ESP8266 Wifi Module

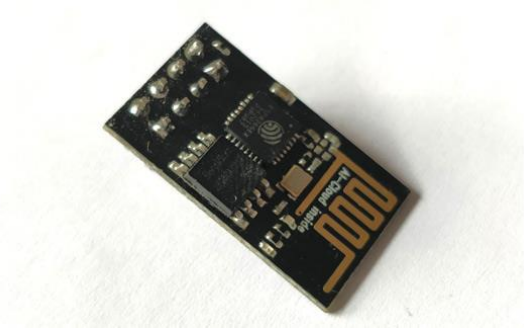


Fig.4. ESP8266 Wifi Module

The ESP8266 is a very user friendly and low cost device to provide internet connectivity to your projects. The module can work both as a Access point (can create hotspot) and as a station (can connect to Wi-Fi), hence it can easily fetch data and upload it to the internet making Internet of Things as easy as possible. It can also fetch data from internet using API's hence your project could access any information that is available in the internet, thus making it smarter. Another exciting feature of this module is that it can be programmed using the Arduino IDE which makes it a lot more user friendly. However this version of the module has only 2 GPIO pins (you can hack it to use upto 4) so you have to use it along with another microcontroller like Arduino, else you can look onto the more standalone ESP-12 or ESP-32 versions. So if you are looking for a module to get started with IOT or to provide internet connectivity to your project then this module is the right choice for you [2].

Ultrasonic Sensor

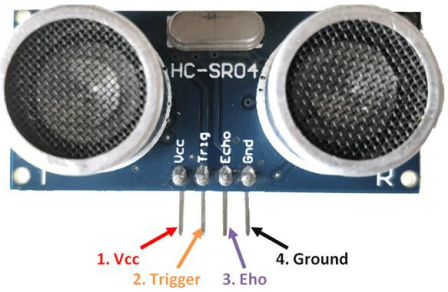


Fig.5. Ultrasonic Sensor

The HC-SR04 ultrasonic sensor includes a transmitter & a receiver. This sensor is used to find out the distance from the objective. Here the amount of time taken to transmit and receive the waves will decide the distance between the sensor and an object. This sensor uses sound waves by using non-contact technology.

- Operating voltage: +5V
- Theoretical Measuring Distance: 2cm to 450cm
- Practical Measuring Distance: 2cm to 80cm
- Accuracy: 3mm
- Measuring angle covered: <15°
- Operating Current: <15mA
- Operating Frequency: 40Hz

LM-35 Temperature Sensor

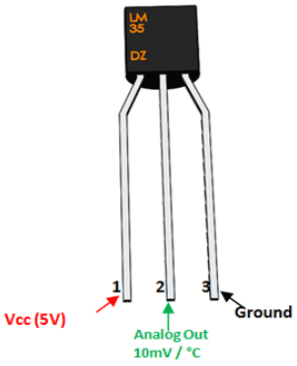


Fig.6. LM-35 Temperature Sensor

LM35 is a precession Integrated circuit Temperature sensor, whose output voltage varies, based on the temperature around it. It is a small and cheap IC which can be used to measure temperature anywhere between -55°C to 150°C. It can easily be interfaced with any Microcontroller that has ADC function or any development platform like Arduino[4].

- Minimum and Maximum Input Voltage is 35V and -2V respectively. Typically 5V.
- Can measure temperature ranging from -55°C to 150°C
- Output voltage is directly proportional (Linear) to temperature (i.e.) there will be a rise of 10mV (0.01V) for every 1°C rise in temperature.
- ±0.5°C Accuracy
- Drain current is less than 60uA
- Low cost temperature sensor
- Small and hence suitable for remote applications
- Available in TO-92, TO-220, TO-CAN and SOIC package

ESP-32 cam Module

Fig.6. ESP32 Cam Module

The ESP32-CAM is a development board with an ESP32-S chip, an OV2640 camera, microSD card slot and several GPIOs to connect peripherals. In this guide, we'll take a look at the ESP32-CAM GPIOs and how to use them. The ESP32-CAM comes with three GND pins (colored in black color) and two power pins (colored with red color): 3.3V and 5V.You can power the ESP32-CAM through the 3.3V or 5V pins. However, many people reported errors when powering the ESP32-CAM with 3.3V, so we always advise to power the ESP32-CAM through the 5V pin.

4. Results

Here we implemented Camera and different sensors to monitor real time condition of the area. Sensors and microcontroller base processing unit for collection of different parameters such as External weather condition, environmental temperature and and real time video streaming for monitoring. So that we can use this Night patrolling Device for security of any society area, college campus, Hospital Areas and Many More.

5. CONCLUSION

From this project we can keep monitoring over an area by regularly moving or travelling a route of the corresponding area. The robot captures the images with the help of camera. These images are then sent to the user in a real time, user will analyses it and if there is any problem observed then alarm is triggered manually. Robot patrolling is mostly used in Military area, Hospitals, Shopping mall, Restricted Zones, Industrial area, Agricultural area etc. The robot uses ESP32 based camera sensor which cuts down the price of using a raspberry pi. This also reduces the instructions and enables programming the robot with a least programmable skills.

Refrences

[1] Monika Agarwal and Akshaypandya, "GSM Based Condition Monitoring of Transformer", IJSRD - International Journal for Scientific Research Development| Vol. 1, Issue 12, 2014 | ISSN (online): 2321-0613
[2]https://components101.com/misc/esp8266module
[3] M.Hussain , M. Salman , Rohit , A.Subhan , H.khalid and S.H.Zaidi, "Condition Based Health Monitoring of Transformers 2018 International conference on Computing Mathematical and Engineering Technology(iCoMET) , Sukkur , 2018,pp.1
[4] https://components101.com/sensors/lm35-temperature-sensor
[5]https://components101.com/sensors/acs712-current-sensor-module
[6]https://components101.com/microcontrollers/arduino-uno

AI Based Voice Assistant Using Python

[1]Mr. Shubham kumar, [2]Nitin Kumar, [3]Dushyant Chauhan, [4]Abhijeet Kumar Ghosh

[1]Assistant professor, Galgotias university

[2,3,4] B tech (Computer Science and Technology), Galgotias University, Greater Noida

Abstract – We are living in a world where we are enclosed by machines in all view and facet. Machines are invented to lesser our efforts of doing hard and difficult works. These machines are getting advance day by day. Especially in the field of computer science these machines are updating every-day. In this rapid upgradation of machines people will feel difficult to work with it. They have to learn that how do it works and what are its functions. And some of them do not know what that particular machine can do. To overcome this difficulty; the idea of talking to machines will be a great help to those people who do not know how to operate them. They just have to give commands or we can say orders to that particular machines to do some task or work without knowing anything technical about the machine and about its background implementation as well about the logic behind it. They have to give orders just as they are ordering their assistants to do some work. They feel comfortable to communicate with the machines if the machines start answering their questions in their language. The appearance and widespread adoption of the Internet of Things has facilitated the active use of artificial intelligence technologies in human life. These independent equipments have became way more smartt in their interactiion with humans as well with 1one another. New skillss lead to the development of different systems, that use new special things into social networks of the IoT. One of the crucial movement in AI is the machanics for identifying a human's natural language

I. Introduction

Artificial intelligence systems that can arrange a natural human - machine interaction in different ways such as speech, conversation, gestures, and so on are gaining popularity at the moment.[1] One of the most researched modes of communication is that focused on a computer understanding human words. It is no longer the case that a software learns to interact with a human by observing his attitudes, behaviour, and movements in order to become his personal assistant, rather than a human learning to communicate with robots. For a long time researches are going on developing refining personalised assistant. These devices are progressing with time, and they can now be used in a range of handheld devices and gadgets in addition to computers. The most common voice assistants are Siri from Apple and "AIVA".

A brief introduction to the architecture and design of voice assistants is presented in section 1 and 2. Section 1 and 2 provide a quick overview of voice assistant architecture and design. Section 3 the

draught work schedule. Section 4 explains the AVIA voice assistant technique. The results of the voice assistant's tests are detailed in section 5. The conclusion and future potential of an assistant based on various artificial intelligence algorithms are identified in sections 6 and 7. The primary purpose of this project is to build a local voice assistant capable of performing human - like task as well as tasks that a human would perform on a daily basis.

The aim of AIVA was to develpoe a voice-controlled personal assistant that could perform a variety of tasks. You may, for example, perform an internet search. It has functions like internet searching. It also includes several new features, such as the ability to post comments on social media sites such as Facebook and Twitter. There are only a few commands that are needed. You can also find out about the weather in your area as well as the environment around you. It will start web-based applications in addition to local storage of the user.

Related works – Every institute that want its own smart assistant has its own demands and expectations that may differ from other one . If some AI subordinate can better synthesise speech, while the other can carry out tasks more efficiently and without the need for additional clarification or corrections. [2] Others can complete a smaller number of tasks more accurately and according to the user's preferences. Obviously, there is no single assistant who is capable of performing all duties equally well. It all depends on which areas of development the developing countries have prioritised. As each and every computer is based upon the idea of machinee learning techniques and make use of enormous data provided by sources before being trained on them, whetherr it is searched systems, various information sources, or social networks, is crucial. The sum of information gathered from different sources decides the assistant's essence. The basic principle of designing such systems remains roughly the same, regardless of the different approaches to learning, algorithms, and methods used. Picture 1 clearly reveals the technique to set up a smart way to interact to an individual by his natural languagee. The main technologies are voice activation, automatic speech recognition, teach - to - speech - voice biometrics, dialogue manager, natural language comprehension, and named entity recognition.

VOICE TECHNOLOGY	BRAIN TECHNOLOGY
Voice Activation	Voice Biometrics
Automatic Speech Recognition (ASR)	Dialog Management
(Teach-To-Speech (TTS)	Natural Language Understanding (NLU)
	Named Entity Recognition NER)

Fig.1 Technologies for creating intelligent systems that communicate with humans using natural language)

II. Required Tools and Framework

Visual Studio Code:

Visual Studio Code it is an open-source code editor made by Microsoft for windows, Linux and macOS. It supports for features like debugging, syntax highlighting, etc. Visual studio code is an advance code editor that supports development task such as debugging, task performance, and version control. It provides tools that a developer needs for a fast code- building.[3]

Basic features of Visual Studio Code:

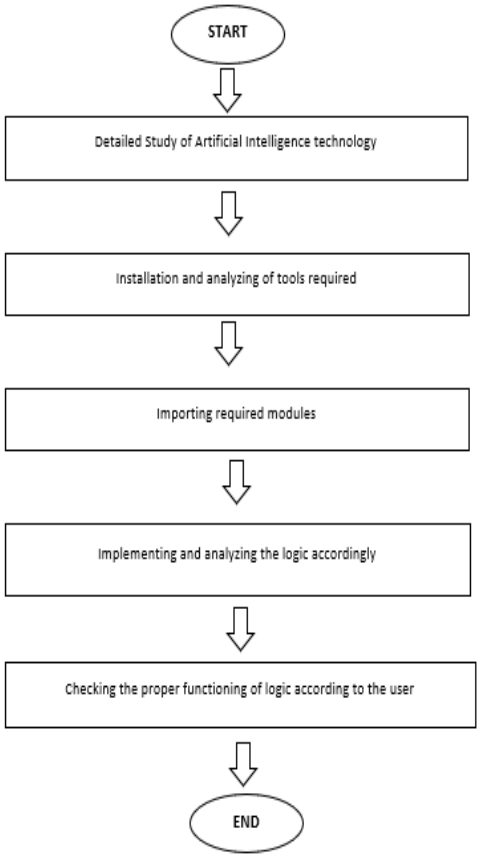
- Support for multiple programming languages
- Intelligence

- Cross – Platform Support

What is the use of Visual Studio Code in making project?

visual studio code is a user friendly code editor that provides various libraries and modules for debugging and compiling of code successfully. This code is written in visual studio code where we import and implement the modules.

III. Implementation Working Flowchart



IV. Merits of Proposed Model

. provides 24/7 customer service Consumers have requested assistance seven day a week, 24 hours a day. They can need assistance at inconvenient times, and when help is unavailable, it can be a stressful situation. Voice assistants may assist you in unavailable, it can be a stressful situation. Voice assistants may assist you in escaping those situations. For a digital talking assistant, there are no sick days or holidays, so customer service and interaction are not interrupted.

. Eradicates Language Barriers When travelling abroad or communicating with content on the internet, the majority of people face language barries. So, what's the solution? Personal assistant technology with automated translation is included to help overcome the language barrier.

. Helps Streamline Operations Another significant benefit of personal voice assistants for companies is that they streamline the processes associated with integrating digital assistants into the company. And with new technology and deep learning, these talking assistants never stop working. They are constantly updating reports, reviewing data, ensuring that critical systems are current.

. Aids Hand-free operation
Customers can use many features without having to use their hands because you can unlock them with only your voice. As a result, certain activities become easier and faster. According to a PwC report, customers often use personal voice assistants for other things such as cooking, watching TV, driving, and so on.

V. Methodology

The workflow of the voice assistant's basic process is shown in figure 2. The input will be given and gets translated to text with the help of module known as speech recognisation . that converted form is then goes to the centrall proccessor, which deduces the essence of the order and executes the appropriate script. The process don't stop here, however. [4]

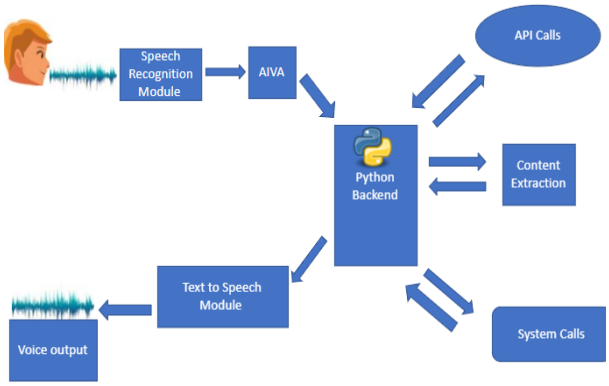


Fig.2 (Basic Workflow)

Even if you've put in very much effort and hard work, other issues will always be there affecting the working of the Assistant. One of the major issue is sounds coming from surroundings that can disturb the working. it can't tell the difference between your voice and other sounds. One more factor to consider is how people change their voice pitch to compensate for noise; speech recognition systems can be sensitive to these pitch changes. [4]

VI. Detailed Workflow

1. Speech Recognition - The computer uses google speech recognition technology to translate speech to text. Users of speech input will receive texts rom a special cora organized on the information center's computer network server, which will be stored for a short period in the system and then goes to google cloud where the speech recognitiion process takes place. The desired text is then send to the central proccessor. [5]
2. Python Backend - The recognition module's output is recevied by the python backend, which decides if the command or speech output is an API request, a context extraction, etc. [6]

3. API calls - API calls apply to the application programming interface. An API is a piece of software that enables two programmes to communicate with one another over the internet. To put it another way, an API is what that provides you the answer for your request.

4. Context extraction - process of extracting structured data from machine – docs that can be read that are unstructured or semi - structured. [7] The majority of the time, this job involves translating human language texts using natural language processing. Extraction of background recent activites in multimedia document processing, such as automated annotation and information extraction from image, audio and video.

5. System calls – It is the process through which we can get service from Kernel of the system on which it is running is known as a system call. Hardware - related services (such as accessing a hard disc drive), the formation and implementation of instant tasks, and connection to core kernel assistance such as tasks planning are all examples.

6. Text - to - Speech - refers to the capacity of a machine for study words audibly. A TTS engine decodes words that are written, which can be heard as sound. TTS engines are provided with a range of language dialects and specialized vocabularies by third - party publishers. [8]

VII. Result

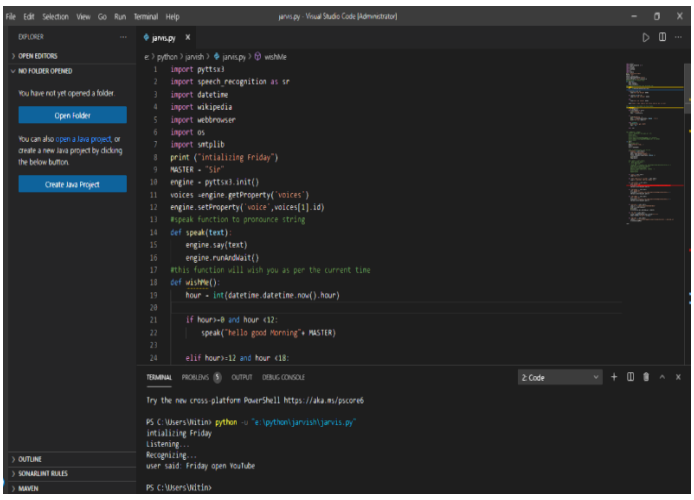


Fig.3

We have successfully imported the required modules and implemented the logic in our code.

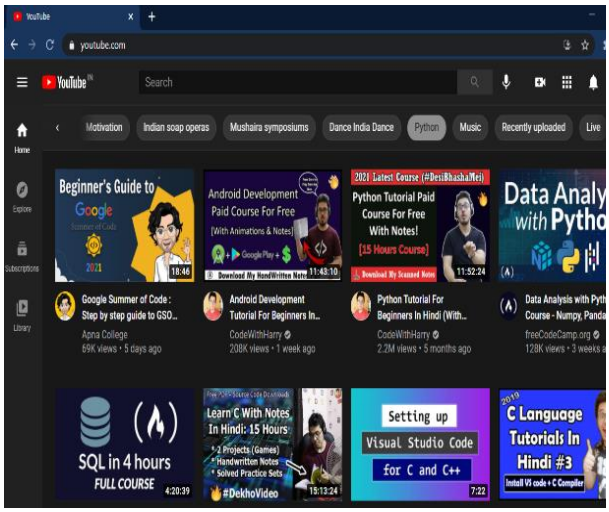


Fig. 4

VIII. Conclusion

In the above here, we revealed the formation and application of an intelligent virtual assistant. The idea is constructed with the help of software modules with Visual Studio code backing that is able to lodge any change in the project if needed in the coming time. The modules that are used in this project makes it more pliable and effortlessly introducing the new modules that will supply it with more features and that also without affecting the current state of the project.

The project not only follows the voice commands and also react to the end user according to the questions put up by the customer or the words spoken by the user such as opening tasks and operations.[9] As soon as it starts it greets the user by saying good morning, good afternoon, good evening depends on the time of usage. This way the user feels good and finds it easy to interact. The applications should also work on verbal input and eliminate any kind of unnecessary manual work required in the user life of performing each and every task. The entire system works on the verbal input rather than the text one.

IX. Acknowledgement

We have taken efforts on this project. However, it would not be possible without the help and kindness of many people and organizations. We would like to extend our gratitude to all. We are very indebted to Mr. Shubham Kumar for their guidance and regular monitoring and for providing us with the necessary details about the project and their support for the completion of the project. We'd like to express our appreciation to a Galgotias University member for their excellent support and motivation in completing this project. We would like to express our heartfelt gratitude to the people in the industry for their time and energy. Our gratitude extends to our project partners as well as those who volunteered to assist us with their skill.

X. References

[1]. Recognition Systems (Microsoft API Google API And CMU Sphinx)", Int. Journal of Research and Application 2017, 2017.

[2]. Artificial intelligence (AI), sometimes called machine intelligence. https://en.wikipedia.org/wiki/Artificial_intelligence B. Marr, The Amazing Ways Google Uses Deep Learning AI. Cortana Intelligence, Google Assistant, Apple Siri.

[3]. Fryer, L.K. and Carpenter, R., 2006. Bots as language learning tools. Language Learning & Technolog

[4]. K. Noda, H. Arie, Y. Suga, T. Ogata, Multimodal integration learning of robot behavior using deep neural networks, Elsevier: Robotics and Autonomous Systems, 2014.

[5]. "CMUSphnix Basic concepts of speech - Speech Recognition process". http://cmusphinx.sourceforge.netlwiki/tutorialconcepts.

[6]. Huang, J., Zhou, M. and Yang, D., 2007, January. Extracting Chatbot Knowledge from Online Discussion Forums. In IJCAI(Vol. 7, pp. 423-428).

[7]. Thakur, N., Hiwrale, A., Selote, S., Shinde, A. and Mahakalkar, N., Artificially Intelligent Chatbot.

[8]. Mohasi, L. and Mashao, D., 2006. Text-to-Speech Technology in Human-Computer Interaction. In 5th Conference on Human Computer Interaction in Southern Africa, South Africa (CHISA 2006, ACM SIGHI) (pp. 79-84).

[9]. Hill, J., Ford, W.R. and Farreras, I.G., 2015. Real conversations with artificial intelligence: A comparison between human– human online conversations and human–chatbot conversations Computers in Human Behavior, 49, pp.245-250.

Plant Disease detection using deep learning

Pratik Mahankal[1], Sumedh Gulvani[1], Shardul Bakare[1], Jahida Subhedar[2]
[1]*Student, Symbiosis skills, and professional university*
[2]*Professor, Symbiosis skills, and professional university*

Abstract: ---- India is a country where still 70% of people are dependent on agriculture as their main source of income, being such a high populous country there is always increasing demand for food supplies. Farmers face approx. 30% losses every year due to crops being affected by some diseases or due to some natural calamity. From past as well as even today may farmers check for diseases in plants through naked eyes most of the time plant diseases on a leaf look the same which can lead to confusion on what kind of fertilizer should we used to overcome it. To overcome this problem, we have created a deep learning-based application to predict the diseases of plants.

Key Words: Deep learning, machine learning, convolutional neural network, plant diseases, Transfer Learning.

I. INTRODUCTION:

Due to the technology boom in modern times, humans can produce enough food that they can feed more than 6 billion peoples. Due to climate changes, there is a shortage of food which threatens the overall population. Diseases in plants caused due to various changes in climate as led to severe shortages of food which has increased the threat of food shortage. Due to global warming, agriculture has been difficult as many small-scale farmers, as well as large-scale farmers, are continuously facing losses due to diseases caused by plants. Generally, small-scale farmers depend on the produced goods as most of the goods are consumed by themselves. Some studies tell that more than 50% of produce is lost due to pests and diseases found in plants. For a country like India, which has high population agriculture is the backbone of this nation, and agriculture being the primary occupation. Due to the shortage of food caused due to diseases in plants, food prices are increasing day by day. Due to the privatization of pesticide and fertilizer companies, there has been a decline in the quality of the crops due to the overuse of fertilizer and pesticides. Several initiatives have been introduced to avoid crop loss due to disease. In the last decade, integrated pest management (IPM) methods have gradually replaced historical approaches of widespread pesticide use. Earlier, local companies/organizations have assisted the farmers to detect such diseases which are sometimes unnoticed.

Introduction of digitalization and IoT and higher bandwidth it is now possible to use the camera of smartphones and other high-resolution cameras to detect the diseases of the plants. Due to the help of this technology, it is comparatively easy to detect disease in plants by the farmers in the rural area. By 2015, 70 % of the global population had access to mobile broadband coverage, with mobile broadband penetration reaching 46 percent in 2015 [6].

1.1 Introduction to machine learning:

Machine learning is the concept of employing algorithms to discover patterns and/or make predictions based on a set of data. There are numerous algorithms available, each with its own set of advantages and disadvantages, as well as varying degrees of complexity. These algorithms are simple to use and available in a variety of programming languages (including R and Python) with varying degrees of coding complexity. They may be able to replace the

need for detailed coding instructions unique to your application with more general instructions [2].

1.2 Introduction to Deep Learning:

Deep learning is one of a subset of machine learning in which a model can learn to perform tasks like classification directly from images, text, or sound. Deep learning can usually be implemented using neural network architecture. The term "Deep" refers to the number of layers in the network—the more the layers, the deeper the network. Usually, Neural networks contain only two or three layers, on the contrary deep networks can have hundreds of layers.

Deep neural network can also combine multiple non-linear processing layers, using simple elements operating in parallel**.** It is inspired by the biological nervous system and consists of an input layer, several hidden layers, and an output layer. The layers are interconnected via nodes, or neurons, with each hidden layer using the output of the previous layer as its input [1].

The recent advances in deep-learning technologies based on neural networks have led to the emergence of high-performance algorithms for interpreting images, such as object detection, semantic segmentation instance segmentation, and image generation. We know that neural networks can learn the high-dimensional hierarchical features of objects from large sets of training data, deep-learning algorithms they can also acquire a high generalization ability to recognize images, ex. they can also interpret images that they have not been shown before, which is also one of the traits of artificial intelligence. Soon after the success of deep-learning algorithms in general scene recognition challenges, attempts at automation began for imaging tasks that are conducted by human experts, such as medical diagnosis and biological-image analysis. However, despite significant advances in image recognition algorithms, the implementation of these tools for practical applications remains challenging because of the unique requirements for developing deep-learning algorithms that necessitate the complete development of hardware, datasets, and software.

1.3 Image Pre-processing:

Picture processing is a method of improving or extracting useful information from images by applying operations to them**.** Image Pre-processing is a form of signal processing in which the input of an image and the output of an image or its characteristics/features. Image processing is one of today's fastest evolving innovations. It is also a crucial research field in engineering and computer science**.** Following three steps are included in Image processing: Importing the image using image acquisition software.

- Analyzing and manipulating the image.
- Output, which may be an altered image or a report based on image analysis.

II. LITERATURE REVIEW:

India is a country where still 70% of people depend on agriculture as their main source of income, being such a high populous country there is always increasing demand for food supplies. Farmers face approx. 30% losses every year due to crops being affected by some diseases or due to some natural calamity [3]. If some plants do get affected sending them to the laboratory for results is a very time taking and risk-taking job as sometimes plants do get affected due to taking healing steps late. This problem can be solved using machine learning and deep learning and using such technology can help the farmers to detect the disease more quickly. In this study, we used the plant village dataset to train the deep learning algorithm [4].

This is where a deep learning network comes in handy. Deep Learning helps us to distinguish between the objects that help to identify if a plant has a disease Convolutional Neural Network (CNN or Convolutional Nets) Deep Learning Model is best for such a scenario.

It is well known for its widely used applications of image and video recognition and also in recommender systems and Natural Language Processing (NLP). However, convolutional is more efficient because it reduces the number of parameters which makes it different from other deep learning models [8].

Authors	Plant name	Diseases/Harvesting identified	Accuracy
Tasneem [11]	Potato	Early Blight and late Bling	94.1%
`B.Klatten[15]	Sugar beet	Cercospora beticola,Ramularia,Phoma Betae	97%
Shovon Paulinus Rozario [18]	Paddy	Bacteria leaf Blight,Brown Spot Leaf, Leaf Scald	NA
Shitala Prasad [13]	Crop	Powdery Mildew,Downy Mildew,Late,Blight	98.96%
Shitala Prasad [14]	Plant	Leaf spots	93%
Rahat Yasir [16]	Crop	Brown leaf spot, Bacterial leaf blight, Brown spot ufra, and rice blast	85%
Alham F [17]	Palm oil	Hawar leaf, Anthracnose, and Pestalotiopsis, Palmarum	87.75%
Monika Bhatnagar [12]	Tomato	Harvesting	NA

Various techniques of image processing and pattern recognition have been developed for the detection of diseases occurring on plant leaves, stem, lesion, etc. By the researchers. Quicker the disease appears on the leaf it should be taken to avoid loss. Therefore, a fast, accurate and less expensive system should be developed. The researchers adopted various methods for the detection and identification of diseases accurately. One such system uses a thresholding and backpropagation network. Input is a leaf image on which thresholding is performed to mask green pixels. Then CNN is used for classification.

1. Off-device image pre-processing using potato plant diseases identified Early blight and Blige using Leaf vein detection and Blob detection algorithm.
2. On-device image pre-processing on sugar beet diseases found Cerospora using Naïve Bayer classifier.

Table 1 Research on Plant Disease Detection

In the previous research, conducted most of them were done on a single plant or two plants. To overcome this limitation, we have used the multiple plant dataset of various types of plants and used the data to train the deep learning model.

III. METHODS AND DATASET:

3.1 Dataset Description:

We have used 87000 images of plants leaves which have been divided into 38 different classes were annotated according to the plant's name and there caused diseases. Each plant is organized in a crop disease such that we were able to predict diseases based only on the plant leaf images. Figure 1 represents one such example from one of the classes from the Plant Village dataset. We changed the size of the images to 256-256 pixels in all of the work and training was done on the resized images. We have completely divided the data set in an 80/20 ratio for training and validating set for preservation of directory structure. A new folder is created for prediction purposes. It contains 14 types of plants and 26 different types of diseases.

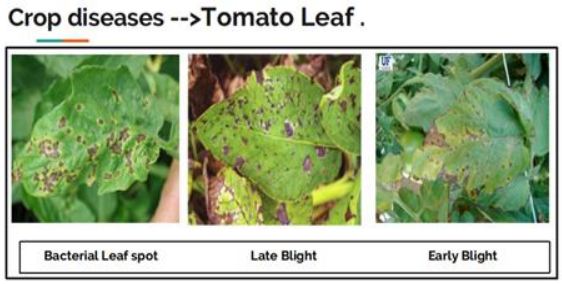


Figure: 1[19]

3.2 Measurement of Performance:
We run all of our experiments through a wide range of train-test set splits, namely 80–20, to see if the model can work on the unseen data or when the app is live.

3.3 Convolutional Neural Network:
Yann LeCun's creation of CNN in 1994 is what propelled the field of applied science and deep learning back to its former glory. The first neural network, called LeNet5, had a terrible validation accuracy of 42 percent. Almost all of the world's largest technology companies now depend on CNN for more effective efficiency. The use of CNN in detecting diseases in mulberry leaves is part of the definition. Before delving into the principle of "functionality and coping with CNN," we must first understand how the human brain identifies an entity despite its varying attributes [5]. We have taken care that there are no missing values in our dataset. The dataset must be uniquely divided into training and testing sets, and there should not be any kind of repetition of the data either training or testing sets. Images that are severely distorted or blurred should be removed from the database before being fed into the neural network. Now that we've mastered the data preprocessing laws, we can plunge straight into the operation of the convolutional neural network.

A. Convolutional Layer:
The pattern is defined by scanning the entire image and preparing it in this layer as a 3x3 matrix. The decorated function of the image is referred to by the matrix kernel. Each value in the kernel is represented by a weight vector. In our study, we have used the 4 convolutional layers with kernel size as 3 and padding as 1 with pooling as true.

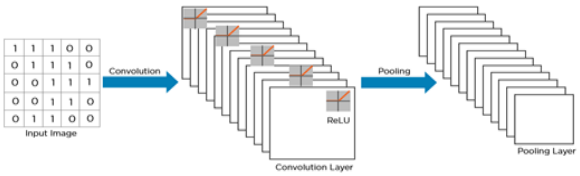


Figure 2- Convolutional Layer [20]

B. Pooling Layer:
Pooling could be a down-sampling operation that reduces the spatial property of the feature map. The corrected feature map currently goes through a pooling layer to get a pooled feature map. In this study, we have used max-pooling with a kernel size of 4.

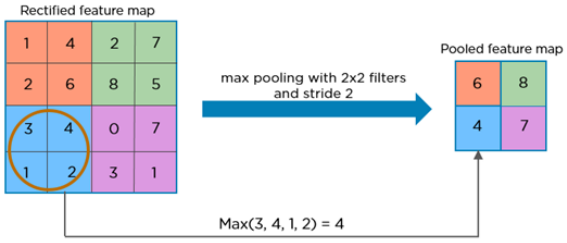


Figure 2- Pooling Layer [20]

C. Activation Layer:
It is part of the CNN where the values are normalized such that they fit within a certain range. ReLU only allows positive values and discards all the negative values [9].

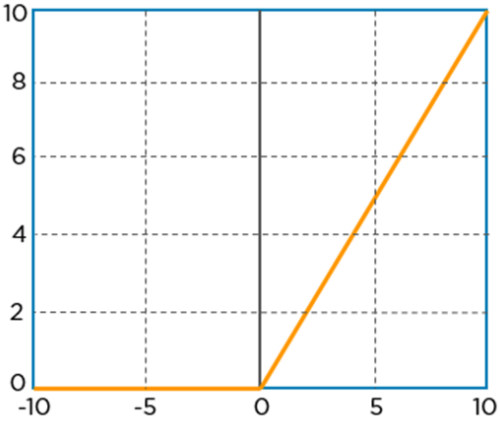


Figure 3 - Activation Layer [20]

D. Fully Joined Layer:
The features are compared to the test image's features, and the same features are applied with the defined mark.

Labels are usually encoded as numbers for numerical convenience; they will be translated to their respective strings later. The fully joined layer is used after the convolutional layer. Which connects the convolutional layer to the output. This is used to see the output in the form of the convolutional layer [9][10]. In this study, we have used 2 fully connected layers one with 128 input neurons and 128 output neurons.

3.3.1 System Architecture:

In our study, we have used the Plant Village dataset, which was trained using convolutional neural networks to predict the diseases of the plants. The images were then preprocessed and cropped. Then the images were trained on the deep learning algorithm to detect the diseases of the plants.

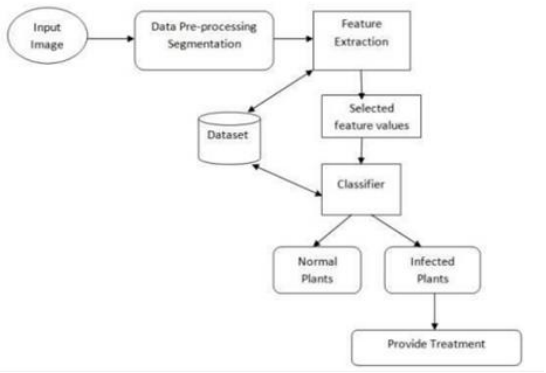


Figure 4- System Architecture

3.3.2 Model:

In our model, we have used 8 convolutional layers with kernel size as 3, padding as 1 and after the first layer after every convolutional layer, we have used the max-pooling operation with the kernel size as 4. In the first convolutional layer, we have used the input size of the 3 and the output channels as 64. All the other convolutional layer has the double the output size of the previous layer. After each of the convolutional layers, we are doing the batch normalization of the output channels. We have used batch normalization because to standardizes the inputs to a layer for each mini-batch. We have used the ReLu as our activation function with the in place as true. In this, we have used two sequential layers, one after the second convolutional layer which has combined two convolutional layers with 128 as the input and the output channels. And second, after the fourth convolutional layer which has two convolutional layers with 512 as the input and output size. The last fully connected layer has an input of 512 neurons and the output as the number of classes (38).

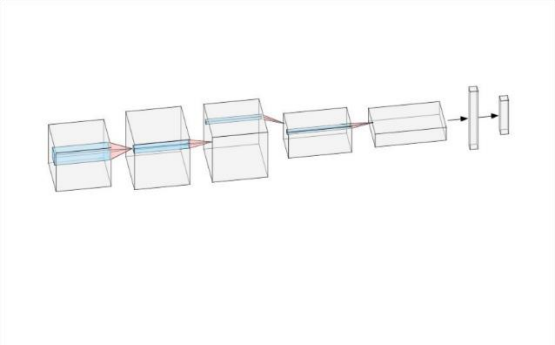


Figure 5: Model Image

3.4 Creating Website using Flask:

Flask makes the process of designing a web application much easier. Flask lets us understand what users need and what sort of response users are requesting back. The app helps us run a basic web application as if it was a real website. Flask also works on basic virtual environments such as HTML and CSS.

We created our website using HTML and CSS. We are sending the images using a POST request to the backend of the website. Where the image is sent to the deep learning model and the model predicts the disease which is again sent to the website.

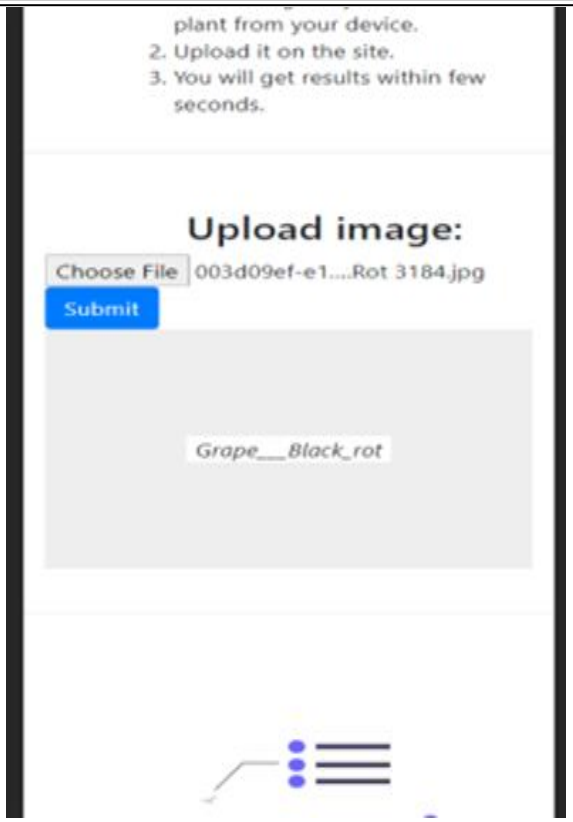


Figure 6: Website

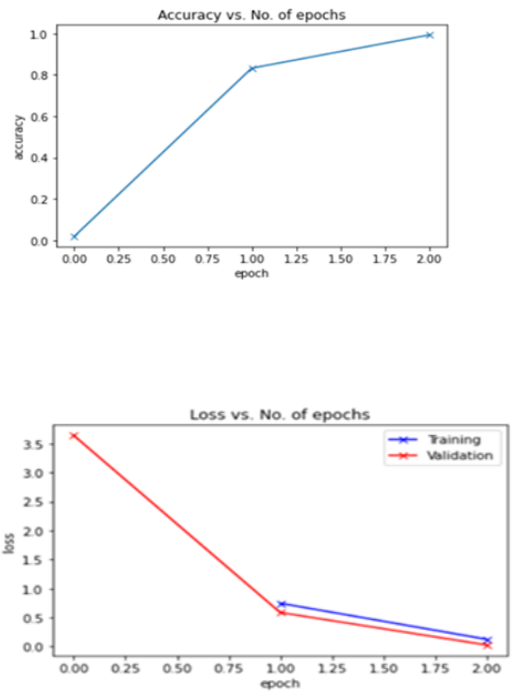


Figure 7: Results

IV. RESULTS:

At the initial stage, we noticed that with help of a dataset of 38 class labels, while testing random samples, we were able to achieve an accuracy of 2.63% on average. Overall, in our experiment we represented configuration with three visual categories of image data, we were able to achieve accuracy up to 85% to 99.18%, by using our village dataset and CNN, see figure 1. Therefore, with help of our developed model of deep learning, we can firmly predict the diseases with an error of around 0.0009 for familiar predicted problems.

To overcome this problem of overfitting, we differ the test set to train set ratio in such a way that we were able to find that even when we trained our models, we found that 80% of our data achieves an overall accuracy of 99.18% and remaining 20% data was used for training the model in the case of ResNet9 architecture. Through observations, we can say that there is not a major drop in performance as we expected when the model was over-fitting.

V. CONCLUSION:

For several years, plant diseases have been a major source of worry in agriculture. Precision agriculture or satellite agriculture has allowed us to see early diseases because crop losses have been reduced by using deep learning methods. Recent advancements in deep learning (DL) provide solutions with highly accurate performance, and available hardware allows for quick

processing. However, the decision-making procedure can be enhanced. Nowadays, existing models find it very difficult to get higher results when they are tested in the current environment with different conditions. Considering these difficulties, the Plant Disease detection model was developed to overcome significant functional limitations. This paper presented a PlantVillage dataset containing images of different plant leaves, at different angles, and in different weather, labeled for detection. Finally, the PlantDiseaseNet, a novel two-stage architecture for plant disease detection, was suggested. The trained model achieved an accuracy of 99.18 percent on the PlantVillage dataset and due to its architectural design, it is considered to be reliable in complex surroundings situations. The use of other information sources, such as location, temperature, and plant age, could theoretically improve accuracy. Future scope for plant disease detection should be focusing more on detecting diseases in various parts of plants and also should be detecting the different stages of the diseases. The developed model could be used as part of a decision support framework, providing optimal decision-making conditions.

REFERENCES:

[1] Vargas, Rocio & Mosavi, Amir & Ruiz, Ramon. (2017). DEEP LEARNING: A REVIEW. Advances in Intelligent Systems and Computing.

[2] Simon, Annina & Deo, Mahima & Selvam, Venkatesan & Babu, Ramesh. (2016). An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering. Volume. 22-24.

[3] Singh, T.V.K. & Satyanarayana, Jella & Peshin, Rajinder. (2014). Crop Loss Assessment in India- Past Experiences and Future Strategies. 10.1007/978-94-007-7796-5_9.

[4]PlantVillageDataset. https://www.kaggle.com/vipoooool/new-plant-diseases-dataset .

[5] O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints .

[6] ITU (2015). ICT Facts and Figures – the World in 2015 GenevaInternational

Telecommunication Union.

[7] Saha, Rohan. (2018). Transfer Learning - A ComparativeAnalysis. 10.13140/RG.2.2.31127.39848.

[8] Jain, Aditya & Kulkarni, Gandhar & Shah, Vraj. (2018). Natural Language Processing. International Journal of Computer Sciences and Engineering. 6. 161-167. 10.26438/IGCSE/v6i1.161167.

[9] Wikipedia. https://en.wikipedia.org .

[10] Rasche, Christoph. (2019). Computer Vision.

[11] Cynthia, Shamse & Hossain, Kazi & Hasan, Md & Asaduzzaman, Md & Das, Amit. (2019). Automated Detection of Plant Diseases Using Image Processing and Faster R-CNN Algorithm. 1-5. 10.1109/STI47673.2019.9068092.

[12] Monika Bhatnagar, Dr. Prashant Kumar Singh - "Choice of Efficient Image Classification Technique using Limited Device", International Journal of Electronics and Computer Science Engineering.

[13] Shitala Prasad, Sateesh K. Peddoju and Debashis Ghosh – " AgroMobile: A Cloud-Based Framework for Agriculturists on Mobile Platform", International Journal of Advanced Science and Technology Vol.59, (2013), pp.41-52 http://dx.doi.org/10.14257/ijast.2013.59.04 ISSN: 2005-4238 .

[14] Shitala Prasad · Sateesh K. Peddoju · Debashis Ghosh – "Multi-resolution mobile vision system for plant leaf disease diagnosis".

[15] B. Klatt, B. Kleinhenz, C. Kuhn, C. Bauckhage, M. Neumann, K. Kersting, E.-C. Oerke, L. Hallau, A.-

K. Mahlein, U. SteinerStenzel, M. Röhrig-"SmartDDS-Plant Disease Detection via Smartphone".

[16] Rahat Yasir and Nova Ahmed- "Beetles: A Mobile Application to Detect Crop Disease for Farmers in Rural Area", Workshop on Human and Technology, 8 – 10 March 2014, Khulna, Bangladesh.

[17] Alham F. Aji, Qorib Munajat, Ardhi P. Pratama, Hafizh Kalamullah, Aprinaldi, Jodi Setiyawan, and Aniati M. Arymurthy- "Detection of Palm Oil Leaf Disease with Image Processing and Neural Network Classification on Mobile Device ", International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013.

.

[18] Shovon Paulinus Rozario- " Krishokbondhu - An automated system for diagnosis of paddy disease, Thesis, SCHOOL OF ENGINEERING AND COMPUTER SCIENCE, Department of Computer Science and Engineering, BRAC University, Submitted on September 1, 2014 .

[19] Plantdisease https://towardsdatascience.com/.

[20] Convolutional neural network https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network

Secured and Entertainment based Techniques by Emotion Recognition using Machine Learning

[1]Mrs.Preethy Jemima P,[2]Mrs.Vishnu Priya N R
[1,2] Department of Computer Science and Engineering
SRM IST, Ramapuram, Chennai
preethy.jemima@gmail.com, nrpriyavishnu31@gmail.com

Abstract-Human beings come across different situations and many environments. And our behavior changes accordingly, in which facial expressions play a vital role. Without text or speech we are highly talented to communicate. Emotional expressions are traced out from the human images from their facial expressions; we can recognize them using the latest machine learning systems. In this paper to derive the high degree of accuracy some methodologies are handled, even though it is difficult and has a high level of complexity. A human-computer interaction system for an emotion recognition has attracted the attention of researchers in psychology, computer science, and related disciplines. The three main components of Emotion detection and are as follows: Image Preprocessing, Feature Extraction, Feature Classification. The basic emotions are further classified under 8 emotion classes. It can detect whether you are angry, happy, sad, etc. The first phase we used skin color detection, extracting facial features like eyes, nose, and mouth. Using CNN fully-connected layers , by increasing the number of layers the complexity is reduced. This concept can be implemented in real time scenario. While driving a car, using sensor the facial expressions of the user are detected and accordingly songs are played. The major advantage is alert can be given to the passengers travelling in the car if the driver feels sleepy. This facial recognition falls under entertainment as well as high level safety and security of human life's.

Keywords: emotion recognition, facial recognition, Preprocessing, Feature Extraction, Feature Classification.

I .Introduction

Emotions are expressed in a variety of ways, such as facial expressions, voices, physiological signals, and text. Under which Facial Expression Recognition(FER) is an emerging and most interesting task in computer vision. Its implementations and advancements in related fields can be seen especially in machine learning and image processing. Nowadays it is been growing in a wide range of applications, including human- computer interaction(HCI). The hardest part is to identify the facial expressions and classify clearly under the emotion classes. As emotions are not static always, it is always a challenging task.

Facial movement features include position and shape changes, are generally caused by the movements of facial elements. Facial expressions are nothing but the change in muscles of the face. The facial elements, constantly change their positions when subjects to emotions. Initially there will be face detection which involves skin color detection and lighting compensation for getting uniformity on face for retaining the required face portion. First extract the facial features like eyes, nose, mouth, etc to segregate the face features and group them into according emotions. Though this topic has a wide coverage of area, the topics major focus is on entertainment and security.

Basic 2 approaches are the detection of action unit and detection of facial point. The framework quantifies facial expression of human by observing the changes in facial muscle when an emotion is triggered. There are about 44 areas in the face which involves in the facial movements so-

called action units (AUs). Hence, facial expression can be recognized through the existence and intensity of several AUs.

Facial expression has two main steps; AU detection and AU recognition. To do such task, we employ CNN Convolutional Neural Network which has an architecture that consists of filter layers and a classification layer. A filter stage involves a convolutional layer, followed by a temporal pooling layer and a soft max unit. Deep learning methods have been proposed to solve the facial semantic feature recognition tasks and to detect facial point. The use of dataset has been annotated and validated by the expert of AUs. Through the dataset with ground truth we can measure the performance of the proposed method. This dataset is helpful of training the system by n number of images as input and preprocessed to segregate under particular emotion class.

II.Related Work

It describes the various speech emotion recognition and detection methods. The classification method has been used to detect several emotions

Zhou Yue et al.,2020[1] suggested a modular multi-channel deep convolutional neural network. The network output uses a global average layer to avoid overfitting. Network has certain advantages over other recognition algorithms. Finally, a real-time facial expression recognition system is constructed by using the trained recognition model. The experimental results show that the system can effectively recognize facial expressions in real time.

Keyur Patel et al.,2020[2] has described the improvements in machine and deep learning algorithms. Deep learning has made facial expression recognition the most trending research fields in computer vision area. This paper presents a systematic and comprehensive survey on current state-of-art Artificial Intelligence techniques (datasets and algorithms) that provide a solution to the aforementioned issues. It also reviews the existing novel machine and deep learning networks proposed by researchers that are specially designed for facial expression recognition based on static images and present their merits and demerits and summarized their approach.

Shekhar Singh et al.,2019[3] introduced Facial Expression Recognition (FER) with Convolutional Neural Networks. In his study he demonstrates the classification of FER based on static images, using CNNs, without requiring any pre-processing or feature extraction tasks. The results shows accuracy of 61.7% on FER2013 in a seven-classes classification task compared to 75.2% in state-of-the-art classification.

Deshmukh, G et al., 2019 [4] proposed a research on the various opinions using the different feature sets like voice Pitch, MFCC (Mel Frequency cepstral co-efficient), STE (short term energy). Classification of the opinions for the region language named as Sanskrit and Marathi was developed. The classification recognition rate for the actual time audio signal regional language Hindi was achieved up to 100%.

Michael Healy et al., 2018 [5] describes a new emotional detection system based on a video feed in real-time. It demonstrates how a personalized machine learning support vector machine (SVM) can be utilized to provide quick and reliable classification. In this paper 68-point facial landmarks features is used. Six different emotions detection training is given to application in order to find the facial expression changes in lab setting. It helps to assess the people emotional conditions so as to play the situational video machine learning

III.Analysis and Design of the Application

A. Existing Work

a. The input was an image which detects the face and extracts the facial features from the detected face and only classified under 5 emotions. The facial expressions are classified in such away like happiness, anger, sadness, fear, and surprise. Face detection is a special case of object detection. It also involves illumination compensation algorithms and morphological operations to maintain the face of the input image.

b. The past work on FER does not take the dynamics of facial expressions many efforts have been made on capturing facial movement features, and almost all of

them are video- based. These efforts try to adopt either geometric features of the tracked facial points (e.g. distance and angular), or appearance difference between holistic facial regions in consequent frames (e.g. differential-AAM), or texture and motion changes in local facial regions (e.g. animation units, and pixel difference). Although achieved promising results, these approaches often require accurate location and tracking of facial points, which remains problem.

Drawbacks:

The system plays a communicative role in interpersonal relations because they can reveal the affective state, cumulative activity, personality, intention and psychological state of a person. The proposed system consists of three modules. The face detection module is based on image segmentation technique where the given image is converted into a binary image and further used for face detection.

B.Proposed Work:

Convolutional neural network (CNN) is a class of neural network, most commonly applied to analyze images. CNNs is of multilayer perceptrons it usually means fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks make them prone to over fitting of data.

It is used to extract higher representations from the image content. Unlike the classical image recognition where you define the image features yourself, CNN takes the image's raw pixel data, trains the model, then extracts the features for better classification.

AAM (Active Appearance Model) method image segmentation algorithm also known as a computer vision algorithm for matching a statistical model of object shape and appearance to a new image. They are built during a training phase. A set of images, together with coordinates of landmarks that appear in all of the images, is provided to the training supervisor.

The sequence of selected images is stored in a database folder. The change in the AAM shape model according to the change in facial expressions measures the distance or the difference between Neutral and other facial expressions. These values are stored and a specific value is assigned for each individual expression for training the data. These difference values are then given as input and the system is trained for the different images for different expressions.

This research focus is on recognizing eight different classes of emotion through facial expression analysis using CNN. It has eight classes: neutral, anger, contempt, disgust, fear, happy, sadness, and surprise. The dataset used in this research is the Extended Cohn Kanade database (CK + database). CK+ consists of 10.708 images from 123 different subjects. During the training time period a set images given as an input is being trained in the system. Before passing the input into the CNN system, reshaping is done that is nothing but converting images into 100x100 pixels. CNN is meant for facial expression recognition.

For learning features and classification the CNN framework is used, and it consists of a feature extraction part and a classification part. The feature extraction part consists of successive convolution layers and pooling layers. The classification part consists of a fully connected layer and an output layer with a softmax function.

Advantages:

The advantage of using color space in images is it is already encoded. Transforming from RGB into any of these spaces is a straight forward linear transformation.

Face detection, Feature extraction and Facial expression recognition. The first phase of face detection involves skin color detection using YCbCr color model, lighting compensation for getting uniformity on face.

IV. Architechture

The architecture consist of basic 4 modules.Before using the data, it is important to go through a series of steps called pre-processing. This makes the data easier to handle.

Module 1

First step is to give the input image that is load the data.This dataset contains the raw pixel values of the images.From which the background

noise is removed. As the raw image occupies large amount of memory of data for processing. It is been resized and redesigned during the preprocessing module. Standardization is the process of putting different variables on the same scale. It rescales data to have a mean of 0 and a standard deviation of 1. This transformation centers the data.

Module 2

After a series of processing has done. The required facial contents alone is extracted from the image. Then it is sent to the module 3 for further scrutinization.

Module 3

In the (convolutional neural networks)CNN framework, Convolutional layers are the major building blocks A application of CNN is nothing but application of filter to an input that results in an activation. Repeated application of the same filter to an input results in a map of activations called a feature map. There are several mask in CNN, the set of input enters into the Pooling in CNN is used mainly for - 1. Dimension Reduction. Pooling in other terms known as subsampling layers. It reduce the spatial size of the representation to reduce the amount of parameters and computation in the network. **Pooling layer** is more familiar to work on feature map independently.

Module 4

Data set is created. By trial and error method, many images are fed to train the algorithm and further test the input data.

A)Training set: the algorithm will read, or 'train', on this over and over again to try and learn its task.

b)Testing set: the algorithm is tested on this data to see how well it works.

The classifier bunches the features that are isolated from the face pictures to the individual attitude classes. Bolster Vector Machine (SVM) is the most generally used classifier. Neural Networks, Nearest Neighbor and so forward are similarly used for grouping.

The image after passing a series of filters that is the number of hidden layers of CNN, the particular emotion is being detected and extracted. Emotion classifier classifies it under 8 basic emotion classes. According to the user's interest the system is trained with the desired output. The output based on entertainment is nothing but playing songs from the database which is tuned from the favorite playlist of the user.

One way to differentiate between two emotions is to see whether the persons mouth and eyes are open or not. If the driver's eyes remains closed for more than 3 seconds a continuous beep sound is played. As in default means that the drivers falls asleep so in order to provide a high level security for the passengers travelling along a beep sound is given.

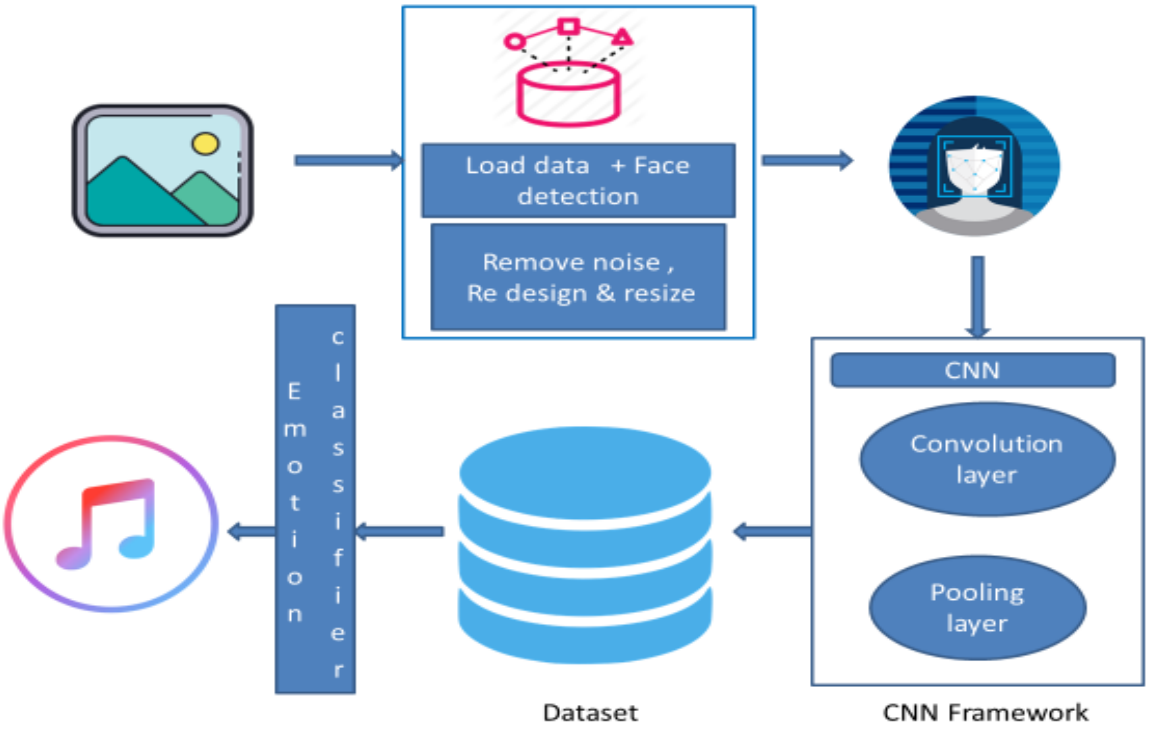


Figure 1. Architecture of the Emotion Recognition System

V. Conclusion and Future work:

The proposed CNN architecture for facial expression recognition is being handled. There are 8 classes of facial expression under emotion classes were classified and used. Using the database different training data size and the result is the mean square error declines as the number of training data increases. From the experiment we can conclude that the mean square error declines as the training data grows. This will gain better result in future work.

People can hide fear by controlling their facial expressions by maintaining a poker face. To overcome this drawback biosignals can be used in future. Biosignals is that they are spontaneous reactions that cannot deceive emotions There are number of physiological signals such as electroencephalogram (EEG), photoplethysmogram

(PPG). This will trap and recognize true emotion as these signals are associated with the central nervous system and helps to overcome the poker face difficulty. Thereby the correct and accurate emotion will be traced out.

References

[1]Zhou Yue, Feng Yanyan,Zeng Shangyou, Pan Bing,"Facial Expression Recognition Based on Convolutional Neural Network"978-1-7281-0945-9/19/$31.00©2019 IEEE
[2] Keyur Patel, Dev Mehta, Chinmay Mistry, Rajesh Gupta, Sudeep Tanwar, Neeraj Kumar, And MamounAlazab, "Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges"date of publication May 11, 2020, date of current version May 26, 2020Digital Object Identifier 10.1109/ACCESS.2020.2993803
[3]Shekhar Singh ,Fatma Nasoz ,"Facial Expression Recognition with Convolutional Neural Networks" 978-1-7281-3783-4/20/$31.00 ©2020 IEEE
[4]Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni "Speech based Emotion Recognition using Machine Learning"Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: CFP19K25-ART; ISBN:978-1-5386-7808-4
[5] Michael Healy,Ryan Donovan, Paul Walsh,Huiru Zheng"A Machine Learning Emotion Detection Platform to Support Affective Well Being" 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)978-1-5386-5488-0/18
[6] S. Sridevi and R. Rachel "An Efficient Approach for Facial Recognition based on AAM and CNN". International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 1681-1692 ISSN: 1314-3395 (on-line version) /$31.00 ©2018 IEEE

[7] MinSeop Lee , Yun Kyu Lee 1, Myo-Taeg Lim ,and Tae-Koo Kang , "Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features"
[8] D Y Liliana , "Emotion recognition from facial expression using deep convolutional neural network", Journal of Physics: Conference Series PAPER 2019 J. Phys.: Conf. Ser. 1193 012004
[9] Aarohi Gupta"Emotion Detection: a Machine Learning Project",A computer vision project about emotion detection,
Dec 28, 2019·
[10] WafaMellouk[a] and WahidaHandouzi[a]" Facial emotion recognition using deep learning: review and insights", Procedia Computer Science
Volume 175, 2020, Pages 689-694

Author Profile:

P.Preethy Jemima received her B.E.(CSE) degree from National College of Engineering , Tirunelveli in 2011 and M.E.(CSE) degree at S.A. Engineering College, Chennai in 2013. Her area of interest are Network Security and Mobile Computing. She is currently working as Assistant Professor in the department of CSE in SRM Institute of science and technology, Ramapuram.

Mrs.Vishnu Priya N R received her B.Tech(IT)-2014-Jeppiaar SRR engineering college m.e(cse)-2016-Jeppiaar engineering college. Her area of interest are Network Security and Mobile Computing. She is currently working as Assistant Professor in the department of CSE in SRM Institute of science and technology, Ramapuram.

A Deep Learning Approach for Automatic Gender Classification using Transfer Learning

Rachana Patel[1], Sanskruti Patel[2], Nilay Ganatra[3], Atul Patel[4]

[1,2,3,4]Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa

Abstract Gender classification has been an active research are for the past few years. Numerous researches applying their knowledge in order to contribute in this area. Gender classification is very important for the various tasks like traffic monitoring, surveillance, computer vision and human-computer interaction. The conventional method applied for the gender classification has various limitations and lack of accuracy with live images. However, deep learning algorithms have been showing their potential in various computer vision application including gender classification. For the purpose of gender classification, in this research paper, pre-trained models VGG19, ResNet50 and MobileNetV2 of Keras deep learning framework are utilized. All the models are already trained on large scale dataset ImageNet. For the experimentation purpose transfer learning concept using pre-trained model has been applied on the Kaggle dataset with 58,700 images.

Keywords: Deep learning, Convolutional neutral networks, Gender classification

1. Introduction

Gender classification was initially the topic of study in psychophysical studies and it focuses on the studying human visual processing. Based on the study, important features are identified which are used to classify human individual either as a male or as a female. Human facial structure plays a significant role in gender classification. Moreover, gender classification can be useful for the applications like surveillance, male and female counting system, biometrics and so on. However, gender classification using facial images is challenging in real time environment as images being affected by the various factors like facial expression, noises, occlusion and varied background [1]. Gender identification from the facial images is considered as a computer vision application.

The figure 1 illustrates the conventional approach that is used in a face-based gender classification. The approach involves various stages that includes image acquisition, image pre-processing, segmentation, feature extraction and classification. However, this approach performs manual feature extraction that requires domain expert who has a prior knowledge of the application domain. Moreover, performance of the classification system is highly dependent on the accuracy of the feature extracted from the input dataset [2]. Though, it is difficult to identify the classifier which is best suitable with chosen features for the optimal classification performance. Moreover, any deviation in the problem domain requires designing system from scratch again.

Advancement in the Deep Learning (DL) architectures in last decade provide significant opportunities for applying it in the various fields and its applications. Deep Convolutional Neural Network is the one of the extensively used techniques for the computer vision application. CNNs are capable of automatic feature generation and provide data insight which is used for efficient classification. It can be used in various face recognition applications like face-based gender identification, age-based detection using the facial images dataset [3]. The figure 2 illustrates the usage of deep CNN approach where Image passed through the deep CNN undergoes to image transformation for further processing, data dimensionality reduction, feature extraction and procurement and classification.

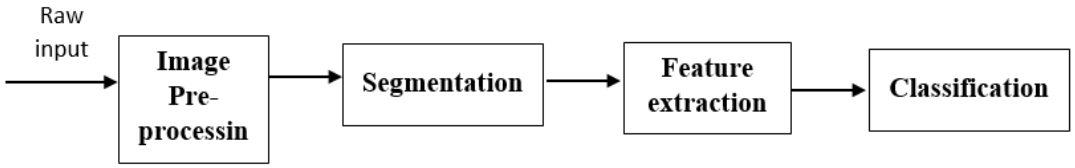


Figure 1. Conventional approach for gender classification

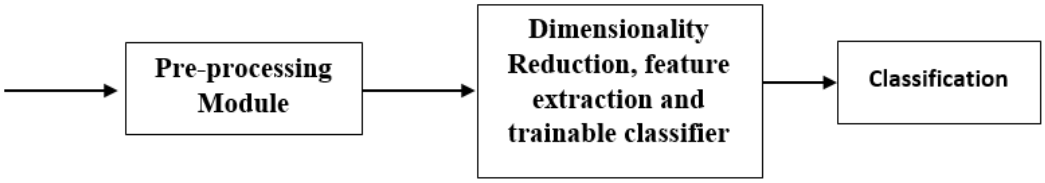


Figure 2. Deep convolutional neural network approach

In this paper, CNN with transfer learning approach has been applied for the gender classification using facial images. The state-of-the-art transfer learning architectures like VGG19, ResNet50 and MobileNetV2 are used for the classification purpose. The architectures extract the features in hierarchical manner from the input images and output the probability of the prediction. Dataset containing 58,700 images is used for the implementation purpose.

2. Literature Review

Authors in the paper [3] proposed their own deep learning architecture for the gender classification using facial images. The architecture consists of ten convolutional layers. These layers are followed by max pooling that is then followed by fully connected layer. For the training and validation purpose 1500 and 1000 images respectively used from CASIA dataset. Also, Gill Levi et al. [4] proposed small network with three convolutional layer and two fully connected layers for the gender classification. This architecture was trained and tested on the Adience dataset which contains approximately 26,000 images. Authors in the paper [5] used Deep Convolutional Neural Network (D-CNN) VGGnet architecture in extreme conditions with dataset with limited image samples. Dataset consists of total 200 images out of which 160 is used for training and 40 is used for the testing purpose. Tiagrajah V et al. [6] used transfer learning concept of deep learning in order to classify gender based on Asian faces. They have created their own dataset consisting 1000 different facial images

of Asian people. Some authors have used traditional machine learning approach for gender classification.

Xiaofeng et al [7] have used SVM classifier on dataset with 310 images for gender classification. They were able to achieve 72.73% highest accuracy with 192 features extracted from the image dataset. Authors in the [8] developed an architecture that has 10 convolutional layers. It has also 4 max pooling layers and one average pooling layer. The backpropagation method was used to train the network. The trained images are classified using KNN classifier. The LFW dataset was used for experiment and the performance of the model was tested on consist of 13,233 sample images.

3. Deep Learning Architectures

Machine Learning (ML) and Deep Learning (DL) are the subdomain of the Artificial Intelligence (AI). However, DL is considered as the hierarchical learning mechanism. Deep learning algorithms are widely used in areas like computer vision, image analysis, machine translation and many more. However, CNN algorithms provide better accuracy compared with traditional machine learning algorithms in the classification task. CNN architecture contains two types of layers; layers for feature extraction and layers for classification. Feature extraction layer uses multiple convolutional layer and activation function for efficient feature extraction. Also, classification layer contains fully connected layers which performs the classification task based on the features extracted by the previous layer. To provide better classification accuracy CNN architecture automatically adjust its input weights at different hidden layers. CNN has been widely used in the various task of computer vision after AlexNet [9] won the ILSVRC in 2012. In this paper, transfer learning approach for gender classification has been applied using three CNN architectures.

3.1 VGG19

The VGG [10] network architecture was introduced and developed by Simonyan and Zisserman in 2014. This simple network was formed by simply stacking 3x3 convolutional layers on top of each other. To reduce the volume size max pooling was used in the network. Two fully connected layers that contains 4096 nodes which are followed by the softmax classifier with 1000 nodes used in the top of the network for the classification. Because of its uniqueness compared with other networks it won the ILSVRC in 2014. It was able to achieve the top 5 accuracy 92.7% The input to the network is fixed size 224x224 RGB image.

3.2 ResNet50

ResNet [11] is considered as the specialist architecture which can train hundreds or even thousands of layers without performance degradation. ResNet architecture introduced skip connections with deep architecture. ResNet50 is the 50-layers deep residual architecture. There are other variants of ResNet architecture with different number of layers like ResNet101 and ResNet152 are also available. It was the winner of ILSRVC 2015 and beaten the human performance on ImageNet dataset. ResNet50 is widely used architecture for transfer learning provides promising result in various computer vision tasks [12]. Skip connection introduced in the network add the output from the previous layer to the next layer. It overcomes the problem of vanishing gradient that arises in training of the neural network.

3.3 MobileNetV2

MobileNet [13] architecture was introduced by Google. It is the lightweight architecture which demand less computational power and well suitable for the handheld and embedded devices for the vision based applications. Unlike traditional convolution method, MobileNet architecture follows depth wise convolution followed by pointwise convolution. It is known as depth-wise separable convolution. Less number of parameters required fewer floating-point multiplication operations. This makes it suitable for the mobile devices and embedded devices that

normally operates on less computational power. Normal CNN architecture uses 3x3 convolution layer followed by batch normalization and ReLu [14]. Whereas, MobileNet splits the 3x3 convolution in the depth-wise convolution and a 1x1 pointwise convolution.

4. Experiments and Results

In this paper, a gender classification using facial images dataset is implemented using transfer learning concept of the deep learning. Transfer learning is the deep learning technique where model trained and developed for a task can be used as the starting point for the model in another similar task. In transfer learning pre-trained model is chosen from available models. Chosen model can be the used as a whole or part of model will be used for the new task is depend on the requirement of the task.

4.1 Dataset

For the experimentation, dataset consisting approximately 58,700 images of male and female is used (KAGGLE, 2019) [15]. The dataset contains cropped images of male and female faces. The dataset is divided into training and validation directories. There are 23,200 images of male faces and 23,800 images of female faces in the training set of datasets. Moreover, validation directory consists of 5841 images of female and 5808 images of male. All images have been resized to 224x224 dimension before feed into the network for training and validation. Moreover, augmentation has been applied while training the network to avoid the problem of network overfitting. The sample images used in this study are shown in following figure 3.

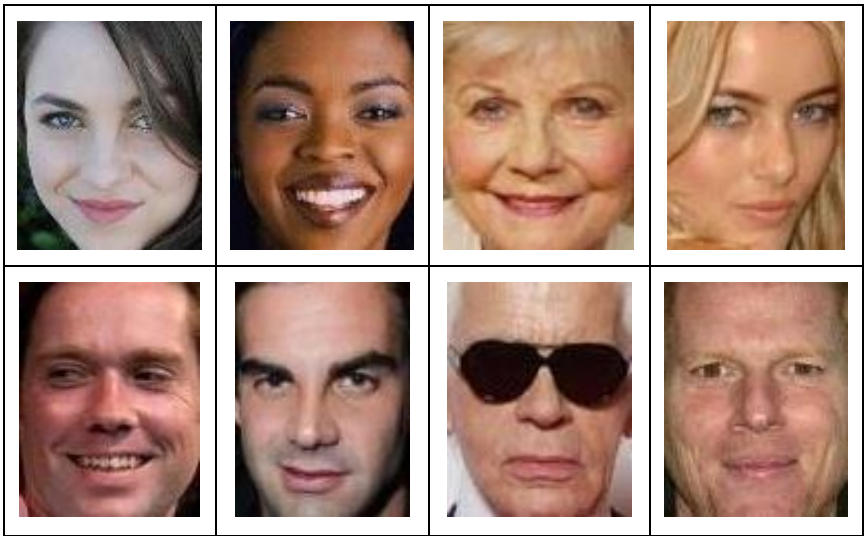


Figure 3. Sample dataset images used for experimentation

4.2 Convolutional Neural Network Models

In this paper, three state-of-the-art pre-trained convolutional neural networks: VGG19, ResNet5- and MobileNetV2 have been used for gender classification based on given dataset. Keras with Tensorflow as a backend used in Google Colab for the implementation. The dataset is divided into training and testing datasets in the ration of 80:20. The general flow of implementation is depicted in the following figure 4.

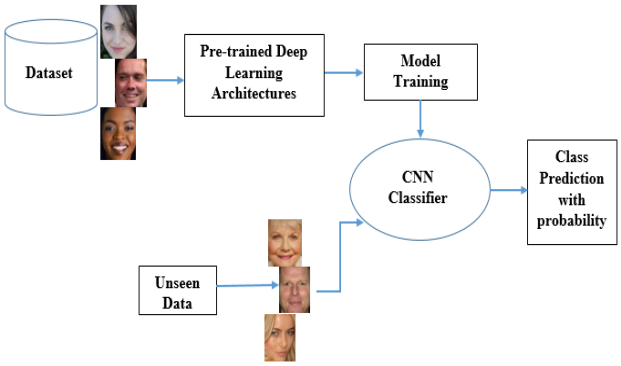


Figure 4. Block diagram of pre-trained model working

Performance comparison of various pre-trained CNN architectures has been done using the accuracy, precision, recall and F1-score. This evaluation metrics are obtained using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [16]. The mathematical equation of each as presented in following equations:

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

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

Table-1 represents the training, validation and testing accuracy for pre-trained networks VGG19, ResNet50 and MobileNetV2 utilized for gender classification. Based on the results obtained it is clear that ResNet50 outperforms compared to other classifiers. However, MobileNetV2 also provides noticeable results. Also, ResNet50 provides better result with accuracy (93%) compare to VGG19 and MobileNetV2 architectures.

Table-1 Obtained training, validation and testing accuracy and loss values

CNN Architecture	Training		Validation		Testing	
	Accuracy (%)	Loss (%)	Accuracy (%)	Loss (%)	Accuracy (%)	Loss (%)
VGG19	90.18	0.073	90.10	0.078	90.14	0.075
ResNet 50	93	0.034	92.90	0.024	92.95	0.048
Mobile NetV2	91.50	0.069	91.30	0.071	91.25	0.068

Table 2. represents the precision, recall and F1-score values for all the pre-trained models. Obtained result clearly states that ResNet50 provides highest values for the all the three performance evaluation matrices applied for gender classification problem using facial imagery.

Table-2 Obtained precision, recall and F1-score values

CNN Architecture	Precision (%)	Recall (%)	F1-Score (%)
VGG19	90.10	90.12	90.10
ResNet50	92.80	92.75	92.77
MobileNetV2	91.35	91.45	91.39

Figure 5 shows the training and testing accuracy obtained by ResNet50 model of gender classification. The graph shown in the figure clearly shows that there is overfitting did not occur during the model training phase. The loss values represent the sum of error occurred for each sample in the training and test subsets of the dataset.

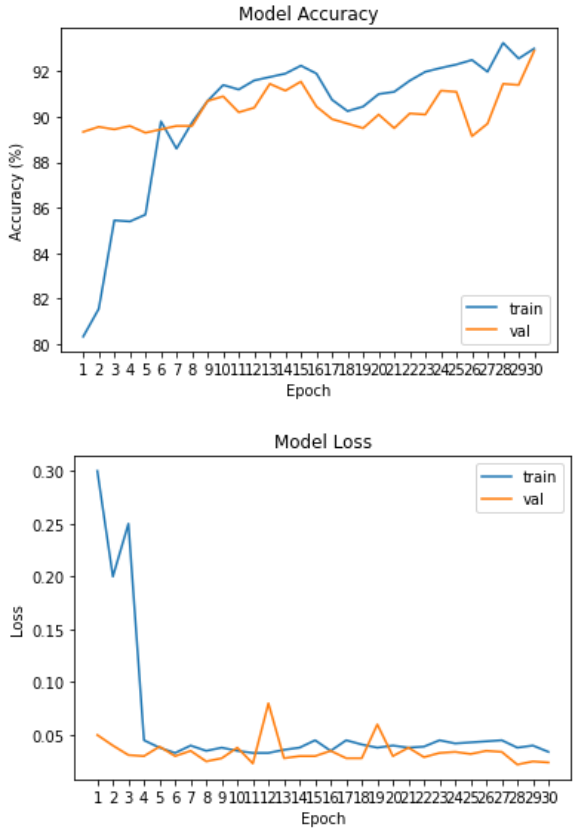


Figure 5. Accuracy and Loss Results for ResNet50 Model

5. Conclusion

In this paper, various pre-trained deep learning models like VGG19, ResNet50 and MobileNetV2 are applied for gender classification using facial images on the dataset with huge amount of images. Moreover, various previous approaches used by the different authors have been discussed during the study. The accuracy and other parameters results obtained during experimentation have been compared for obtaining the best pre-trained gender classification model. All pre-trained models were trained with 30 epochs and by considering batch size of 16. ResNet50 model delivered best accuracy on training, validation and testing sets. Then, MobileNetV2 and VGG19 also provided significant accuracy in gender classification.

References

[1] F. H. C. Tivive and A, "Bouzerdoum(Sep 2006) "A gender recognition system using shunting inhibitory convolutional neural networks," pp. 5336–5341.

[2] R. E. Eidinger and T, "Hassner (Dec 2014) "Age and gender estimation of unfiltered faces"IEEE," *Transactions on Information Forensics and*, vol. Security,9(12):21702179.

[3] S. Arora and M. P. S. Bhatia, "A robust approach for gender recognition using deep learning," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2018.

[4] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

[5] A. Dhomne, R. Kumar, and V. Bhan, "Gender recognition through face using deep learning," *Procedia Comput. Sci.*, vol. 132, pp. 2–10, 2018.

[6] T. V. Janahiraman and P. Subramaniam, "Gender classification based on Asian faces using deep learning," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, 2019.

[7] X. Wang, A. Mohd Ali, and P. Angelov, "Gender and age classification of human faces for automatic detection of anomalous human behaviour," in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, 2017.

[8] S. Haseena, S. Bharathi, I. Padmapriya, and R. Lekhaa, "Deep LearningBased Approach for Gender Classification," 2018, pp. 1396–1399.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton,

"ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv [cs.CV]*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv [cs.CV]*, 2015.

[12] A. Verma, H. Qassim, and D. Feinzimer, "Residual squeeze CNDS deep learning CNN model for very large-scale places image recognition," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2017.

[13] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv [cs.CV]*, 2017.

[14] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo, "A novel image classification approach via dense-MobileNet models," *Mob. Inf. Syst.*, vol. 2020, pp. 1–8, 2020.

[15] A. Chauhan, "Gender Classification Dataset." .

[16] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm," *Towards Data Science*, 24-Feb-2018. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. [Accessed: 06-Jun-2021].

Security Enhancement Model for Intrusion Detection System using Classification Techniques

[1]Rakhi Shukla, [2]Dr. Aarti Kumar
[1]Research Scholar, Rabindranath Tagore University, Bhopal, Madhya Pradesh,
rakhi.shukla@mp.gov.in
[2]Head-MOOC & Content ,Rabindranath Tagore University, Bhopal, Madhya Pradesh,
aartikumar@aisect.org

ABSTRACT
With the growing use of computer networks across different fields and applications, network security is becoming increasingly important. Today, with the rapid growth and the broad application of the Internet and Intranet, computer networks have brought great convenience to people's life and work. Many researchers have introduced more innovative techniques to detect intrusions in recent years, such as machine learning, data mining, evolutionary approaches, and optimization techniques. Intrusion detection is considered one of the emerging research areas nowadays. This paper presents the overview for detecting intrusion using different techniques and also discusses the other intrusion detection techniques.

Keywords:- Intrusion detection, Machine learning, Deep learning, Classification, KDDCUP.

INTRODUCTION

An intrusion detection system (IDS) monitors anomalous activities and differentiates between normal and abnormal behaviors (intrusion) in a host system or a network. Intrusion Detection Systems are a mechanism, which protects resources and data from unauthorized access, misuse, and malicious intrusions in a distributed computing environment. Machine learning techniques, such as Neural Networks, Support Vector Machines, Naïve Bayesian Classifiers, etc., are standard techniques for intrusion detection. IDSs constantly monitor and analyze the system, which allows the machine learning model to recognize everyday/normal behavior. This allows the model to detect abnormal/anomalous behavior and react with the appropriate response. The standard dataset used for IDS developments and testing is the KDD99 dataset. The IDS is tasked with monitoring and analyzing network activity to differentiate between normal and anomalous activities. If the anomalous activity goes undetected, this could potentially cause severe damage to the infrastructure and reliability of a computer system. Therefore, the detection rate of anomalous activity must be maximized. Simultaneous to anomalous activity detection, the IDS must minimize the false positive rate to avoid undue hassle and confusion. False positives do not put the system at risk. However, they can become a problem if the rate at which they occur is high, limiting the IDS's ability to provide reliable and precise results. The balance between detection rate and false-positive rate is the key for an effective IDS. The balance between detection and false-positive rates becomes more challenging when regular activity and anomalous activity are not static.

The activity on the network can change, and the IDS must be aware of this change and adapt accordingly. If not, the ability of the IDS to provide accurate and reliable results is greatly diminished. Therefore, an IDS must adapt to different environments, potentially bringing different activities and behavior unseen by the IDS.

LITERATURE REVIEW

Here we present the literature survey for the network-based intrusion detection system, as we know that the demand of computer network is increasing day by day, so we have to protect our network from the intruder or attacker, in this section we review the various researcher's paper for the efficient and attack free network or system. Here we present the author review with their respective details is mentioned in reference sections.

Anomaly-based IDS makes the detection of the data packets in the network traffic analyze the packets of data that unfit the typical profile that has been created. Ripon Patgiriet. Al [1]applied machine learning algorithms to detect intrusions effectively. Machine Learning is a statistical method for handling regression and classification tasks. These methods Moreover, They analyze the performance of the model in binary classification and multiclass classification. The number of neurons and different learning rate impacts the performance of the proposed model. Here, they compare it with J48, artificial neural network, random forest, support vector machine, and other machine learning methods proposed by previous researchers on the benchmark data set.

An intrusion detection system (IDS) monitors anomalous activities and differentiates between normal and abnormal behaviors (intrusion) in a host system or a network. The IDS must maintain a high intrusion detection rate (DR) while simultaneously maintain a low false alarm rate (FAR).James Brownet al.[4] focus on detecting anomalous network packet instances using an Evolutionary General include Support Vector Machines (SVM) for regression and classification, Naive Bayes for classification, and k-Nearest Neighbors(KNN) for regression and classification.

In recent years, one of the main focuses within NIDS research has been the application of machine learning and shallow learning techniques such as Naive Bayes, Decision Trees, and Support Vector Machines. Shone et al. [2] proposed a novel deep learning model to enable NIDS operation within modern networks. The model they propose is a combination of deep and shallow learning, capable of correctly analyzing a wide range of network traffic. More specifically, they combine the power of stacking their proposed non-symmetric deep auto-encoder (NDAE) (deep-learning) and the accuracy and speed of Random Forest (RF) (shallow learning). They have practically evaluated their model using GPU-enabled Tensor Flow and obtained promising results from analyzing the KDD Cup '99 and NSL-KDD datasets.

Chuanlong et al. [3]explored how to model an intrusion detection system based on deep learning, and we propose a deep learning approach for intrusion detection using recurrent).

Regression Neural Network (E-GRNN). They use simulated network data obtained from the UNB ISCX Intrusion Detection Evaluation Dataset. They extracted features from the application layer protocols (e.g., HTTP, FTP, SMTP, etc.) used in network activities. The E-GRNN takes the standard GRNN by evolving the sigma value used for training and a feature mask, which extracts salient features from the dataset by removing irrelevant and redundant features. The E-GRNN model reduces the computational complexity of the network anomaly detection and increases the accuracy as well. The E-GRNN reduced the feature set by an average of 60% while maintaining an average detection rate of 93.63% and a false positive rate of 2.82%. This shows the efficacy of the EGRNN model for network anomaly.

Longjing Liet al. [5] proposed the GINI GBDT-PSO method, a novel hybrid intrusion detection model to improve the performance of network intrusion detection systems. The proposed model first extracts the optimal subset of features from the whole dataset using the Gini index. Then, the GBDT algorithm, a gradient boosting approach, is adopted to detect abnormal connections. In addition, the PSO algorithm is employed to optimize the parameters of the GBDT algorithm in the proposed model. This model can enhance the overall performance for network intrusion detection effectively and improve the detection performance for each type of attack.

Machine learning plays an essential role in building intrusion detection systems. However, with the increase of data capacity and data dimension, shallow machine learning is becoming more limited. Yanqing Yanget al. [6] proposed a fuzzy aggregation approach using the modified density peak clustering algorithm (MDPCA) and deep belief networks (DBNs). To reduce the size of the training set and the imbalance of the samples, MDPCA is used to divide the training set into several subsets with similar sets of attributes. Each subset is used to train its sub-DBNs classifier. These sub-DBN classifiers can learn and explore high-level abstract features, automatically reduce The evaluation results demonstrate that the proposed classifier outperforms other models in the literature data dimensions, and perform classification well. According to the nearest neighbor criterion, the fuzzy membership weights of each test sample in each sub-DBNs classifier are calculated. The output of all sub-DBNs classifiers is aggregated based on fuzzy membership weights. Experimental results on the NSL-KDD and UNSW-NB15 datasets show that our proposed model has higher overall accuracy, recall, precision, and F1-score than other well-known classification methods. Furthermore, the proposed model achieves better accuracy, detection rate, and false positive rate than state-of-the-art intrusion detection methods.

Network Intrusion Detection System (NIDS) constitutes an essential security tool for monitoring network traffic and identifying network attacks. NIDS can be categorized into three main categories based on the detection method they use in identifying potential attacks as signature-based, anomaly-based, or specification-based NIDS. Malek Al-Zewairiet. al[7], they explore a deep learning binomial classifier for Network Intrusion Detection System is proposed and experimentally evaluated using the UNSW-NB15 dataset. Three different experiments were executed to determine the optimal activation function, select the essential features, and test the proposed model on unseendata.

with 98.99% accuracy and 0.56% false alarm rate on unseen data.

Ref. No.	Publication Details	Methods	Dataset	Performance Parameters	Limitations
[1]	IEEE Symposium Series on Computational Intelligence, 2018.	Random forest & Support vector machines	NSL-KDD	Accuracy, Precision, Recall	May used some more classification techniques
[2]	IEEE Transactions On Emerging Topics In Computational Intelligence, 2017.	Deep learning	KDDCUP 99, NSL-KDD	Accuracy, Precision, Recall, F-Score, False alarm	Real-world network traffic Datasets
[3]	IEEE Access, 2017.	Recurrent neural networks	NSL-KDD	Actual positive rate, False positive rate, the Detection rate	To reduce the training time using GPU acceleration
[4]	IEEE, 2016.	Evolutionary General Regression	UNB ISCX Intrusion	Accuracy, True positive rate, False	Need more trained dataset with

		Neural Network	Detection Evaluation Dataset.	positive rate, Detection rate, False-negative rate	feature reduction techniques
[5]	Journal of Sensors, 2018.	Gini index with Particle swarm optimization	NSL-KDD	Accuracy, Detection rate, Precision, F1-Score, and False alarm	Performance may increase using some machine learning model
[6]	Applied Science Journal, 2018.	Modified Density Peak Clustering Algorithm and Deep Belief Networks	NSL-KDD	Accuracy, Detection rate and False positive rate	Plan to use the adversarial learning method to synthesize U2R and R2L attacks
[7]	International Conference on New Trends in Computing Sciences, IEEE 2017.	Multilayer feed-forward artificial neural network using back-propagation	UNSW-NB15 dataset	Accuracy, False positive rate, Precision, F1-Score, and Recall	The performance will also measure in some other data types
[8]	International Conference on Soft-computing and Network Security, IEEE 2018	Decision Tree model	KDD99 intrusion dataset	Accurate positive, False positive, and Processing time	The performance will also measure in some other performance parameters value.

Table 1: Comparative study for intrusion detection techniques.

KDD Cup99

The KDD Cup '99 dataset was used in DARPA's IDS evaluation program. The data consists of 4 gigabytes-worth of compressed tcp dump data resulting from 7 weeks of network traffic. This can be processed into about 5 million connection records, each with about 100 bytes. It consists of approximately 4,900,000 single connection vectors, each of which contains 41 features. These include Basic features (e.g., protocol type, packet size), Domain knowledge features (e.g., number of failed logins), and timed observation features (e.g.%, of connections with SYN errors). Each vector is labeled as either normal or as an attack.

It is good to note that the KDD CUP dataset has been widely used by the researchers, especially for IDS studies. This database contains a standard set of data to be audited, including a wide variety of intrusions simulated in a military network environment. Moreover, it is pretty tricky to collect such a vast amount of data with a Lab's set-up environment to obtain long-term raw TCP dump data for a network. Thene twork was operated as if it were a natural environment but sprinkled with multiple attacks.

Sr. No.	Name of Features	Types
1	duration	continuous
2	protocol_type	symbolic
3	service	symbolic
4	flag	symbolic
5	src_bytes	continuous

6	dst_bytes	continuous
7	land	symbolic
8	wrong_fragment	continuous
9	urgent	continuous
10	hot	continuous
11	num_failed_logins	continuous
12	logged_in	symbolic
13	num_compromised	continuous
14	root_shell	continuous
15	su_attempted	continuous
16	num_root	continuous
17	num_file_creations	continuous
18	num_shells	continuous
19	num_access_files	continuous
20	num_outbound_cmds	continuous
21	is_host_login	symbolic
22	is_guest_login	symbolic
23	count	continuous
24	srv_count	continuous
25	serror_rat	continuous
26	srv_serror_rat	continuous
27	rerror_rate	continuous
28	srv_rerror_rate	continuous
29	same_srv_rate	continuous
30	diff_srv_rate	continuous
31	srv_diff_host_rate	continuous
32	dst_host_count	continuous
33	dst_host_srv_count	continuous
34	dst_host_same_srv_rate	continuous
35	dst_host_diff_srv_rate	continuous
36	dst_host_same_src_port_rate	continuous
37	dst_host_srv_diff_host_rate	continuous
38	dst_host_serror_rate	continuous
39	dst_host_srv_serror_rate	continuous
40	dst_host_rerror_rate	continuous
41	dst_host_srv_rerror_rate	continuous

Table 2: List of KDDCUP features.

DATASET DESCRIPTION

To get to know about the data and find relations between data, it is necessary to discuss data objects, data attributes, and data attributes. All the features present in KDD datasets have two types, either **symbolic or continuous**.

Continuous data have an infinite number of states. Continuous data is of float type. There can be many values between any numbers. For example, the height of any person may vary between 5.2 to 6.2.

Symbolic data are distinctive in their own right on any sized data sets, small or large. For example, it is not unreasonable to have data consisting of variables, each recorded in a range. Likewise, we can formalize a computer security-based engineering company as having a knowledge base consisting of the files and data. Such data and files are more aptly described as concepts rather than standard data, and as such, are also examples of symbolic data.

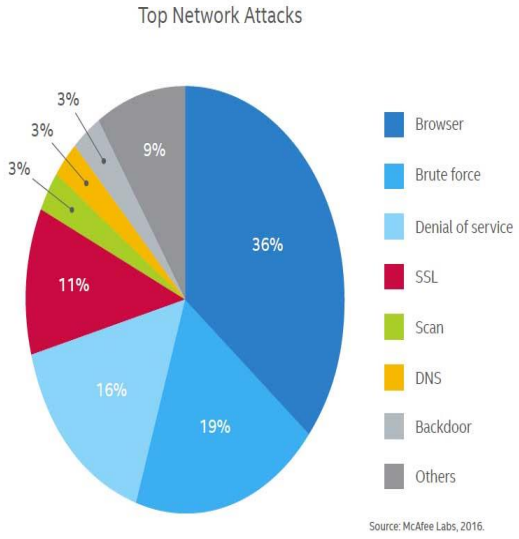


Figure 1: Top network Attacks [McAfee Labs, 2016].

Intrusion detection techniques used by the researchers are:

RANDOM FOREST

Random Forest is a supervised classification algorithm. A random forest can be used for regression and classification tasks [1]. Random forest classifier forms a bunch of several decision trees from randomly selected features. Then, it calculates votes from the different decision trees for each predicted target, and the highest voted class is considered the final prediction. Let training set is provided as: [A1, A2, A3,A4] with their corresponding label as [B1, B2, B3, B4] random forest can generate three decision tree taking a subset of input,for example
1.[A1, A2, A3]
2.[A1, A2, A4]
3.[A2, A3 A4]

Finally, it predicts based on the majority of votes from each decision made by the decision trees. The random-forest algorithm brings extra randomness into the model while growing the trees. Instead of searching best feature while splitting the node, it searches for the best features among a random subset of features.

SUPPORT VECTOR MACHINE

Support vector machine is a supervised classification algorithm.SVM is a discriminative classifier that separates defined by separating hyper-plane. More clearly, SVM takes training data and separates data into categories divided by a clear gap called the hyper-plane. SVM tries to find the best or optimal hyper plane, which has the most significant distance from the nearest point, in high dimensions, separating the training set into categories. Support vectors are those vectors that are nearest to the hyper-plane. The goal is to select a hyper plane with a margin as much as possible between hyper plane and any vector within the training set, giving a greater chance of newdata being classified correctly.

RECURRENT NEURAL NETWORKS

Recurrent neural networks include input units, output units, and hidden units, and the hidden unit completes the most important work. The RNN model essentially has a one-way flow of information from the input units to the hidden units. The synthesis of the one-way information flow from the previous temporal concealment unit to the current timing hiding unit is shown in the figure below. Here can regard hidden units as the storage of the whole network, which remembers the end-to-end information. When we unfold the RNN, we can find that it embodies deep learning. A RNNs approach can be used for supervised classification learning. Recurrent neural networks have introduced a direction all oop that can memorize the previous information and apply it to the current output, which is the essential difference from traditional Feed-forward Neural Networks (FNNs). The preceding output is also related to the current output of a sequence, and the nodes between the hidden layers are no longer connectionless; instead, they have connections. The output of the input layer and the output of the last hidden layer act on the input of the hidden layer.

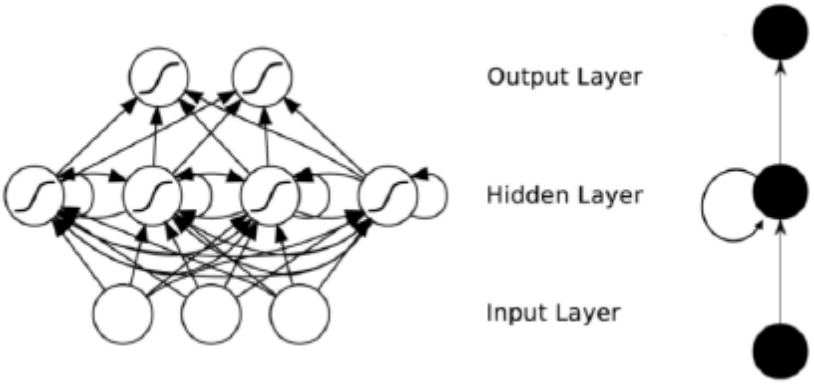


Figure 2. Recurrent Neural Networks (RNNs) [3].

GINI INDEX

Usually, the dataset for network intrusion detection contains many features. However, not every feature contributes to the task of detecting intrusion. Feature selection ,which can remove redundant or irrelevant features, isa crucial step for intrusion detection. Based on the optimal feature space here, we can enhance the speed of training a classifier for network intrusion detection and improve its detection performance. The goal of feature selection is to get a group of significant features from the whole dataset, such that these selected features are essential for training a classification model. Gini index to undertake the mission of feature selection. The Gini index, which Corrado Gini, an Italian statistician, developed, and sociologist, in 1912, was used initially to measure the statistical dispersion of income distribution across different population sectors [5].

DEEP LEARNING

Deep Learning algorithms are a modern update to artificial neural networks that exploit abundant, affordable computation. Deep learning permits an algorithm to learn a representation of data with various levels of generalization. These methods have been applied to visual object recognition, object detection, detecting network intrusion, and many other domains. A deep learning algorithm can be trained in a supervised and unsupervised way. Deep Learning algorithm in a supervised way: Convolutional neural network (CNN) usually is trained in a supervised way. CNN is now the benchmark model for computer vision purposes. The CNN architecture is used to structure 2D images, and the most crucial acknowledgment of CNN is face recognition and intrusion detection [13].

C4.5 DECISION TREE MODEL

A decision-making tree is a decision advocate system represented as a tree graph. Decision tree learning takes it as a predictive model to observe an item (represented in the branches)to conclude the item's target value (represented in the leaves).Target variable helps to classify the tree models; tree structure contains leaves and branches representing class labels and junctions. A decision tree specifically performs indecision analysis, decisions, and decision-making. To generate a decision tree, an algorithm is being used in C4.5 evolved, and It also alluded asa statistical classifier.C4.5 construct decision trees by getting data from training set as
in ID3, using the approach of information entropy [8].

PROBLEM IDENTIFICATION

Among these mechanisms, intrusion detection systems (IDSs) play a vital role in protecting computing infrastructures from attackers and intruders. The balance between detection and false-positive rates becomes more challenging when regular activity and anomalous activity are not static. The activity on the network can change, and the IDS must be aware of this change and adapt accordingly. If not, the ability of the IDS to provide accurate and reliable results is greatly diminished. Therefore, an IDS must adapt to different environments, potentially bringing different activities and behavior unseen by the IDS.

CONCLUSION

Network Intrusion Detection Systems (NIDS) have been developed rapidly in academia and industry in response to the increasing cyber-attacks against governments and commercial enterprises globally. The annual cost of cybercrime is continuously rising. The most devastating cyber crimes are caused by malicious insiders, denial of services, and web-based attacks. With the increasing variety and complexity of IDSs, the development of IDS evaluation methodologies, techniques, and tools has become a key research topic.

Few things are explored for system and systematize standard practices in the area of evaluation of such systems.

1. Researchers In this article, the author applied the different machine learning techniques for intrusion detection techniques like the support vector machine as a classification and random forest classification technique [1]. In future work, we may also use some other classification techniques and optimization techniques for optimal results. We also used feature selection techniques, reduced the feature value, and enhanced the existing system's performance.

2. The author presents the deep learning model [2] for the intrusion detection system using the KDD datasets, and these datasets are divided into two categories normal and abnormal. Abnormal categories include Dos, probe, U2R, and R2L; with the deep learning model author used here, the number of hidden layers is 8, 16, and 24; in the future, we may increase the number of hidden layers and also reduce the computation time.

3. Here[4],the authors present the regression techniques for the intrusion detection system and compute the value of performance parameters like accuracy, false-negative rate, actual positive rate, and detection rate. The balance between detection and false-positive rates becomes more challenging when regular activity and anomalous activity are not static. The activity on the network can change, and the intrusion detection system must be aware of this change and adapt accordingly. If not, the ability of the intrusion detection system to provide accurate and reliable results is greatly diminished. Therefore, an intrusion detection system must adapt to different environments, which potentially bring different activity and behavior unseen by the intrusion detection system; in future work, we may increase the performance of an existing system to reduce the false alarm rate.

In the future, we plan to implement an intrusion detection model based on machine learning algorithms and improve the existing system's performance.

REFERENCES

[1]Ripon Patgiri, Udit Varshney, Tanya Akutota, and Rakesh Kunde, "An Investigation on Intrusion Detection System Using Machine Learning," IEEE 2018, pp 1684-1691.

[2] Shone, N, Tran Nguyen, N, Vu Dinh, P and Shi, "A Deep Learning Approach to Network Intrusion Detection," IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, pp 1-11.

[3] Chuanlong Yin, Yuefei Zhu, Jinlong Fei, Xinzheng He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," IEEE 2017, pp 21954-21961.

[4] James Brown, Mohd Anwar, Gerry Dozier, "An Evolutionary General Regression Neural Network Classifier for Intrusion Detection," IEEE 2016, Pp 1-5.

[5] Longjing Li, Yang Yu, ShenshenBai, Jianjun Cheng, Xiaoyun Chen, "Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO,"Journal of Sensors, 2018, Pp 1-10.

[6] Yanqing Yang, Kangfeng Zheng, Chunhua Wu, XinxinNiu, Yixian Yang, "Building an Effective Intrusion Detection System Using the Modified Density Peak Clustering Algorithm Deep Belief Networks," Applied Science Journal, 2019, Pp 1-25.

[7] Malek Al-Zewairi, Sufyan Almajali, Arafat Awajan, "Experimental Evaluation of a Multi-Layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System," International Conference on New Trends in Computing Sciences, IEEE 2017, pp 167-173.

[8] M. Mazhar Rathore, Faisal Saeed, Abdul Rehman, Anand Paul, Alfred Daniel, "Intrusion Detection using Decision Tree Model in High-Speed Environment," International Conference on Soft-computing and Network Security, IEEE 2018, Pp 1-5.

[9] L. Khalvati, M. Keshtgary, N.Rikhtegar, "Intrusion Detection based on a Novel Hybrid Learning Approach,"Journal of AI and Data Mining, 2018, Pp 157-162.

[10] Ali Safaa Sadiq, BasemAlkazemi, SeyedaliMirjaliliNoraziah Ahmed, Suleman Khan, Ihsan Ali, Al-Sakib Khan Pathan, KayhanZrar Ghafoor, "An Efficient IDS Using Hybrid Magnetic Swarm Optimization in WANETs," IEEE Access 2018, Pp 29041-29052.

[11] Aleksandar Milenkoski, Marco Vieira, Samuel Kounev, Alberto Avritzer, Bryan D. Payne "Evaluating Computer Intrusion Detection Systems: A Survey of Common Practices," ACM Computing Surveys, Vol. 48, September 2015, pp 1-41.

[12] Rafah Samrin, D Vasumathi," Review on Anomaly-based Network Intrusion Detection System," ICEECCOT 2017, pp141-147.

[13] Nasrin Sultana,Naveen Chilamkurti,Wei Peng,RabeiAlhadad," Survey on SDN based network intrusion detection system using machine learning approaches," Springer Nature 2018, pp 1-9.

AUTHORS PROFILE

AUTHOR : **RAKHI SHUKLA**

Rakhi Shukla is a Ph.D. student of the university Rabindranath Tagore University. Her current research interests include Intrusion Detection Systems. Apart from that, She is serving as Asst. IT Officer at the Directorate of Technical Education, Government of Madhya Pradesh.

CO- AUTHOR: **Dr AARTI KUMAR**

Dr. Aarti Kumar, Professor, Computer Science, is a Doctorate in Computer Application from National Institute of Technology, Bhopal, whose research area is"Information Retrieval." She has excelled in her educational career by always being in the merit list, and a national awardee in Teaching and has been lucky to receive this award by the then President of India, His Excellency Dr. A. P. J. Abdul Kalam in Vigyan Bhavan, New Delhi. She has a complete teaching and administrative experience of 27 years. She has served in various capacities, from Head of Department to Director IODE. At present, she is managing MOOC and Online Content for five universities. She has been an active member of the IT cell in the university, the SPOC for Local Chapter, SWAYAM NPTEL, IIT, and Faculty Organizer, IIT Bombay Spoken Tutorial MOOC Training Program. She has also been the Associate Member of the Information Retrieval Society of India and a Professional Member of the Association for Computer Machinery (ACM). She is also the Editorial Board Member of the American Journal of Data Mining and Knowledge Discovery. Her areas of interest are Information Retrieval, Operating Systems, Data Mining, and Artificial Intelligence. She has written various research papers for reputed journals and has also participated as a resource person in a few conferences and workshops. She has been a Radio counselor of CIC and MCA courses of IGNOU, translated Computer Science courses of IIT Madras into Hindi. They have complimented her for her excellent translation. She is also inclined towards acting and singing and has the chance to act in the Bollywood movie "Aarakshan" and participate in Antakshari with Annu Kapoor and Pallavi Joshi.

Systematic Review of Deep Learning Techniques for Visual Feature Representation and Learning

[1]Rupali Tabakade , [2]Dr. Varsha Jotwani
[1]Research Scholar, [2]Associate Professor
[1,2,] Rabindranath Tagore University, Bhopal

Abstract :- Visual features representation along with deep learning techniques have a great area of research today as perspective of industries like facebook AI research. These industries aimed to focus on deep features learning with dataset and self-learning model. Most recent efforts in unsupervised feature learning already existed on either small or highly created datasets like ImageNet, whereas using non-curated raw datasets was found to decrease the feature quality when evaluated on a transfer task. From past 5-10 years lot of experimental research has been published and growing path of new ideas to improve accuracy in proposed system is not yet stopped. Paper has a systematic conclusion, the work already done in area related to visual feature representation, similarity computation methods and experimental results comparison. Our goal with paper is to bridge the performance gap with lot many existing techniques of deep leaning.

Index Terms :- Deep features, Deep leaning, Self Learning model.

1. Introduction:-

Computer vision has been revolutionized by high capacity Convolution Neural Networks (ConvNets) and large-scale labeled data. Recently weakly-supervised training on hundreds of millions of images and thousands of labels has achieved state-of-the-art results on various benchmarks. Interestingly, even at that scale, performance increases only log linearly with the amount of labeled data. Thus, sadly, what has worked for computer vision in the last five years has now become a bottleneck: the size, quality, and availability of supervised data [11].

Unsupervised representation learning is highly successful in natural language processing, e.g., as shown by GPT and BERT [2]. But supervised pre-training is still dominant in computer vision, where unsupervised methods generally lag behind. The reason may stem from differences in their respective signal spaces. Language tasks have discrete signal spaces (words, sub-word units, etc.) for building tokenized dictionaries, on which unsupervised learning can be based. Computer vision, in contrast, further concerns dictionary building, as the raw signal is in a continuous, high-dimensional space and is not structured for human communication (e.g., unlike words). Several recent studies present

promising results on unsupervised visual representation learning using approaches related to the contrastive loss. Though driven by various motivations, these methods can be thought of as building dynamic dictionaries. The "keys" (tokens) in the dictionary are sampled from data (e.g., images or patches) and are represented by an encoder network. Unsupervised learning trains encoders to perform dictionary look-up: an encoded "query" should be similar to its matching key and dissimilar to others. Learning is formulated as minimizing a contrastive loss.

Unsupervised learning has been widely studied in the Machine Learning community [9], and algorithms for clustering, dimensionality reduction or density estimation are regularly used in computer vision applications. For example, the \bag of features" model uses clustering on handcrafted local descriptors to produce good image-level features [4]. A key reason for their success is that they can be applied on any specific domain or dataset, like satellite or medical images, or on images captured with a new modality, like depth, where annotations are not always available in quantity. Several works have shown that it was possible to adapt unsupervised methods based on density estimation or dimensionality reduction to deep models.

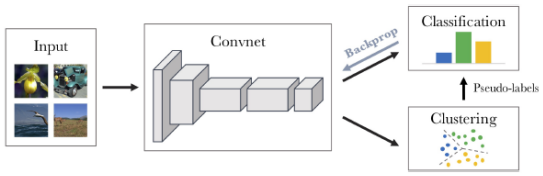


Figure 1: Illustration of the Convent method with clustering and classification.

1.1 Deep Neural Network:

Deep neural networks excel at perceptual tasks when labeled data are abundant, yet their performance degrades substantially when provided with limited supervision (In below fig, red). In contrast, humans and animals can learn about new classes of images from a small number of examples. What accounts for this monumental difference in data-efficiency between biological and machine vision? While highly structured representations may improve data-efficiency, it remains unclear how to program explicit structures that capture the enormous complexity of real world visual scenes, such as those present in the ImageNet dataset. An alternative hypothesis has therefore proposed that intelligent systems need not be structured a priori, but can instead learn about the structure of the world in an unsupervised manner. Choosing an appropriate training objective is an open problem, but a potential guiding principle is that useful representations should make the variability in natural signals more predictable. Indeed, human perceptual representations have been shown to linearize (or 'straighten')the temporal transformations found in natural videos, a property lacking from current supervised image recognition models, and theories of both spatial and temporal predictability have succeeded in describing properties of early visual areas.

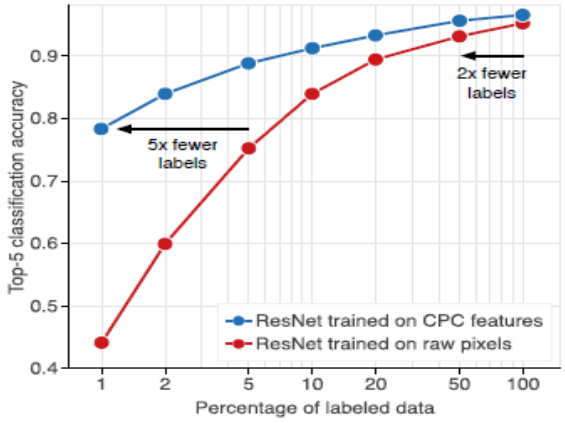


Figure 2: Data-efficient image recognition with Contrastive Predictive Coding. With decreasing amounts of labeled data, supervised networks trained on pixels fail to generalize (red). When trained on unsupervised representations learned

with CPC, these networks retain a much higher accuracy in this low-data regime (blue) [3].

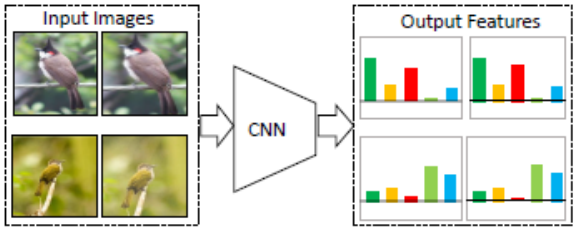


Figure 3: Illustration of our basic idea. The features of the same instance under different data augmentations should be invariant, while features of different image instances should be separated [7].

1.2 Convolutional Neural Network:

Pre-trained convolutional neural networks, or Convnets, have become the building blocks in most computer vision applications. They produce excellent general-purpose features that can be used to improve the generalization of models learned on a limited amount of data. The existence of ImageNet [6], a large fully-supervised dataset, has been fueling advances in pre-training of convents However, As a matter of fact, ImageNet is relatively small by today's standards; it \only" contains a million images that cover the specific domain of object classification. A natural way to move forward is to build a bigger and more diverse dataset, potentially consisting of billions of images. This, in turn, would require a tremendous amount of manual annotations, despite the expert knowledge in crowd sourcing accumulated by the community over the years. Replacing labels by raw metadata leads to biases in the visual representations with unpredictable consequences [4]. Learning a deep neural network together while discovering the data labels can be viewed as simultaneous clustering and representation learning. The latter can be approached by combining cross-entropy minimization with an off-the-shelf clustering algorithm such as K-means. This is precisely the approach adopted by the recent DeepCluster method, which achieves excellent results in unsupervised representation learning. However, combining representation learning, which is a discriminative task, with clustering is not at all trivial. In particular, we show that the combination of cross-entropy minimization and K-means as adopted by DeepCluster cannot be described as the optimization of an overall learning objective; instead, there exist degenerate solutions that the algorithm avoids via particular implementation choices [9].

2. Discussion:

2.1 Detailed Analysis of different approaches applied:

Human observers can learn to recognize new categories of images from a handful of examples; yet doing so with artificial ones remains an open challenge. The efficient recognition of data is enabled by representations which make the variability in natural signals more predictable. Therefore revisit and improve Contrastive Predictive Coding is a better solution, as unsupervised objective for learning such representations. The given table-1 express an analysis of similarity measured in previous year for features of leafs. Some new implementation produces features which support state-of-the art linear classification accuracy on the ImageNet dataset. When used as input for non-linear classification with deep neural networks, this representation allows us to use 2–5x fewer labels than classifiers trained directly on image pixels.

Table 1: Analysis of Feature Extraction Techniques based on similarities (leaf data) [17]

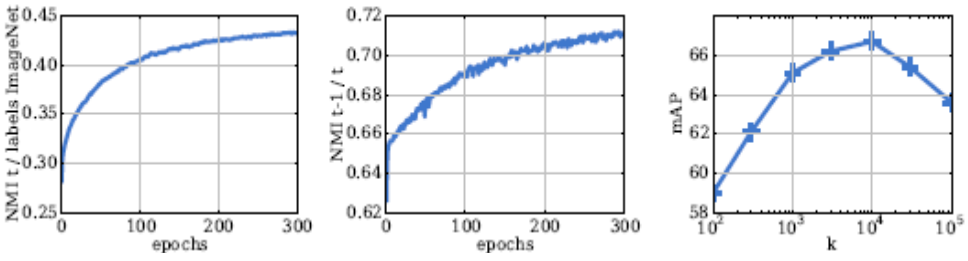
SI. No.	Authors & Year	Methodology/ Approach	Description
1	De Chant S. et.al. 2017, [12]	Deep CNN model have been used for extraction of local and global	They applied the method on maize leaves and used for prediction of

		features of the input image.	disease Northern leaf blight (applied for binary classification only).
2.	Yang Lu et.al. , 2018, [13]	6 layer CNN network proposed for feature extraction with the use of 3 convolution layers , 1 for extraction of low level features other two for extraction of high level features.	16 features are extracted by using 3 convolution and 3 max pooling filters and applied on rice plant diseases of 10 classes. Classification accuracy of 95.48% achieved.
3	Nikos Petrellis, 2019, [14]	Color, area and the number of the lesion spots featured have extracted. Then feature have been put	Novel method of automated and manual feature extortion has been proposed,
		along with additional information like weather metadata, to create disease signature.	where, user's input can also considered as a feature. They do this task using smart phones.
4.	Manso L. el.al. , 2019 [15]	Mathematical equations have been used to extract texture features (like contrast, entropy, homogeneity) and color features (like mean, variance etc.)	15 features extracted, which includes texture and color features only but not considered the affected area of leaf.
5.	Saradhambal. G. et.al. , 2019, [16]	k means clustering for segmentation have been used for segmentation, and shape and texture features are considered as a main	Total 10 features extracted. 5 of which are shape features (like area, perimeter, no of component etc.) and 5 are texture

		features which are calculated by mathematical equations.	features (like contrast, entropy, co relation etc.)

The unsupervised representation substantially improves transfer learning to object detection on the PASCAL VOC dataset, surpassing fully supervised pre-trained ImageNet classifiers. A main purpose of unsupervised learning is to pre-train representations (i.e., features) that can be transferred to downstream tasks by fine-tuning. Authors present Momentum Contrast (MoCo) for unsupervised visual representation learning, a perspective on contrastive learning as dictionary look-up, they build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linear protocol on ImageNet classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. MoCo can outperform its supervised pre-training counterpart in 7 detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, sometimes surpassing it by large margins.

Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. In this work, author [4] presents DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k- means, and uses the subsequent assignments as supervision to update the weights of the network. They apply DeepCluster to the unsupervised training of convolution neural networks on large datasets like ImageNet and YFCC100M. The primary study and experiment on visual corps image features provides given effect as in figure 4 with respect of clustering goodness [4]. Effect of the experiment is to train DeepCluster on ImageNet [6] unless mentioned otherwise. It contains 1:3M images uniformly distributed into 1; 000 classes.



(a) Clustering Quality (b) Clustering Reassignment (c) Influence of k

Figure 4: (a): evolution of the clustering quality along training epochs, (b): evolution of cluster reassignments at each clustering step;(c): validation mAP classification performance for various choices of k.

2.2Findings and Comparisons Summary for different network & dataset:

As a clustering the unsupervised learning and deep clustering provides a good performance and continuously engaged by many research. It showed the trust of deep clustering algorithm by many researchers. Table 2: shows the work carried by many

researches with improves affection on using more convolution layers for ImageNet. DeepCluster outperforms the state of the art from conv3 to conv5 layers by 3 to 5%. The largest improvement is observed in theconv4 layer, while the conv1 layer performs poorly, probably because the Sobel filtering discards color. Linear classification on ImageNet and Places using activations from the convolutional layers of an AlexNet as features. We report classification accuracy on the central crop. Numbers for other methods are from Zhang.

Table 2: Analysis by different authors on linear classification on ImageNet and Place

	ImageNet					Place				
Method	Conv1	Conv2	Conv3	Conv4	Conv5	Conv1	Conv2	Conv3	Conv4	Conv5
Place Lables	Na	Na	Na	Na	Na	22.1	35.1	40.2	43.3	44.6
ImageNetLables	19.3	36.3	44.2	48.3	50.5	22.7	34.8	38.4	39.4	38.7
Random	11.6	17.1	16.9	16.3	14.1	15.7	20.3	19.8	19.1	17.5
Pathak[17]	14.1	20.7	21	19.8	15.5	18.2	23.2	23.4	21.9	18.4
Doersch[18]	16.2	23.3	30.2	31.7	29.6	19.7	26.7	31.9	32.7	30.9
Zhang[19]	12.5	24.5	30.4	31.5	30.3	16	25.7	29.6	30.3	29.7
Donahue[20]	17.7	24.5	31	29.9	28	21.4	26.2	27.1	26.1	24
Noroozi and Favaro[21]	18.2	28.8	34	33.9	27.1	23	32.1	37.5	34.8	31.3
Noroozi[22]	18	30.6	34.3	32.5	25.7	23.3	33.9	36.3	34.7	29.6
Zhang[23]	17.7	29.3	35.4	35.2	32.8	21.3	30.7	34	34.1	32.5
Deep Cluster	12.9	29.2	38.2	39.8	36.1	18.6	30.8	37	37.5	33.1

Consistently the conclusive Comparisons on ImageNet linear classification with different techniques have been judged on 100- 400 epoc. All are based on ResNet-50 pre-trained with two 224 x 224 views. The transfer leaning is a new area where we find almost similar kind results statistics. Transfer Learning. All unsupervised methods are based on 200-epoch pre-training in ImageNet. VOC 07 detection: Faster R-CNN fine-tuned in VOC 2007 trainval, evaluated in VOC 2007 test; VOC 07+12 detection: Faster R-CNN fine-tuned in VOC 2007 trainval + 2012 train, evaluated in VOC 2007 test; COCO detection and COCO instance segmentation: Mask R-CNN [18] fine-tuned in COCO 2017 train, evaluated in COCO 2017 value has been shown in table 3:

Table 3: Comparisons on ImageNet linear classification on some latest networks

Method	Batch Size	Negative Pairs	Momentum Encoder	100 ep	200 ep	400 ep	800 ep	VOC 07 detection			VOC 07+12 detection			COCO detection			COCO instance seg		
								AP50	AP	AP75	AP50	AP	AP75	AP50	AP	AP75	AP50	AP	AP75
SimCLR (repro.+)	**4096**	**Yes**	No	66.5	68.3	69.8	70.4	**75.9**	**46.8**	**50.1**	**81.8**	**55.5**	**61.4**	**57.7**	**37.9**	**40.9**	**54.6**	**33.3**	**35.3**
MoCo v2 (repro.+)	**256**	**Yes**	**Yes**	67.4	69.9	71	72.2	**77.1**	**48.5**	**52.5**	**82.3**	**57**	**63.3**	**58.8**	**39.2**	**42.5**	**55.5**	**34.3**	**36.6**
BYOL (repro.)	**4096**	No	**Yes**	66.5	70.6	73.2	74.3	**77.1**	**47**	**49.9**	**81.4**	**55.3**	**61.1**	**57.8**	**37.9**	**40.9**	**54.3**	**33.2**	**35**
SwAV (repro.+)	**4096**	No	No	66.5	69.1	70.7	71.8	**75.5**	**46.5**	**49.6**	**81.5**	**55.4**	**61.4**	**57.6**	**37.6**	**40.3**	**54.2**	**33.1**	**35.1**

Unsupervised image representations have significantly reduced the gap with supervised pre-training, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pair wise feature comparisons, which is computationally challenging. The online algorithm, SwAV, takes advantage of contrastive methods without requiring computing pair wise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or "views") of the same image, instead of comparing features directly as in contrastive learning. Simply put, they use a "swapped" prediction mechanism where they predict the code of a view from the representation of another view. The method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network.

One core objective of deep learning is to discover useful representations, and the simple idea explored here is to train a representation-learning function, i.e. an encoder, to maximize the mutual information (MI) between its inputs and outputs. This work investigates unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder. Importantly, [6] they show that structure matters: incorporating knowledge about locality in the input into the objective can significantly improve a representation's suitability for downstream tasks. They further control characteristics of the representation by matching to a prior distribution adversarial. Their method, which they call Deep InfoMax (DIM), outperforms a number of popular unsupervised learning methods and compares favorably with fully-supervised learning on several classification tasks in with some standard architecture. Siamese networks are general models for comparing entities. Their applications include signature and face verification, tracking, one-shot learning, and others. In conventional use cases, the inputs to Siamese networks are from different images, and the comparability is determined by supervision. Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions [1]. In this paper, they report surprising empirical results that simple Siamese networks can learn meaningful representations even using none of the following: (i) negative sample pairs, (ii) large batches, (iii) momentum encoders. Their experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. They also provide a hypothesis on the implication of stop-

gradient, and further show proof-of-concept experiments verifying it.

Conclusion:

Learning visual representations with self-supervised learning has become popular in computer vision. The auxiliary tasks models where labels are free to obtain have most tasks end up providing data to learn specific kinds of invariance useful for recognition. In many articles the exploitation of different self-supervised approaches to learn representations invariant to (i) inter-instance variations (two objects in the same class should have similar features) and (ii) intra-instance variations (viewpoint, pose, deformations, illumination, etc.). Instead of combining two approaches with multi-task learning, they argue to organize and reason the data with multiple variations. Specifically, they propose to generate a graph with millions of objects mined from hundreds of thousands of videos.

Competiveness of minimalist method suggests shape is an important core reason for effectiveness. Representation learning focuses on modeling invariance by different network. Lot many survey and statistics represents that unsupervised learning in variety of computer vision task give and shows better results. MoCo's improvements are considerable and noticeable for small dataset and suggest that it may not be used for large scale data. MoCo can be used with pretext task for constructive learning.

References:

[1] Xinlei Chen, Kaiming He, "Exploring Simple Siamese Representation Learning", IEEE 2020, pp. 1-10.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, SainingXie,Ross Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning", IEEE 2020, pp. 1-12.

[3] Olivier J. Henaff, AravindSrinivas, Jeffrey De Fauw, Ali Razavi,Carl Doersch, S. M. Ali Eslami, Aaron van den Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding", 2020, pp. 1-13.

[4] Mathilde Caron, PiotrBojanowski, Armand Joulin, MatthijsDouze, "Deep Clustering for Unsupervised Learningof Visual Features", 209, pp. 1-30.

[5] Mathilde Caron, Ishan Misra, JulienMairal, Priya Goyal, Piotr Bojanowski, Armand Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", 34th Conference on Neural Information Processing Systems, 2020, pp. 1-23.

[6] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio, "Learning Deep Representations By Mutual Information Estimation And Maximization", Published as a conference paper at ICLR 2019, pp. 1-24.

[7] Mang Ye, Xu Zhang, Pong C. Yuen, Shih-Fu Chang, "Unsupervised Embedding Learning via Invariant and Spreading Instance Feature", 2019, pp. 1-11.

[8] Mathilde Caron, PiotrBojanowski, JulienMairal, Armand Joulin, "Unsupervised Pre-Training of Image Features on Non-Curated Data", 2019, pp. 1-14.

[9] Yuki M. Asano, Christian Rupprecht, Andrea Vedaldi, "Self-LabellingVia Simultaneous Clustering And Representation Learning", Published as a conference paper at ICLR 2020, pp. 1-22.

[10] Xiaolong Wang, Kaiming He, Abhinav Gupta, "Transitive Invariance for Self-supervised Visual Representation Learning", 2018, pp. 1329-1338.

[11] PriyaGoyal, DhruvMahajan, Abhinav Gupta, IshanMisra, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", IEEE 2018, pp. 6391-6400.

[12] DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E.L., Yosinski, J., Gore, M.A., Nelson, R.J., Lipson, H., "Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning". Phytopathology (2017).https://doi.org/10.1094/PHYTO-11-16-0417-R.

[13] Lu, Yang, et al. "Identification of rice diseases using deep convolutional neural networks." Neuro computing 267 (2017): 378-384.

[14] Petrellis, Nikos. "Plant disease diagnosis for smart phone applications with extensible set of diseases." Applied Sciences 9.9 (2019): 1952.

[15] Giuliano L. Mansoa, HelderKnidel, Renato A. Krohlinga, José A. Ventura, "A smart phone application to detection and classification of coffee leaf miner and coffee leaf rust", Preprint submitted to Journal of LATEX Templates, arXiv:1904.00742v1 [cs.CV] 19 Mar 2019.

[16] Saradhambal.G, Dhivya.R, Latha.S, R.Rajesh," Plant Disease Detection And Its Solution Using Image Classification", International Journal of Pure and Applied Mathematics, Volume 119 No. 14 2018, 879-884, pp. 879-884.

[17] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context en-coders: Feature learning by in painting. In: CVPR. (2016)

[18] Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. (2015)

[19] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. (2016)

[20] Donahue, J., Kr¨ahenb¨uhl, P., Darrell, T.: Adversarial feature learning. arXivpreprint arXiv:1605.09782 (2016)

[21] Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solvingjigsaw puzzles. In: ECCV. (2016)

[22] Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning tocount. In: ICCV. (2017).

[23] Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. arXiv preprint arXiv:1611.09842 (2016)

Author's Profile :

1. Rupali Tabakade
is Research Scholar in Rabindranath Tagore University, Bhopal. I have also done my M.Phil.in CS from Rabindranath Tagore University, Bhopal. I have published my paper in International Conference on Intelligent Computing and Information System (ICICIS-2012) with paper id 452 on Data Mining .This conference was held in Pachmarhi, Pipriya (M.P.) during October 27-28, 2012.

2 .Dr. Varsha Jotwani
is currently working as Associate Professor with Rabindranath Tagore University . She is PhD in Computer Applications. She has vast teaching and academic developments at leading institution of Bhopal, India. She has published various International and National Research papers in the high quality journals. She is also well versed in developing curriculum for UG and PG Students under the field of Information Technology.

Language Translation by Stand-Alone Voice Cloning: A Multispeaker Text-To-Speech Synthesis Approach based on Transfer Learning

[1]Sakshi Bhajikhaye , [2]Dr Sonali Ridhorkar , [3]Vidhi Gautam , [4]Mamta Soni , [5]Mayank Badole , [6]Adarsh Kant , [7]Pranjali Rewatkar

[1,3,4,5,6,7] HOD, Department of Computer Science, G.H. Raisoni Academy of Engineering and Technology, RTMNU, Maharashtra, India 440033.

[2] Student, Department of Computer Science, G.H. Raisoni Academy of Engineering and Technology RTMNU, Maharashtra, India 440033.
sakshi.bhajikhaye.cs@ghraet.raisoni.net,vidhi.gautam.cs@ghraet.raisoni.net.

Abstract— Stand Alone Language Translator is a speech to speech translation application for android mobile phone, which enables the translation of speech signals in a source language to the target language in the human voice, which is the same as the source voice. Stand Alone Language Translator includes three modules, Speech Recognition, Language Translation, and Speech Synthesis. The speech recognition module captures the voice or speech from the mobile user through Microphone, identifies then converts speech into text, and then the text is sent to Language Translation along with a sample voice for further process. Language Translation module does the process of translation, i.e., this module consists of a library for both languages, and when text is received by this module, it converts the text of one language to another selected by a user, and thus it sends the translated text to the last module. The speech Synthesis module acts as the text-to-speech translator, i.e., when it gets the translated text. This module processes translated text which converts it into speech and will provide the output to the user in the same voice as the source human voice. Thus, this language Translation application works by integrating all these three modules and gives the user the best output.

Keywords— End-to-end speech-to-speech translation, Speech Recognition, Language Translation, Speech Synthesizer, Multilingual speech

I. INTRODUCTION

Owing to this era, the global scenario adds to the demand for communication among speakers of different languages. Stand Alone Language Translator (SALT) enables the communication between people speaking in different languages. Stand Alone Language Translator being able to speak and have one's words translated automatically into the other person's language. This translator is used to convey the original tone and intent of a source language to the target language.

Automatic speech to speech translation technology consists of three separate technologies: technology to recognize speech (speech recognition), technology to translate the recognized words (language translation), and technology to synthesize speech in the other person's language (speech synthesis). Speech to Speech Translation systems is often used in a specific situation which includes supporting conversations in non-native languages. The demand for trans-lingual conversations triggered by IT technologies has boosted research activities on Speech to Speech Translation technology. The work proposed for Speech to Speech Translation is a mobile application for an android platform that translates the real-time speech of one language into another required targeting language. A good speech-to-speech translation system can be characterized by its ability to keep intact the fluency and meaning of the original speech input.

II. REVIEW

CASMACAT is a modular, web-based translation workbench that offers advanced functionalities for computer-aided translation and the scientific study of human translation. MateCat is a tool whose objective is to improve the integration of machine translation and human translation within the so-called computer-aided translation framework. It provides translators with text editors that can manage several document formats and suitably arrange their content into text segments ready to be translated [3]. Curriculum learning (CL) might help avoid bad local minimums, hasten training convergence, and improve generalization. These advantages

have been empirically demonstrated in various tasks, including shape recognition, object classification, and language modeling [1]. The advantage of SMT is that one does not require a deeper syntactic understanding of Source and Target languages [8]. Voice Translator is an android mobile application that helps the user to translate one language to another by using a Bluetooth environment which makes it possible to talk with every human being indifferent language [7]. The main goal of stand-alone voice conversion is to modify an utterance from the source speaker while keeping the linguistic contents unchanged in order to match the frequency of the target speaker [4].

III. PROBLEM DEFINITION

Let us consider a dataset of utterances grouped by their speaker where we denote the jth utterance of the ith speaker as uij. Utterances are indicated in the waveform domain. Then, we denote by xij -the log-mel spectrogram of the utterance uij. A log-mel spectrogram is a deterministic, non-invertible function that extracts the features of speech from waveform in order to handle speech in machine learning.

The encoder E computes the embedding in the following equation given below:

eij = E (xij; wE)

that corresponding to the utterance uij,

where wE are the parameters of the encoder.

Additionally, the authors define a speaker embedding as the centroid of the embeddings of the speaker's utterances:

-------(1)

The synthesizer is indicated by S, parametrized by wS, approximate to xij

Given,

ci and tij are the transcripts of utterance uij.

We have ,

xˆij = S (ci, tij; wS).

In our implementation, we use an utterance embedding rather than the speaker embedding, giving

xˆij = S (uij, tij; wS).

In the end, the vocoder V, parametrized by wV, is tasked to approximate uij given xˆij.

We have, uˆij = V (xˆij; wV).

One could train this framework in an end-to-end fashion with the following objective function:

Min wE, wS, wV LV(uij , V(S(E(xij ; wE), tij ; wS); wV))

Where LV is a loss function in the waveform domain.

IV. METHODOLOGY

SALT is mainly divided into three main modules- Speech recognition module (speech-to-text conversion), language translation module, Speech synthesis module (text-to-speech conversion) as shown in the following Figure [a]:

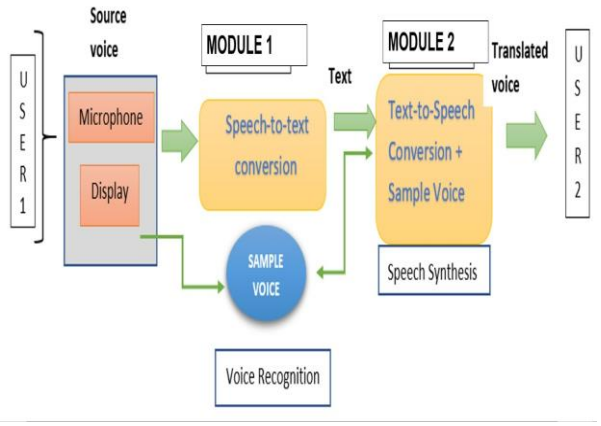


Fig [a]. Block Diagram of SALT framework

A. Speech-to-text conversion

The first module of our work is based on the Speech-to-text conversion, which is the Voice Recognition module. This also includes Speech Translation. Speech Translation is a function that instantly translates a spoken language into the selected foreign language. Here, we are using Google Translation API using a python script to record and convert the recognized Speech into Text format. A Speech-to-Text API is used to convert voice from the Microphone and generate English TEXT format from multiple languages.

This works on the Cloud platform, where it converts audio to text by applying powerful neural network models. In Speech-to-Text conversion, Google translates API's enables easy integration of Google Speech recognition technologies into developer applications. Cloud Translation can translate text between thousands of language pairs dynamically. Translation lets websites and programs integrate programmatically with the translation service. It also provides Natural Language Processing involves various technologies, such as sentiment analysis, entity recognition, entity sentiment analysis, and other text annotations, to developers. The speech-to-text conversion uses machine learning to reveal the structure and meaning of the text. The powerful, trained models of the Natural Language API empower developers to easily apply natural language understanding to their applications. It combines Natural Language with Speech-to-

Text API to extract insights from audio conversations. During the initial stage of Cloud Translation, you need a project that has the Cloud Translation API enabled and credentials to make authenticated calls.

Google Neural Machine Translation (GNMT) translates a whole sentence at a time, rather than just letter by letter. It uses broad context to help it figure out the most relevant translation, which is then rearranged and adjusts with proper grammar. It uses deep learning techniques, in particular, long short-term memory networks, in order to translate a whole sentence at a time, which has been measured to be more accurate between English and French, German, Spanish, and Chinese. GNMT improves the quality of translations because it uses an example-based machine translation (EBMT) method. In this method, the system learns from millions of examples. It uses this broad context to help it figure out the most relevant translation. The GNMT network attempts interlingual machine translation. Most common words in English have at least two senses, which produces equal odds in the likely case that the target language uses different words for those different senses. The odds are similar in other languages to English.

B. Text to speech conversion

The second module of our work is based on Text-to-speech (TTS), which is the Speech Synthesis module, the process of synthesizing an artificial speech from a text prompt. Most of the research focus has been since gathered around making these deep models more efficient, sound more natural, or training them hastily. The current Text To Speech tools and functionalities provided by Google, such as the Google assistant6 or the Google cloud services, make use of these same models. The complete framework is a three-stage pipeline.

The framework is divided into three stages are as follows:

• A speaker encoder is the stage that derives an embedding from the short utterance of a single speaker. The embedding represents a meaningful representation of the voice of the speaker, such that similar voices are closed in latent space.

• A synthesizer that controls the embedding of a speaker and generates a spectrogram from the text.

• A vocoder is the stage that infers an audio waveform from the spectrograms generated by the synthesizer. It generates an embedding which is used to control the synthesizer, and a text is processed as a phoneme sequence is given as input to the synthesizer. The vocoder then takes the output of the synthesizer to generate the speech waveform. The following Figure [b] is illustrated below:

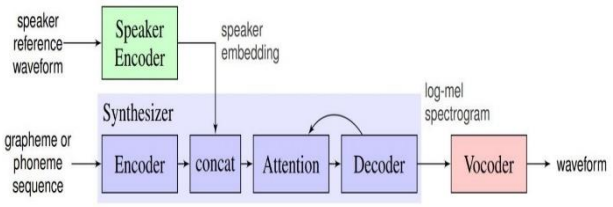


Fig b. Block diagram of TTS

The blue blocks in the above Figure represent a high-level view of the Tacotron architecture modified to allow conditioning on a voice. The above Figure is extracted from (Jia et al., 2018). This model can be trained separately and on distinct datasets. For the encoder, one looks for a model that is robust to noise and able to capture many characteristics of the human voice. A large corpus of many speakers would be preferred to train the encoder for the datasets of the synthesizer and the vocoder. The Transcripts are required, and the quality of the generated audio can be equally good as that of the data. Higher quality and annotated datasets are required. This shows that they are smaller in size.

Speaker Encoder: The speaker encoder needs to generalize well enough to produce meaningful embeddings on the dataset of the synthesizer, and even when trained on a common dataset, it still has to be able to operate in a zero-shot setting at inference time. We reproduced this model with a PyTorch implementation of our own. We synthesize the parts that are pertinent to SV2TTS as well as our choices of implementation.

A template is created for a person by deriving their speaker embedding from a few utterances. This process is called enrolment. The Generalised End to End loss simulates this process to optimize the model. To avoid segments that are mostly silent when sampling partial utterances from complete utterances, we use the webrtcvad8 python package to perform Voice Activity. Detection (VAD).

A last pre-processing step applied to the audio waveforms is normalization to make up for the varying volume of the speakers in the dataset. Figure [d] shows the steps to silence removal with VAD, from top to bottom. The orange line is the binary voice flag, where the upper value means that the segment is voiced and unvoiced when lower. In all our tests, the UMAP projections perfectly separate utterances from the test set of each of the three datasets, with large inter-cluster distances and small intra-cluster variance. The following Figure [c]UMAP projections of utterance embeddings from randomly selected batches from the train set at different

iterations of our model. Utterances from the same speaker are represented by a dot of the same color.

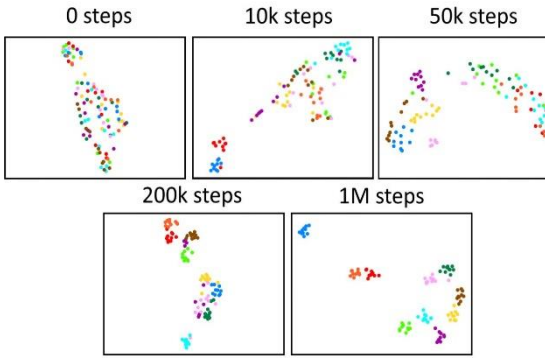


Figure [c]. UMAP projections of utterance embeddings from randomly selected batches from the train set at different iterations of our model.

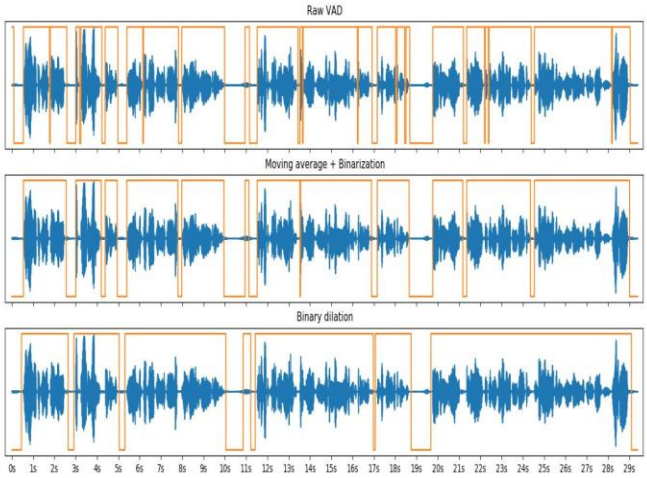
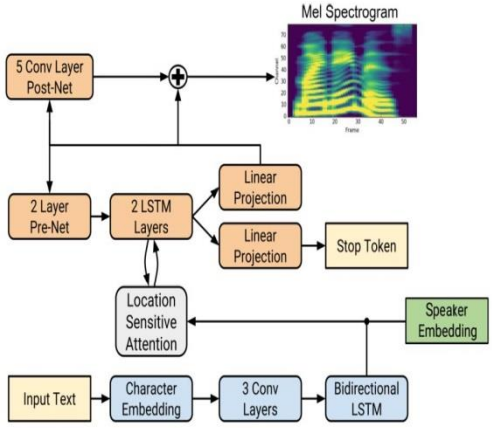


Figure [d]. Silence removal with VAD, from top to bottom

1) *Synthesizer:* The synthesizer is Tacotron without Wavenet. Tacotron is a recurrent sequence-to-sequence model. It predicts a mel spectrogram from the text. Tacotron features an encoder-decoder structure (not to be mistaken with the speaker encoder of TTS) that is bridged by a location-sensitive attention mechanism.

The entire sequence of frames goes through a residual post-net before it becomes the mel spectrogram. This architecture is represented in Figure [e]. We have used the automatic speech recognition (ASR) model to force-align the LibriSpeech transcripts to text. Tacotron usually operates faster than in real-time. We use a python implementation. We adapt this implementation to profile the noise and to clean the speech.



Figure[e]. Architecture of Synthesizer

2) *Vocoder:* The vocoder model which we use is an open-source PyTorch implementation that is based on WaveRNN.

Figure [f], which is given below, depicts the WaveRNN architecture. At each training step, a mel spectrogram and its corresponding waveform are divided into an equal number of segments. The spectrogram segment and the waveform segment t − 1 are given as input to the model. The mel spectrogram goes through an up-sampling network to match the length of the target waveform. As a result, the number of mel channels remains unchanged. A Resnet-like model uses the spectrogram as input to generate features that will control the layers throughout the transformation of the mel spectrogram to a waveform. The resulting vector is repeated to match the length of the waveform segment. This conditioning vector is then split equally four ways along the channel dimension, and the first part is concatenated with the upsampled spectrogram and with the waveform segment of the previous timestep. As a result, the vector goes through a diverse transformation with skip connections which include the first two GRU layers

and a dense layer. The conditioning vector is concatenated between each step with the intermediate waveform. In the end, two dense layers produce a distribution over discrete values that correspond to a 9-bit encoding of mu-law commanded audio.

.

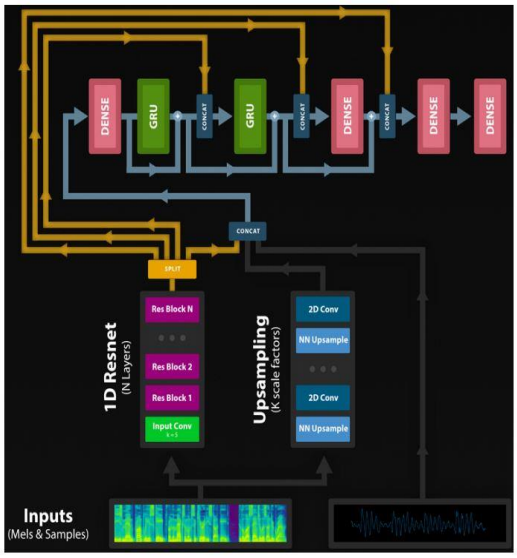
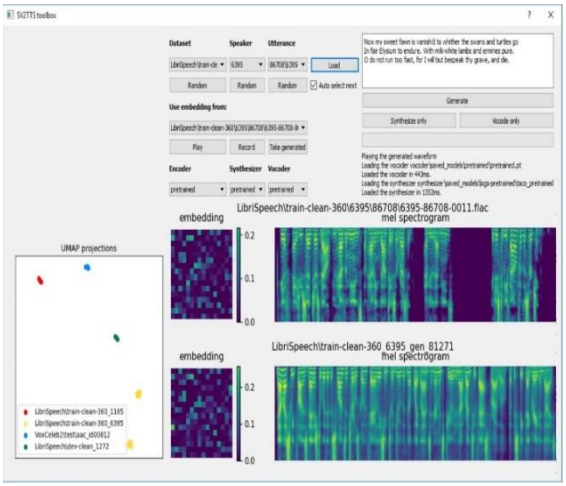


Figure [f]. WaveRNN architecture

3) *Text to speech tool box interface:* The interface of the toolbox can be seen in Figure [g]. The WaveRNN architecture is written in Python with the Qt4 graphical interface and is, therefore, a cross-platform. The image is best viewed on digital support. A user begins the process by selecting an utterance audio file from any of the datasets available on their disk. The toolbox handles many popular speech datasets and can be customized to add up-to-date ones. Further, the user can also record utterances so as to clone one's own voice. Once an utterance is loaded, embedding will be computed, and the Uniform Manifold Approximate Projections will be updated reflexively. The mel spectrogram of the utterance is pulled out, but it is only for reference. We draw an embedding vector with a heatmap plot. This embedding is unidimensional vectors. The embeddings are drawn give us visual cues as to how two embeddings differ from each other. When an embedding has been enumerated, it can be used to generate a spectrogram. The user can write any arbitrary text in order to be synthesized. Note that the punctuation is not supported by our model and will be discarded. The user has to insert line breaks between parts that should be synthesized individually to tune the prosody of the generate utterance. The complete spectrogram is the concatenation of those parts. After synthesizing, the spectrogram will be displayed on the bottom right of the interface. While synthesizing the same sentences multiple times will yield contrasting outputs.

In the end, the user can generate the segment corresponding to the synthesized spectrogram with the vocoder. The progress of the generation is displayed by a loading bar. After the progress is displayed, the embedding of the synthesized utterance is generated on the left of the synthesized spectrogram. This embedding will also be projected with Uniform Manifold Approximate Projections. The user is unbound to take that embedding as a reference for further generation. The following Figure [g] depicts the toolbox interface.



Figure[g]. TTS Toolbox Interface

C. User Interface

We are using Flutter technology to develop our user interface. Flutter is an open-source UI software development kit created by Google. The major components of Flutter include the Dart platform, Flutter engine, Foundation library, Design-specific widgets, Flutter Dev Tools.

In Python, we have used a microwave framework. In order to connect the Flutter with Python, we have installed pip and flask. We then worked on two inputs- flask and just sonify. We have created a flask instance and assigned it to the app variable, and run it in debug mode. We have installed the HTTP(dart) package, which would be used in the later stage. We have used the setup empty string variable assigned to display the data. We have also framed a button with a final response which, on clicking it, will get us back to our python script to fetch the data and display it on the screen. We had done this by sending an HTTP get request to get the request from the data server. In return, we will get an URL from the flask application. In the flask, we have framed the routes and route path and used methods. We have defined a function that is executed after we hit that specific route and then return to sonify and execute it. Once we get the data from the python script, we decode it back to key-value format and store it in response and use the set state. We receive the data in the Flutter sent by the python script.

The above modules can be summarized using a web API which is developed on the cloud computer to reduce the cost-effectiveness of the whole project.

V. CONCLUSIONS

A stand-alone language translator is a device that bridges the gaps of language barriers with the intelligent system in real-time by translating the source language to the target language. At present, we need a real-time system that can translate multilingual speech with the matched frequency of the human voice with proper fluency and accuracy.

This system fulfills all the requirements. This device works on three main modules-speech recognition, language translation, and speech synthesis.

This is an effective device which is easily accessible to any of the Bluetooth device anywhere without the use of internet with minimum cost and is very easy to use. Therefore, in this situation, the system proposed will suffice the purpose reasonably well and minimize the communication inefficiencies.

REFERENCES

1) Takatomo Kano, "End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs," IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28), Year:2020
2) Ye Jia, Yu Zhang, Ron J. Weiss," Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," Google Inc., 2019
3) Rashid Ahmad, Priyanka Gupta, Nagaraju Vuppala," Transzaar: Empowers Human Translators, "18th International Conference on Computational Science and Applications (2018)
4) M. Teduh Uliniansyah, Hammam Riza, Agung Santosa, Gunarso, Made Gunawan, Elvira Nurfadhilah," Development of Text and Speech corpus for an Indonesian Speech to Speech Translation System," Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique, South Korea (2017)
5) Mrinalini K and Vijayalakshmi P," Hindi-English Speech-to-Speech Translation System/or Travel Expressions," International conference on communication. Year:2015
6) M.D. Faizullaha Ansari, R.J. Shaji," Multilingual speech to speech translation system in Bluetooth environment," International Conference on Control, Communication and Computational technology (2014)
7) Akshay Suresh Deshpande, Keshav Sheshrao Ambulgekar, Kedar Raghunath Joshi," Voice to Voice Language Translation System," International Journal of Engineering Research & Technology (10, October-2014)
8) Karunesh Arora, Mukund Kumar Roy," Speech to Speech Translation: a communication boon," CSIT (7 June 2013)
9) Yuxuan Wang, RJ Skerry-Ryan. Tacotron: Towards End-to-End Speech Synthesis, Google, Inc-2017.
10) Leland McInnes and John Healy. UMAP: Uniform manifold approximation and projection for dimension reduction. 02 ,2018.

Pest Detection in Agricultural Plantation of Cotton Crops using Convolutional Neural Network

Sandhya Potadar, Aakanksha Khare, Shalaka Buche, Aksha Khairmode
Department of electronics and telecommunication Engineering
MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra

Abstract— India is a diverse country with 70% of its people depending upon agriculture. With the ever-increasing need for food and shelter, there is a drastic increase in the need for clothing as well. For this, the production of cotton crops also needs to be increased. Pests and crop diseases play a vital role in reducing the yield. Identification and classification of crop pests are essential to aid farmers. Manual detection of pests and the use of economical friendly pesticides may result in a reduction of the destruction caused to crops. The extra use of pesticides may reduce yield and will pollute the air and the soil of that region. As Maharashtra is the Manchester of cotton production, the proposed system focuses on pest detection on cotton crops. To address this issue, a convolutional neural network architecture is being applied to the mixture of images collected from farmer's fields and open-source platforms. The data augmentation techniques are applied such as shear, rotation, brightening, translation, and zooming to avoid the network from overfitting. The CNN architecture learns and extracts high-level complex features in image classification applications instead of extracting handcrafted features by traditional methods. The results obtained from the proposed CNN model classifies the visuality of pests on cotton crops with a model accuracy of 91.54% and can be applied in the agriculture field.

Keywords—Agriculture, Early pest detection, Image processing, CNN, Capsule Network.

I. INTRODUCTION

India is a diverse country with 70% of its people depending upon agriculture production. Production of crops in the agricultural sector is important for humans. With the ever-increasing need of food and shelter there is a drastic increase in need for clothing. For this, production of crops should also be increased. The pests and crop diseases play a vital role in reducing the yield. Agriculture deposits quite a handful in the income and annual budget of country. Countries trade raw material for industries which comes from the primary source and agriculture is one of them. A nation's export depends on the agricultural sector. The biological parameters that affect agricultural production are the presence of pests in plants. Pests are one of the leading causes of agricultural losses. Insects and pests are harmful for plants as they feed on them and affect photosynthesis, and sometimes may lead them to cause disease. There are many chemical and biological methods for pest control. But to achieve the maximum effectiveness of certain methods, careful monitoring of the entire property, which farmers can see with the naked eye, is generally recommended. They go about their daily activities. The problem with this is only relying of human observation sometimes leads to missing on some new pest or that may lead to damage the plant more if the pest or disease is unknown. The early detection of pests is an important task for successful agriculture. Detecting pest earlier requires a systematic approach to be followed, especially on large farms and plantations. In contemporary global its miles dealt with as alternate. So, loss of manufacturing in the end impacts alternate in addition to the human society. In this evolving age of technology, use of era is completed everywhere and everywhere feasible. Agriculture is one of the sectors in which use of technology can be performed to grow productivity in addition to assist the farmers in gaining higher yields. To simplify this cycle of dwelling and contributing inside the discipline of agriculture, a machine vision system is developed to enforce this venture.

II. CONVOLUTIONAL NEURAL NETWORK

The algorithm used in this paper is Convolutional Neural Network. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which uses set of rules, to extract features from images in each layer of ConvNet and given on to the next layer that is it performs convolution of 2-D image and a weight matrix. The process of convolving image and filter is continued till the last layer, till all the features are extracted from an image. The feature learning layers and classification layers of CNN is shown [1] in fig 1.

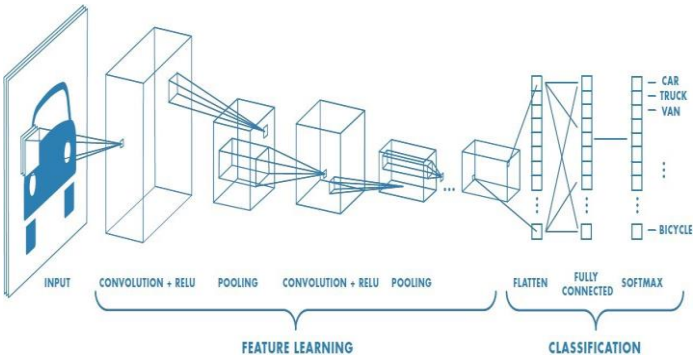


Fig 1: CNN Architecture

A. CONVOLUTIONAL LAYER

A kernel is passed over a 2-D RGB image, a dot product is performed between input image and a kernel in each layer of network to. The process is continued in each layer to extract complex features of the image by various filters. Padding of zeros inside the rows and columns is done to ensure the same length of image matrix.

B. ACTIVATION LAYER

The convolution layer generates a matrix, comparatively smaller in dimension than the original image. This matrix is administered through an activation layer, which introduces non-linearity. The activation function is usually ReLu (Rectified linear unit) to speed up the training process.

C. POOLING LAYER

"Pooling" is the technique of down sampling and reducing the matrix size. A filter is exceeded over the consequences of the preceding layer and selects one range out of each institution of values (generally the maximum, this is referred to as max pooling). This lets in the network to train a whole lot faster, focusing on the most critical records in every image feature. The max pooling computation will reduce the extension of the statistics.

D. FLATTEN LAYER

In between the convolutional layer and the completely related layers, there's a 'Flatten' layer. Flattening transforms, a 2-dimensional matrix of capabilities right into a vector that may be fed into a fully connected layer. Default fully related classifier is fully connected NxN with N neurons. Matrix has to have dimension at least equal to the number of neurons in the last layer of this network. Neurons in the last layer can connect to a lot of parameters because they already had learnt about the features in order to classify it.

E. FULLY CONNECTED LAYER

A fully connected layer is a neural layer that combines the results of the convolutional layers to generate a prediction. In network-based learning, the fully connected layer is an important part of neural networks because it allows us to apply back propagation. Back-propagation is a process of tuning the weights of multiple layers that are used to get input and predict output. By definition, fully connected means that all nodes in one layer connect to all nodes in the next layer.

III. LITERATURE REVIEW

In [1], the authors used an in-depth study method to detect diseases and pests accurately in the leaves and other parts of the crop. The proposed method consists of the following steps: - the first step in the acquisition of the image; The second step is to process the image. After the dataset is ready it is given the CNN model. In the next step, feature removal and classification are performed using the transfer learning method. CNN with transfer learning method is used here, which is very effective in processing large amounts of data and provides high accuracy. In [2], the authors focused on detecting insects using the Support Vector Machine. This paper uses image processing techniques to propose an automatic pest identification

system. Images of insects were collected and stored in a database. Depending on the features released the Support Vector Machine is trained. Here the color element is used to train the Support Vector Machine to distinguish pixels from leaves and pixels from insects. Morphological action is used to remove unwanted objects from a separated image. The purpose of this paper is to identify this bug and provide information on the number of bugs found. The separation process here is done with the help of SVM.

In [3], the authors summarize the progress that has been made so far in the use of digital images and machine learning to detect and detect insects, thus providing a complete picture of the subject in a single source. It provides an in-depth discussion of research opportunities and major ongoing weaknesses, emphasizing the technical aspects that hinder the actual adoption of the program. It suggests some possible indications for future research on this topic. In [4], they focused on detecting early insects. Pictures of diseased plants were obtained using scanners or cameras. The resulting image was later processed to interpret the contents of the images by means of image processing. The partition algorithm alone cannot provide a good quality output, it required a pre-processing step. Preparation in advance has a variety of steps such as extracting sound and enhancing the image. Due to the complexity and uncommonness in the images of insects, it is imperative to include pre-processing step before the process of quality separation and precise extraction. The proposed system relies on rapid detection of insects. This paper also introduced the integration of k-means as a method of segmentation.

In [5], the authors have proposed a technique for pest detection using the segmentation algorithm-means clustering in MATLAB. The prime focus was examination of small spots on plant leaves which are usually left undetermined by naked eye. Image acquisition and image preprocessing was carried out followed by K-means clustering which made the technique reliable in presence of extensive data. The proposed algorithm was successful in counting the pests on the leaves and displaying the infected area. In [6], an algorithm was proposed for automatic detection and estimation of Whitefly on cotton crops using image processing. The acquired leaf image was set through a set of procedures which included the steps color space conversion, background subtraction, thresholding followed by morphological operation. The technique proposed was successful in counting the number of whiteflies on the cotton leaf.

In [7], machine vision and image processing techniques were used in MATLAB for the detection of pests and disease on cotton crops. The color models RGB, HSI, YCbCr were implemented for extraction of damaged parts from the input cotton leaf image. The ratio of damage was chosen as the key aspect for determination of degree of damage caused by pests and diseases. Further the paper also discussed and compared the color models on the basis of the accuracy of the results obtained.

In [8], the authors explored the combination of different Image processing algorithms that can be used on a pre- processed image for plant disease detection. The paper discussed various segmentation techniques for partitioning of images along with the various feature extraction and classification techniques that can be implemented for extracting features from infected parts. Furthermore, the paper implored the usage of different ANN methods for classification of plant diseases.

IV. PROPOSED METHODOLOGY

The proposed workflow is as shown in Fig 2. First step is image acquisition followed by Image preprocessing, data augmentation and finally a CNN model for classification.

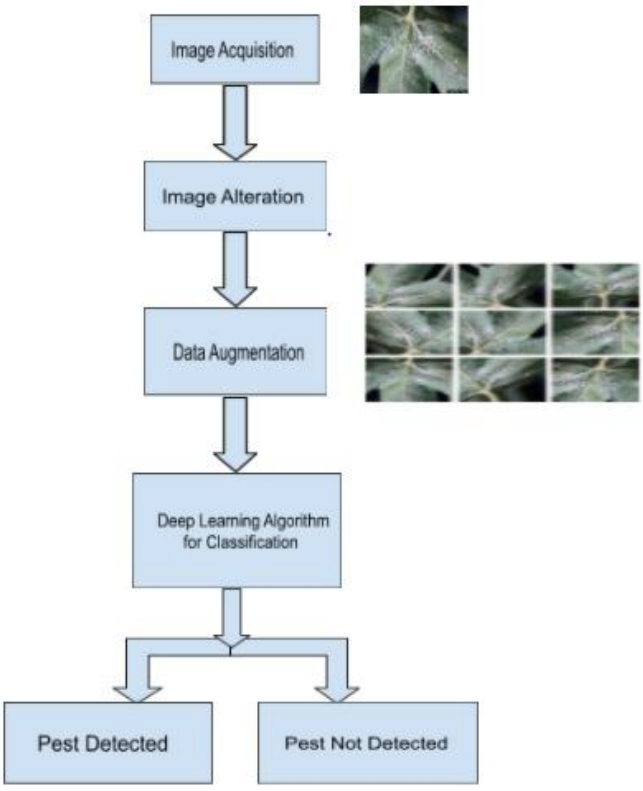


Fig 2: Basic flow of Pest detection

The further steps are mentioned below in detail:

A. DATASET GATHERING

Image acquisition or the gathering of images is done from many resources. The dataset so formed contains the real time images which were captured from a farm by a phone camera of 13 megapixels. Healthy plant images and diseased plant images both are required for training and testing. Keeping in consideration the need for varied pixel images, some images were taken from freely available open sources [9]. The collection of real time images collected from the field was a strenuous task as the cotton crops are seasonal.

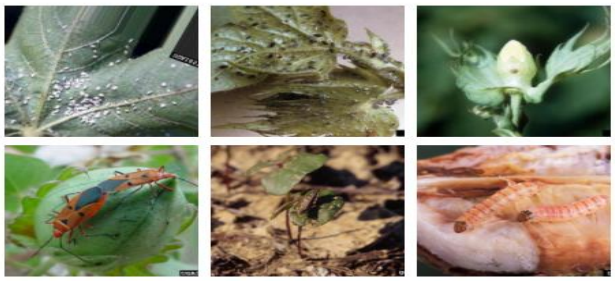


Fig.3: from top left a) Whitefly b) Aphid c) Boll-weevil second row from left d) Red cotton bug e) grasshopper f) Pink bollworm.

Whiteflies are sap- sucking bugs which typically lead to reduced plant vigor, upward curling of leaves, lint contamination with honey dew and related fungi. Cotton aphids, Aphis gossypii are extraordinarily variable in length and shade, varying from mild yellow to darkish inexperienced or nearly black. The maximum common symptoms of damage consist of leaf crumpling and downward curling of leaves. Boll Weevil is a kind of a beetle that broadly speaking feeds on cotton buds, it's early degree assault signs may be diagnosed with the aid of small feeding punctures on the facet of the bud. Red cotton bugs feed on developing and mature seeds, they generally stain the lint to ordinary yellow color. Grasshoppers usually lead to partial or full defoliation of leaves while the crimson bollworm larvae burrow into cotton bolls to feed on the cotton seeds.

B. IMAGE ALTERATION

The images are collected from different sources, so the images have different size, format and resolution. For smooth operation of the CNN model, it needs to identify the pest images collected to be in the same format all over. To acquire more accurate and precise results the resolution and resizing of images to 256 x 256 pixels is done.

C. DATA AUGMENTATION

The prediction accuracy of the CNN model hugely depends upon the diversity and the amount of dataset available during training. To attain this goal and ensure smooth running of CNN, data augmentation was performed. Execution of data augmentation led to an increase of dataset where initial images that were up to 1000 developed into a dataset of nearly 2000 images. The main operations performed on the images were, brightening, rotation, shearing, height shift, width shift, flipping, zooming and many more permutations which allowed to create a data set with good challenges.

D. ALGORITHM FOR CLASSIFICATION

This stage starts with generating data from the location followed by target size, batching of images and categorizing it. Creating a CNN model from scratch with each line of code compiling convolution followed by pooling and then second convo-pooling. Which then after takes us to flattening and finally the output layer of the model. The first convolutional layer has 64 filters with the kernel size of 3 with same padding. The same padding has both the output tensor and input tensor have the same width and height of 256. The activation function used is ReLu. The pooling computation reduces the dimension of feature map. We are using max pooling 2-D with a size of 2x2 and stride of 2. Flattening is done with units equal to 128. The unsupervised model is trained at several different values of epochs. Epoch means training the model with training data for one cycle. The accuracy achieved after 20 such epochs is 91.54% as showed in fig.9. The accuracy increases with increase in the number of dataset images.

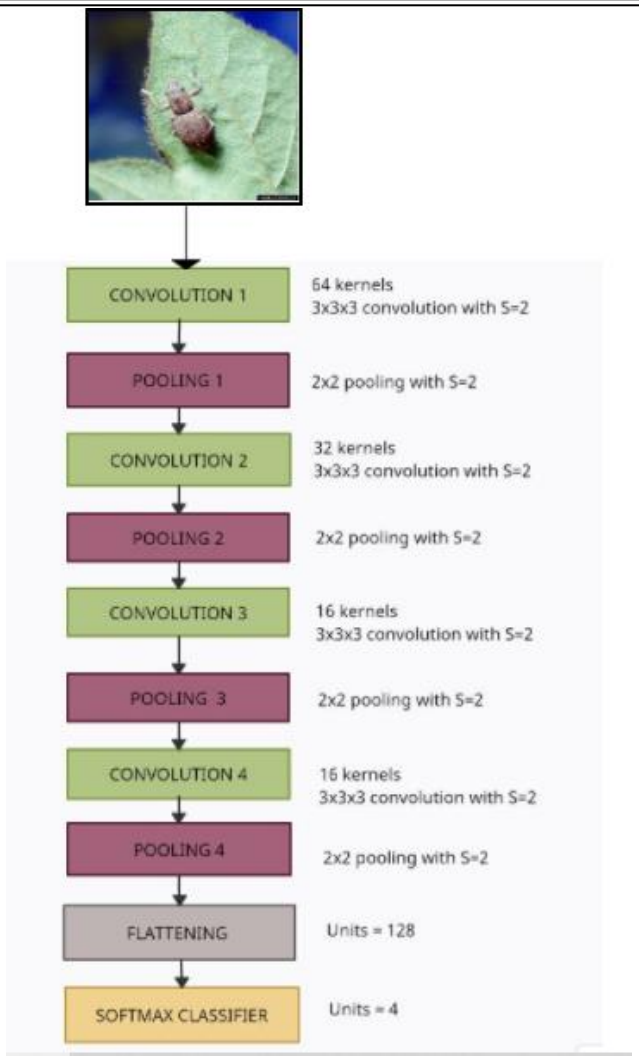


Fig. 4: Model Architecture

V. RESULTS

After compiling and running the code over the testing and validation data set, an accuracy of 91.5% was obtained after 20 epochs. The data set which was generated after augmentation created replicated images of some original images.

fig.5(original) fig.6(original)

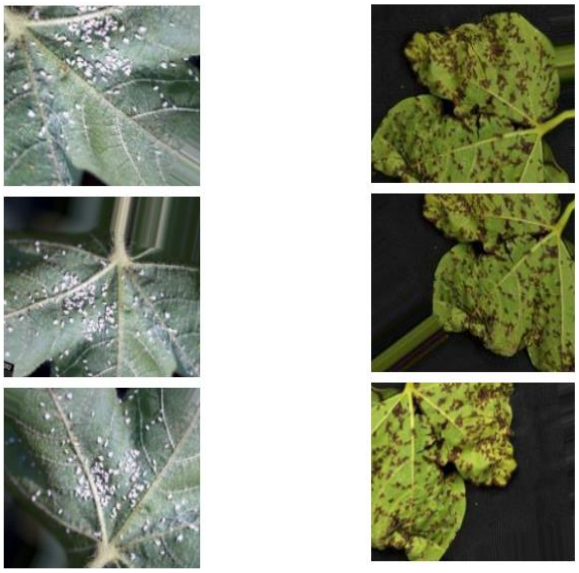


Fig. 7: Augmented output for both the original images

Metrics epoch	Accuracy	Loss
10	84%	0.30
20	91.54%	0.23

fig. 8 Training output

Here the accuracy calculations are done on testing data using the formula,

$$Accuracy = \frac{No\ of\ correctly\ predicted\ pest}{Total\ no\ of\ pests}$$

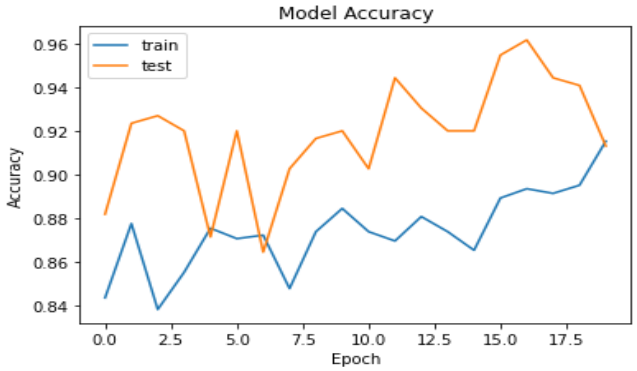


Fig. 9 Model Accuracy Graph

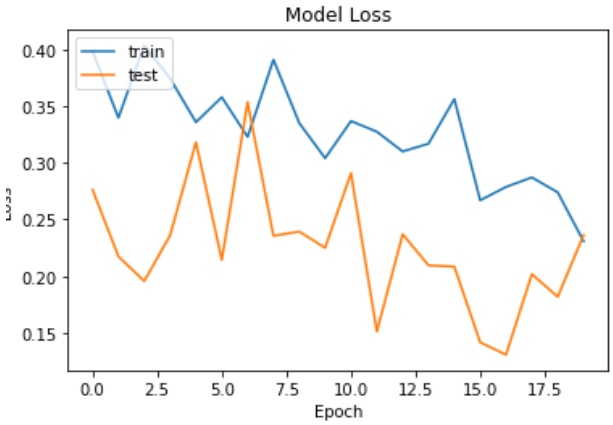


Fig. 10 Model Loss Graph

VI. CONCLUSION & FUTURE SCOPE

The proposed system is an advancement in the field of agriculture, which will be a backbone for detection of the pest. Detection of pests at an early stage will help to prevent the crop damage to the farmers. It will also aid in reducing the use of unwanted and extra usage of pesticide and in return increase the natural value of soil. Implementation of Data augmentation results in a larger dataset in turn yielding a better accuracy after CNN implementation. The CNN model presented here gives an accuracy 91.54%. Capsule Network can be implemented instead of CNN to eliminate the need of a larger dataset, avoid loss due to max pooling and also addresses the prediction of an object, ConvNet fully loses pose and orientation, and therefore the same data is distributed to a same neuron that isn't capable of handling such quite data in an object.

VII. REFERENCES

[1] Pruthvi P. Patel, Dinesh Kumar B. Vaghela "Crop Diseases and Pests Detection Using Convolutional Neural Network",2019 IEEE ,978-1-5386-8158-9/19

[2] Preetha Rajan, Radhakrishnan. B, Dr.L. Padma Suresh "Detection and Classification of Pests from Crop Images Using Support Vector Machine" ,2016 IEEE DOI: 10.1109/ICETT.2016.7873750

[3] Jayme Garcia Arnal Barbedo, "Detecting and Classifying Pests in Crops Using Proximal Images and Machine Learning: A Review" ,2020 Embrapa Agricultural Informatics, Campinas, SP 13083-886, Brazil, AI 2020, 1, 312–328; doi:10.3390/ai1020021

[4] Murali Krishnan, Jabert.G, "Pest Control in Agricultural Plantations Using Image Processing",2013 IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), e-ISSN: 2278-2834, p- ISSN: 2278-8735.Volume 6, Issue 4(May. - Jun. 2013).

[5] K. Ashish Reddy, N. V. Megha Chandra Reddy and Sujatha. S, "Precision Method for Pest Detection in Plants using the Clustering Algorithm in Image Processing", International Conference on Communication and Signal Processing, July 28 - 30, 2020, India.

[6] Monica N. Jige, Varsha R. Ratnaparkhe, "Population Estimation of Whitefly for Cotton Plant Using Image Processing Approach", 2017 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India.

[7] Qinghai He, Benxue Ma, Duanyang Qu, Qiang Zhang, Xinmin Hou, Jing Zhao, "Cotton Pests and Diseases Detection based on Image Processing", TELKOMNIKA, Vol. 11, No. 6, June 2013, pp. 3445 ~ 3450 e-ISSN: 2087-278X

[8] Sachin D. Khirade, A. B. Patil, "Plant Disease Detection Using Image Processing", 2015 International Conference on Computing Communication Control and Automation, DOI 10.1109/ICCUBEA.2015.153.

[9]. www.ipmimages.org

AUTHOR'S PROFILE

Mrs. Sandhya Potadar
Assistant.Professor. Electronics and Telecommunication Engineering
MKSSS's Cummins College of Engineering for Women, Pune. Affiliated to Savitribai
Phule Pune University.

Aakanksha Khare
B. Tech in Electronics and Telecommunication Engineering
MKSSS's Cummins College of Engineering for Women, Pune. Affiliated to Savitribai Phule Pune University.

Shalaka Buche
B. Tech in Electronics and Telecommunication Engineering
MKSSS's Cummins College of Engineering for Women, Pune. Affiliated to Savitribai Phule Pune University.

Aksha Khairmode
B. Tech in Electronics and Telecommunication Engineering
MKSSS's Cummins College of Engineering for Women, Pune. Affiliated to Savitribai Phule Pune University.

Attendance Management system using Face Recognition

[1] Sandhya Potadar, [2] Riya Fale, [3] Prajakta Kothawade, [4] Arati Padale
Department of Electronics & Telecommunication Engineering,
MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra, India

Abstract - ***Managing attendance can be a tedious job when implemented by traditional methods like calling out roll calls or taking a student's signature. To solve this issue, a smart and authenticated attendance system needs to be implemented. Generally, biometrics such as face recognition, fingerprint, DNA, retina, iris recognition, hand geometry etc. are used to execute smart attendance systems. Face is a unique identification of humans due to their distinct facial features. Face recognition systems are useful in many real-life applications. In the proposed system, initially all the students will be enrolled by storing their facial images with a unique ID. At the time of attendance, real time images will be captured and the faces in those images will be matched with the faces in the pre-trained dataset. The Haar cascade algorithm is used for face detection. Local Binary Patterns Histogram (LBPH) algorithm is used for face recognition and training the stored dataset, that generates the histogram for stored images and the real time image. To recognize the face, the difference between histograms of real time image & dataset images is calculated. Lower difference gives the best match resulting in displaying the name & roll number of that student. Attendance of the student is automatically updated in the excel sheet.***

Keyword-Face detection & recognition, Haar Cascade Classifier, Local binary pattern histogram (LBPH).

I. INTRODUCTION

The person in photos and in real time videos can be identified using facial recognition systems. It is a tier of biometric security. Other types of biometric security are voice, fingerprint and eye recognition. In real time we can use it to unlock phones, find missing persons, aid forensic investigations, help the blind, etc. The main aim of this project is to build an attendance system based on face recognition.

Different systems can be used for face recognition like attendance systems with fingerprint scanners, RFID tags and readers, facial recognition and location-based attendance systems. The attendance system with a fingerprint scanner minimizes the issue of proxy attendance. The system by using an RFID reader is much faster but high chances to get proxy attendance. Attendance systems using face recognition provide authenticated data also proxy data entering chances are much less as compared to other systems.

In the proposed system, at the time of enrolment, video is captured and images of students are stored through facial detection, recognition and recorded in a database. In real time, video of a student entering a classroom will be captured, face will be detected and matched with the dataset images, name & roll number of the present student will be displayed along with updating the attendance.

II. RELATED WORK

Face is a unique identity of any person. It is used in many domains and is the fastest growing research area.

Many systems are being proposed for attendance management. One of the systems [1], generates a smart attendance system which uses Quick Response (QR) code to track & record the attendance. Students and professors are given a unique QR code, at the beginning of the course, they are required to scan their QR code using a QR reading device. Attendance of students whose QR code is scanned will be recorded. This system is responsive to mobile phones and different computer systems.

A reliable attendance monitoring system based on biometric is developed [2], which is used to monitor the presence of students in a more effective way. It reduces the chances of marking proxy attendance and also reduces the problems like missing papers of attendance, which occur during marking attendance manually. Teachers have a small fingerprint scanner with them and students will press their finger on it to mark their attendance. Attendance management systems using Iris recognition [3], are more reliable and accurate because of the inner characteristics of iris like uniqueness, time invariance, immovability etc. The Iris pattern of each student is used for attendance. By using the camera live images of student iris are captured and stored in a database. Gray coding algorithm is used for measuring radius of iris and then that radius is matched with the radius of each student in the database and attendance of that student will be marked.

In one of the proposed models [4], two databases (face database & attendance database) are used. During enrolment, facial images of students are stored into the face database. The camera captures the images of the classroom, the images get enhanced and the attendance is marked in the attendance database after face detection & recognition. AdaBoost algorithm and Principal Component Analysis (PCA) are used for face detection and face recognition respectively.

The LBPH algorithm [5], can recognize the front face as well as side face with approximate accuracy of 90%. The flow of this algorithm starts with dividing the image into blocks and calculating the histogram of each block, then combining the histogram of all the blocks into a single histogram. This histogram has some value which is used for comparing later with the real time image histogram for identification. Multiple faces can be detected in a single detection hybrid process of Haar cascade and Eigenfaces method are used [6]. This process is able to detect multiple faces with an accuracy of 91.67%. By using this method, we can recognize faces during day and night time and are also able to detect 15 degrees side facing faces. By using a webcam this process can successfully perform at more than 200 cm.

One of the methodologies [7], considers accuracy rate, stability of system in actual time video processing, truancy of system and interface setting of the face recognition system. Face detection and recognition are two main parts of face recognition. Feature extraction is done by the LDA (Linear Discriminant Analysis) method. This model takes help of methods such as geometric Feature method, Subspace analysis method, Neural Network methods, Support Vector Machine (SVM) method to develop their face recognition algorithm. Experimentally this model of video face recognition system gives an accuracy rate up to 82%.

III. METHODOLOGY

The proposed methodology starts with the registration of students into the system. Following methodology has few main stages such as capturing images, pre-processing of the images, Haar Cascade classifier is used for face detection, developing a dataset of images, the further process of face recognition is done with the help of LBPH algorithm as shown in fig 1.

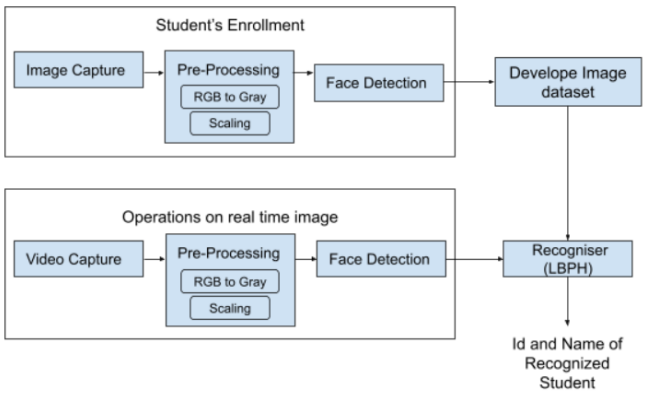


Fig 1: Proposed Methodology for Attendance monitoring.

1. Image Capture

The high-resolution camera which is used for capturing video is used to take frontal images of the students.

2. Pre-processing

The images are converted from RGB to Grayscale and are scaled down by a factor of 1.2.

3. Face Detection

Face Detection is composed of four stages as shown in fig 2.

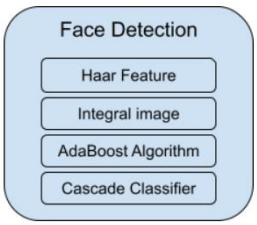


Fig 2: Face Detection

A. Haar Features

Haar features are the same as convolutional kernels and are used to detect the features in a given image. There are different kinds of Haar features such as line feature, edge feature, four - rectangle feature etc. A single value is used to represent each feature which is calculated by subtracting the sum of pixels under the white rectangle from the sum of pixels under the black rectangle as shown in fig 3. Haar cascade algorithm makes use of 24*24 windows which ends up calculating 160000+ features in a window. To simplify the work of calculating the feature values, an Integral image algorithm is introduced.

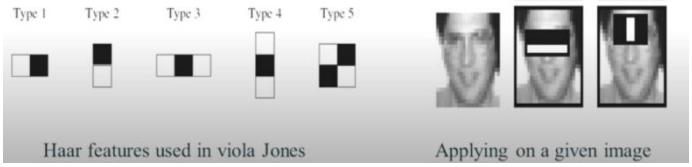


Fig 3: Haar Feature Selection.

B. Integral Image

To reduce the computation of the pixels to find the feature values, Viola Jones introduced a technique called Integral Image. As shown in fig 4, the value of the pixel at (x, y) in an Integral image is calculated by adding the values of pixels above and to the left of (x, y) pixel.

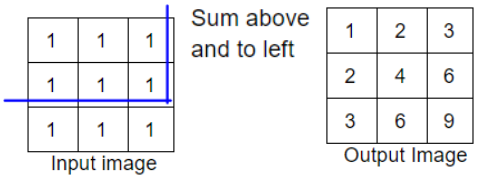


Fig 4: Integral Image.

Calculation of the sum of all pixels inside any given rectangle can be done by using only four values at the corner of that rectangle with the help of Integral image as shown in fig 5.

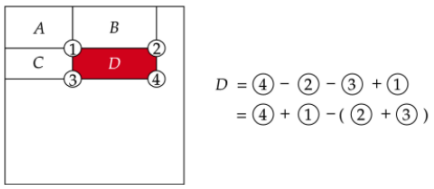


Fig 5: Calculation of sum of pixels using Integral Image.

C. AdaBoost Algorithm

As Haar cascade uses 24*24 windows there can be 160000+ features within a detector which needs to be calculated. AdaBoost is an algorithm based on machine learning. This algorithm finds the best features among all the possible features and eliminates the irrelevant features. To evaluate and decide whether a given window has a face or not, a weighted combination of all the found features are used. Features are included only if they can at least work better than random estimation. These features are known as weak features. AdaBoost linearly combines these weak classifiers to design a strong classifier. The weak classifier gives 0 or 1 as an output depending on its performance in image. Output is 1 when it performs well and able to identify features applied on image and Output is 0 when no pattern of the feature is present in image.

The steps involved in the AdaBoost algorithm are as follows.

1. It distributes the uniform weights over training examples, positive weights for faces and negative weights for non-faces.

Fig 6.1: Uniform distribution of weights

2. Selects the weakest classifier i.e., with lowest weighted error. For e.g., x=0.5 if x>0.5 then they are faces otherwise non-faces, but due to this some of the non-faces comes under faces category.

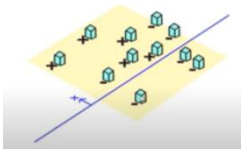


Fig 6.2: Selection of weakest classifier

3. To overcome this drawback the algorithm increases the weights on the misclassified training examples because we need a new classifier which concentrates more on misclassified features.

Fig 6.3: Increase weights on misclassified examples

4. These steps are repeated and at the end, all the weak classifiers obtained at all iterations are combined linearly and help to define a perfect boundary region.

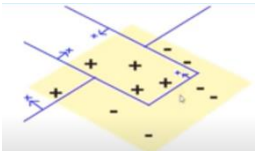


Fig 6.4: Linear combination of weak classifiers

A single rectangular feature is used to classify a single rectangular feature and with respect to the weighted neighbor, the positive and negative images are separated. Gaussian weak classifiers are also used for this purpose.

D. Cascading

In Paul Viola and Michael Jones detection algorithm, one single image is scanned many times by the detector with a new size every time. When multiple faces appear in an image the algorithm concentrates on removal of non-faces and brings out the most feasible face area. Since the computation cost is very high for each window when a particular strong classifier is a linear combination of all the best features and is not so appropriate for evaluation. Hence a cascade classifier is used.

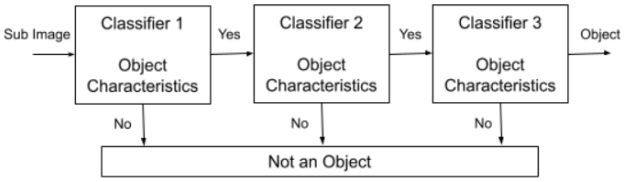


Fig 7: Cascading Classifiers

Cascade classifiers comprise various stages. All the features are grouped into different stages and each stage contains a strong classifier and a feature. To design these stages, AdaBoost is used. As shown in fig 7, each stage decides whether a sub- window is a face or not. The window is discarded if it does not contain a face. While training a classifier the number of features, stages and thresholds are taken into account.

4. Developing a dataset

The faces detected in images are stored in the database after pre-processing and detection. A minimum of 20 images are captured per individual student along with a unique ID. The dimensions of these stored images are 212x212 pixels. These images are later used to train the recognizer.

5. Face Recognition

Local Binary Pattern (LBP) is a smooth & adequate operator, which operates by setting the pixels of an image by thresholding the neighborhood of each pixel and examines the outcome as a binary number. Histogram of Oriented Gradients (HOG) descriptor increases the detection performance when combined with LBP. Therefore, a combination of LBP & HOG which gives LBPH algorithm is used for face recognition.

Steps involved in LBPH are as follows.

1. LBPH considers four parameters for face recognition which are as follows

- Radius: To set up a circular local binary pattern radius is used. Generally, it is set to 1.
- Neighbors: To set the circular local binary pattern neighbors are used. Normally, set to 8.
- Grid X: Gives cells count which are in horizontal direction. Normally, it is set to 8.
- Grid Y: Gives cells count which are in vertical direction. Normally, it is set to 8.

2. Training the Algorithm: A database of the face images of students which are to be recognized is used to train the algorithm. The unique ID which is set while developing a dataset is useful for recognizing the student.

3. Applying the LBP operation: By intensifying facial characteristics, create an intermediate image that describes the original image. Based on the parameters like radius and neighbour, the algorithm uses a sliding window concept. Fig 8 describes this operation.

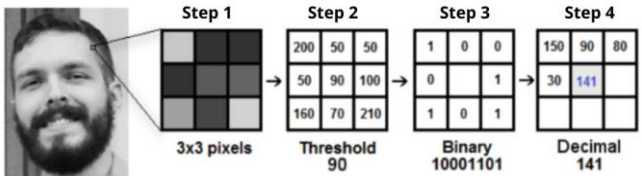


Fig 8: LBP Operation

- Assume that we have a grayscale facial image, take a part of images as a 3 x 3 matrix containing pixel intensities in the range (0 – 255), as shown in step1 and 2 in fig 8.
- In step 3, consider central pixel intensity as threshold and change the values of 8 neighbors with respect to the threshold value. (Set it to 1 if neighboring pixel intensity is greater than or equal to the threshold value, otherwise set to 0.)
- In step 4, convert the binary value into decimal value. The central pixel value of the image matrix is replaced by a decimal value. This central pixel is actually a pixel of a primary image.
- Applying these steps to all the parts of the image, we get a new image (result of LBP operation) that describes the features of the primary image.

4. Extracting the Histograms: Grid X and Grid Y parameters are used to divide images into multiple grids.

Each histogram holds only 256 positions (0-255) that shows the existence of each pixel intensity as the image is in grayscale. Histogram of each cell is to be concatenated to generate a bigger and new histogram. The final histogram shows the characteristics of the primary image as shown in fig 9.

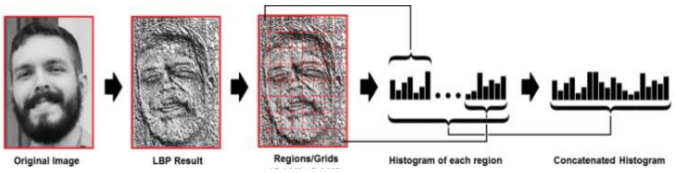


Fig 9: Extracting Histogram of image

5. Performing the face recognition: The algorithm for creating histogram is initially trained. Each image from the training dataset is represented by each histogram. To generate a histogram for the input image the above steps are performed again on that image. The histogram of input image is compared with the histograms of dataset images, selecting the closest histogram gives the matching image from the dataset. Various methods like Absolute value, Euclidean distance, etc can be used to compare the histograms. The Euclidean distance can be calculated using equation 1 to compare the histograms:

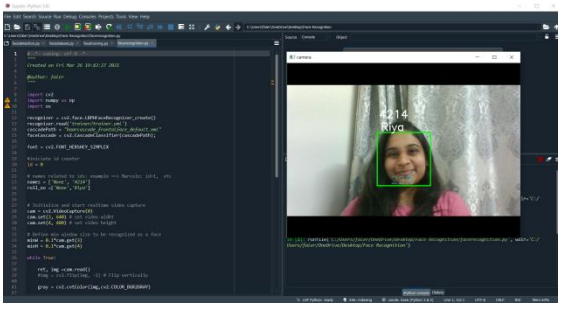
$$D = \sqrt{\sum_{1}^{n} (HistD - HistR)^2} \quad \text{-----Eq1}$$

Where HistD - Histograms of dataset images
HistR - Histogram of real time image

The algorithm returns a unique Id of the student with the minimum difference in the histograms of the student's image and dataset images. It also returns the calculated distance, which can be used as a 'confidence' measurement. Lower the confidence measurement, more is the precision of the recognizer.

IV. RESULTS

The result of the proposed system is shown in fig 10. The unique Id and name of the students are displayed with the confidence number.



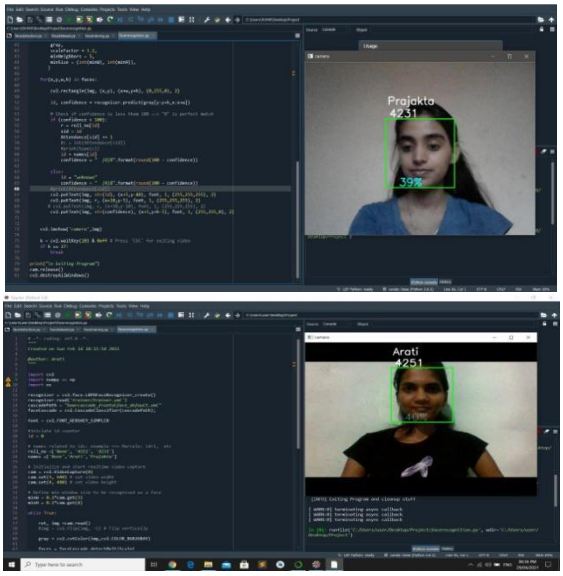


Fig 10: Results

V. Conclusion

The proposed method uses face detection and face recognition that helps to maintain the automated attendance system. For detection, Paul–Viola Jones algorithm is used and for face recognition Linear Binary Pattern Histogram (LPBH) algorithm is applied.
In the result, the unique ID and name of the student is displayed along with the confidence percentage. Confidence percentage represents the distance between the histogram of the stored image and histogram of the real time image and is calculated by using Euclidean distance. Lower is the distance, higher is the recognition rate.

VI. FUTURE SCOPE

The future scope of the project can be integrated with the hardware components for example GSM through which a monthly list of the defaulter students can be sent to the mentor.
Additionally, an application can be developed to help students to maintain a track of their attendance. It can also be used in offices where a large group of employees sit in a hall and their attendance will be marked automatically by capturing a video but for this the accuracy of the recognition needs to be improved.

VII. REFERENCES

[1] Asri Nuhi, Agon Memeti, Florinda Imeri, Betim Cico, "Smart Attendance System using QR code", 9th Mediterranean conference Embedded Computing, Budva, Montenegro, 2020

[2] M.A. Meor, M.H. Misran, M.A. Othman, M.M. Ismail, H.A. Sulaiman, A. Salleh, N. Yusop Centre for Telecommunication Research and Innovation FakultiKej. ElektronikdanKej. Komputer Universiti Teknikal Malaysia Melaka Hang Tuah Jaya, Durian Tunggal 76100, Melaka, Malaysia ,2014

[3] Amena Khatun, A.K.M Fazlul Haque, Sabbir Ahemad, Mohammad Rahman, "Design and Implementation of Iris Recognition Based Attendance Management System". ICEEICT Jahangirnagar University, Bangladesh, 2015.

[4] Shreyak Sawhney, Karan Kacker, Samyak Jain, Shailendra Narayan Singh, Rakesh Garg, "Real-Time Smart Attendance System using Face Recognition Techniques" in Amity University Uttar Pradesh, Noida, 2019.

[5] Awais Ahmed, "LBPH based Improved face recognition at low Resolution" UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA 2018 IEEE

[6] "Multi-Faces Recognition Process Using Haar Cascades and Eigenface Methods" Teddy Mantoro, Media A. Ayu, Suhendi Sampoerna University, Jakarta, Indonesia, 2018

[7] HAO YANG AND XIAOFENG HAN "Face Recognition Attendance System Based on Real-Time Video Processing" is supported in part by the Basic Public Welfare Research Project of Zhejiang Province under Grant LGF20H180001, 2020

VIII. AUTHOR'S PROFILE

Mrs. Sandhya Potadar
Assistant Professor, Electronics and Telecommunication Dept.
MKSSS's Cummins College of Engineering for Women, Pune.
Affiliated to Savitribai Phule University

Riya Fale
B.Tech in Electronics and Telecommunication Engineering.
MKSSS's Cummins College of Engineering for Women, Pune.
Affiliated to Savitribai Phule Pune University.

Prajakta Kothawade
B.Tech in Electronics and Telecommunication Engineering.
MKSSS's Cummins College of Engineering for Women, Pune.
Affiliated to Savitribai Phule Pune University.

Arati Padale
B.Tech in Electronics and Telecommunication Engineering.
MKSSS's Cummins College of Engineering for Women, Pune.
Affiliated to Savitribai Phule Pune University

AODV-QSRP a QoS based Routing Protocol for Mobile Ad-hoc Networks

Dr. Sanjeev Kumar Sharma[1], Dr. Komal Tihiliani[2],Anupriya Singh[3]
[1,2,3] Assistant Professor,SIRTT (Bhopal), India

Abstract
MANETs are very advantageous when there is no infrastructure or it has been fully or partially destroyed due to some hurdels like earthquake, floods and so on. Routing in MANETs are very challenging issue due to limited resources like limited battery power, limited bandwidth, mobility, routing and QoS and so on. Due to these challenges and scarcity of resources in MANETs, it's very difficult to achieve high quality of service. QAODV, AODV, OLSR, DSDV and ZRP and so on. are some of very popular routing protocols for MANETs. Minimum no. of hops is the route selection criteria used by most of the routing protocols. This makes it necessary to consider QoS parameters to the routing protocols. Above mentioned protocols are not QoS QoS routing protocol and insufficient to achieve high Qos, because they do not consider parameters which will affect the QoS, Instead they only consider hop count as a route selection criteria. In this paper we present a new QoS based routing algorithm AODV-QSRP (AODV based Quality of Service Routing Protocol) for MANETs. AODV-QSRP is based on existing AODV and is reactive (On-Demand) in nature. AODV-QSRP attempts to provide high QoS for the real time applications while considering the various important QoS parameters like Bandwidth, Delay, Link Quality and Battery power for route selection. We have evaluated and compared the performance of AODV-QSRP with QAODV, AODV,OLSR, DSDV and ZRP. To implement and simulating the result Network Simulator-2 (NS-2.32) on Fedora platform is used, which is a event driven and real-time simulator.
Keywords: MANET, QoS, AODV, OLSR, DSDV, ZRP

I. Introduction

MANETs operates on IEEE 802.11(b) standard. According to IEEE 802.11(b) specification that any device equipped with wireless links can communicate directly in absence of any infrastructure. MANETs are independent and autonomous system of mobile devices i.e. laptops, PDA, Smart phones and so on. These devices are equipped with wireless links and can connect directly to each other and forming a short live on-the-fly network. These nodes are self configurable and rapidly deployable, so we can form a network to communicate whenever there is no infrastructure or infrastructure has been fully or partially destroyed. Earth quack, military operations, floods, video conferencing are some area where we can take the advantages of MANETs. There is no central coordinator or access point, each device acts as a router and host. Nodes. Though MANETS are very advantageous but have various routing critical issues to be deal with carefully. Nodes are mobile so topology exactly unpredictable at any time and routing is at its best when the exact topological information's are available. So dynamic topology makes routing a very typical task in MANETs. Limited resources is again a big issue. As the nodes are very light weight devices with limited resources like battery power, storage, processing power makes MANETS to provide better quality of service to real-time applications is a challenging task. QoS is biggest challenging issue in MANETs [1] [2].

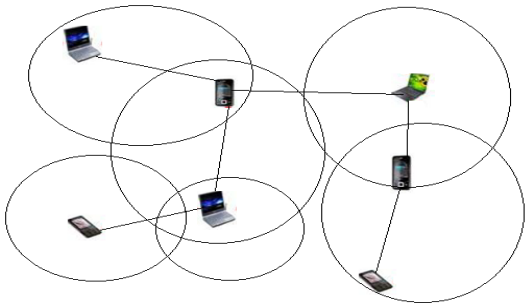


Figure 1: Mobile Adhoc Network

II. QoS Challenges in MANETs

MANETs was originally proposed for military operations and disaster management. However evolution of multimedia technology like IPTV, VOIP and so on. Quality of service in mobile ad-hoc networks has become an area of interest. Because of various requirements of different real time applications. As according to these applications required QoS parameters, quality of service for an application can be identified as a set of measurable pre-defined QoS constraints like delay, bandwidth, packet loss, and jitter, required bandwidth and so on which an underlying network needs to provide as and when needed while transmitting a packet from a source to its destination.

Quality of Service (QoS) refers to the capability of a network to provide better service to selected network traffic. Where quality covers data loss, delay or latency, jitter, efficient use of network resources and service means end-to-end communication between applications (i.e. Audio, video, E-mail).Overall goal of quality of service is to provide priority including dedicated bandwidth, reliability controlled jitter and latency [4].

in wired networks for multimedia traffic, QoS parameters are easy to handle and improved because of the availability of resources like processing power of routers and bandwidth.. But in case of MANETs QoS requires some new constraints due to high mobility, dynamic network topology and limited bandwidth, limited processing and power capacity than wire-based network. So it's very tough ask to deal with QoS in MANETs. Error prone natures of the wireless links (attenuation, multipath fading, noise and so on) are another important issue. Control overhead limits the bandwidth as the bandwidths available in MANETs are very less than the wired one. Security is another challenging issue due to broadcasting behaviour of MANETs. Limited battery power is another issue because any node can sink at any time [1][2].

The provision of quality of service (QoS) guarantees is challenging in mobile ad hoc network comparing with that in wire line networks, because of the frequently unpredictable topology changes due to mobility and power depletion.To reserve bandwidth and guarantee the specified delay , jitter for real-time application is a challenging task in MANETs.All of the existing QoS models require accurate link state such as available bandwidth, packet loss rate, delay, and topology information. i.e. For example, in a scheme involving resource reservation, if the sender moves, a new route to the destination must be established. Traditional internet QoS protocols like RSVP is not suitable for wireless environment due to the error-prone nature of wireless links and the high mobility of mobile devices in MANETs. Therefore, providing QoS in MANETs is more challenging than in fixed and wireless networks .A number of research have been conducted on required QoS in, but current results are not appropriate for MANETs and still quality of service for MANETs is an open and challenging issue in MANETs [7][1][2].

III. Quality of Service (QoS) parameters

Overall running applications on the networks can be divided in to two categories: Real-time or time dependent and Non-real time or ordinary. Real-time

applications require strong high QoS.. On the other hand applications like traditional mail, file transfer are the non real time applications and do not require high QoS values. Goal of quality of service is to provide priority to the applications including dedicated bandwidth, reliability controlled jitter and latency. QoS for a network is the value of various parameters that decides the quality of a service or application running on the network. When we talk about QoS parameters, End to End delay, jitter, Bandwidth and throughput are the primary concern,. In most research these parameters are the primary concern for the researcher to improve the quality of service. As the MANETS consists of mobile nodes they have limited resources like processing power, transmission capabilities and battery are the most critical issue that must be handled carefully, while designing a new routing protocol[1][2].

A Quality of Service Parameters

Before designing a routing protocol we must consider some network parameters which will affect overall performance of the network. Following are the some of the most important one.

- **End-to-End delay or Latency:** Total time taken by a packet to reach from source node to destination node is the End-to-End delay or in simple terms delay. End-to-End delay is the summation of transmission delay, propagation delay, queuing delay and processing delay from source node to destination node. For real time applications delay is very critical factor because as the delay increase, the quality of communication decreases.

End-to End Delay=Packet Arrival time- Packet Start time.....(1)

- **Packet delay variance:** Packet delay variance or jitter is the average variation in the arrival times between consecutive packets. It is due of congestion in the network or different packet arrives from different paths.

End-to End delay= $\sum_{i=1}^{n} \frac{Delay_i - Delay_{i-1}}{n-1}$ **.....(2)**

- **Throughput:** Throughput is the avg. amount of data (bits/sec.) transferred in a specific amount of time from source to the destination.

Throughput=Total packets sent*packet size*8/Data Duration.....(3)

- **Packet delivery ratio:** PDR is the total number of packets successfully delivered to the destination.

PDR= (Total packets Received / Total Packets sent)*100.....(4)

- **Packet loss Ratio:** It tells the ratio of packet lost against the number of packet sent. Packet loss ratio is reciprocal to PDR and throughput.

PLR= (Total packet send- Total packet received)/send *100.....(5)

- **Control overhead:** These packets are used to manage and maintain topological information in the network. But some of the protocol based on link state and distance vector routing algorithms uses these messages and create extra overhead in the network. Control overhead limits the available bandwidth for communication and leads the network into congestion. To efficient use of available bandwidth the control overhead must be kept minimized.
- **Bandwidth:** Bandwidth tells the amount of data can be sent on the network at a time. Amount of data transfer can't be equal to the available bandwidth because of control traffic. As the control traffic increase, it will limit the available bandwidth for communication.
- **Average energy consumed and Battery power:** due to several operations like sending and receiving data by nodes will consume the battery of a mobile node and this energy is limited for mobile nodes.

Avg_Energy_Consumption= total energy/nodes.....(6)

IV. QoS and Related Work In MANETs:

Providing quality of service (QoS) guarantee is challenging in mobile ad hoc network as compared to networks, because of Dynamic topology due to mobility and power depletion. To reserve bandwidth and guarantee the specified delay, jitter for real-time application is a challenging task in MANETs.Because of dynamic topology and limited resources QoS support in MANETs is a very challenging task. Routing is a very important mechanism for QoS guaranteeing in network. QoS routing must find a path from a source to a destination which satisfies the QoS requirements [8] [9]. In wired networks Integrated services (IntServ) and Differentiated services (DiffServ) are the two models proposed and used. DiffServ is a per service basis approach and all the traffic is classified into several classes according to their requirements. Since IntServ is a per flow basis approach means that each node in the network must have enough capacity for storing, processing and forwarding data, it is not well suited for networks like MANETs because of scarcity of resources. Any how these approaches are not as beneficial because of the hurdles in MANETs [10].

When it comes to QoS routing, the routing protocols have to ensure that the QoS requirements are met. Like wired networks, it is a tough task to ensure Quality of Service (QoS) provisioning in MANETs [11]. Almost all of the routing protocols for MANETs, such as AODV, ZRP, OLSR, DSR are designed without considering QoS parameters. The number of hopes is only criteria to select a route by these routing protocols. It is clear that such routing protocols are insufficient for real-time and multimedia applications, which require high and guaranteed QoS [18]. AODV is most popular and widely used due to the advantages of it i.e. small computation, self repair, reactive routing protocol. However AODV focuses on Number of hops to select a route and does not deal with other QoS parameters. So it isn't a pure QoS routing protocol [12]. There are several approaches and proposals for QoS provisioning in the existing AODV. Perkins et al. revised the QoS AODV internet drafts by adding more field extensions and new formats. De Renesse, Ghassemian, Friderikos and Aghvami proposed a QoS-AODV enhancement by referring to QoS-AODV IETF proposal while implementing only the extension of minimum available bandwidth into AODV route messages for QoS provisioning purpose.[11][12]. Nur Idawati Md Enzai, Farhat Anwar, Omer Mahmoud extended the work done by de Renesse QoS-AODV while adding minimum available bandwidth, maximum delay. Performance for above variants of AODV is better but increase routing and calculation overhead along with increase in avg. energy consumed [13] [11].

Chen and Heinzelman proposed a QoS protocol based on admission control scheme and a feedback scheme to meet the QoS requirements of real-time applications. [14]

Zheng Sihai, Li Layuan, Guo Lin proposed a QoS-based multicast routing protocol QMMRP. Entropy of nodes is treated as an important parameter to find a stable path; new protocol uses bandwidth reservation mechanism to achieve certain QoS requirements [15]. C.Wu et al. [16] presented an ad hoc on-demand multipath routing (Q-AOMDV), which provides Quality of Service (QoS) support. In this work they have used bandwidth, hop count and end-to-end delay in mobile ad hoc networks (MANET). The protocol gives path preference to the metric calculated by delays, bandwidth, and hop-count[1][2].

DSDV

Destination sequence Distance Vector is a proactive routing protocol. It is based on classical Bellman-Ford routing algorithm. It overcomes with the looping problem in case of broken links. Bellman-Ford routing protocol works successfully works in wired network, but in MANETs due to topological changes it creates the problem of count to infinity and can't work efficiently. To overcome this problem DSDV adapts a new attribute, sequence number in the routing table. Using this attribute in routing table DSDV can differentiate between stale route and new route, and overcome the problem of count to infinity. Each node using DSDV maintains routing table, which contains all available destination, metric and

next hope to reach them along with the sequence number generated by destination. Each node advertises the routing table periodically and updates its routing table with newly received information[1][2][9][24][30].

AODV

Perkins, Belding-Royer and Das introduced a novel scheme of Ad-Hoc on Demand Distance Vector (AODV) routing protocol, which shows optimization of route by flooding RREQ packets[31]. Ad-hoc On Demand Distance vector routing protocol is a reactive routing protocol. AODV provides unicast and multicast both kind of communication. AODV eliminates the periodic broadcast of Hello packets and also minimizes number of active routes between an active source and destination. AODV can determine multiple routes between a source and a destination, but implements only a single route. Route discovery and Route maintenance are the two basic operations performed by AODV Using RREQ, RREP, RERR and Hello Packets [10] [11] [12]. Route discovery is done by broadcasting RREQ packets to all destinations. When a node receives the RREQ packet then it cheks weather it is destination, if it is so then it generates Route reply to the destination. If it knows the route to destination than it uses RREP packet to inform the destination about the route. If neither its a destination nor its having route towards route it simply broadcast the route request message further. In case of any error it uses RERR message to inform the source about the Error[1][2][14][30].

ZRP

Zone based routing protocol is a hybrid routing protocol. It takes the benefit of both the proactive and reactive routing protocols. The main idea behind the ZRP is to get rid of routing overhead and long route request delay of reactive and proactive routing protocols.. Whole network is divided into small routing zones on the basis of no of hops from. Inside a particular zone it works as proactive and outside the zone it works as reactive. Within the zone routes are available as and when needed same as proactive routing concept. If the destination is not in its zone means than reactive approach same as AODV is used to find the route. Routing zone for each node is separate from each other. Routing zone is defined in terms of hopes. A zone for a node is the number of nodes that lies within the N hops away from this node i.e. N=2, means the nodes coming within the radius of 2 are in the zone of this node. Number of nodes in a zone depends on the transmission power of nodes[1][2][15][23].

V. The proposed QoS based Routing Methodology and Algorithm

Applications can specify required QoS constraints, and the routing protocol will search for a route that satisfies them during its route discovery process. In order to provide QoS, We can revise the conventional AODV (reactive) routing protocol. Adding the QoS information to each node in its routing table i.e. available bandwidth, delay, link quality and remaining battery power. Extensions can be added to the RREQ message during the route discovery process. .When a path discovery process is initiated, calculating the corresponding QoS values and finally on the basis of QoS value (minimum) we can find the path based on the best QoS value.

Route Optimization

Finding a route in MANETs is an optimization problem, while taking several Qos parameters together. Each and every parameter must be optimized means either maximize or minimize. For example available bandwidth may vary on each and every link throughout the routing path but we have to consider the minimum one. If there are too many paths and need to select one than on a path minimum available bandwidth must be considered and so on for the other paths and finally chose maximum available bandwidth from different paths. Similarly the same

strategy is used to select a path on the basis of delay, battery power and link quality.

Let G (V, E) a graph, where V is a set of vertex (Mobile nodes) and E is set of edges (all communication link) E = {(i, j)/ i, j ∈ V, i ≠ j}. We consider Delay, Bandwidth, Link quality and remaining battery power as our main criteria for route selection. As we have discussed in the beginning of this paper to achieve high QoS these parameters must be considered. Where bandwidth, link quality and battery power must be maximum from available minimum bandwidth throughout the path and delay must be minimum. For route selection we can take bandwidth, link quality and battery power as a single objective by taking all together and assigning them a weight say P_i, as they must be maximized. Delay is taken separately as it must be minimum. So we define out new QoS metric as follows.

QoS routing metric=

$$\boldsymbol{MaxMin}_{(i,j)\in E}(P_1 B_{ij} + P_2 Lq_{ij} + P_3\, BP_{ij}) \quad \text{.....(7)}$$

$$\boldsymbol{Min}_{(i,j)\in E}\ D_{ij} Cij \quad \text{.....(8)}$$

Where- P is weight, Cij Decision variable (C_{ij}=1, if (i,j) is on the routing path) , Lq link quality of the link (i,j), Bij Bandwidth available on (i,j), Dij End to End delay on link (i,j), BPij is available battery power on the link (i,j).

Bp>=B_{min} (greater or equal to 0.5e6)
Dp<=D_s (less than or equal to 0.3)
Lqp>=L_{qs} (greater or equal to 1e-6)
BPP>=BP_s (minimum 0.2 Joules)

B_{min}, D_s , L_{qs} and BP_s are the constraints defined over the objective function. These constraints must be satisfied during the route selection procedure to optimize the route.

VI. Simulation Evaluation and Result Discussion

1.1 Effect of Mobility

Mobility is one of the most critical issues in MANETs that must be deal carefully. As the mobility increase the probability of route break also increases. It also increases control overhead. Due to route break the packets may get loss and also increase the delay .In our scenario we have taken values 0,5,10 meters per second mobility. As we can observe by fig. No. 6,7,8,9 and 10, AODV-QSRP performs better as compare to other protocols and its PDR and throughput is approximately (near about) 100 percent. Though QAODV performs better than AODV, OLSR, DSDV and ZRP, but it does not reach up to 100 percent performance as PDR. Performance of QAODV is also poor compare to AODV-QSRP. As we have set our QoS requirements at application level, AODV-QSRP performs accordingly. Maximum Avg. delay for AODV-QSRP is under the required delay by application (0.3) that is 0.2 in for our protocol. Avg. delays of DSDV and OLSR protocols is less than AODV,LPPMM and ZRP because these are proactive protocols and have routes available at any time. Avg. Jitter for DSDV, OLSR and ZRP gets increase as the mobility increases. Avg. Jitter for AODV-QSRP (<0.2) is very less as compared to the base protocol that is AODV and other protocols. Though avg. Jitter for AODV is less than other protocols greater to AODV-QSRP.

Table 1

SIMULATION SETUP PARAMETERS FOR MOBILITY

Simulation Parameters	Values
Tool	NS-2
Network Type	MANET(IEE802.11(b))
Simulation Area (in meters)	1000*1000
Simulation time (In sec.)	500s

Data duration (in sec.)	100
No. Of Nodes	60
Node placement strategy	Random distribution
Mobility Pattern	Random Waypoint
Mobility speed (meter/sec.)	0,5,10
Type of source	CBR
No. Of source	3
Packet size(Bytes)	1100
Interval between packets	0.1 sec.
Routing Protocols	QAODV, AODV, DSDV, OLSR, AODV-QSRP, ZRP

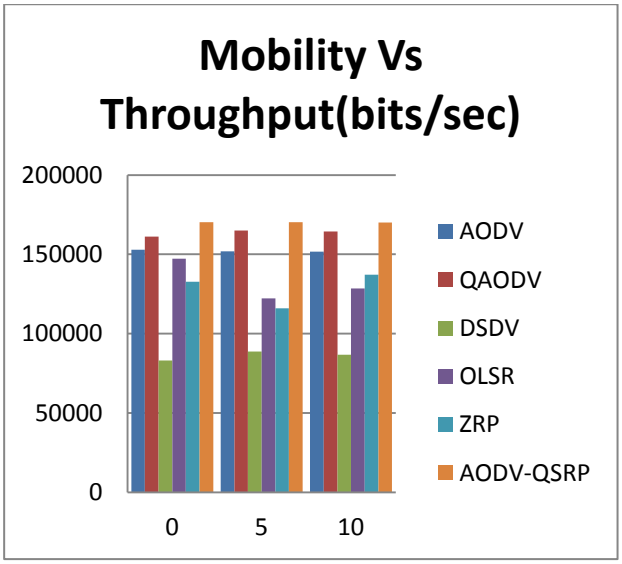


Figure 2: Mobility Vs Throughput(bits/sec)

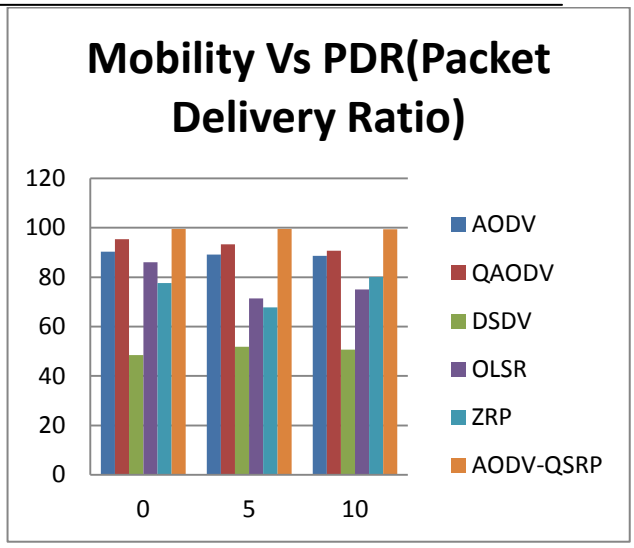


Figure 3: Mobility Vs PDR(Packet Delivery Ratio)

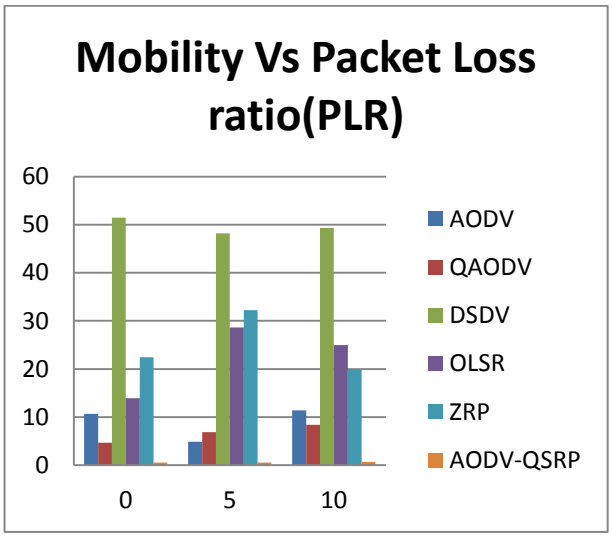


Figure 4: Mobility Vs Packet Loss ratio(PLR)

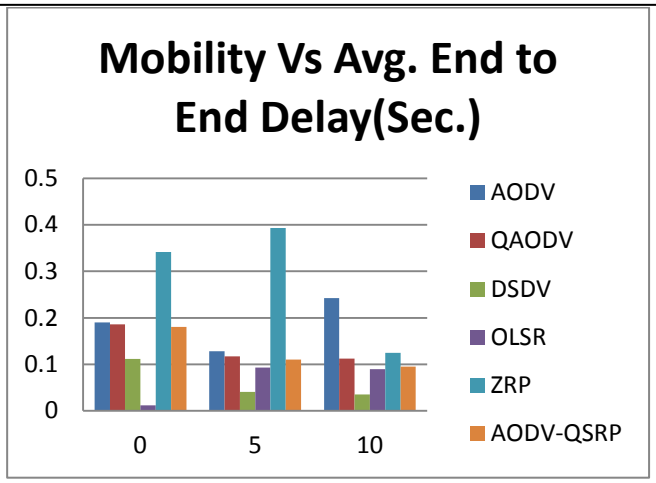


Figure 5: Mobility Vs Avg. End to End Delay(Sec.)

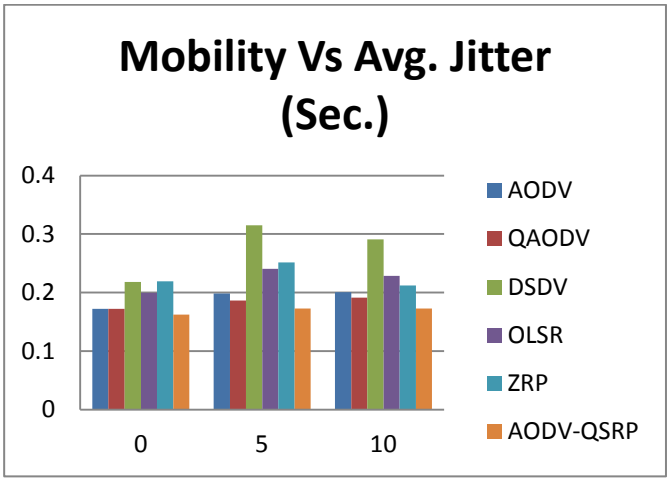


Figure 6: Mobility Vs Avg. Jitter (Sec.)

1.2 Effect of No. of Nodes

Increasing the nodes means the corresponding network will get more dense and the probability of data delivery gets increase. It leads into cost of congestion in the network as the number of communication gets increase. in our scenario we have taken 50,60,70,80,90,100 as the node values. As we can observe by fig. No.21, 22, 23, 24 AODV-QSRP performs better than other protocols and its PDR and throughput is approximately 100 percent. Though QAODV performs better than AODV, OLSR, DSDV and ZRP, but it does not reach up to 100 percent PDR. It Performs poor as compare to AODV-QSRP.As we have set our QoS requirements at application level, AODV-QSRP performs accordingly. Maximum Avg. delay for AODV-QSRP is under the required delay by application (0.3) that is 0.2 in our result. Avg. delays of DSDV and OLSR protocols is less than QAODV, AODV-QSRP and ZRP because these are proactive protocols and have routes available at any time. Delay for ZRP gets increase as the size of network grows because its hybrid in nature. Avg. Jitter for DSDV, OLSR and ZRP gets increase as the size of network increases. Avg. Jitter for AODV-QSRP (<0.2) is very less as compared to other protocols. Though avg. Jitter for QAODV is less than other protocols but greater to AODV-QSRP. So overall performance of AODV-QSRP is better than the other protocol and meets the specified QoS requirements. As our protocol is based on AODV and it performs better than AODV as can be seen by comparison graphs

Table 2: simulation setup parameters for nodes

Simulation Parameters	Values
Tool	NS-2
Network Type	MANET(IEE802.11(b))
Simulation Area (in meters)	1000*1000
Simulation time (In sec.)	500s
Data duration (in sec.)	100
No. Of Nodes	50,60,70,80,90,100
Node placement strategy	Random distribution
Mobility Pattern	Random Waypoint
Type of source	CBR
No. Of source	3
Packet size(Bytes)	1100
Interval between packets	0.1 sec.

Routing Protocols	AODV,QAODV, DSDV, OLSR, AODV-QSRP, ZRP

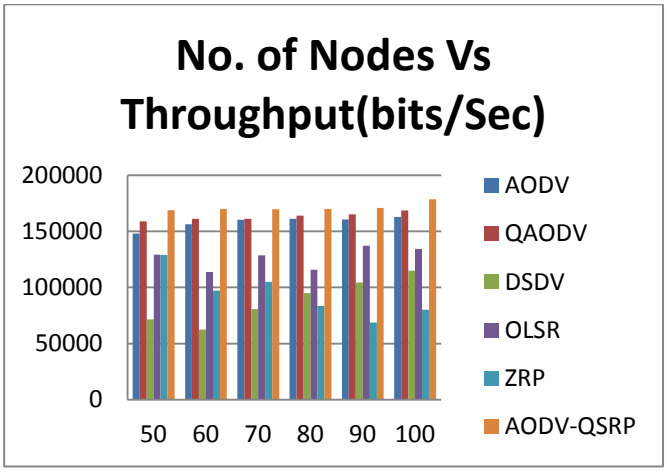


Figure 7: No. of Nodes Vs Throughput (bits/Sec)

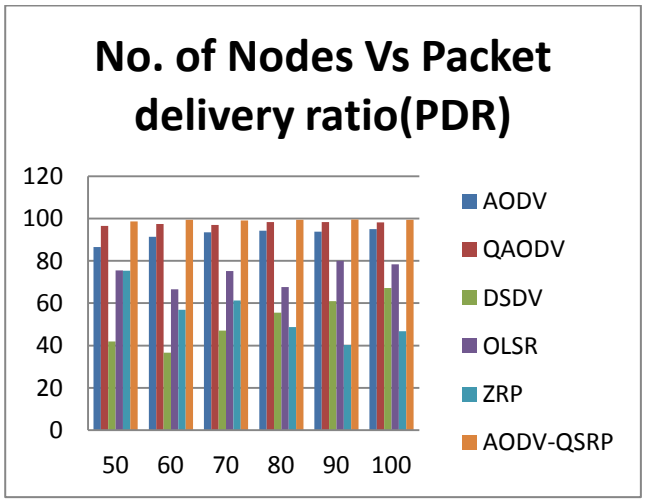


Figure 8: No. of Nodes Vs Packet delivery ratio (PDR)

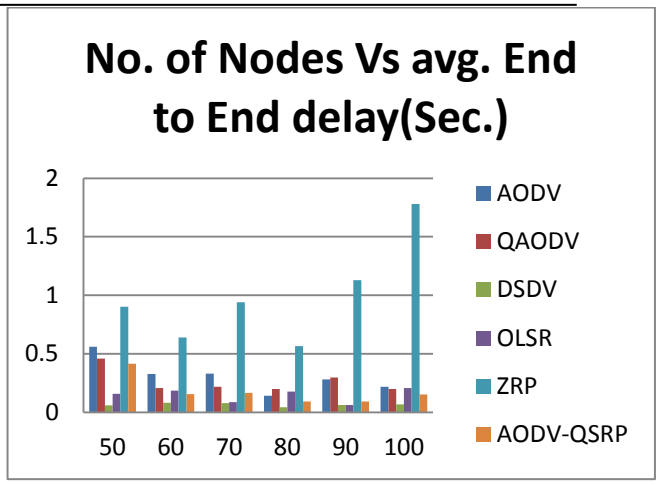


Figure 9: No. of Nodes Vs avg. End to End delay(Sec.)

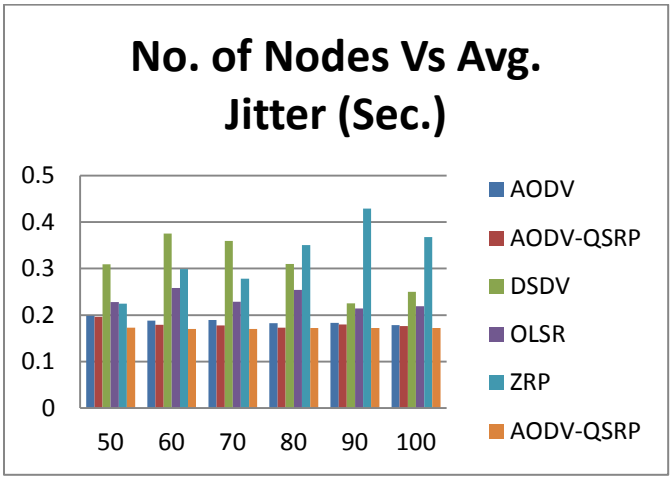


Figure 10: No. of Nodes Vs Avg. Jitter (Sec.)

VII. Conclusion and Future Work

To achieve high QoS in MANETs for real time communication is a very important and issue because of various issues faced in MANETs. Therefore in this paper, we proposed an efficient QoS based routing algorithm to improve the performance and to support real-time applications in MANETs. Our Protocol recognise the problem of QoS in MANETs. To achieve the batter performance for real-time applications we have used several QoS parameters in our protocol, Which has not been taken in the QoS protocols like

QAODV etc, AODV-QSRP also energy aware routing protocol because it consider remaining battery power. To support high QoS we also considered delay, bandwidth and link quality as route selection criteria. As according to simulation study overall performance of AODV-QSRP is better than AODV and QODV and other protocol and meets the specified QoS requirements as we need. As our protocol is based on AODV and it performs better than QAODV, AODV as can be seen by comparison graphs. Due to several calculations overhead the avg. energy consumption is more than AODV, OLSR and DSDV. In future study we will try to minimize the overhead and reduce the battery consumed by this protocol.

References:

[1] Sanjeev Kumar Sharma,Sanjay Sharma" Performance evaluation of routing in MANETs based on QoS parameters", International Journal of Modern Computer Science and Applications (IJMCSA), Volume No.-4, Issue No.-1, January, 2016,, PP 49-54, ISSN: 2321-2632

[2] Sanjeev Kumar Sharma, Sanjay Sharma "IPv4 Vs IPv6 QOS: A challenge in MANET ", IJSAIT, Vol. 3 , No.4, Pages : 07 - 11 (2014), ISSN 2278-3083.

[3] Masoumeh Karimi, Deng Pan, **"**Challenges for Quality of Service (QoS)in Mobile Ad-Hoc Networks (MANETs)", PP. 978-1-4244-4565-3/09/2009, IEEE.

[4] Ayman Mansour Murad, Bassam Al-Mahadeen, Nuha Mansour Murad "Adding Quality of Service Extensions to the Associativity Based Routing Protocol for Mobile Ad Hoc Networks (MANET)", Asia-Pacific Services Computing Conference, 2008,IEEE.

[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services", 1998, IETF RFC 2475.

[6] L. Zhang, S. Deering, and D. Estrin, "RSVP: A New Resource Reservation Protocol", IEEE Network, Vol. 7, No.5, Septmberl993, pp. 8-18.

[7] Masoumeh Karimi, Deng Pan,"Challenges for Quality of Service (QoS) in Mobile Ad-Hoc Networks (MANETs)", IEEE, 2009.

[8] Sudha Singh,S. C. Dutta, D. K. Singh "A study on Recent Research Trends in MANET",IJRRCS,Vol3, No. 3, PP. 1654-58,june-2012.

[9] RFC 2205 "Resource ReSerVation Protocol (RSVP)"

[10] L. Chen. "QoS-Aware routing based on bandwidth estimation for mobile Ad hoc networks" *J+. IEEE Journal on Selected Areas in Communications, 2005, 23(3):561-572.

[11] Banwari, Deepanshu Sharma, Deepak Upadhyay. "Routing Algorithms for MANET: A Comparative Study ." IJEIT , Volume 2, Issue 9, (2013).

[12] Beherouz A. Forouzan, "Data communication and networking" 4th Edition, McGraw-Hill Forouzan Networking Series.

[13] http://vlssit.iitkgp.ernet.in/ant/ant/7/theory

[14] L.Chen and B. Heinzelman, "QoS-aware routing based on bandwidth estimation for mobile Ad Hoc networks", IEEE Journal on Selected Areas in Communications, 2005.

[15] Zheng Sihai , Li Layuan, Guo Lin ,"QoS-Based Multicast Routing Protocol in MANET", International Conference on Industrial Control and Electronics Engineering,2012,IEEE Computer Society, 978-0-7695-4792-3/12 /2012, IEEE DOI 10.1109/ICICEE.2012.

[16] Mandeep Kaur Gulati, Krishan Kumar," A Review of QoS Routing Protocols in MANETs", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 – 06, 2013, Coimbatore, INDIA, PP. 978-1-4673-2907-1/13/2013, IEEE

[17] https://en.wikipedia.org/wiki/Wireless_ad_hoc_network

[18] V. Vidhyasanker, B.S. Manoj, and C. Siva Ram Murthy,"Slot Allocation Schemes for Delay Sensitive Traffic Support in Asynchronous Wireless Mesh Networks," The International Journal of Computer and Telecommunications Networking, Vol. 50, Issue 15, pp. 2595-2613, 2006.

[19] N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: A survey," IEEE Wireless Commun. Mag., vol. 11, no. 6, pp. 6–28, Dec. 2004.

[20] https://en.wikipedia.org/wiki/List_of_ad_hoc_routing_protocols.

[21] Perkins,C.E., and Bhagwat.P. (1994). Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers.ACM, pp.234 – 244.

[22] P. Muhlethaler, T. Clauser, A. Laowti, A. Qayyum,L. Viennot, Hipercom Project INRIA Rocquentcourt, BP 105,78153bLe Chesnay cedx, France "Optimized Link State Routing Protocol (OLSR) Routing protocol"

[23] Charles Perkins, Elizabeth M. Royer, Samir R. Das, "Ad-Hoc On-Demand Distance Vector (AODV) Routing", draft-ietf-manet-aodv-08.txt.

[24] S. Das, C. Perkins, E. Royer, "Ad hoc on demand distance vector (AODV) routing", Internet Draft aodv protocol, work in progress,2002.

[25] P. Muhlethaler, T. Clauser, A. Laowti, A. Qayyum, L. Viennot, Hipercom Project INRIA, "Optimized Link State Routing Protocol (OLSR) Routing protocol", RFC 3616, www.ietf.org.

[26] Nicklas Beijar, Networking Laboratory, Helsinki University of Techno "Zone Routing Protocol(ZRP)", https://tools.ietf.org/html/draft-ietf-manet-zone-zrp-00

[27] Masoumeh Karimi Technological University of American (TUA) USA, "Quality of Service (QoS) Provisioning in Mobile Ad-Hoc Networks,

Intelligent Crop Monitoring System Using Decision Tree Approch

S.Veenadhari [1], Pratima Gautam 2
[1]Associate Professor,Rabindranath Tagore University, Bhopal, Madhya Pradesh, India
s.veenadhari@rntu.ac.in
[2] Dean Of Computer Science,Rabindranath Tagore University,Bhopal
Pratima.gautam@aisectuniversity.ac.in

ABSTRACT
The externalities of climatic parameters adversely affected the crop production as well as their prediction models, despite the large quantities of data are collected both in agriculture as well as in meteorological departments. Use of advanced computing techniques like data mining will help in the extraction of hidden predictive information from large databases. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems helps in quick and accurate prediction models.. This advanced computing tools and techniques were not used in India to identify the influence of input parameters on the agricultural production. An attempt has been made in the present study to predict the crop productivity as influenced by climatic parameters in Madhya Pradesh State using data mining techniques like C 4.5 algorithm and Decision tree. The prediction accuracy of the developed model varied from 76 to 90 per cent for the selected crops and selected districts. Based on these observations the overall prediction accuracy of the developed model is 82.00 per cent. With a high prediction accuracy the developed model can be used by the farmers and policy makers in arriving at expected productivity trend well in advance i.e., before the harvest of the crop.

Keywords: Data Mining, Crop production, Climatic parameters, C4.5. Observed productivity, Predicted productivity

INTRODUCTION

Several methods of predicting and modeling crop productivities have been developed in the past with varying rate of accuracy, as these don't take into account the characteristics of the weather. A regression analysis was used for predicting the input interaction on crop productivity [8]. They studied the major agricultural inputs and their effect on crop productivity and also cost of cultivation affected by different inputs in selected States of India. A model for real time assessment of the direction and quantum of variability in wheat productivities was developed [5]. A simple technology trend model in conjunction with crop simulation model (CERES-Wheat in DSSAT environment) was used for early wheat productivity prediction at six locations representing the six major wheat-growing states, which contribute about 93% of national wheat production. A simulation model when run on a common set of soil properties, genetic coefficients and agronomic practices, is supposed to capture inter-annual productivity variability due to year-to year varying weather conditions. The study has significance in issuing an early 'national wheat' production forecast using in-season weather data up to February and normal weather data for the rest of the period.

A process model for analyzing data, and described the support that Waikato Environment for Knowledge Analysis (WEKA) provides for this model[1]. The domain model learned by the data mining algorithm can then be readily incorporated into a software application. This WEKA based analysis and application construction process was illustrated through a case study in the agricultural domain i.e., in mushroom grading. Thematic information related to agriculture which has spatial attributes was reported in a study [4]. Their study aimed at discerning trends in agriculture production with reference to the availability of inputs. The Predicted and Real vs. Counter graph illustrates how closely the Poly-Analyst prediction follows the actual value of the attribute over the range of the dataset. The study demonstrated the scope for application of

spatial mining tools for a utility study and analysis. The specific application of Poly analyst gave a clear scope for evaluation and comparison of predicted and real values.

A decision tree classifier for agriculture data was proposed, though some data set includes missing values, the experiment showed the performance of the proposed method [2]. This new classifier uses new data expression and can deal with both complete data and incomplete data. In the experiment, 10-fold cross validation method is used to test abalone data set, horse-colic data set and soybean data set. Their results showed the proposed decision tree is capable of classifying all kinds of agriculture data.

Data of rural labor, arable land area and the gross output value of agriculture about 30 cities of China based on the decision tree was analyzed [11], and adopted clustering analysis method to discretize continuous data during the process of data mining in order to subjectivity comparing to the traditional classification methods. Forecasting the rainfall events using data mining techniques were adopted for understanding nature of rainfall[3]. The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduction in crop productivity. India is an agricultural country and its economy is largely based upon crop productivity. Thus rainfall prediction becomes a significant factor in agricultural countries like India. Crop productivity based on climate parameters by using the machine learning techniques was predicted in a study conducted in the state of Madhya Pradesh [10].

The accurate prediction of different species of crop productivities across several districts could help a lot of farmers and others alike [7]. Data mining models were described to improve crop productivities prediction from previous agriculture information [9]. It also used to select a best crop by farmer, to plant depending on the weather situation and provides required information to prefer the suitable season to do better farming. This paper presents new research possibilities for the application of new classification methodologies to the problem of productivity prediction. Using these techniques, the crop productivity can be improvised and increase the income level of the farmer, will be increased.

Close perusal of reviewed literature indicated that the application of data mining techniques in the field of agriculture in Indian subcontinent is very limited. Some of the studies carried out in India, statistical tools were used to develop regression equations to find out the various parameters influence. The limitation of the earlier studies was either they are carried out in a smaller area or that the influence of climatic parameters were not considered. Therefore, the present study was carried out to develop innovative applications of data mining techniques in predicting influence of agro-climatic factors on crop production in Madhya Pradesh State.

METHODOLOGY

The present study was aimed to find out the influence of climatic parameters on crop production in selected districts of Madhya Pradesh. The selection of district has been made based on the area under that particular crop. Based on this criteria first top five districts in which the selected crop area is maximum were selected. The crops selected in the study is based on the predominant crops in the state. They are maize, paddy, soybean and wheat.

From the literature reviewed, it was found that the crop productivities are affected by many natural and artificial parameters. In this study only the natural parameters i.e., climatic parameters only were considered to arrive at its influence on crop growth assuming all other parameters that influence the crop productivity were not considered due to non availability of data and they vary with individuals and locations. Based on the availability of crop as well as climatic data on a continuous twenty years chronological series the parameters collected in the study are: rainfall, maximum temperature, minimum temperature, potential evaporation transpiration, cloud cover, wet day frequency and corresponding crop productivity data. For each crop, data of climatic and crop productivity were collected from top five districts in which the area of that particular crop is maximum and are presented in table 1

Table 1.Crops and the corresponding district identified in the analysis

Name of the crop	Districts selected				
Maize	Chindhwara	Dhar	Jhabua	Rajgarh	Shajapur
Paddy	Bhalaghat	Mandla	Rewa	Sidhi	Shadol
Soybean	Dewas	Hoshangabad	Raisen	Shajapur	Ujjain
Wheat	Guna	Sagar	Satna	Tikamgarh	Vidhisha

For each selected crop, a district was selected to find out whether a relationship can be arrived between crop productivity and climatic parameters selected in the study for understanding the trend and ease in presentation. Regression analysis was first carried out by considering crop productivity as a dependent variable and other variables such as rainfall, maximum temperature, minimum temperature, potential evaporate transpiration, wet day frequency and cloud cover as independent variables. Considering influence of one, two, three, four, five and six independent variables selected in the study carried out regression analysis. Comparing each independent variable with dependent variable in order to find out the improved R-square value was carried out. The R-square value was found highest when six independent parameters were considered, therefore, relationship between each independent variable with dependent variable was not reported.

The relationship between the independent and dependent variable can be written as:

$$Y = \text{Intercept} + b1x1+b2x2+b3x3+b4x4+b5x5+b6x6 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ 1$$

where, Y : Productivity, kg/ha ; x1: Rainfall, mm, ; x2: Maximum Temperature, oC, ;

x3 : Minimum Temperature, oC ; x4: Potential Evaporation transpiration, mm

x5: Wet Day Frequency, days, ; x6: Cloud Cover, % ;

b1, b2, b3, b4, b5, and b6 : are parameter estimates

In the second method of developing prediction model, ID3 and C4.5 are algorithms introduced for inducing Classification Models, also called Decision Trees were used[6]. From the data set of records were prepared which consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. A decision tree was determined on the basis of non-category attributes predicting correctly the value of the category attribute. Usually the category attribute takes only the values {true, false}, or {success, failure},{high,low} or something equivalent.

In the decision tree each node corresponds to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf. In the decision tree at each node is associated with the non-categorical attributes, which are most informative among the attributes not yet considered in the path from the root. This establishes what is a "Good" decision tree. Entropy is used to measure how informative is a node. C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the issues not dealt with by ID3: Avoiding over fitting the data, determining how deeply to grow a decision tree, reduced error pruning, rule post-pruning. handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with

differing costs, improving computational efficiency etc.

C 4.5 uses information gain as its attribute selection measure. This measure is based on information theory. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or impurity in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple tree is found.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i=1,…,m). Let si be the number of samples of S in class Ci. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where pi is the probability that an arbitrary sample belongs to class Ci and is estimated by si/s.

Let attribute A have v distinct values {a1, a2, … av}. Attribute A can be used to partition S into v subsets, {S1, S2, …,Sv} , where Sj contains those samples in S that have values aj of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let sijbe the number of samples of class Ci in a subset Sj. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + .. + s_{mj}}{s} I(s_{1j}, ..., s_{mj})$$

$$\frac{s_{1j} + ... + s_{mj}}{s}$$

The term acts as the weight of the jth subset and is the number of samples in the subset (i.e., having value aj of A) divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset Sj

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij})$$

where pij is the probability that a sample in Sj belongs to class Ci.

Encoding the information that would be gained by branching on A is

Gain (A) = I (s1, s2,…,sm) – E(A)

In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

In summary, decision tree induction algorithms have been used for classification in a wide range of application domains. Such systems do not use domain knowledge. The learning and classification steps of decision tree induction are generally fast.

RESULTS AND DISCUSSIONS

The commonly followed regression model approach was used initially to explain the limitations of such models in crop productivity predictions. Therefore, regression analysis was used to develop a regression model for predicting crop productivity

from the climatic input parameters. Though, the analysis was carried out for all the crops selected in each district, however, analysis of one crop i.e., soybean in Dewas district was presented for simplicity in understanding.

The input data of rainfall, ambient temperature, potential evapo-transipiration, wet day frequency, cloud cover and productivity were used in the regression analysis are presented in the following table 2.

Regression analysis of Soybean crop in Dewas district

Table 2: Climatic parameters and soybean productivity in Dewas district for continuous 20 years

S.No	Rainfall, mm	Max Temp, ^{o}C	Min Temp, ^{o}C	PET, mm	Wet Day Frequency, days	Cloud Cover, days	Productivity, kg/ha
1	972	31.9	19.8	6.39	4.5	38.3	1898
2	1545	31.5	19.9	6.44	5.7	38.3	1680
3	1167	31.8	19.3	6.43	5.0	38.5	1615
4	911	31.4	19.9	6.53	4.2	35.1	1651
5	1069	31.4	19.8	6.45	4.6	39.9	2721
6	1128	32.0	19.5	6.42	4.8	39.6	2261
7	1204	31.8	19.2	6.40	4.6	40.3	1640
8	872	32.0	19.4	6.50	3.8	36.2	2262
9	997	32.2	19.7	6.44	4.4	39.6	1984
10	924	32.6	20.0	6.48	4.4	38.4	2665
11	1220	31.6	19.1	6.43	4.8	38.8	1620
12	949	32.8	19.2	6.44	4.2	37.9	2100
13	1275	32.0	19.5	6.46	4.9	38.0	1630
14	953	32.0	19.4	6.47	4.4	38.0	2205
15	912	32.2	19.6	6.47	4.3	38.0	1890

16	1006	32.4	19.9	6.53	4.5	39.0	1925
17	831	31.8	19.3	6.43	3.9	38.9	1651
18	1484	31.5	19.0	6.39	5.2	39.6	1720
19	823	32.1	19.5	6.46	3.9	38.3	2021
20	907	32.3	19.7	6.47	4.1	38.3	2154

Regression analysis was carried out by calculating analysis of variance of the variable with six degrees of freedom. The analysis of variance of different climatic parameters and maize productivity and the parameter estimate of multiple regression equation are presented in following tables 3 and 4.

Table 3: Analysis of variance of different climatic parameters and soybean productivity

Source	Degrees of Freedom	Sum of squares	Mean square	F- Value	Pr> F
Model	6	789548	131591	1.25	0.3454
Error	13	1372385	105568		
Corrected total	19	2161933			

Table 4. The parameter estimates of multiple regression equation developed for soybean crop in Dewas district

Variable	Parameter estimate	Standard Error	Type II, SS	F-Value	Pr> F
Intercept	-21062	19243	126470	1.20	0.2936
Rainfall	-1.92	1.24	252100	2.39	0.1463
Maximum Temperature	148.94	245.25	38932	0.37	0.5541
Minimum Temperature	-239.49	299.14	67662	0.64	0.4378
Potential Evapotranspiration	2910.69	2724.06	120529	1.14	0.3047
Wet Day Frequency	655.39	547.49	151277	1.43	0.2527

Cloud Cover	83.96	89.45889	93001	0.88	0.3651

Using the equation no.1, the regression equation for predicting soybean productivity in Dewas district can be written as: Productivity = -21062-1.92 x1+148.94 x2-239.49 x3+2910.69 x4+655.39x5+83.96 x6.Using the above equation crop productivity data was predicted and was compared with observed productivity and are presented in table 5.

Table 5: Predicting soybean productivity in Dewas district using multi regression equation

S.No.	Observed productivity, kg/ha	Predicted productivity, kg/ha	Deviation, mm
1	1898	1841.0	56.9
2	1680	1582.7	97.2
3	1615	2034.9	-419.8
4	1651	1792.3	-141.3
5	2721	2227.9	493.0
6	2261	2038.6	222.4
7	1640	1782.4	-142.3
8	2262	1842.6	419.3
9	1984	2056.9	-72.8
10	2665	2188.5	476.4
11	1620	1828.0	-208.0
12	2100	2116.5	-16.5
13	1630	1798.0	-168.0
14	2205	2152.9	52.1
15	1890	2117.9	-227.9
16	1925	2278.5	-353.5
17	1651	2041.7	-390.6
18	1720	1554.5	165.5
19	2021	2009.2	11.8
20	2154	2049.9	104.1

The above developed empirical model has a R square of 0.365 indicating about 37% of variation in the productivity can only be explained by these variables and remaining 63% of variation is from other variables which were not included in this model. Though, the average absolute deviation between predicted and observed productivity for 20 observations was found as 216.7 mm, culling out the most influencing parameter among the selected parameters on the crop productivity is not possible. Like this for the other districts also we cannot find the most influential parameter.

Prediction of crop productivity using C4.5 algorithm

ID3 and C4.5 are algorithms introduced for inducing Classification Models, also called Decision Trees were used in developing web based user friendly software [6]. The algorithms for calculation of Entropies of each independent variables and gains were developed. The web based software has been developed in C# language in .net platform as an user interface page, where the registered users can give their input data pertaining to their geographical area of study. On the home page of the web site (www.cropadvisor.in) the methodology adopted in the study, contact information of the administrator,

new user registration and registered users login are appeared.

This website is designed as an interactive software tool for predicting the influence of climatic parameters on the crop productivities. C 4.5 algorithm is used to find out the most influencing climatic parameter on the crop productivities of selected crops in selected districts of Madhya Pradesh. This software provides an indication of relative influence of different climate parameters on the crop productivity, other agricultural input parameters responsible crop productivity were assumed uniform across the study area due to paucity of desired information across the study area. Based on the C 4.5 algorithm, decision tree and decision rules have been developed, which are displayed when icon decision tree is selected.

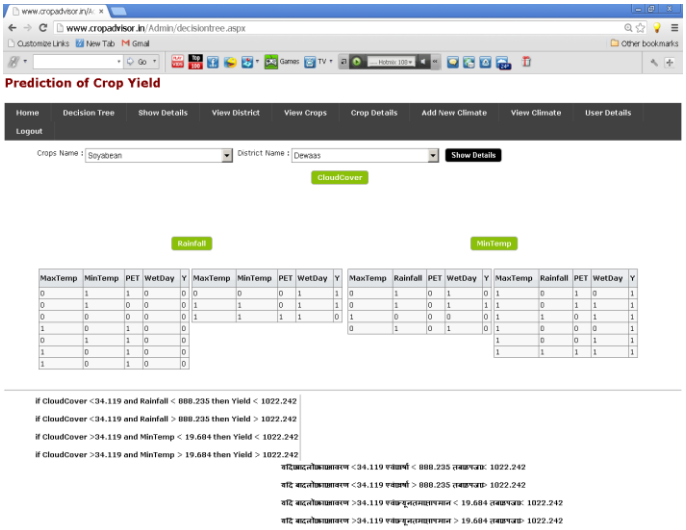


Figure 1: Screen shot of Soybean crop productivity prediction with decision tree and rules in Dewas district

Using the developed software the influence of climatic parameters on crop productivity in selected districts of Madhya Pradesh was carried out for predominant crops. For Soybean crop in all the selected districts, the most influencing parameter was found to be cloud cover, for paddy crop it was found as rainfall, for maize crop it was maximum temperature and for wheat crop the minimum temperature.

Validation of any developed software is essential for proving its efficacy in giving out the desired output from the given set of inputs. The software developed in this study is a web based application which provides user friendly output from the given set of inputs. Therefore, the decision rules that were framed from the developed software were used for validation of the software by predicting the productivities of selected crops in all the selected district with the observed values. The prediction accuracy was also worked out by comparing the predicted productivity with the observed productivities. For each crop the validation of the developed software has been carried out. However, in this paper, the validation of soybean crop in Dewas crop is explained detailed and for other districts and corresponding crops, overall prediction accuracies were presented for better understanding of the work.

Validation carried out for Dewas district for soybean crop is presented below:

The decision rules developed based on the model developed are:

if cloud cover is < 34.1 days and rainfall is < 888.0 mm then soybean productivity is < 1022 kg/ha

if cloud cover is < 34.1 days and rainfall is > 888.0 mm then soybean productivity is > 1022 kg/ha

if cloud cover is > 34.1 days and min.temp is < 19.7^{o}C then soybean productivity is < 1022 kg/ha

if cloud cover is > 34.1 days and min.temp is > 19.7^{o}C then soybean productivity is > 1022 kg/ha

Based on the above decision rules the observed values of the most influencing parameters of this district were tabulated along with the observed productivity of selected crop (Table 6)

Table 6: Prediction accuracy of developed model for soybean crop in Dewas district

Cloud cover, days	Rainfall, mm	Min. temp, oC	Observed Productivity, kg/ha	Predicted Productivity, kg/ha	Is prediction accurate
35.0	833	19.7	1093	> 1022	Yes
34.0	1161	19.2	1061	> 1022	Yes
34.0	93	19.8	1061	> 1022	Yes
32.6	1033	20.0	1008	> 1022	No
34.9	1068	19.3	1022	< 1022	Yes
34.7	833	20.2	1247	> 1022	Yes
35.2	962	19.8	1147	> 1022	Yes
31.8	694	19.9	1010	< 1022	Yes
35.2	694	19.7	1160	> 1022	Yes
33.1	692	20.1	986	< 1022	Yes
34.3	1168	19.0	1099	< 1022	No
33.8	824	19.2	925	< 1022	Yes
34.1	878	19.8	1275	>1022	Yes
33.9	851	19.5	912	< 1022	Yes
33.8	802	20.1	907	< 1022	Yes
34.5	915	20.1	1023	> 1022	Yes
34.5	687	19.5	880	< 1022	Yes
35.0	1293	19.3	780	< 1022	Yes

34.0	709	19.6	920	< 1022	Yes
34.0	728	19.7	930	< 1022	Yes

Out of 20 years of data the predictions were correct in 18 years and were incorrect in two years indicating the prediction accuracy of the developed model at 90 per cent in case of soybean in Dewas district.

Overall accuracy of prediction by developed model for all the crops selected and district selected is presented below:

For soybean crop for all the selected districts, cloud cover was found to be most influencing parameter on crop productivity followed by other climatic parameters. In Dewas district the soybean productivities were influenced by cloud cover, rainfall and minimum temperature. Based on the decision rules developed, the prediction accuracy of the model was worked out as 90 per cent for soybean crop in Dewas district. Similarly in Hoshangabad, Raisen, Ujjain and Shajapur districts the developed decision rules predicted the soybean productivities at an accuracy of 85, 95, 90 and 75 per cent respectively. The overall prediction accuracy of the developed model for soybean crop in selected districts was found to be 87 per cent.

For paddy crop for all the selected districts, Rainfall was found to be most influencing parameter on crop productivity followed by other climatic parameters. In Balaghat district the paddy productivities were influenced by Rainfall, Wet Day frequency and Minimum Temperature. Based on the decision rules developed, the prediction accuracy of the model was worked out as 90 per cent for paddy crop in Balaghat district. Similarly in Mandla, Rewa, Sidhi, and Shadol districts the developed decision rules predicted the paddy productivities at an accuracy of 90, 70, 90, and 85 per cent respectively. The overall prediction accuracy of the developed model for paddy crop in selected districts was found to be 85 per cent.

For Maize crop for all the selected districts, Maximum Temperature was found to be most influencing parameter on crop productivity followed by other climatic parameters. In Chindhwara district the maize productivities were influenced by Maximum Temperature , Potential Evapo Transpiration and Rainfall. Based on the decision rules developed, the prediction accuracy of the model was worked out as 85 per cent for maize crop in Chindhwara district. Similarly in Jhabua, Rajgarh, Shajapur and Dhar districts the developed decision rules predicted the maize productivities at an accuracy of 85, 70, 70 and 70 per cent respectively. The overall prediction accuracy of the developed model for maize crop in selected districts was found to be 76 per cent.

For Wheat crop for all the selected districts, Minimum Temperature was found to be most influencing parameter on crop productivity followed by other climatic parameters. In Guna district the wheat productivities were influenced by Minimum Temperature , Rainfall and Wet Day Frequency. Based on the decision rules developed, the prediction accuracy of the model was worked out as 90 per cent for wheat crop in Guna district. Similarly in Sagar, Satna, Vidhisha and Tikamgarh districts the developed decision rules predicted the wheat productivities at an accuracy of 75, 85, 70 and 80 per cent respectively. The overall prediction accuracy of the developed model for maize crop in selected districts was found to be 80 per cent.

The web based software developed for predicting the crop productivity from the given input of climatological parameters indicated a clear trend of each crop being predominantly influenced by a particular climatic parameter. The average of accuracy obtained under a particular crop in different district were averaged and the prediction accuracy of developed model for different crops are presented in table 7.0.

Table 7: Prediction accuracy of developed model for different crops.

S.No.	Name of the Crop	Average prediction accuracy, %
1	Soybean	87
2	Paddy	85
3	Maize	76
4	Wheat	80

The prediction accuracy of the developed model varied from 76 to 90 per cent for the selected crops and selected districts. Based on these observations the overall prediction accuracy of the developed model is 82.00 per cent. With a high prediction accuracy the developed model can be used by the policy makers in arriving at a policy decision well in advance i.e., before the harvest of the crop. The study demonstrated the use of developed software for extracting useful information from existing secondary data of climatic parameters to predict the crop productivity. It can also be concluded from the study for fast and easy prediction of crop productivities from the given set of climatic data, the developed software is a very useful tool.

CONCLUSIONS

Data mining tools predict the future trends and behavior allowing the businesses to make proactive knowledge driven decisions. Due to global climatic changes, the agricultural production is being influenced. Application of data mining techniques in predicting the agricultural production helps the policy makers to arrive at prior strategies, however, such applications are limited in India. In the present study, using the data from secondary sources application of data mining techniques for crop production was attempted. The continuous twenty years climatic parameters and the major crops productivity data of selected districts were utilized in the present study. Several methods of predicting and modeling crop productivities have been developed in the past with varying rate of accuracy, as these don't take into account the characteristics of the weather, and are mostly empirical, in the present study a software tool named 'Crop Advisor' has been developed as an user friendly web page for predicting the influence of climatic parameters on the crop productivities. C4.5 algorithm is used to find out the most influencing climatic parameter on the crop productivities of selected crops in selected districts of Madhya Pradesh. This software provides an indication of relative influence of different climatic parameters on the crop productivity, other agro-input parameters responsible for crop productivity are not considered in this tool, since, application of these input parameters varies with individual fields in space and time. The present study showed that the process of application of data mining techniques were successful in extracting the knowledge from the existing information. The study is based on decision tree to assess the influence of climatic factors on the crop productivity. The prediction accuracy of the developed model varied from 76 to 90 per cent for the selected crops and selected districts. Based on these observations the overall prediction accuracy of the developed model is 82.00 per cent. With a high prediction accuracy the developed model can be used by the farmers and policy makers in arriving at expected productivity trend well in advance i.e., before the harvest of the crop.

The conclusions of the study includes: i) the decision tree analysis indicated that the productivity of a crop productivity is influenced by climatic parameters. These parameters are different for different crops, however, they are common for the selected district of a particular crop, ii) the decision tree and decision rules of the present study are helpful in understanding and much to be desired as representations of knowledge interpretations, iii) rules formed from the decision tree are helpful in predicting the conditions for the above average or low average of crop productivity under the given

climatic parameters, iv) C4.5 algorithm can be successfully used in handling the continuous data of climatic parameters , v) user friendly web based software (www.cropadvisor.in) can be used by the end user by giving the desired input for arriving at productivity prediction.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support rendered by Dr. Bharat Misra, Associate Professor, Mahatma Gandhi Chitrakoot Gramodya Vishwavidyala, Satna, India and Dr. CD Singh, Principal Scientist, ICAR-Central Institute of Agricultural Engineering. She also would like to extend her gratitude to colleagues at AISECT University, Bhopal for their ever lasting encouragement in finalizing this manuscript.

REFERENCES

[1] Cunnigham S.J. and Holmes G. Developing innovative application in agriculture using data mining. Proceedings of 3rd International Symposium on Intelligent Information Technology in Agriculture. Beijing, China. October: 20-25, 2005.

[2] Jun Wu, Anastasiya Olesnikova, Chi-Hwa Song, Won Don Lee. The Development and Application of Decision Tree for Agriculture Data. 2009.IITSI: 16-20.

[3] Kannan, M., Prabhakaran S.and.Ramachandran, P. Rainfall forecasting using data mining technique. International Journal of Engineering and Technology Vol.2 (6), 2010, 397-401.

[4] Kiranmai, C., Murali Krishna, I.V., and Venugopal Reddy A..Data Mining Of Geospatial Database for Agriculture Related Application. Proceedings of Map India. New Delhi, 2006. 32-36.

[5] Nain A. S., Dadhwal,V. K. and T, Singh, T.P. Real time wheat productivity assessment using technology trend and crop simulation model with minimal data set. Current Science. 2002. 82(10): 1255-1258.

[6] Quinlan, J.R.. Learning efficient classification procedures and their applications to chess and games. In R.S.Michalski J.G et.al., Machine learning: an artificial intelligence approach. Palo Alto: Tioga Publishing Company, 1983a, 30-35.

[7] Shakil Ahamed A. Navid, T.M, and Tanzeem Mahamood. Applying data mining techniques to predict annual productivity of major crops and recommend planting different crops in different districts in Bangladesh. Software Engineering, Artificial Intelligence, Network and Parallel /Distributing Computing(SNPD). 1-3 June. 2013. Japan

[8] Singh G, Chandra Hukum. Production and economic factors growth in Indian agriculture. Technical Bulletin, Pub. by Central Institute of Agricultural Engineering. 2003, 1-25.

[9] Sujatha R and Isakki P. A study on crop productivity forecasting using classification techniques. Computing technologies and Intelligent Data Engineering. 7-9 Jan. 2013

[10] Veenadhari S, Mishra B, Singh C.D. Machine learning approach for forecasting crop productivity based on climatic parameters. Computer Communication and Informatics(ICCCI). 3-5 Jan. 2014India

[11] Yi-Yang, Gao and Ren Nan-Ping. Data mining analysis of our agriculture based on decision tree. ISECS International Colloquium. 8-9th August.2009 China: 134-138.

Enhancing the Efficiency of image and video forgery detection using convolutional neural networks

Ms. Sonal Pramod Patil [1*], Shital Shivnarayan Jadhav [2,] Hiralal Bhaskar Solunke[3]

[1] Assistant Professor, Computer Science and Engineering Department, G H Raisoni Institute of Business Management, Jalgaon

2 Assistant Professor, Computer Science and Engineering Department, G H Raisoni Institute of Business Management, Jalgaon

3 Assistant Professor, Computer Science and Engineering Department, G H Raisoni Institute of Business Management, Jalgaon

Abstract: Convolutional neural networks (CNNs) have become a de-facto technique for classification of multi-dimensional data. Activation functions like rectified linear unit (ReLU), softmax, sigmoid, etc. have proven to be highly effective when doing so. Moreover, standard CNN architectures like AlexNet, VGGNet, GoogLenet, etc. further assist this process by providing standard and highly effective network layer arrangements. But these architectures are limited by the speed due to high number of calculations needed to train and test the network. Moreover, as the number of classes increase, there is a reduction in validation and testing accuracy for the networks. In order to remove these drawbacks, we propose a hybrid CNN architecture, that adds a bio-inspired layer to the existing CNN architecture in order to improve the accuracy and speed of forgery classification. The developed system was tested on both images and videos for different kinds of forgeries, and it was observed that the proposed system obtains more than 98% accuracy for both testing and validation sets.

Keywords: Image, video, forgery, convolutional, neural, bio-inspired, hybrid

1. Introduction

Convolutional neural networks take into consideration a large number of features in order to classify image and video data into different classes. These features are evaluated using features masks, which include, but are not limited to,

- Horizontal feature mask
- Vertical feature mask
- Diagonal feature mask
- Color feature mask
- Edge feature mask, and many more

These masks work in different strides. Each stride covers a particular part of the image, and is selected such that there is limited over-sampling and maximum feature coverage. The following figure 1 indicates the feature coverage of CNNs. Which indicates that even the simplest of CNNs have a large number of features for comparison,

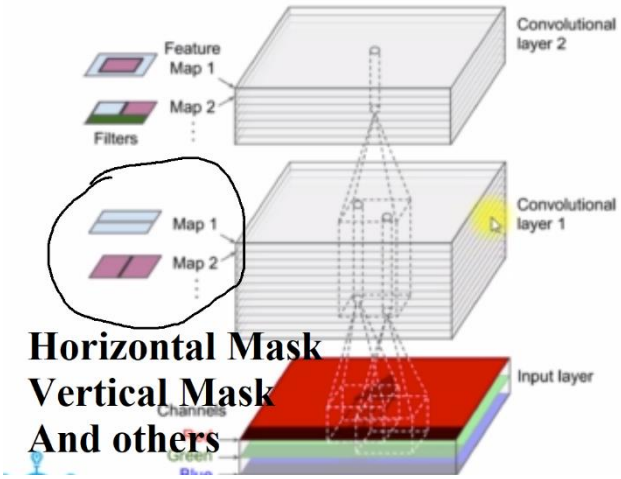


Figure 1. CNN Feature maps

Moreover, each of the feature maps is cascaded with other feature maps in order to increase the number of

final features for classification. The classification process combines multiple neural nets in order to obtain the final class. The combination requires different layers to be connected in a manner that achieves higher accuracy. Standard architectures like AlexNet, VGGNet, GoogLenet, etc. have been proposed for the same. The following figure 2 indicates the architecture of GoogLeNet which was proposed by Google for CNN-based classification systems,

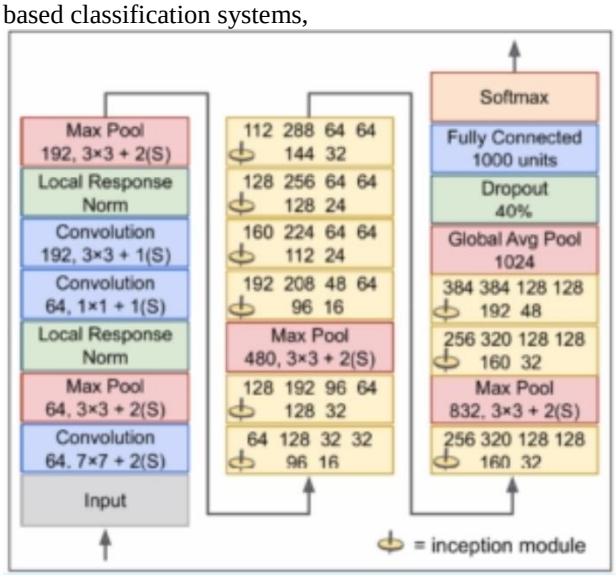


Figure 2. GoogLeNet CNN architecture

Using these architectures image and video forgery can be detected and classified. In this paper, we have used a standard VGGNet architecture and combined it with a self-designed bio-inspired layer. This not only improves the accuracy of VGGNet, but also increases the speed of operation of the overall system. The design details for the proposed system is described in section 3 of this paper. The next section describes works done by different researchers over the years in image and video forgery detection, followed by the proposed network design, and finally we conclude this text with some interesting observations about the proposed system, and future research that can be done for further checking the system performance.

2. Literature review

Since the inception of CNNs in the late 90s, there has been much research in the field of tuning the networks for better accuracy performance. Moreover, CNNs play a major role in finding out the best feature vectors suited for accurate classification, and once a sufficient level of validation accuracy is obtained, then the selected features are given to a flat neural network for accurate classification. Various CNN architectures have been proposed over the years for evaluation of forgeries, and each of them has their own unique network layer design. For instance, the work in [1] uses an improved mask regional CNN, that adds a Sobel-based filter to masks of recurrent CNNs. Adding this simple filter reduces the complexity of the CNN by more than 20%, and thereby increases the speed of training and evaluation. They have worked on copy-move and splicing forgeries, and are able to classify them with an accuracy of 95%. This accuracy is the training-phase accuracy, and can be increased by using standard CNN architectures like the ones mentioned in [2]. Here researchers have utilized the standard AlexNet CNN architecture and combined that with wavelet transforms for better feature evaluation. The CNN is further attached with a classic support vector machine (SVM) for better classification performance on splicing forgeries. The system is able to achieve 95% accuracy on validation sets, which is a very good performance considering that CASIA 1 and CASIA 2 datasets have been used. JPEG forgeries have always been difficult to classify using standard classifiers, but CNNs make the task achievable with good accuracy. The work in [3] uses deep CNNs for detecting non-aligned JPEG forgeries. The system works with same and different quantization matrices, thereby increasing the complexity of the CNN. Their work is tested in UCID and RCID and RAISE datasets, and showcases an accuracy of above 90%, which can be further improved by using standard architectures like GoogLenet and VGGNet. Another work similar to [1] is mentioned in [4], wherein researchers have used masked CNN in order to improve the accuracy of splicing classification. In this work, a new architecture called as ResNet-conv is proposed, which removes the feature pyramid network in ResNet-FPN and replaces them with a set of convolutional layers. The accuracy on a computer-generated dataset was found out to be 96%, which is sub-optimal considering the fact no standard dataset was used for evaluation. We would recommend researchers to further evaluate this system with standard datasets in order to obtain the final resulting accuracy.

Video forgery detection has been inspired from image forgery detection, and has similar algorithms for performing the tasks. The video forgery is further divided into inter-frame and intra-frame forgeries. The inter-

frame forgery is same as image forgery, and uses the exact same algorithms as used for image forgery detection. But inter-frame forgery detection algorithms must take into consideration the information from different sets of frames in order to detect forgeries. In order to do so, the work done in [5] can be considered as a defacto standard. In this work, the researchers have used CNNs for detecting forgeries from H.264/AVC sequences. The CNN is used to estimate compression parameters, then a frame-delta calculation layer combined with key-frame identification is used in order to localize tempered areas. The proposed algorithm obtains an accuracy which is slightly more than 70%, and thus a lot of improvement can be done in this area. Trigonometric transforms and standard CNN architectures can be used to improve its accuracy further. This can be shown from [6], wherein these techniques have been used for a better system performance. They have used different transforms and concluded that Fourier transforms give the best performance for detection of forgeries. Using DCT and DFT in 2D, a training set accuracy of 100% can be achieved. This has to be tested on the testing set, but as per the evaluations given in the text, it can be predicted than an accuracy of more than 95% can be achieved with proper selection of training and testing sets. Video forgeries can be detected by combination of more than one algorithm for classification. This can be proven from [7], wherein researchers have used CNNs with moth search optimization in order to improve the accuracy of video forgery detection. It uses multiple operations like differential arrays in horizontal, vertical and diagonal directions; thresholding and Markov transition probability analysis. The proposed system is able to classify video forgeries with more than 90% accuracy, and be further improved by introduction of AlexNet, and VGGNet architectures.

Fast shallow CNNs from [9] are applied by researchers in [10] for boundary-based image and intra-frame video forgeries. They have used the CASIA 1.0 and CASIA 2.0 datasets in order to perform this task. The evaluations on low resolutions images using fast shallow CNN (FSCNN) is commendable, but can further be improved. Usually the accuracy on low resolution images is between 80% to 90%, while for high resolution images it is between 90% to 98%. Similar to [2], the work done in [11] also uses CNNs for better classification performance for forgery detection in images. But they use a spatial rich approach which improves the weight calculations for the CNNs, thereby the CNNs train quicker, and have better validation performance. An accuracy of more than 94% is achieved on both CASIA 1 and CASIA 2 datasets. Another work on copy-move forgery detection is done in [12], wherein pre-processing layers are added that can be applied to both images and videos for effective forgery detection. The system is able to achieve an accuracy of more than 90% on validation sets, but can be further improved with the VGGNet architecture. Similar to [3], the work done in [13] uses deep-nets for classifying forgeries in the JPEG images, it further localizes these forgeries using a post-processing CNN. The results obtained using this 2 stage CNN outperform other methods by more than 20%. The proposed method is able to identify the region forgeries with more than 90% accuracy, and localize them with more than 85% accuracy. Due to such overwhelming figures, this network can be utilized for giving highly effective forgery detection system designs. Another copy-move forgery detection algorithm is devised in [14], which uses a combination of Generative Adversarial Network (GAN) and Convolutional Neural-Network (CNN). This combination yields in detecting forgeries with an accuracy of more than 95%. Due to such a high performance, this system is used by many researchers for real-time forgery detection. The GAN produces different masks which are supported by the CNN, and the combination of outputs of both of these masks is given to a copy-move forgery detection (CMFD) classifier, that takes the best of both the networks in order to find out the final forgery status of the image/video. The concept of GAN is very unique and must be further explored in other systems. The work in [15] uses a hybrid system that includes a frame absolute difference layer to cut down temporal redundancy between video frames, a max pooling layer to reduce computational complexity of image convolution, and a high-pass filter layer to enhance the residual signal left by video forgery. This system is highly effective in identifying the forgery locations in the video, and thereby improving the overall system efficiency. The proposed system achieves more than 95% training and testing set accuracies on different video datasets. Our algorithm is also inspired by the works given in [7] and [15], and uses a self-adaptive ML algorithm to obtain high system performance. This model is described in the following section, followed by the performance analysis of the same.

3. Proposed hybrid CNN Algorithm

The proposed hybrid machine learning based CNN algorithm used to detect forgeries in images videos, works in two phases,

- Intensive + Incremental learning phase
- CNN evaluation phase

The intensive learning phase works in the following steps,

i. Initialize the learning parameters, such as,

Number of learning rounds = Nr

Number of learning solutions = Ns

Learning rate = Lr

Max features per solution = Fmax

Max classifiers per solution = Cmax

Max number of frames/images per solution = Imax (Imax is 1 for image forgery detection)

ii. For each round, for each solution which has to be changed in this round, perform the following to find a new solution,

a. Select random but forged frames from the video/image. Make sure that the number of frames is exactly Imax

b. Evaluate adaptive key-points from each of the frames

c. Select Fmax key-points from the given set

d. The selection of these Fmax key-points must be done based on the distance of key-points from one another. Due to the fact that closely connected key-points generally have the same information, while key-points which are far from each other have more information

e. Select Cmax number of random classifiers, from the following list of classifiers,

i. k-nearest neighbours (kNN) [16]

ii. Support vector machine (SVM)[17]

iii. Neural network with different layer configurations (NN)[18]

iv. Quadratic linear classifier (QL)[19]

v. Mahalanobis classifier (MH)[20]

vi. Random forest classifier (RF)[21]

vii. Naïve Bayes classifier (NB) [22]

f. Apply classifier learning for all the Cmax classifiers using these Fmax features on each of the Imax frames

g. Evaluate the accuracy of the classifier system, and mark it as the learning convergence for this solution, using the following formula,

$$Lc = \frac{\sum_{i=1}^{Imax} Ai/Ndi}{Imax} \dots (1)$$

$$Ai = Accuracy\ for\ i^{th}\ image$$

$$Ndi = Normalized\ Delay\ needed\ for\ processing\ the\ i^{th}\ image$$

h. The normalized delay is evaluated using the following formula,

$$Ndi = \frac{di}{\sum di} \dots (2)$$

$$where, di = delay\ needed\ to\ process\ the\ i^{th}\ image$$

i. Observe this solution, and keep it for ready reference

iii. Evaluate the learning convergence for each of the solutions, and then evaluate the learning threshold

$$Lth = \frac{\sum_{i=1}^{Ns} LCi}{Ns} * Lr \ \ldots (3)$$

$$where, Lr \ is \ the \ learning \ rate$$

iv. For each solution which satisfies equation 4, pass it onto the next round, else, discard the solution and replace it with the help of step (ii)

$$LCi > Lth \ \ldots (4)$$

v. Repeat steps (ii) to (iv) for Nr rounds, and prepare the following table at the end of the Nr round,

Sol. Num	Sel. CFs	Selected KPs	LC val	Accuracy (%)

Table 1. The intensive learning-table

vi. From the table 1, select the solution with highest value of LC and highest value of accuracy and use it for the execution phase. In the table, CF stands for classifier, KP stands for key points

Due to the intensive learning phase, we get a large number of solutions, which are kept for further evaluation in the actual execution phase. The following steps are performed in the actual execution phase,

i. Select the best accuracy entry from the learning table 1

ii. For each of the input frame in the video, apply the feature selection as mentioned in the 3rd column

iii. Apply the classifiers as mentioned in the 2nd column of table 1, and evaluate if the image is forged or not

iv. Inject random training set entries for evaluation, and repeat steps (i) to (iii) for these random entries

v. Evaluate the accuracy for these random entries, and evaluate the value of Lc for each of these entry sets

vi. If the value of Lc for a given set is lower than the one selected from table 1, then update table 1 with this value

vii. Select the next best entry from table 1, and repeat the process for each of the frames

viii. In case more than half of the entries of table 1 are replaced, then retrain the algorithm with the help of the pre-execution step, and re-create table 1 with better entries of Lc

Due to the continuous learning process which takes place in this algorithm, the overall system's accuracy improves, and we get a better forgery detection rate than any of the individual algorithms. The adaptive key-point based algorithm for feature selection is designed using the following steps,

- Key point extraction
- Training the system
- Evaluation of forgery using adaptive key point selection

Initially the input images are given to a key point extraction algorithm, which evaluates Maximal Stable Extremal Regions [23] (also known as MSER), and applies Speed up Robust Features (SuRF) [24] method on each of these regions. The SuRF method evaluates key-points for each of the regions, the distribution and values of these key points assists in evaluation of the image properties. Once the key-points are evaluated, they are tagged using the training phase. The training phase basically stores information about the image key-points along with the tag whether the image is forged or not, and if it is forged then the type of forgery (splicing, copy-move, etc.) with which the image is tampered. This information helps in identification of the key-points which are later used for comparison by the adaptive key-point selection algorithm and identify the forgery type. The tagged database along with the full key-point features of the image to be tested are given to the adaptive key-point classifier. The adaptive key-point classifier works via the following process,

- The query image features are distinguished with the help of the feature length for a given region
- For each region, the number of key-points which are equal to the key-points in the database are compared using standard key-point matching and a score value

S1 is calculated based on the number of features matched

- For other unequal length features, the following process is followed,
    - If the number of features of the query image are less than the number of features of the database entries, then the database entries are trimmed and comparison is done to find out the score S2
    - If the number of features of the query image are more than the number of features of the database entries, then the input image feature entries are trimmed and comparison is done to find out the score S2
- The total score of the image is evaluated using, S=S1+S2+S3
- This is done for each database entry and scores are evaluated for all of them
- Finally, the entry with maximum score is selected and classification is done
- Based on this classification, we obtain whether the input image is forged or non-forged

Once the system gives the result about the forgery status of the image/video sequence, then the selected features from the sequence are given to the VGGNet CNN model. The VGGNet model can be shown from the following figure,

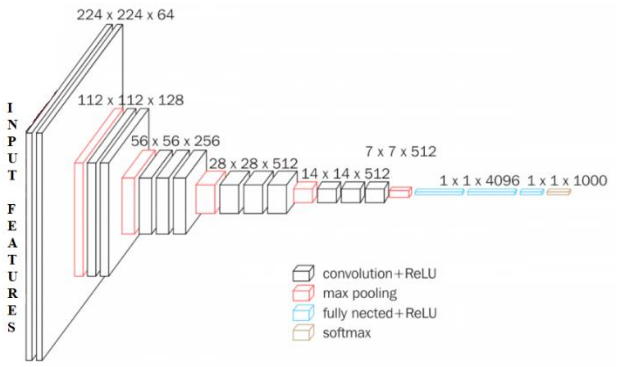


Figure 3. The VGG Net model [25-28]

The input image features are given to a 64-mask set unit, that evaluates 64 different feature masks from the input features. These feature masks include horizontal, vertical, diagonal, and other components. Each of these features is then given different ReLU based layers, where feature optimization is done, and finally the input features are classified into one of N classes. This addition of the ML layer to the VGG net model helps in reducing the training delay, and thereby improves the overall performance of the CNN. The result and evaluation of the proposed model is performed in the next section.

4.Results and observations

We compared the performance of the proposed model using CASIA 1 and CASIA 2 datasets. A training: testing: validation ratio of 7:2:1 was used while performing the evaluation. Different CNN architectures were compared, and the final performance evaluation is done. The results on a combined dataset for different algorithms can be seen from table 1,

Img. Tested	Acc. (%) DWT + LBP [2]	Acc. (%) NN with Cpix [3]	Acc. (%) RCNN [4]	Acc. (%) Bio-CNN [7]	Acc. (%) Multi CNN [8]	Acc. (%) Proposed CNN
10	90.00	90.00	100.00	100.00	100.00	100.00
25	95.00	93.00	95.00	100.00	100.00	100.00
40	95.20	96.00	96.80	97.10	100.00	100.00
50	95.60	96.50	97.30	97.60	98.60	100.00
75	95.60	96.70	97.60	97.80	98.70	100.00
100	95.70	96.80	97.80	97.90	98.70	99.30
200	95.70	96.70	97.90	98.10	98.80	98.90
500	95.70	96.80	97.90	98.20	98.80	99.40

Table 1. Accuracy of different classifiers

From the results it is evident that the accuracy of the proposed method outperforms other methods. But our method also outperforms other CNN implementations w.r.t. the speed of operation. This can be observed from the following table 2,

Img. Tested	Delay (ms) DWT + LBP [2]	Delay (ms) NN with Cpix [3]	Delay (ms) RCNN [4]	Delay (ms) Bio-CNN [7]	Delay (ms) Multi CNN [8]	Delay (ms) Proposed CNN
10	2.36	6.90	8.42	5.52	11.60	1.24
25	2.87	7.60	9.52	6.25	13.12	1.51
40	3.69	7.90	10.54	6.91	14.52	1.94
50	5.98	12.90	17.16	11.26	23.65	3.15
75	5.99	18.20	21.99	14.43	30.31	3.15
100	6.98	35.70	38.80	25.46	53.47	3.67
200	7.59	72.60	72.90	47.84	100.47	3.99
500	8.60	125.20	121.64	79.82	167.63	4.53

Table 2. Delay v/s number of images tested

The reduction in delay is due to the pre-processing done during the machine learning phase. Due to the pre-processing the number of feature sets needed for classification reduce drastically which reduces the overall training and evaluation delays. Of course, there is a delay in feature extraction, but that is infinitesimal when compared to the final delay of evaluation for the networks. The results can be observed with the help of the delay and accuracy graphs as shown in the following figures.

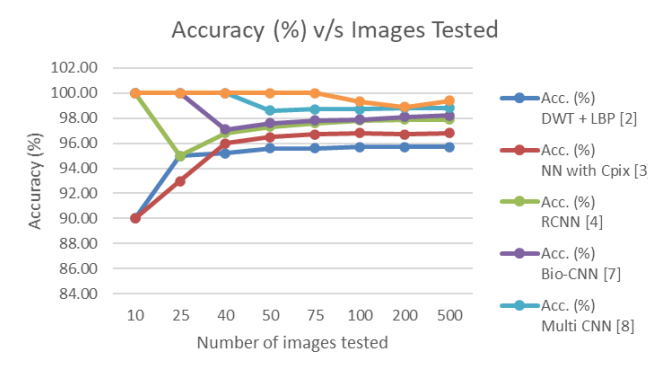


Figure 4. Accuracy of different algorithms

From figure 4 it is seen that there is a tough competition in this domain, as CNN performs considerably well, but the delay graph shown in figure 5 showcases the superiority of the proposed technique,

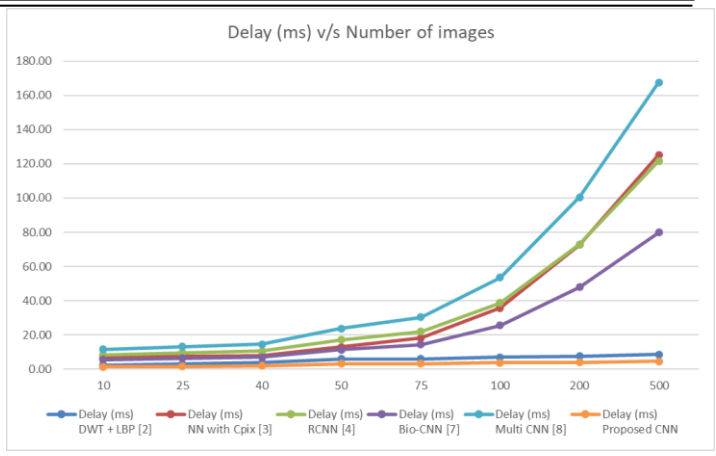


Figure 5. Delay performance of the proposed CNN

We developed the proposed system using Python 3.5 using the Tensor Flow and Keras libraries, and tested the same on the image datasets. The results from the proposed systems can be showcased in the figures 6, 7 and 8. It is clear from these figures that the system is able to quickly and effectively identify the number of forged images/video sequences from the set of input images/videos. Thus, it can be used for real-time scenarios where both speed and accuracy constraints are to be satisfied.

Figure 6. Forged image

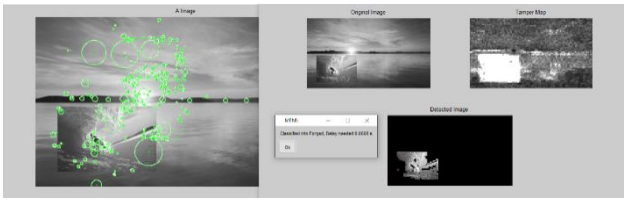


Figure 7. Forged video sequence

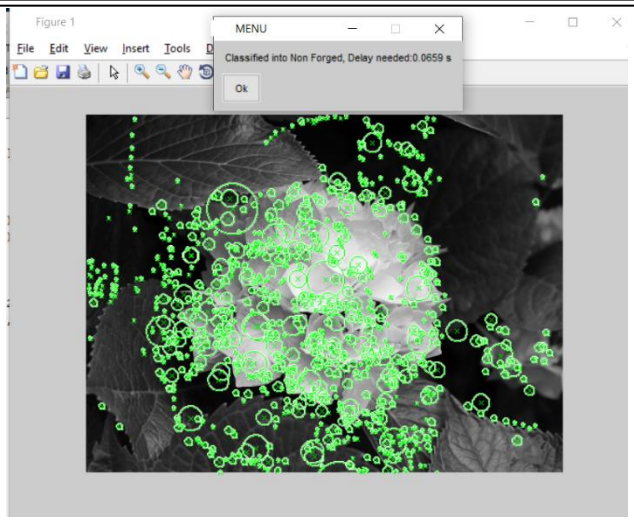


Figure 8. Non forged image

From these results we can observe that the developed system has high accuracy, and low delay of operation and thereby can be used in any kind of image/video forgery system.

5. Conclusion

The results indicate that the delay and accuracy have both been optimized by the proposed classifier. In order to showcase the improvement in both delay and accuracy, we can observe a comparison of mean delay and mean accuracy over different image/video frames is done, and it is seen that the delay is reduced by more than 40%, while the accuracy is improved by more than 5% than the most effective classifiers. These results encourage us to use the proposed system for real-time forgery detection. In future, GAN-based networks can be integrated with the proposed system in order to further improve its performance.

6. References

[1] A. Roy, R. Dixit, R. Naskar, and R. S. Chakraborty, "Copy-Move Forgery Detection with Similar But Genuine Objects," in *Digital Image Forensics*, ed: Springer, 2020, pp. 65-77.

[2] M. S. a. a. R. Hwaitat A. , "AN ENHANCED PARTICLE SWARM OPTIMIZATION USING FREQUENCIES WAVE SOUND (FPSO)," *Journal of Theoretical and Applied Information Technology*, vol. 97, 2019.

[3] B. Yang, X. Sun, H. Guo, Z. Xia, and X. Chen, "A copy-move forgery detection method based on CMFD-SIFT," *Multimedia Tools and Applications*, vol. 77, pp. 837-855, 2018.

[4] K. M. Hosny, H. M. Hamza, and N. A. Lashin, "Copy-for-duplication forgery detection in colour images using QPCETMs and sub-image approach," *IET Image Processing*, vol. 13, pp. 1437-1446, 2019.

[5] S. Kumar, J. Desai, and S. Mukherjee, "A fast keypoint based hybrid method for copy move forgery detection," *arXiv preprint arXiv:1612.03989*, 2016.

[6] J. Zheng, Y. Liu, J. Ren, T. Zhu, Y. Yan, and H. Yang, "Fusion of block and keypoints based approaches for effective copy-move image forgery detection," *Multidimensional Systems and Signal Processing*, vol. 27, pp. 989-1005, 2016.

[7] M. Puri and V. Chopra, "A survey: Copy-Move forgery detection methods," *International journal of computer systems*, vol. 3, 2016.

[8] Sonal Patil, Dr, K. N. Jariwala, "Improving the performance of image and video forgery detection using hybrid convolutional neural networks ," *Procedia Computer Science*, vol. 78, pp. 61-67, 2016.pp101-117

[9] S. Sadeghi, H. A. Jalab, K. Wong, D. Uliyan, and S. Dadkhah, "Keypoint based authentication and localization of copy-move forgery in digital image," *Malaysian Journal of Computer Science*, vol. 30, pp. 117-133, 2017.

[10] H. A. Alberry, A. A. Hegazy, and G. I. Salama, "A fast SIFT based method for copy move forgery detection," *Future Computing and Informatics Journal*, vol. 3, pp. 159-165, 2018.

[11] Y. Sun, R. Ni, and Y. Zhao, "Nonoverlapping Blocks Based Copy-Move Forgery Detection," *Security and Communication Networks*, vol. 2018, 2018.

[12] S. P. Patil and K. Jariwala, "Digital Image Forgery Detection Using Passive Techniques by Means of Keypoint Classification," International Journal of Uncertainity, Fuzziness and Knowledge BAse Systems,Vol.29,

[13] M. Hashmi and A. Keskar, "Fast and Robust Copy-Move Forgery Detection Using Wavelet Transforms and SURF," *INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY*, vol. 16, pp. 304-311, 2019.

[14] P. Mukherjee and S. Mitra, "A review on copy-move forgery detection techniques based on DCT and DWT," *International Journal of Computer Science and Mobile Computing*, vol. 4, pp. 702-708, 2015.

[15] J.-C. Lee, "Copy-move image forgery detection based on Gabor magnitude," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 320-334, 2015.

[16] M. Mohan and V. Preetha, "Gabor filter—HOG based copy move forgery detection," *Journal of Electronics and Communication Engineering*, vol. 2, pp. 41-45, 2017.

[17] A. Hilal and S. Chantaf, "Uncovering copy–move traces using principal component analysis, discrete cosine transform and Gabor filter," *Analog Integrated Circuits and Signal Processing*, vol. 96, pp. 283-291, 2018.

[18] M. M. Ardakan, M. Yerokh, and M. A. Saffar, "A New Method to Copy-Move Forgery Detection in Digital Images Using Gabor Filter," in *Fundamental Research in Electrical Engineering*, ed: Springer, 2019, pp. 115-134.

[19] K. Asghar, Z. Habib, and M. Hussain, "Copy-move and splicing image forgery detection and localization techniques: a review," *Australian Journal of Forensic Sciences*, vol. 49, pp. 281-307, 2017

Deep Machine Learning Tracker for Real Time Objects Detection

[1]Mrs Sonal Tiwari, [2] Dr. Shailja Sharma, [3]Third Dr Sanjeev K Gupta
[1] First Research Scholar ,[2] Second Associate Professor, [3]Third Professor and Dean Academics
[1]sonal_infonet@yahoo.com,[2] shailja.sharma@aisectuniversity.ac.in, [3]sanjeevgupta07@yahoo.com

Abstract—***This paper proposes a novel tracker which is controlled by sequentially actions learned by deep reinforcement learning agent. The basic idea is to localize objects in a scene by finding dense regions in the scene and guide the visual attention to that part. To achieve this goal, the problem is formulated as a sequence of decision making tasks. Formulating the problem as a decision making task led to applying a variant of RL [3] algorithms called DQL [1] to solve it. Using Deep Learning (DL) and Q-learning, an attempt was made to learn representation for objects from different categories and guide an intelligent agent to focus on an object in a scene.***

Index Terms— ***Computer Vision, Deep learning, Reinforcement learning, Visual Tracking.***

I. INTRODUCTION

Visual tracking is one of the fundamental problems in the computer vision field. Finding the location of the target object is difficult because of several tracking obstacles such as motion blur, occlusion, illumination change, and background clutter. Conventional tracking methods [17, 42, 7, 15, and 13] follow target objects using a low-level handcrafted feature. Although they achieve computational efficiency and comparable tracking performance, they are still limited in solving the above-mentioned obstacles because of their insufficient feature representation.

Popular algorithms in object detection have mostly concentrated on methods where localization and detection are conducted separately. The basic idea of all these methods is first to localize objects in an image and then classify them. While the main idea is similar, different approaches have been used to increase efficiency and effectiveness [4, 5, 6]. However, recent methods have emphasized the approaches where end-to-end learning is applied. In this way both localization and classification can be done with a single pass through an image. These models merge the two basic steps into one model to conduct localization and classification simultaneously [7, 8, 9].

However, to find objects in an image,these algorithms cover every sub-region in an image. Unlike these algorithms, where every patch in an image is processed, human vision doesn't search for objects in each region. It is shown in [10] that the human vision system localizes objects by perceiving the whole scene and successively searching dense regions by sequentially turning visual attention to the important local areas.

We also propose a learning algorithm with reinforcement learning (RL) to train the model. We train our network to select actions to track the position of the target using samples extracted from training videos. In this step, the network learns to track general objects without sequential information. In the RL stage, the pre-trained network is used as an initial network. We perform RL via tracking simulation using training sequences composed of sampled states, actions, and rewards. The network is trained with deep reinforcement learning based on policy gradient [38], using the rewards obtained during the tracking simulation. Even in the case where training frames are partially labeled the proposed framework successfully learns the unlabeled frames by assigning the rewards according to the results of tracking simulation.

II. BACKGROUND AND RELATED WORK

The recent object detection algorithms that apply deep learning methods to overcome the problem, each method has used deep learning in a different way to enhance effectiveness and efficiency. As the methods evolve, they move closer to the idea of end to- end learning.

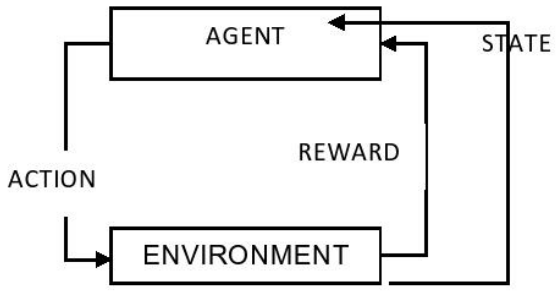


Fig 1 Reinforcement learning

Boykov and Huttenlocher [9] used a Kalman filter to track vehicles in an adaptive framework. In [9], the object parameters include position and configuration of non-occluded features. At each frame, a maximum a posteriori (MAP) estimation of the object parameters will be found. However, evolution and observation functions cannot always be modeled as linear functions. If fk and gk are non-linear functions,an extended Kalman filter (EKF) or Unscented Kalman Filter (UKF) can be used for optimization. Rosales and Scaroff [59] used the EKF to estimate 3D object trajectories from 2D image motion. And then the trajectory, occlusion and segmentation information are utilized in extracting stabilized views of the moving objects. The most commonly used method is a particle filter, which is based on the stochastic sampling method. A particle filter was first introduced by Isard and Blake [31] in computer vision. The advantage of a particle filter is the ability to handle arbitrary densities. If state space is discrete and the number of states is finite, hidden Markov models (HMM) can be used for tracking. Chen et al. [14] used the HMM formulation for tracking. In [14], a joint data association filter (JPDAF) was used to compute the HMM's transition probabilities, taking into account the intercorrelated neighboring measurement. Besides, multiple hypothesis filter (MHF) provides another way to evaluate the probability of measurement sequence. MHF can be use to track the modes of the state density. Cham and Rehg [11] utilized a variant of MHF for tracking highly articulated objects.

To improve efficiency, Comaniciu et al. [17] proposed a mean shift algorithm for nonrigid object tracking. The feature histogram-based target representations are regularized by spatial masking with an isotropic kernel. The masking induces spatially-smooth similarity functions suitable for gradient-based optimization. The target localization problem was formulated using the basin of attraction of the local maxima, and the mean shift procedure was used to perform the optimization. Furthermore, Dore et al. [18] proposed multicue adaptive particle filter-based tracker (MAPT) algorithm to track deformable objects. Shape and color cues were exploited to handle deformable objects.

As for tracking algorithms, the classical particle filter work well for tracking moving objects by steady cameras. But, these methods are not applicable for moving cameras because they are not robust to global appearance changes and sudden camera motion. Some existing methods for moving cameras are as follows. Meuter et al. [52] used UKF to estimate the movement of people. They assumed that targets and the camera are moving on the same plane and that some camera parameters are known. The moving host and target movements can be modeled as 2D movements on a flat ground-plane. The motion and the measurement model are combined by an UKF. Thus, this method can only be applied to specific environment because it has several constraints. Ess et al. [20] proposed a multi-person tracking algorithm.

People were detected using shape information and were tracked by a tracking-by detection approach. The method integrated stereo depth and visual odometry in the hypothesize-and-test model selection framework. However, this method needs a calibrated stereo rig to obtain depth information and can only track specific trained objects. For tracking algorithms used in moving cameras, the major problem is sudden camera motion because sudden camera motion not only largely changes object position, but also causes image blur. Image blur will cause object appearance changes.

Those situations happen frequently, especially for a camera in a car. To solve this kind of problem, the method proposed by Behrad et al. [5] used object detection to re-track the missing target. Kumar et al. [43] proposed an integrated method to overcome the sudden motion problem, but this method made use of both top-down and bottom-up approaches. Both approaches need foreground segmentation information. If foreground segmentation is required to assist tracking, there is less flexibility in embedded computing. Our goal is to develop tracking algorithms using an uncalibrated monocular camera from a moving platform. Moreover, the algorithm is expected to overcome the problem caused by sudden camera motion without using detection information.
The goal of reinforcement learning (RL) is to learn a policy that decides sequential actions by maximizing the cumulative future rewards [30]. Recent trends [23, 32, 28, 27] in RL field is to combine the deep neural networks with RL algorithms by representing RL models such as value function or policy. By resorting of the deep features, many difficult problems such as playing Atari games [23] or Go [27] can be successfully solved in semi-supervised setting. Also, several methods were proposed to solve the computer vision problems, such as

object localization [3] or action recognition [16], by employing the deep RL algorithms.

III. SCHEME FOR TRACKING CONTROLLED BY RL ACTIONS

Visual tracking solves the problem of finding the position of the target in a new frame from the current position. The proposed tracker dynamically pursues the target by sequential actions. The proposed networks predict the action to chase the target from the position of the current tracker. The tracker is moved by the predicted action from current state, and then the next action is predicted from the moved position. By repeating this process over the test sequence, we solve the object tracking problem.

3.1 Problem Statement

Basically our tracking strategy follows Markov Decision Process (MDP). The MDP is defined by states s € S, actions a €A, state transition function s= f(s, a), and the reward r(s, a). In our MDP formulation, the tracker is defined as an agent of which goal is to capture the target with a bounding box shape. The action is defined in a discrete space and a sequence of actions and states is used to iteratively pursue the resulting bounding box location and size in each frame. The agent decides sequential actions until finalizing the target's position, and then, goes to the next frame.

3.2 Action

The action space A consists of eleven types of actions including translation moves, scale changes, and stopping action as shown in Figure 2. The translation moves include four directional moves, {left, right, up, down} and also have their two times larger moves. The scale changes are defined as two types, {scale up, scale down}, which maintain the aspect ratio of the tracking target. Each action is encoded by the 11-dimensional vector with one-hot form.

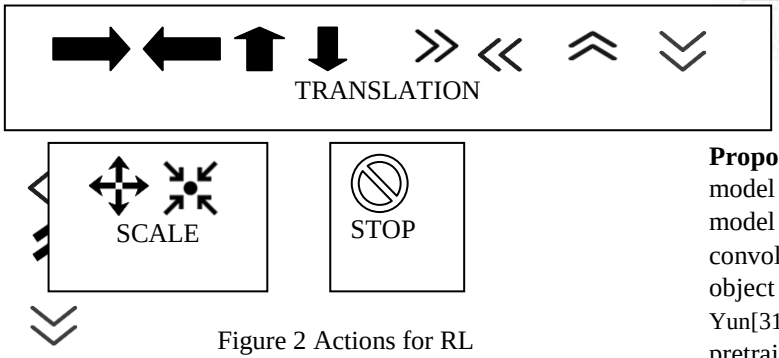


Figure 2 Actions for RL

State. The state st is defined as a tuple (pt, dt), where pt € $R^{112\times112\times3}$ denotes the image patch within the bounding box (we call simply "patch" in the following) and dt € R^{110} represents the dynamics of actions denoted by a vector (called by "action dynamics vector" in the following) containing the previous k actions at t-th iteration The patch pt is pointed by 4-dimensional vector bt = [x(t), y(t),w(t), h(t)], where (x(t), y(t)) denotes the center position and w(t) and h(t) denote the width and height of the tracking box respectively.

State transition function. After decision of action a_t in state s_t, the next state s_{t+1} is obtained by the state transition functions: patch transition function fp(·) and action dynamics function $f_d(\cdot)$. The patch transition function is defined by b_{t+1} = fp(bt, at) which moves the position of the patch by the corresponding action.

Reward. The reward function is defined as r(s) since the agent obtains the reward by the state s regardless of the action a. The reward $r(s_t)$ keeps zero during iteration in MDP in a frame. At the termination step T, that is, aT is 'stop' action R(s,s') is assigned by,

$$R(s,s') = \begin{cases} 1 & if\ IoU(b,g) \geq 0.7 \\ -1 & otherwise \end{cases}$$

where IoU(b ,g) denotes overlap ratio of the terminal patch position bT and the ground truth G of the target with intersection-over-union criterion. The tracking score zt is defined as the terminal reward, zt = r(sT), which will be used to update model in reinforcement learning.

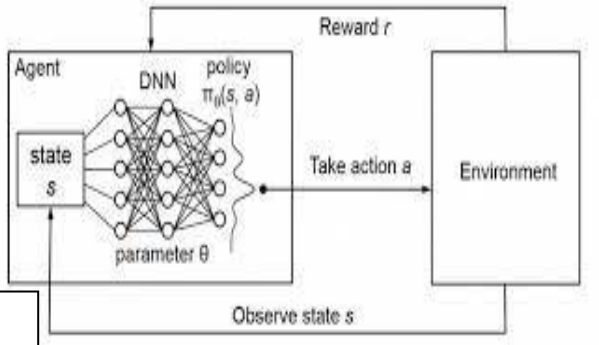


Figure 3 Theoretical Model

Proposed Network We proposed an improved and efficient model which predicts fast object detection. as shown in figure model contains three convolution layer {convol1, convol2, convol3} which are identical to model using features for object identification like VGG-M net[30] etc. Sangdoo Yun[31] used to initialize their model with VGG-M net as pretrained model also they take the different layer as motivated by this model for giving a tracker. The next FC4

fully connected layer combined with dropout Relu and has 512 nodes when we calculate output. The output of this layer is combining with action dynamics having 110 dynamics. The final layer FC5 having all calculated score with action probabilities which decide what action must be applied if agent repeat action and come again to previous than oscillation occurs already we give the way how to deal the situation .agent updates states while selecting sequential action until it reaches final state and stop the process .

IV. TRAINING PROPOSED NETWORK

In the reinforcement learning stage, network parameters WRL, ({w1, ...,w6}), except fc7 layer are trained. Training ADNet with RL in this section aims to improve the network by policy gradient approach [38]. The initial RL network WRL has the same parameters of the network trained by SL (WSL). The action dynamics dt is updated in every iteration by accumulating the recent k actions and shifting them in first-come-first-out strategy. Since the purpose of RL is to learn the state-action policy, we ignore the confidence layer fc7,which is needed in tracking phase.

We are using Q network which takes input as state representation and produce as output eleven action value for object identification as we using shallower faster RCNN which is trained in a way that is acts as pretrained network for Q network it having advantages that learning of Q function is relatively faster and RCNN will acts as feed forward for Q network.

To perform a training step with experience replay method, having taken action at at time t, the agent's experience is stored in a buffer with the format of $e_t = (s_t; a_t; r_t; s_{t+1})$. The buffer, $D_t = \{e_1; e_2; \ldots\ldots\ldots et\}$, is a list of experiences that the agent has faced during its interaction with the environment. In this way, during the learning process, instead of using one experience per step, which is an inefficient use of data, a batch of experiences is sampled from the buffer and the network is trained by those experiences. Despite efficient use of data, experience replay also separates correlation between sequences of repeated experiences in order to prevent bias training towards repeated experiences which is the result of the agent being trapped in a part of the environment. It is proposed to use one network and dataset to obtain a feature extractor and show how tuning the model for one category can improve the performance of the network for other classes.

V. EXPERIMENTS

We evaluated our method on the popular visual tracking benchmarks, Object Tracking Benchmark (OTB) [39, 40], comparing with existing trackers. Also, we validated the effectiveness of network by demonstrating various self comparisons. The experiments were conducted on the following specifications: i7-4790K CPU, 32 GB RAM, and cpu using python tensor flow.

Research Name	Prec.(20px)	IOU(AUC)	FPS	GPU/CPU
Proposed Method	91.6	0.712	< 1	O
MDNet [9]	90.9%	0.678	< 1	O
C-COT [1]	90.3%	0.673	< 1	O
DeepSRD CF [8]	85.1%	0.635	< 1	O
HDT [10]	84.8%	0.564	5.8	O
MUSTer [8]	76.7%	0.528	3.9	X
MEEM [11]	77.1%	0.528	19.5	X
SCT [2]	76.8%	0.533	40.0	X
KCF [6]	69.7%	0.479	223	X
DSST [4]	69.3%	0.520	25.4	X
GOTURN [5]	56.5%	0.425	125	O

We evaluated our method on two OTB datasets: OTB-50 [39], which has 50 video sequences, and OTB-100 [40], which has 100 video sequences including OTB-50. In order to pre-train ADNet, we used 360 training videos from VOT2013 [19], VOT2014 [20], VOT2015 [18], and ALOV300 [29], in which videos overlapping with OTB- 0 and OTB-100 were excluded. The tracking performance was measured by conducting a one-pass evaluation (OPE) based on two metrics: center location error and overlap ratio [39]. The center location error measures the distance between the center of the tracked frame and the ground truth and the bounding box overlap ratio measures the Intersection-over-Union (IOU) ratio between the tracked bounding box and the ground truth.

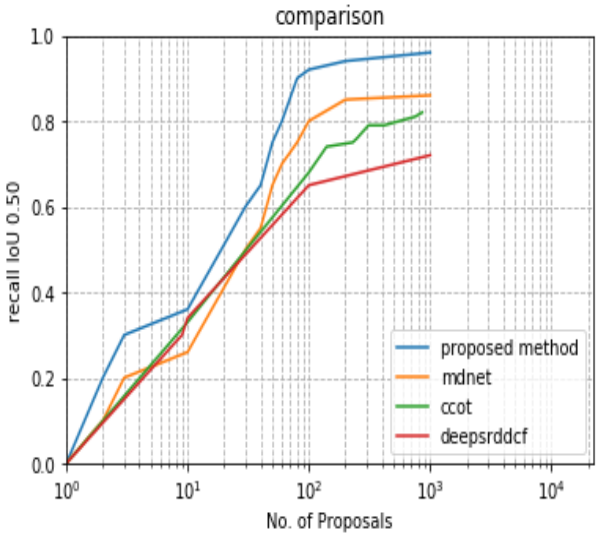


Figure 4 Precision and recall comparison with other methods

In the experiment, the ratio of the frames using re-detection to the whole frames was around 9%, and the ratio of the frames requiring more than five actions to capture the target to the whole frames was only around 4%, that is, most of the frames require fewer than five actions to pursue the target in each frame and this is the strength of this research.

We compared proposed research in a comprehensive comparison with diffrent state-of-the-art trackers including MDNet [24], CCOT[9], GOTURN2 [12], HDT [25], DeepSRDCF [8],SINT [31], FCNT [34], SCT [5], MUSTer[15], CNNSVM [14], MEEM [42], DSST [7], and KCF [13]. Figure 3 shows the plots of precision and success rate based on center location error and overlap ratio respectively and Table 1summarizes the comparison of tracking performance with computational speed (fps). The proposed method shows comparable performance with the state-of-the-art trackers, MDNet [24] and C-COT [9] in both precision and success rate. The proposed method is much efficient in computation, and is about three times faster than MDNet and C-COT. ADNet-fast, the fast version of ADNet, has a 3% performance degradation but runs in real-time (15 fps) and shows performance comparable with those of other CNNbased trackers such as DeepSRDCF [8] and HDT [25]. As shown in Table. 1, method achieved the best performance among the real-time tracking algorithms. The bounding box flow from the initial position to the captured target position is shown in the left-most columns and the sequential transitions of the state are represented by the image patches and the selected actions.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel tracker restricted by Reinforcement learning based novel model, which pursues the target object by sequential actions iteratively. To the greatest of our knowledge, it is the first effort to adopt the tracking strategy controlled by pursuing actions trained by deep reinforcement learning. Action-based tracking makes a considerable contribution to the reduction of computation difficulty in tracking. In addition, reinforcement learning makes the use of to some extent labeled data possible, which could greatly gain actual applications. According to the evaluation results, the proposed tracker achieves a state-of the-art performance in 3 fps, which is three times faster than the existing deep network-based trackers adopting a tracking-by-detection strategy. Furthermore, for some video sequence proposed tracker achieves a real-time speed (15fps) with an accuracy that outperforms state-of-the-art real time trackers.

In addition, adding a termination action to find all objects in an video brings the agent close to the idea of having a complete object detection algorithm. The agent would be able to count the number of objects and continues to search until it has localized all. It is noteworthy that adding a new action will introduce a new error and it is predicted that much more training would be needed in order to achieve acceptable results. Regarding counting the number of objects in video, a new variant of neural networks called Neural Arithmetic Logic Units (NALU) [27] have been recently introduced. While previously neural networks weren't able to generalize outside the range of numerical values confronted during training, NALU can learn systematic numerical computation. It can be used in partnership with Conv or RNN to learn counting the number of objects in sequence.

REFERENCES

[1] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In ECCV, 2016.

[2] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, andJ. Young Choi. Visual tracking using attention-modulated disintegration and integration. In Proceedings of the IEE Conference on Computer Vision and Pattern Recognition, pages 4321–4330, 2016.

[3]] J. Choi, H. Jin Chang, S. Yun, T. Fischer, Y. Demiris, and J. Young Choi. Attentional correlation filter network for adaptive visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[4] M. Danelljan, G. H¨ager, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[5] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. arXiv preprint arXiv:1604.01802, 2016.
[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Highspeed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3):583–596, 2015.
[7] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. arXiv preprint arXiv:1502.06796, 2015.
[8] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 749–758, 2015.
[9] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. arXiv preprint arXiv:1510.07945, 2015.
[10] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, and J. L. M.-H. Yang. Hedged deep tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[11] J. Zhang, S. Ma, and S. Sclaroff. Meem: robust tracking via multiple experts using entropy minimization. In European Conference on Computer Vision, pages 188–203. Springer, 2014.
[12] Hedi Harzallah, Fred. et al "Combining efficient object localization and image classification" Proc. IEEE Conference 12 Computer Vision, Sep 2009, pp. 237244.
[13] Endres et al. "Category independent object proposals" Computer Vision ECCV 2010, pp 575.588.
[14] I. Laptev, Gyuri Darko "Improvements of object detection using boosted histograms" in Proc BMVC, volume 3. 2006, pp. 949,958.
[15] P. Dollar,Zithik et al "Edge boxes Locating object proposals from edges" Proc. Eur. Conference Computer Vision, 2014, pp. 391,405.
[16] A. Zisserman, M. Everingham, et al "The pascal visual object classes (voc) challenge" IJCV, 88(2):303–338, 2010.
[17] Rajkumar Goel , Vineet Kumar, Saurabh Srivastava,A. K. Sinha "A Review of Feature Extraction Techniques for Image Analysis" ICACTRP 2017 Vol. 6, Special Issue 2, February 2017.
[18] D. G. Lowe et al "Object recognition from local scale-invariant features" Proc. IEEE Conference Computer Vision., Sept 1999, pp 115.1157.
[19] S. Ridella, et al "Learning algorithm for nonlinear support vector machines suited for digital VLSI" electron Lett., vol. 35 pp 1349.1350, August 1999.
[20] W. Dong, R. Socher, Fei-Fei et al "ImageNet A large-scale hierarchical image database" Proc. IEEE Conference Computer Vision Pattern Recognition (CVPR), Jun. 2009.
[21] J. Donahue ,R. Girshick et al "Rich feature hierarchies for accurate object detection and semantic segmentation" CVPR IEEE, 2014.
[22] J. Sun, R. Girshick, K. He et al "Faster R-CNN- Towards real-time object detection with region proposal networks" preprint arXiv 1506.01497, 2015.
[23] K. Simonyan et al "Very deep convolutional networks for large-scale image recognition" arXiv 1409, 1556 2014.
[24] D. McAllester, R. B. Girshick et al. "Object detection with discriminatively trained part based models" IEEE Trans on Pattern Analysis and Machine Intelligence,1627–1645 2010.
[25] A. A. Efros, A. Gupta et al Ensemble of exemplar-SVMs for object detection and beyond. In ICCV. IEEE, 2011.
[26] D. Silver, M. Riedmiller, V. Mnih et al. "Human-level control through deep reinforcement learning" Nature, 518,529-533, 2015.
[27] A. A. Efros., A. Gupta et al "Unsupervised discovery of mid-level discriminative patches" ECCV 2012 Springer.
[28] B. Schiele, J. Hosang et al "How good are detection proposals, really" arXiv 1406-6962, 2014.
[29] T. Hofmann, M. B. Blaschko et al "Beyond sliding windows: Object localization by efficient subwindow search" CVPR IEEE, 2008.
[30] A Vezhnevets, V Ferrari and A Gonzalez-Garcia "An active search strategy for efficient object class detection" CVPR IEEE conference, 2015.
[31] Aleksis Pirinen and Cristian Sminchisescu "Deep Reinforcement Learning of Region Proposal Networks for Object Detection" IEEE conference CVF 2018.
[32] Q. Dong,G. Bai et al "Cube CNN SVM- A novel hyper spectral image classification method" Proceeding IEEE 28 Conference ICTAI, November 2016, pp. 1027-1034.
[33] R. Girshick, K. He et al "Faster R-CNN Towards real time object detection with region proposal networks" Proceeding Advance Neural Info. Process System, 2015, pp. 91-99.
[34] Sutton R S "Introduction to reinforcement learning" volume – 135 book on RL.
[35] D. Silver, D. Wierstra et al "Playing atari with deep reinforcement learning" ar X iv preprint arXiv 1312.-5602, 2013.
[36] G. Lever,T. Degris,M. Riedmiller et al "Deterministic policy gradient algorithms" International Conf. on Machine Learning ICML,2014.
[37] D. Silver,A. Guez and H. Van Hasselt "Deep reinforcement learning with double q-learning" CoRR, 1509.06461,2015.
[38] D. Silver,K. Kavukcuoglu et al "Playing atari with deep reinforcement learning" arXiv preprint arXiv-1312-5602,2013.
[39] M. Lanctot, N. de Freitas et al "Dueling network architectures for deep reinforcement learning" arXiv 1511-06581, 2015.
[40] L. Itti et al. "Computational modeling of visual attention" Nature reviews Neuro science 194-203, April 2001.

LSTM based mobility prediction in Ad-Hoc Network

[1]Subrata Debbarma, [2] Dr. Rakesh Kumar
[1]Department of Computer Science, Assam University, Silchar, India, [2] Department of Computer Science, Assam University, Silchar, India
[1]Email:er.sub83@gmail.com, [2] Email: rakesh_rbl@rediffmail.com

Abstract- Because of the easiness to deploy and to extend networks, now day's ad-hoc network trends to integrate with fix infrastructure network to extend internet services and to make available of M-IoTs. However, most challenging problem of ad hoc network is how to adapt the movement of nodes consisting of a network the adaptability to the mobility impact on the network performance. In this paper, we propose a Long Short-term Memory-RNN for mobility prediction. It predicts node location trajectory in ad-hoc network in order to improve the quality of services (QoS) by improving network connectivity. Experimental results on node location trajectory using model and real-world based dataset, we demonstrate that the proposed prediction method has significant improvements on generalization ability and reduces prediction error.

Keywords- Ad-Hoc Network, Mobility Prediction, Long Short Term Memory (LSTM)

I. Introduction

Mobility prediction becomes more and more important in ad hoc networks to make available the network resources and to provide QoS. Ad-hoc network is an infrastructure-less wireless network dynamically reconfigurable and moves in randomly manner. Ad hoc network are easy and flexible to install deploy in application such as emergency response, community, home, vehicle and sensor network, where are each node act as self-router and self-host in collaboration with others node within the network. Mobility presents a challenging issue and most challenging problem in ad hoc network is how to adapt the mobility of the nodes consisting of a network, the adaptability to the mobility affects the performance of the network. So mobility prediction of a node in advance is the solution to overcome these challenges. Mobility prediction of a node is the estimation of their future locations. In ad hoc network location estimation means its geographical coordinate estimation, its main advantage is to estimate link connectivity with time in order to improve network performances [5]. Location prediction is a particular case of time series prediction, time series is a set of observations from past until present, time series prediction is to estimate future observations. The use of Long Short-term Memory in time series prediction is an increasing trend. Mobility prediction problem using LSTM can be resolved by exploiting long-term location time series prediction which is obtain from GPS (Global Positioning System) receiver.

The rest of this paper is organized as follows. Section II presents some related works. Section III formulates the mobility prediction problem and describes the dataset. Section IV introduces the LSTM architecture and learning algorithm for prediction. In Section V, we give the experimental results and discussion. We finally some concluding remarks are presented in Section VI.

II. Related Works

Mobility prediction is a particular case of time series prediction of node trajectory. Artificial Neural Network (ANN) can be resolved by exploiting short or long term location time series prediction. With the advantages of mobility prediction in ad-hoc network early method have been attempt to model mobility prediction using node location trajectory.

In [1] - [4] were proposed location prediction by several Markov model. R. Nagwani and D. Singh Tomar [1] proposed Hidden Markov Model (HMM) based node location trajectory prediction, to reduce connection interruption and to upgrade QoS in MANET. Authors claim that the proposed method route rediscovery time is lesser than Dynamic Source of Routing (DSR) protocol. Mieso K.

Denko [2] presented a technique of mobility management scheme based on Markov model and found accurate prediction for lower number of prediction steps and higher order-Markov models. Furthermore better prediction can be achieved for smaller network sizes and lower randomness.
Shaojie Qiao et al [3] proposed trajectory prediction algorithm to discover transition rules from one location to another and present new strategies for solving the discontinuous hidden state chain and the state retention problems using HMM. Qiujian Lv et al [4] presented the user mobility prediction system using HMM. Effective user living habits are analyzed on the models of spatio-temporal predictor and next place predictor, select one to apply as per user interest points. However, these Markov model based prediction methods are discovering transition rules from one location to another and it is computationally cost. Apart from Markov Model mobility prediction methods and modeling widely used based on ANN were proposed in [5] - [11]. H. Kaaniche and F. Kamoun [5] proposed link duration estimation of nodes by predicting trajectory node location using RNN and for long time series prediction. To minimize the error function Backpropagation through time (BPTT) algorithm was used for training the network. Model based RWP were used to calculate the efficiency of prediction. Mohamed Elleuh et al [6] proposed Adaptive Neuro-Fuzzy Inference System (ANFIS) based prediction techniques for node location trajectory. The network is learned using back propagation algorithm to minimize a set or measure a defined error. The prediction accuracy depend on the number of input neurons and membership function (MF). Shang Y., Guo W., Cheng S [7] used clustering approach that makes use of intelligent mobility prediction based on the wavelet neural network to estimate more stable clusters. Lahouari Ghouti et al [8] presented the extreme Learning Machine (MLE) single layer feed-forward architecture for mobility prediction in MANET. The proposed prediction method examine the prediction efficiency using two synthetic data generated by BonnMotion tool based on Gauss-Markov and Random Walk mobility model. ELM outperforms then MLPs. Ghouti L [9] compared MLP and ELM based mobility prediction in MANET, to examine the prediction efficiency used several synthetic dataset as well as real-world dataset and the comparison result presents the ELM based predictor has attained very low prediction errors in the terms of MSE or MAE. Nermin Makhlouf [10] proposed three layer feedforward networks to calculate link duration to predict the future mobility of a node position. The backpropagation algorithm is used to learn the neural network. RWM model create location pattern for the propose system. Y. Yayeh, H. Lin et al [11] deep learning technique a five-layered neural network has been proposed for mobility prediction to predict based on the node movement history to know mobile stations current mobility information based on pause time, speed and movement direction. RWM model create location pattern for the propose system. However, these methods faces problem for long term time series prediction and these methods are discovering node trajectory sequences and suffer from the data sparsity problem. In addition, other conventional machine learning techniques such as DT, NN and SVR also been applied for mobility prediction in [12] to predict coordinates as continues variables regression based three machines learning i.e. DT, NN and SVR were applied. The main metrics for comparison were accuracy (MSE) and duration (Second). In terms of accuracy NN is best followed by DT and in terms of duration or speed DT is best followed by NN. However, these methods need discrete location, thus not applicable to node trajectories composed of sequence coordinates with frequent time intervals. C. Wang, L. Ma, et al. [13] proposed a basic LSTM framework for human mobility prediction focus on single-user trajectory prediction and experiment on a model-based mobility dataset. Afterward for multi-user multi-step prediction propose a region oriented prediction scheme and put forward an LSTM-based Seq2Seq framework. Experiments on a realistic dataset show that the proposed framework outperforms from other competing approaches

As we described above the researchers were done mobility prediction with different mobility methods using ANN in order to increase the network performance. This paper deals LSTM based mobility prediction in Ad-Hoc network.

III. Problem Formulation and location data description

In this section, the location prediction problem is formulated followed by the location dataset description.

A. Problem formulation

Defining a problem. Mobility prediction of a node is the estimation of their future locations and the node location means its geographical coordinates. We assume that each node are able to learn its location using GPS receiver so it can periodically record its geographical location, all the recorded location define the node trajectory. We denote a

trajectory as $X = \{l_1, l_2, \ldots\}$ where $l = (x_t, y_t)(t = 1, 2, \ldots)$ is a two dimensional coordinate representing the node location (figure 1) at time t. The time interval between each two adjacent points is fixed.

Our main objective is to predict the sequence of the next *N* step location points. Knowing its previous location, $X = \{l_1, l_2, \ldots, l_T\}$, $\{(i = 1, 2, \ldots T)$ where T is the length of location trajectory}, a mobile node can predict their future location points, $Y = \{(l_{T+1}\ l_{T+2} \ldots l_{T+N-1} \ldots l_N)\}$, where $l = (x_T, y_T)\ (T = 1, 2, \ldots T_N)$ is a two dimensional coordinate representing the predicted node location at time T, N is the length of predicted location trajectory.

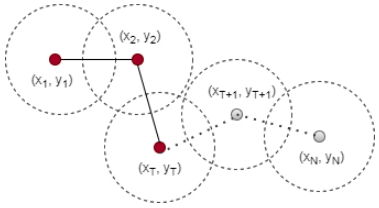


Fig.1 Location Coordinates

B. *Location Data Description*

To evaluate the performance of the mobility prediction we adapt both model-based and real-world mobility dataset. Model based mobility traces are generated in BonnMotion tool and the real-world data traces are collected from CRAWDAD Data Set.

Model-based mobility traces data. We used BonnMtion tool in this study to generate model based dataset. BonnMtion tools is a java based simulation environment which is use to generate model based mobility dataset, invented in the University of Bonn, Germany to investigate wireless ad hoc network characteristics[16]. Several mobility models can be used in this tool to generate synthetic mobile dataset. However, in this paper we used Gauss-Markov mobility model, the simulation parameters are shown in the table 1.

Table 1 Simulation parameters

Parameters	Values
Mobility Model	Gauss Markov
Simulation Area (X & Y)	1000m X 1000m
Number of Nodes (n)	10
Duration (d)	4000 sec
Ignore (i)	10
Minimum node speed (l)	0 m/sec
Maximum node speed(h)	20 m/sec
Update frequency()	2.5s
Angle Standard Deviation()	0.392
Speed Standard Deviation	0.5
Uniform speed()	False

Real-World mobility traces data. We have collected real-world dataset which is contributed by Ana Aguiar [17]. The data traces sequence of geographical coordinates are latitude, longitude and timestamp. For recording geographical coordinates used android phone GPS traces and placed into the four fire fighters during a forest fire exercise.

IV. LSTM based prediction

In this section we put forward a Long Short Term Memory based mobility predictor model. LSTM is an advanced Recurrent Neural Network that can learn long-term dependencies; it has been used in advanced machine learning practices such as time-series prediction. The efficient of the prediction technique depends on the choice of the network architecture and learning algorithm.

A. *LSTM background*

LSTM is a special type of RNN which is motivated to deal long term dependencies data to overcome exploding and vanishing gradients problem faces by RNN to process long term data. In contrast to the RNN, the RNN is simple and the structure is loop the network modules but LSTM have more complicated structure then RNN extended structure with gates and cell state. In each hidden layer LSTM consist of three gates to control the cell state. The cell state store the learned information with time steps for future references which is control by the gates namely input gate, forget gate and output gate. The cell state update information control by input gate, the input gate decides the activation to entry into the cell state. An output gate determines the output activation to output new cell state and flow to next network. The forget gate help to forget the old information and reset the network [13]. As illustrated in Figure 2 LSTM memory blocks, the figure shows function of three gates and cell state.

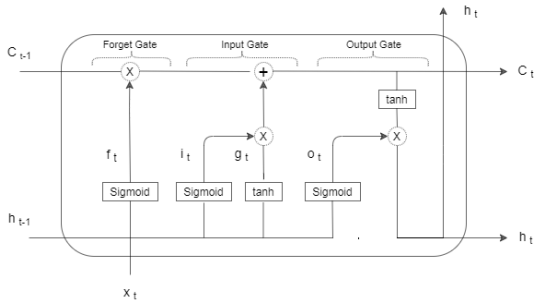


Fig.2 LSTM Memory blocks

Assume that x_t and h_t represent the input and output at the current time-step, h_{t-1} and c_{t-1} is the output and cell state at the previous time-step respectively. The cell state at time step t is given by $c_t = f_t \odot c_{t-1} + i_t \odot g_t$ the output (hidden) state at time step t given by $h_t = o_t \odot \tanh(c_t)$, *Where* $\odot$ denotes the Hadamard product (element-wise multiplication of vectors). The key component at time step t of the LSTM scheme is given below [18] [19].

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f)$$
$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o)$$
$$g_t = \tanh(W_g x_t + R_g h_{t-1} + b_g)$$

Where W, R, b, i, f, g, o and σ denotes the input weights, recurrent weights, bias, input gate, forget gate, layer input, output gate and sigmoid function respectively. The activation function given by $\sigma(x) = (1 + e^{-1})^{-1}$

B. LSTM architecture for prediction

Here we used LSTM to predict the future location of a node. Figure.3 present the

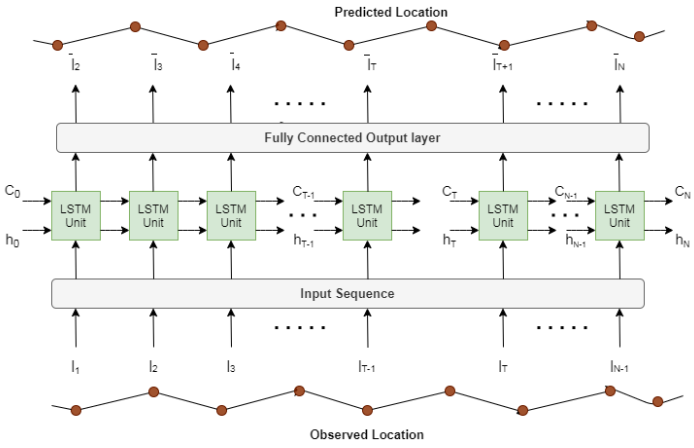


Fig.3 LSTM prediction model

Proposed LSTM based single user mobility predictor model, the architecture composed of three layers, first layer is the input layer with one feature sequence values shifted by one time step, and the given input sequence is processed by the input layer so that each two-dimensional coordinate sequence is mapped to LSTM block. Then, the processed sequences is sent to the second layer name LSTM layer and specify the LSTM layer to have 200 hidden units which is the major part of the predictor model. The LSTM network learns the value each time step to predict the value of next time step, for training the network we specify the training options adam as a solver, training epoch 250, learning rate piecewise learning schedule, initial learning rate 0.005 and learning drop rate 125 with factor 0.2 respectively.

At time step t, the LSTM network by using previous time-step output h_{t-1} and cell state c_{t-1} compute the output of the next time step and updated cell state c_t (shown in figure 2). LSTM layer takes the output of the previous layer as input and feeds its output to the next layer. Finally, the third layer an fully connected output layer maps the output of the LSTM layer at each time-step to a two-dimensional coordinate and compute the prediction sequences, assume l_{T+1} as the predicted location of the next time-step, and thereby we get the prediction sequences Y= {(l_{T+1} $l_{T+2} \ldots$ $l_{T+N-1} \ldots$ l_N).

C. Training Algorithm

The main aim of the training is to minimize the error between the predicted location and the actual location. Thus, we choose the Root Mean Square Error (RMSE) as the loss function and adopt Backward Propagation through Time (BPTT) algorithm[18][19] to update the network parameters.

Let the input at time t in the LSTM cell be, the cell state from time $t-1$ and t be c_{t-1} and c_t and the output for time $t-1$ and t be h_{t-1} and h_t the initial value of c_t and h_t at $t = 0$ will be zero.

Step 1: Forward Pass through different gates

Inputs: x_t, h_{t-1} and c_{t-1} are given to the LSTM cell.

Passing through input gate

$$Z_g = W_g x + R_g h_{t-1} + b_g$$
$$g = tanh(Z_g)$$
$$Z_i = W_i x + R_i h_{t-1} + b_i$$
$$i = \sigma(Z_i)$$

Input gate out = $g * i$

Passing through forget gate

$$Z_f = W_f x + R h_{t-1} + b_f$$
$$f = \sigma(Z_f)$$
$$Forget\ gate\ out = f$$

Passing through the output gate

$$Z_o = W_o x + R_o h_{t-1} + b_o$$
$$o = \sigma(Z_o)$$
$$Out\ gate\ out = o$$

Calculating the current cell state c_t

$$c_t = c_{t-1} \odot f + g \odot i$$

Calculating the output gate h_t

$$h_t = o \odot \tanh(c_t)$$

Step 2: Backward Pass Calculating the gradient through back propagation through time at time stamp t using chain rule.

Let us consider E to be the error function, and is generally referred to the gradient pass down by the above cell be $E_\delta = \frac{\partial E}{\partial h}$, if we are using MSE (mean square error) for error then,

$E_\delta = (x - h(x))$

Here x is the orignal value and $h(x)$ is the predicted value.

Gradient with respect to gates and cell sate as below

Output gate:

$$\frac{\partial E}{\partial o} = \left(\frac{\partial E}{\partial h_t}\right) * \left(\frac{\partial h_t}{\partial o}\right) = E_\delta * \left(\frac{\partial h_t}{\partial o}\right)$$
$$\frac{\partial E}{\partial o} = E_\delta * \tanh(c_t)$$

Current cell sate c_t :

$$\frac{\partial E}{\partial c_t} = \left(\frac{\partial E}{\partial h_t}\right) * \left(\frac{\partial h_t}{\partial c_t}\right) = E_\delta * \left(\frac{\partial h_t}{\partial c_t}\right)$$
$$\frac{\partial E}{\partial c_t} = E_\delta * o * (1 - \tanh^2(c_t))$$

Input gate $\frac{\partial E}{\partial i}$, $\frac{\partial E}{\partial g}$

$$\frac{\partial E}{\partial i} = \left(\frac{\partial E}{\partial c_t}\right) * \left(\frac{\partial c_t}{\partial i}\right)$$
$$\frac{\partial E}{\partial i} = E_\delta * o * (1 - \tanh^2(c_t)) * g$$

Similarly

$$\frac{\partial E}{\partial g} = E_\delta * o * (1 - \tanh^2(c_t)) * i$$

Forget gate:

$$\frac{\partial E}{\partial f} = E_\delta * \left(\frac{\partial E}{\partial c_t}\right) * \left(\frac{\partial c_t}{\partial f}\right)$$
$$\frac{\partial E}{\partial f} = E_\delta * o * \left(1 - tanh^2(c_t)\right) * c_{t-1}$$

Previous cell state c_{t-1}

$$\frac{\partial E}{\partial c_{t-1}} = E_\delta * \left(\frac{\partial E}{\partial c_t}\right) * \left(\frac{\partial c_t}{\partial c_{t-1}}\right)$$
$$\frac{\partial E}{\partial c_{t-1}} = E_\delta * o * \left(1 - tanh^2(c_t)\right) * f$$

Gradient with respect to learnable weights as below

Output gate weights:

$$\frac{\partial E}{\partial W_o} = \left(\frac{\partial E}{\partial o}\right) * \left(\frac{\partial o}{\partial W_o}\right) = E_\delta * \tanh(c_t) * \sigma(z_o) * (1 - \sigma(z_o) * x_t$$
$$\frac{\partial E}{\partial R_o} = \left(\frac{\partial E}{\partial o}\right) * \left(\frac{\partial o}{\partial R_o}\right) = E_\delta * \tanh(c_t) * \sigma(z_o) * (1 - \sigma(z_o) * h_{t-1}$$
$$\frac{\partial E}{\partial b_o} = \left(\frac{\partial E}{\partial o}\right) * \left(\frac{\partial o}{\partial b_o}\right) = E_\delta * \tanh(c_t) * \sigma(z_o) * (1 - \sigma(z_o)$$

Forget gate weights:

$$\frac{\partial E}{\partial W_f} = \left(\frac{\partial E}{\partial f}\right) * \left(\frac{\partial f}{\partial W_f}\right) = E_\delta * o * (1 - tanh^2(c_t)) * c_{t-1} * \sigma(z_f) * (1 - \sigma(z_f) * x_t$$

$$\frac{\partial E}{\partial R_f} = \left(\frac{\partial E}{\partial f}\right) * \left(\frac{\partial f}{\partial R_f}\right) = E_\delta * o * (1 - tanh^2(c_t)) * c_{t-1} * \sigma(z_f) * (1 - \sigma(z_f) * h_{t-1}$$

$$\frac{\partial E}{\partial b_o} = \left(\frac{\partial E}{\partial f}\right) * \left(\frac{\partial f}{\partial b_o}\right) = E_\delta * o * (1 - tanh^2(c_t)) * c_{t-1} * \sigma(z_f) * (1 - \sigma(z_f)$$

Input gate weights:

$$\frac{\partial E}{\partial W_i} = \left(\frac{\partial E}{\partial i}\right) * \left(\frac{\partial f}{\partial W_i}\right) = E_\delta * o * (1 - tanh^2(c_t)) * g * \sigma(z_i) * (1 - \sigma(z_i) * x_t$$

$$\frac{\partial E}{\partial R_i} = \left(\frac{\partial E}{\partial i}\right) * \left(\frac{\partial i}{\partial R_i}\right) = E_\delta * o * (1 - tanh^2(c_t) * g * \sigma(z_i) * (1 - \sigma(z_i) * h_{t-1}$$

$$\frac{\partial E}{\partial b_i} = \left(\frac{\partial E}{\partial i}\right) * \left(\frac{\partial i}{\partial b_i}\right) = E_\delta * o * (1 - tanh^2(c_t) * g * \sigma(z_i) * (1 - \sigma(z_i)$$

$$\frac{\partial E}{\partial W_g} = \left(\frac{\partial E}{\partial g}\right) * \left(\frac{\partial g}{\partial W_g}\right) = E_\delta * o * (1 - tanh^2(c_t)) * i * (1 - tanh^2(z_g)) * x_t$$

$$\frac{\partial E}{\partial R_g} = \left(\frac{\partial E}{\partial g}\right) * \left(\frac{\partial g}{\partial R_g}\right) = E_\delta * o * (1 - tanh^2(c_t)) * i * (1 - tanh^2(z_g) * h_{t-1}$$

$$\frac{\partial E}{\partial b_g} = \left(\frac{\partial E}{\partial g}\right) * \left(\frac{\partial g}{\partial b_g}\right) = E_\delta * o * (1 - tanh^2(c_t)) * i * (1 - tanh^2(z_g))$$

Using all gradient, we can easily update the weights associated with input gate, output gate, and forget gate. The gradients of the weights are calculated based on the Equation 1 - 3, $\delta = \frac{\partial E}{\partial}$ and T is total number of weights in learnable weights.

$$\delta W_{(i,g,f,o)} = \sum_{t=0}^{T} \{\delta_{(i,g,f,o)}, x_t\} \quad (1)$$

$$\delta R_{(i,g,f,o)} = \sum_{t=0}^{T-1} \{\delta_{(i,g,f,o)}, h_t\} \quad (2)$$

$$\delta b_{(i,g,f,o)} = \sum_{t=0}^{T-1} \{\delta_{(i,g,f,o)}\} \quad (3)$$

V. Results and Discussion

In this section, we evaluate the prediction performance of the proposed method on node location trajectories from *model-based mobility traces data* as well as from real-world mobility traces data as detailed in section III (B). In *model-based mobility traces data*, we have considered an Ad-hoc mobile node which moves according to Gauss Markov mobility model with a varying speed in [0...20]. Its coordination is recorded each 2.5s, from initial time 0s to till 4000s. In real-world mobility traces data, GPS traces collected from a team of firefighters during a forest fire exercise. The traces were generated Latitude, Longitude and Timestamp by Android phones placed in each of four firefighters, its coordination (Latitude and Longitude) is recorded in one second interval of time almost six hours, initial recorded time is 47:03.8, in this study we used only one hour recorded data sequences from the both datasets starting from 0s to till 3600s. So we obtain two location time series x and y. In order to predict the mobile node trajectory, we have tested the predictor on two location time series, duration one hour (i.e. 3600s). The first 70% data sequences are used for training and the rest 30% for test or generalization. In order to fully learn the node mobile trajectory, during the training process, we take the complete trajectory except the last point (i.e. $\{l_1, l_2, \ldots, l_T\}$) as an input and sliding the time-series forward to one step ahead as standard output (i.e. $\{l_2, l_3, \ldots, l_{T+1}\}$) till end of data sequence.

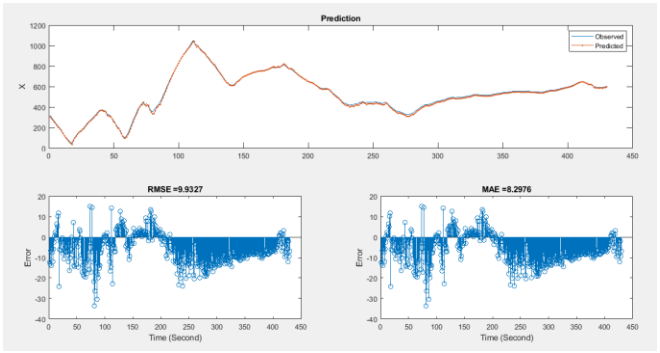
We used Matlab2018a to build and train our model. The training is done on an Intel core i3 machine with 3GB RAM. To quantitatively assess the overall performance of LSTM model, Root Mean Square error (RMSE) and Mean Absolute

Error (MAE) is used to estimate the prediction accuracy which is in equation 4 and 5.

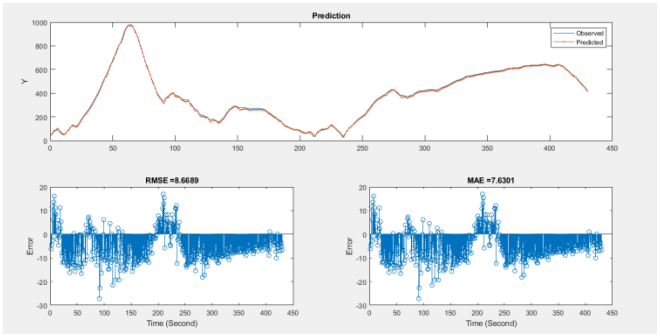
$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (l_i - \hat{l}_i)^2} \quad (4), \quad \text{and}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |l_i - \hat{l}_i| \quad (5)$$

Where l_i is the observed coordinate location value, $\hat{l}_i$ is the predicted coordinate location value and N represents the total number of predictions. Figures 4(a), (b) and 5 (a), (b) show the test of the predictor on the two series x and y of model-based and real-world based mobility traces data respectively, and in figure: 6(a), (b) present the prediction of the mobile node trajectory based on both datasets. Finally, in table 2 compares the prediction error of different prediction methods (i.e. FFNN and RNN) and show the superiority of LSTM. The prediction methods of FFNN and RNN are referring from [10] and [5] respectively.

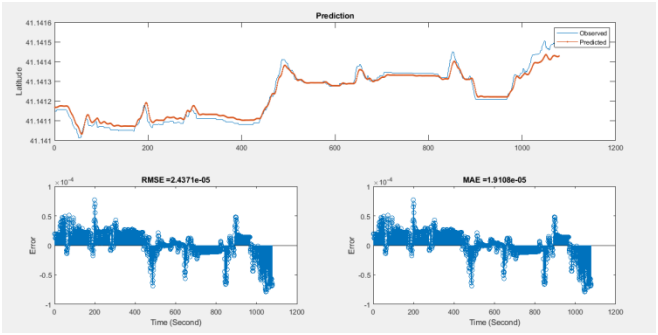


4(a)

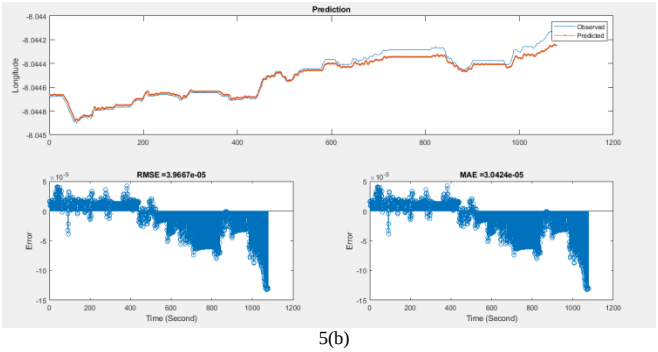


4(b)

Fig.4 Comparison of single user mobility prediction performance on a model based dataset over one hour observation. (a), (b) show the observed and predicted positions of x-coordinate time series and y-coordinate time series respectively.

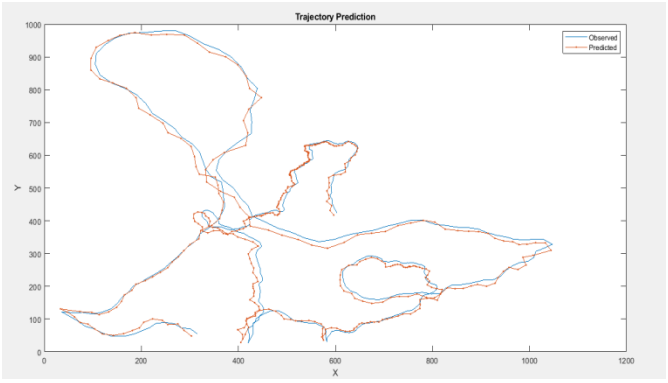


5(a)

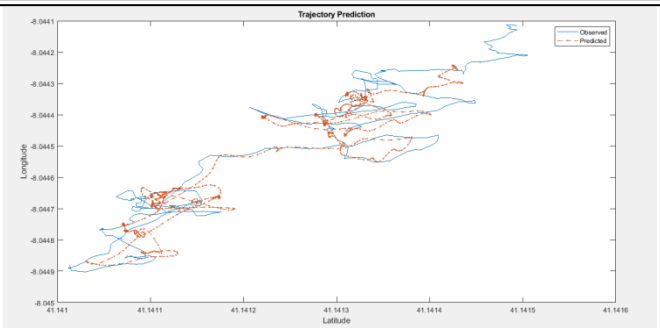


5(b)

Fig.5 Comparison of single user mobility prediction performance on a real-world based dataset over one hour observation. (a), (b) show the observed and predicted positions of latitude time series and longitude time series respectively.



(a)



(b)

Fig.6. Mobility prediction, (a) Prediction on model (Gauss Markov) based dataset, (b) Prediction on real-world dataset [17]

Table 2 Comparison of prediction methods

Performance Metrics	Methods	Model based dataset		Real-world based dataset	
		X	Y	Latitude	Longitude
RMSE	FFNN	14.0309	14.1608	3.3523e-05	7.0949e-05
	RNN	25.3233	11.6738	7.1006e-05	6.9883e-05
	LSTM	9.9327	8.6689	2.4371e-05	3.9667e-05
MAE	FFNN	10.4461	11.2633	1.7349e-05	3.2912e-05
	RNN	20.7876	8.3955	3.8474e-05	3.1110e-05
	LSTM	8.2976	7.6301	1.9108e-05	3.0424e-05

VI. Conclusion

A good accuracy mobility prediction increase network performance and QoS in ad hoc network, keeping this aim in this paper, we present LSTM model to apply mobility prediction both on model based and real-world based dataset. The evaluation result shown that LSTM architectures are well suited for mobility prediction. The evaluation between two dataset the model based dataset has lower error then real-world dataset. The results of our evaluation show that LSTM outperforms then basic FFNN and RNN in terms of performance accuracy. In future work, the LSTM Network model will be extended to predict multiuser-multistep mobile node trajectory in ad hoc network.

REFERENCES

[1] Nagwani, R., & Singh Tomar, D. (2012). Mobility Prediction based Routing in Mobile Adhoc Network using Hidden Markov Model. *International Journal of Computer Applications*, *59*(1), 39-44. doi: 10.5120/9516-3919

[2] Denko, M. Mobility Prediction Schemes in Wireless Ad Hoc Networks. *Multimedia Systems And Applications Series*, 171-186. doi: 10.1007/0-387-22792-x_9

[3] Qiao, S., Shen, D., Wang, X., Han, N., & Zhu, W. (2015). A Self-Adaptive Parameter Selection Trajectory Prediction Approach via Hidden Markov Models. *IEEE Transactions On Intelligent Transportation Systems*, *16*(1), 284-296. doi: 10.1109/tits.2014.2331758

[4] Lv, Q., Qiao, Y., Ansari, N., Liu, J., & Yang, J. (2017). Big Data Driven Hidden Markov Model Based Individual Mobility Prediction at Points of Interest. *IEEE Transactions On Vehicular Technology*, *66*(6), 5204-5216. doi: 10.1109/tvt.2016.2611654

[5] Heni Kaaniche, F. K (2010). Mobility Prediction in Wireless Ad hoc networks using neural networks, Journal of Telecommunications, 95 - 101.

[6] Elleuch, M., Kaaniche, H., & Ayadi, M. (2015). Exploiting Neuro-Fuzzy System for Mobility Prediction in Wireless Ad-Hoc Networks. *Advances In Computational Intelligence*, 536-548. doi: 10.1007/978-3-319-19222-2_45

[7] Shang Y., Guo W., Cheng S. (2005) Clustering Algorithm Based on Wavelet Neural Network Mobility Prediction in Mobile Ad Hoc Network. In: Wang J., Liao XF., Yi Z. (eds) Advances in Neural Networks – ISNN 2005. ISNN 2005. Lecture Notes in Computer Science, vol 3498. Springer,Berlin,Heidelberg. https://doi.org/10.1007/11427469_63.

[8] Ghouti, L., Sheltami, T., & Alutaibi, K. (2013). Mobility Prediction in Mobile Ad Hoc Networks Using Extreme Learning Machines. *Procedia Computer Science*, *19*, 305-312. doi: 10.1016/j.procs.2013.06.043

[9] Ghouti, L. (2016). Mobility prediction in mobile ad hoc networks using neural learning machines. Simulation Modelling Practice And Theory, 66, 104-121. doi: 10.1016/j.simpat.2016.03.001

[10] Makhlouf, N. (2016). Exploiting Neural Networks for Mobility Prediction in Mobile Ad Hoc Networks. International Journal of Electro revue, 66 - 67.

[11] Yayeh, Y., Lin, H., Berie, G., Adege, A., Yen, L., & Jeng, S. (2018). Mobility prediction in mobile ad-hoc network using deep learning. *2018 IEEE International Conference On Applied System Invention (ICASI)*. doi: 10.1109/icasi.2018.8394504

[12] Cadger F., Curran K., Santos J., Moffett S. (2012) MANET Location Prediction Using Machine Learning Algorithms. In: Koucheryavy Y., Mamatas L., Matta I., Tsaoussidis V. (eds) Wired/Wireless Internet Communication. WWIC 2012. Lecture Notes in Computer Science, vol 7277. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30630-3_15.

[13] Wang, C., Ma, L., Li, R., Durrani, T., & Zhang, H. (2019). Exploring Trajectory Prediction Through Machine Learning Methods. *IEEE Access, 7*, 101441-101452. doi: 10.1109/access.2019.2929430

[14] Altche, F., & de La Fortelle, A. (2017). An LSTM network for highway trajectory prediction. *2017 IEEE 20Th International Conference On Intelligent Transportation Systems (ITSC)*. doi: 10.1109/itsc.2017.8317913

[15] Gite, P. (2017). Link stability prediction for mobile Ad-hoc network route stability. *2017 International Conference On Inventive Systems And Control (ICISC)*. doi: 10.1109/icisc.2017.8068651

[16] BonnMotion - A mobility scenario generation and analysis tool. (2018). Retrieved 8 June 2018, from http://sys.cs.uos.de/bonnmotion

[17] Ana Aguiar, CRAWDAD dataset it/vr2marketbaiaotrial (v.2019-09-16), downloaded from https://crawdad.org/it/vr2marketbaiaotrial/20190916, Sep 2019.

[18] Narayan, A., & Hipel, K. (2017). Long short term memory networks for short-term electric load forecasting. *2017 IEEE International Conference On Systems, Man, And Cybernetics (SMC)*. doi: 10.1109/smc.2017.8123012

[19] Azzouni, A., & Pujolle, G. (2017). *A Long Short-Term Memory Recurrent Neural Network Framework for Network Traffic Matrix Prediction*. https://arxiv.org/abs/1705.05690

WEBCAM BASED REAL TIME PRINTED TO SCANNED TEXT DOCUMENT CONVERSION

[1]Nishant Kumar, [2]Swarnika Verma, [3]Sumit Singh Rajput
[1,2,3] Btech (Computer Science and Engineering), Galgotias University, Greater Noida,U.P,India

Abstract: Optical detection is based on a web camera according to the test, this is what the system is used to recognize the signs, and letters in the text, by comparing the two-the alphabet pictures. The purpose of this prototype is to develop a software for optical character recognition (OCR) system, which uses a proprietary algorithm, which is consistent with the select button. It has its own applications, which use the Pattern of the Matching, the algorithm is used in order to recognize what is, and the big and the small: (A – Z) and numbers (0 -9), the use of a new font, courier services, with the help of a bit-mapped image format, a photo, the size is 240 × 240), and to recognize the alphabet a comparison between the images that are already stored in the database. The purpose of this prototype is to solve the problem of the blind, the people who are reading this, the character recognition is so difficult to recognize without the use of any of the technologies, and the Pattern is compared with a solution to fix this problem.
Keywords-Optical Character Recognition, Algorithm, According To The Template.

I. INTRODUCTION

Optical character recognition using a webcam with the help of the example is that the system is useful in order to recognize the characters or letters in a text, by comparing the two images of the characters. The purpose of this prototype is to develop a software for optical character recognition (OCR) system, which uses a proprietary algorithm, which is consistent with the select button . It has its own applications, which use the Pattern of the Matching, the algorithm is used in order to recognize what is, and the big and the small: (A – Z) and numbers (0 -9), the use of a new font, courier services, with the help of a bit-mapped image format, a photo, the size is 240 × 240), and to recognize the alphabet a comparison between the images that are already stored in the database.

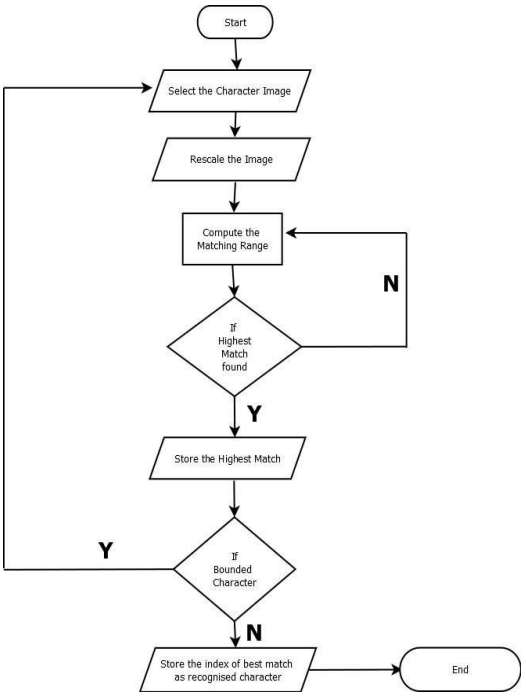
The purpose of this prototype is to solve the problem of the blind, the people who are reading this, the character recognition is so difficult to recognize without the use of any of the technologies, and the Pattern is compared with a solution to fix this problem. The field of Optical character recognition from image, and it only requires a matlab-supported notebook or desktop PC .

It can also be carried out with the help of a smartphone on the Android platform. In this topic, as it combines the features of optical character recognition and refers to thedevelopment of ideas is a useful application that will convert the pictures to the words with the help of matlab . OCR-image to which the input data, get the text out of images, and then convert it to the word document .

This can be useful in a variety of applications, such as banking, law, business, industry, other industries, as well as the computer, and the OCR software.

.

PROPOSED MODEL



(1) THE PROPOSED METHOD :

In a typical speech recognition system is composed of the following components:

1. In the Scanned Image
2. Primary treatment
3. Download,, and, calculate the mall
4. Match The Template

(a) In the spring of photos and videos with the help of a WEBCAM

In order to get a picture of you both. In the FIRST case, the photo was taken by a web cam in order to be machine-editable. The artwork can be in any specific format, such as, jpeg, tiff, bmp, etc.

(b) In the PRIMARY TREATMENT :

There are a number of businesses in the scanned image so that the image will be more convenient and easy to use than that of sub-categories. In essence, the primary goal of treatment is to improve the quality of the scanned image is of the input. The noise, the mathematical verification of the activities can be used in this section, the primary treatment. There are binarization, the limit of detection, segmentation, and incelm. It operates a number of businesses in the scanned image data is in the data.

(c) The LOADING, AND the DESIGN of the TEMPLATE

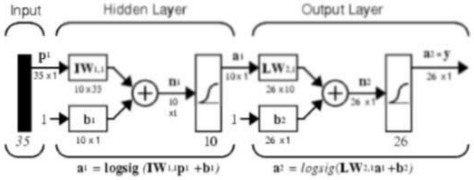
This process involves the use of signs, or to the template database. There is a model for all of the possible characters are input. For recognition to occur, the current input character is compared with a pattern, or an exact match, or the pattern, with the closest representation of the input character.

(d) MATCH THE TEMPLATE

Where I(x, y) is the input of the character, and Tn(x, y) is a model of n, a matching function, etc (I, Tn) will return the value of the business, the statement is, the better the pattern to get to the match, the sign of X.

(2) ARCHITECTURE

A neural network requires a 35-input and 26 of its neurons at the output layer of a specific letter of the alphabet. The network is a two-layer log-sigmoid network. The logarithmic sigmoid network, and is a favourite, and his introduction to vary between 0 and 1.

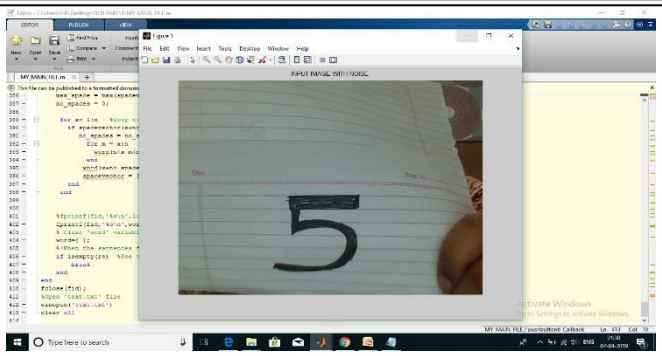


(3) ABBREVIATIONS

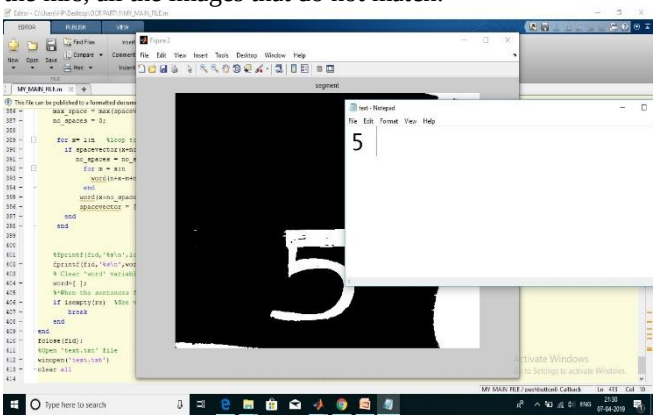
OCR	Optical Character Recognition
DIA	Document Image Analysis
DFD	Data Flow Diagram
IEEE	Institute of Electrical and Electronics Engg.
GUI	Graphical User Interface
HMM	Hidden Markov Model

(4) INPUT TO THE SYSTEM AND EXTRACTING DESIRED RESULT

The user will give the input as an image, with the help of a webcam . The image is uploaded to the program for initial processing and binarization of the input data, and then accept the user input with the help of the template corresponding to the alogorithm.



After going through all of the steps, the system segment, and the scale of the input data, and then compare it to their overlay to achieve the desired results, when compared with the mrs, all the images that do not match.



(5) RESULT ANALYSIS

It is expected that, in the print area will be reduced. The emphasis will be on recognizing the casual writing, the so-here is the font and writing style. This is a complex issue, which requires the improvement of the speech-to-text techniques. The potential of speech-recognition algorithms, which seems to be a combination of different methods and use cases that can be used in the context of a wider extent than the modern methods.

Independent confirmation text

An optical character recognition system can be designed using more than one font in mind at all times. This method is very useful for the text-independent case. Because of the font or the font size of the finds is a string, the string is interpreted as a single character. The only symbol it can be the equivalent of the symbol can be obtained by means of an effective editor. Efforts were made to develop, according to the editor of the Tamil and in English language. .

OCR to any other Indian language

With the exception of the Language and of the Language, all the other Indian languages, and requires the development of OCR for the printing of the character, and writings, of the characters of OCR işlənilməlidir for all the languages, including Bengali and Hindi. Of course, recognizing printed characters are lightweight, as compared with those of the manuscript. Also, when you print an OCR document, it must be able to perform all of the functions of the other characters, such as checking the spelling, sentences and grammar, as well as a co-editor with a keyboard, the code and font encoding is also required.

OCR, Italics,

There is a great need for a text recognition system to recognize the hand-writing, and manuscript on palm leaves. This is a really tries to avoid typing at the keyboard, and the font encoding.

The language of the converter with OCR

For the development of a full-OCR language of the font encoding, functions, and purpose to check the grammar, as well as that, it might not be possible to implement a converter to convert words from one language to another with the help of the translator and translation system.

Speech-to-text OCR

It is the most popular program of the day is the speech-to-text. Recognized as the Printed or hand-written in nature, the documents can be saved, and after that, the speech can be generated to output the results. This will help the blind people to send and receive data.

Speech-to-Text OCR Converter,

Just like the previous program, you'll be able to convert text-to-speech.

No of images	No of detectable words	No of false detections	No of positive detections	Accuracy
90	200	43	157	78.5 %

ACKNOWLEDGEMENTS

We respect and appreciate, Miss Sonya Kukereja by providing us with the opportunity to undertake the project, for the Transformation of the print or scan of the document on our project .With all support and guidance which helped us to complete the project in the right way. We are very grateful for all of it, and so, so good in terms of the guidance, even if it was a one-time schedule.

We are thankful to her and successful enough to continue to receive the support, encouragement, and guidance to all teachers in the Department of computer science and engineering at the University of Galgotias University's Faculty of Btech,who helped us to successfully complete this job. In addition, we would like to express our sincere respect to all the laboratory staff for their support at the right time.

(6) REFERENCES

[1] X. Zhai, F. Bensaali and R. Sotudeh,(2015) 'Real-time optical character recognition on eld programmable gate array for automatic number plate recognition system', IET Circuits, Devices Systems, vol. 7, no. 6, pp. 337-344.

[2] L. Eikvil, Optical Character Recognition(2012), 1st ed. Oslo pp 10-20.

[3] H. Granger, R. Chanchad, and S. Razak(2011). 'License Plate *Number Detection Algorithm for Qatari License Plates.' [Online] Available:http://www.qatar.cmu.edu/ srazak/courses/15112f13/hw/LicenseDetection.pdf pp 12-21.*

[4] G. Vamvakas, B. Gatos, N. Stamatopoulos and

S. Perantonis,(2014) 'A Complete Optical Character Recognition Methodology for Historical Documents', The Eighth IAPR International Workshop on Document Analysis Systems pp 1-6.

[5] Hiral Modi, M.C Parikh(2017), 'A Review on Optical Character

Recognition Techniques' International Journal of Computer Applications pp 7-19

Data analytical aspects of scale development with reference to EFA and CFA

[1] Dr. Umesh Ramchandra Raut, [2] Dr. Prafulla Arjun Pawar

[1] Associate Professor, Department of Management Sciences, Savitribai Phule Pune University, Sub Centre Nashik, Nashik, Maharashtra, India

,[2] Professor, Department of Management Sciences (PUMBA), Savitribai Phule Pune University, Pune, Maharashtra, India

[1]ur.raut20@gmail.com ,[2] mypumba@gmail.com

***Abstract*—** *Measurement scales are a crucial instrument in social science research for measuring latent variables such as attitudes, opinions and beliefs. Several guidelines and procedures are suggested in the conceptualization, development and implication of multi-item scales, which ensure the validation of measurement scales. These procedures have been defined in the psychometric literature since the late 1970s. Conventionally, with unusual peculiarities, the literature supported the process outlined by various researchers, who identified several actions to take in developing a measurement scale. These steps refer to construct and domain definition and scale validity, reliability, dimensionality and generalizability. Various statistical instruments are used in the scale-developing steps, almost always referring to variables measured on a metric scale such as correlation coefficients, factorial analysis, and regression models. Instead, items forming scales are almost always calculated on a level that is not metric; items are often ordinal and, in some rare cases, nominal. The present research paper highlights the data analytical importance of EFA and CFA in the scale development procedure. The study also presents the significant and procedural aspects of the various validity and reliability tools in scale development. The study provides an important implication for future research perspective in the context of measurement scale development.*

Index Terms—*Scale development, measurement scale, EFA, CFA, validity and reliability.*

I. INTRODUCTION

Multi-item measurement measures are largely employed in marketing research for various reasons (Churchill 1979). Single-item measures (Bergkvist & Rossiter, 2009) are unique in that each item tends to have only a low correlation with the attribute being measured. Another single item tends to categorise personalities into a comparatively small number of groups; third, personal items typically possess substantial measurement deviation and, last but not least, many aspects related to marketing investigation are multidimensional and not directly observable. It is unreliable to include attitudes towards complicated objects with single-item scales. A large variety of multi-item scales have been proposed in the marketing literature to measure a sample of beliefs about attitude objects (such as agreement or disagreement with a number of statements) and to connect the answers in some form of average score. The most commonly applied are the Likert and semantic differential scales. Likert scales require respondents to indicate a level of agreement and disagreement, including a type of items or statements associated with the attitude or thing. Five ordered answer levels are frequently used, but there are also Likert scales with seven or nine ordered responses.

II. THEORY: ROLE OF MEASUREMENT SCALE IN SOCIAL SCIENCE RESEARCH:

Developing a multi-item scale is an intricate procedure and requires much expertise. Many measures are available in the existing literature, such a Brand Loyalty (Pawar & Raut, 2012; Raut, 2015), Brand Resonance (Raut, Brito, & Pawar, 2020). Many papers in the marketing literature are devoted to this topic. The first papers appeared in the 1970s; in particular, two seminal works were published, to which almost all the later literature on the topic refers. (Peter, 1979) reviewed traditional reliability theory and measurement, discussing basic concepts and evaluating assessment procedures for use in marketing research. Peter also introduced generalizability theory, providing a unified conceptual and operational approach for addressing reliability issues. Lastly, the author applied reliability assessment to the area of marketing, specifically consumer behaviour. Churchill (1979) proposed a framework, a kind of protocol, by which measures of constructs of interest to marketers having desirable reliability and validity properties could be developed.

This mentioned structure implemented by the many studies and published the significant literature base that proposes fresh or refined instruments to measure marketing constructs. The procedure proposed by Churchill has been confirmed by the series of steps. We need to specify the domain and the definition of the construct considered as the first step. A detailed review of the present literature and expert viewpoints are usually helpful.

The second step consists of generating items that taking the domain as specified; the following steps aim at purifying the measure, which means obtaining an instrument that is valid and reliable. Items should exhibit content validity – that is, they must be consistent with the theoretical domain of the construct. Statements should therefore undergo several pilot tests on samples from the relevant population and can also be screened by expertise.

Items are also judged on their readability, clarity and redundancy. On the basis of these criteria, unnecessary items are eliminated, and unclear items are rewritten. In this phase, it is also possible for items relevant to the measure but ignored in a preceding step to be included in the scale. The procedure continues by assessing reliability with first-hand data. A measure is considered reliable to the extent that independent but comparable measures of the same trait or construct of a given object match. Examining whether the measure behaves as expected in relation to other constructs evaluates criterion validity. A final step consists of determining norms, i.e. assessing the position of the individual with reference to the measured characteristics by comparing that person's score with the scores achieved by others.

III. METHODOLOGY FOR SCALE DEVELOPMENT:

Statistically speaking, in the scale development process, the researcher generally used two important statistical tools such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Raut, & Pawar, 2019). Also, the analysis of reliability and validity statistics are a very important part of the development of measures.

3.1. Exploratory Factor Analysis:

Exploratory factor analysis is implemented to identify the underlying determinants or latent variables for a set of variables. The study accounts for the relationships (i.e., correlations, co-variation, and variation) between the items. Exploratory factor analysis is based on the common factor model (Haig, 2005). Also, many times social sciences researchers use this technique for the reduction of factors; that is not relevant for further analysis (Field, 2009). The EFA and CFA are based on the common factor model, so they are related procedures. Exploratory factor analysis may be

practised as an exploratory first step while the development of a measure, and later CFA may be applied as a secondary step to test whether the structure classified in the EFA works in a different representation.

Sample Size and EFA:

Sample size affects many statistical indices such as correlation, covariance, reliability and validity statistics. Considering the EFA, the sample size has many rules. The basic rule is to recommend that a researcher has at least 10–15 participants per variable (Nunnally, 1978). While Kass& Tinsley (1979) recommended having between 5 and 10 participants per variable up to a total of 300 sample size is acceptable.

KMO and Bartlett's test:

The KMO can be calculated for single and multiple variables and describes the ratio of the squared correlation linking variables to the squared partial correlation between variables. Kaiser (1974) suggests accepting values larger than 0.5 as acceptable. Moreover, values within 0.5 and 0.7 are standard. The values greater than 0.9 considered excellent; the value lies between 0.8 to 0.9, considered as great, and 0.7 to 0.8, considered as a good (Hutcheson & Sofroniou, 1999; Field, 2009).

Rotated Component Matrix (Factor Loadings)

It is likely to evaluate the statistical importance of a factor loading with the help of the Rotated Component Matrix (Stevens, 2002). Typically, researchers use loading of an entire value of higher than 0.3 to be significant. However, the importance of factor loading will depend on the sample size. Stevens (2002) provided a table of critical values corresponding to which loadings can be compared. He suggests that for a sample size of 50, loading of 0.722 can be deemed vital. For 100 sample size, the factor loading must be higher than 0.512. For the 200 sample size, it must be greater than 0.364. For 300, it must be higher than 0.298. For 600, it should be higher than 0.210. Hence, in very large samples, small loadings can be considered statistically significant. Apart from this, commonly, 0.7 consider as good factor loadings (Field, 2009).

How Many Factors:

To decide how many factors depend on many things, even the researcher can decide how many factors he wants from all observed variables. There are many assumptions regarding how many factors have extracted, but in general, researchers consider eigenvalues. The recommendation of eigenvalues over 1, but you could change this other value we want. It is apparently fine to run a fundamental analysis with the Eigenvalues over one and scree plot and compare the results. If the two assumptions give different results, then examine the commonalities, and we have to decide for ourselves which of the two criteria to believe (Haig, 2005; Field, 2009).

Correlations Coefficient:

In the correlation matrix, we can check the singularity, Multicollinearity and convergent validity of variables. Although moderate Multicollinearity is not a difficulty for factor analysis, it is essential to circumvent severe Multicollinearity (i.e. variables that are quite highly correlated) and singularity (variables that are absolutely correlated). If the value of correlation is higher than .90, it is an indication of Multicollinearity (Harrington, 2009).

3.2. Confirmatory Factor Analysis:

Structural equation modelling is a common and extended family of analyses applied to examine measurement models (Confirmatory Factor Analysis), i.e., relations among indicators and latent variables, and to test the structural model of the associations among latent variables. There are various highly useful software units for conducting confirmatory

factor analyses, and many of them can be used to conduct CFA, SEM, and other analyses.

Sample Size and SEM

Ten samples for every one observed variable consider as great, seven sample sizes are good, and five is acceptable. In another way, less than 100 is considered "small" and 100 to 200 is "medium" -greater than 200 is "large" (Kline, 2005). Produces accurate parameter estimates and reliable goodness of fit test" when the ratio of a sample size to parameters/statements or items is 4:1 or 5:1 (Lee & Song, 2004).

Estimation Methods:

There is different estimation methods explained in the literature, such as weighted least squares (WLS), maximum likelihood (ML), weighted least squares (ULS) and generalised least squares (GLS).

If the SEM model adds one or many categorical indicators, which ultimately shows the enormously non-normality nature of the data and in such conditions, maximum likelihood (ML) is not appropriate methods.

Model Fit Indices:

In CFA, there different types of model fit indices such as Absolute Fit Indices, Parsimony Correction Indices, Comparative Fit Indices, Predictive Fit Indices.

Absolute Fit Indices:

The most common absolute fit index is the model chi-square ($\chi 2$), which tests whether the model fits precisely in the population or not. Another absolute fit index is the Root Mean Square Residual (RMR). Because the RMR is affected by the metric of the input variables, it can be challenging to interpret. The Standardised Root Mean Square Residual is based on the difference among the correlations in the information matrix and the correlations prognosticated by the SEM model, which is standardised and therefore easier to interpret and is generally preferred over time RMR (Brown, 2006).

Parsimony Correction Indices

The root means the square error of approximation (RMSEA) tests the extent to which the model fits reasonably well in the population; it is sensitive to model complexity, but unlike the model chi-square, it is comparatively inconsiderate to sample size. Close fit (CFIT) indicates the probability (p) that RMSEA is less than or equal to 0.05 (Brown, 2006).

Comparative Fit Indices

Comparative fit indices are used to evaluate the fit of a model relative to a more restricted, nested baseline model. Examples include the close fit index (CFI) and the Tucker-Lewis index (TLI) or non-Normed fit index (NNFI).

Predictive Fit Indices

Predictive fit indices "assess model fit in hypothetical replication samples of the same size and randomly drawn from the same population as there searcher's original sample … these indexes may be seen as population-based rather than sample-based" (Kline, 2005). The Akaike information criterion (AIC) is used with maximum likelihood (ML) estimation and "favours simpler models" so, in some senses; it is also a parsimony fit index (Kline, 2005). The Akaike information criterion (AIC) is commonly used to distinguish between two or more non-nested models examined on the identical data set. A smaller Akaike information criterion (AIC) propose that the model is more likely to replicate, has some parameters, and fits better; hence, when contrasting models, the one with the less AIC is accepted as the "more valid" model. The expected cross-validation index (ECVI) is additionally practised when relating models and will result in the similar rank ordering of models as the AIC (Kline, 2005). Furthermore, to the AIC, the ECVI is population-based and parsimony adjusted. The predictive fit indices are used for

comparing models, so unlike the other categories of fit indices, there are no guidelines for what represents acceptable fit.

Recommendations for Assessing Acceptable Model Fit

Multiple guidelines are suggested for the acceptable model fit indices the RMSEA close to 0.06 or less; CFI close to 0.95 or more; SRMR also close to 0.08 or less and TLI near to 0.95 or greater is considered acceptable fit (Brown, 2006). It also needs to understand that there are no very rigid guidelines, and whatever suggested by (Brown 2006) and Kline (2005), the use of close to is based on the purpose of the study. Kline (2005), suggest "RMSEA is less than or equal to .05 symbolises close approximate fit, values within .05 and .08 advice reasonable error of approximation, and RMSEA greater than .10 suggests poor fit". CFI "greater than roughly .90 may indicate a reasonably good fit of the researcher's model" (Kline, 2005) and SRMR value "less than .10 is generally considered favourable" (Kline, 2005).

Modification Indices:

Modification indices (MI) are they are data-driven indicators of changes to the model that is likely to improve model fit. MI is analogous to single DF/ $\chi 2$ tests; therefore, Modification indices higher than 3.84 (or roughly 4) shows a change that will probably result in a significant improvement in model fit. Modification indices can recommend changes to any characteristics of the model, including adding paths within latent variables, adding paths through latent variables to observed variables not originally specified as indicators of that latent variable, adding error covariance between observed variables, and so forth. MI for covariance suggests adding error covariance either between two errors or between an error and a latent variable. Many of the modifications suggested by the MI may not make sense given theory and prior research; such nonsensical modifications should not be made regardless of how large the parameter change would be.

IV. RELIABILITY AND VALIDITY:

Generally, in terms of EFA reliability, coefficients around .90 are considered "excellent," values around .80 are "very good," and values around .70 are "adequate" (Kline, 2011).

When there is difference exist between the items of different construct it means that the Discriminant validity is demonstrated, it's nothing but there should low correlation between the items of the different construct (Bagozzi, Yi, & Phillips, 1991). According to Brown (2006), the correlation value close to 0.85 shows poor discriminant validity. If there is a high correlation between the items of similar construct, which ensure the evidence of convergent validity (Bagozzi, Yi, & Phillips, 1991), If we find a correlation of 0.36 between latent variables, then we have evidence of discriminant validity, as predicted by the theory. However, if we find a correlation between observed variables under the same construct greater than 0.87, it would suggest good convergent validity. The important thing to note here is that the underlying theory is the basis on which decisions about construct validity are built.

V. IMPLICATION:

A present study is providing the platform to the academician and research in the prospect of scale development. Developing the scale is a very complex phenomenon, but while following an appropriate statistical approach, it easy to conceptualise, develop and implementation of the measurement scales, and the entire procedural thing will provide the present study. The present study highlights the methodological aspects of exploratory factor analysis and confirmatory factor analysis, which are crucial in the procedure of the development of the scale. Importance of

validity and reliability is very important, and when we talk about scale development, it reaches its immense level, considering this fact the present study significantly highlights the various ways to test the validity and reliability of the measurement scale, which will highly beneficial to the academician and researcher in the development of new measurement scale or revalidation of the existing measurement scale.

VI. DISCUSSION AND CONCLUSION:

The researcher needs to think about various aspects while going for the scale development, including Multicollinearity, missing data and outliers. If there is a strong correlation exists between the items, it's an indication of Multicollinearity. Multicollinearity poses a problem only for multiple regressions because simple regression requires only one predictor. Low levels of Collinearity pose little threat to the models generated (Field, 2009). The research many times fails to indicate the volume of missing data, which ultimately severely affect the measurement model. It is also important to mention how the researcher handled the missing data issue during the test of the measurement scale. (Tabachnick & Fidell, 2007). Outliers are severe or extremely unusual circumstances that can bias estimators and importance tests (Yuan & Bentler, 2001). Outliers are problematic; they may produce non-normality and may result in Heywood cases (Brown, 2006). Questionable outliers can be withdrawn from the investigations (Meyers, Gamst, & Guarino, 2006) if the sample size is adequately big to allow that as an option; nevertheless, if outliers are dropped, then we should examine how this influences the generalizability of your outcomes. At the same time, developing the measurement scale needs to be very careful in terms of Multicollinearity, missing data and outliers. The past research evidence that the development of the measurement scale or revalidation of the existing measurement scale is not that tough if research following all analytical data paths appropriately and efficiently.

REFERENCES

[1] Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. Administrative Science Quarterly, 36, 421–458.
[2] Bergkvist, L., & Rossiter, J. R. (2009). Tailor-made single-item measures of doubly concrete constructs. International Journal of Advertising, 28 (4), 607-621.
[3] Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.
[4] Churchill, J. G. (1979, Feb). A paradigm for developing better measures of marketing constructs. Journal of Marketing Research, 64-73.
[5] Comrey, A. L., & Lee, H. B. (1992). A first course in factor analysis (2 Ed.). Hillsdale: NJ: Erlbaum.
[6] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.
[7] Field, A. (2009). Discovering Statistics Using SPSS (Introducing Statistical Methods Series) (3 Ed.). London: Sage Publications Ltd.
[8] Fuchs, C., & Diamantopoulos, A. (2009). Using Single-Item Measures for Construct Measurement in Management Research: Conceptual Issues and Application Guidelines. Business Administration Review, 69 (2), 195-210.
[9] Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. Multivariate Behavioral Research, 40, 303–329.
[10] Harrington, D. (2009). Confirmatory Factor Analysis. New York: Oxford University Press.
[11] Hutcheson, G., &Sofroniou, N. (1999). The multivariate social scientist. London: Sage.
[12] Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39, 31–36.
[13] Kass, R. A., & Tinsley, H. E. (1979). Factor analysis. Journal of Leisure Research, 11, 120–138.
[14] Kline. (2005). Principles and practice of structural equation modeling (2 Ed.). New York: Guilford Press.
[15] Kline, R. B. (2011). Principles and Practice of Structural Equation Modeling (3 ed.). New York: The Guilford Press.
[16] Koeske, G. F. (1994). Some recommendations for improving measurement validation in social work research. Journal of Social Service Research, 18 (3/4), 43–72.
[17] Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. Multivariate Behavioral Research, 39, 653–686.
[18] MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. Psychological Methods, 4 (1), 84–99.
[19] Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). Applied multivariate research: Design and interpretation. Thousand Oaks: Sage.
[20] Nunnally, J. (1978). Psychometric Theory (2 ed.). New York: McGraw Hill.
[21] Pedhazur, E., &Schmelkin, L. (1991). Measurement, design and analysis: an integrated approach. Hillsdale NJ: Erlbaum.
[22] Peter, J. (1979). Reliability: A Review of Psychometric Basics and Recent Marketing Practices. Journal of Marketing Research, XVI, 6- 17.

[23] Raut, U. R. (2015). A Study on Brand Loyalty and Its Association with Demographics of Consumers: Evidence from the Cellphone Market of India. IUP Journal of Brand Management, 12(3), 30-44.
[24] Raut, U.R., Brito, P.Q., & Pawar, P.A. (2020). Analysis of Brand Resonance Measures to Access, Dimensionality, Reliability and Validity. 21(1), 162-175.
[25] Pawar P.A., and Raut, U. R., (2012), "Analysis of Cell-Phone Market in India for Extracting New Dimensions of Consumer Brand Loyalty Measurement", Zenith International Journal of Multidisciplinary Research, 2(7), 114-130.
[26] Raut, U.R., & Pawar, P.A. (2019) Measurement model of factors affecting consumer purchase decision of private label brands, ZENITH International Journal of Multidisciplinary Research, 9(5), 74-83
[27] Satorra, A., &Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. Psychometrika, 66 (4), 507-514.
[28] Stevens, J. P. (2002). Applied multivariate statistics for the social sciences (4 Ed.). Hillsdale: NJ: Erlbaum.
[29] Tabachnick, B. G., &Fidell, L. S. (2007). Using multivariate statistics (5 th ed.). Boston: Allyn and Bacon.
[30] Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. Journal Of Counseling Psychology, 34 (4), 414-424.
[31] Yuan, K., &Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. British Journal of Mathematical and Statistical Psychology, 54, 161–175.

AUTHORS PROFILE

Dr. Umesh Raut pursued his MBA (Marketing) from Savitribai Phule Pune University. He awarded joint PhD (Marketing) from the Department of Management Sciences (PUMBA) and Savitribai Phule Pune University of Pune- India and Faculty of Economics (FEP) University of Porto- Portugal. He has worked as Post-Doctoral Research Fellow (Marketing) with Institute of Media and Marketing, Corvinus University of Budapest, Hungary. He published several empirical research papers in national and international journals and also presented various empirical research papers at the esteemed international conference including countries such as Italy, France, Spain, Portugal, Hungary, Indonesia and India. He delivered the SPSS training program at various National and International University.

Dr. Prafulla Pawar is engineering graduate [BE – Electronics] from University of Pune and has done MBA in Marketing and MBA in Systems management. He holds a PhD degree in strategic management. During academic career of 19 years and industrial career of 4 years, he has published 39 research papers in reputed national and international journals. He is the author of two books in management. Dr. Pawar was a visiting Professor for University of Deusto, Spain and University of Bologna, Italy and Masaryk University, Czech Republic. His interest areas are Strategic Management and emerging economies. He has pioneered Innovation & Incubation Centre on the SPPU Campus and he is also the Director of SPPU Alumni Association.

Automated System for Zero Downtime Database Migration using Scripting

[1]Unnikrishnen Nampoothiry, [2]Prof. V.M.Lomte, [3]Samruddhi Pund, [4]Siddhesh Patil, [5]Atharv Relekar
[1][3][4][5] Student R.M.D.Sinhgad School of Engineering ,[2] HoD, Department Comp.Engg. RMDSSOE
[1]unnikrishnan0810@gmail.com,[2] comphod.rmdssoe@sinhgad.edu, [3]samruddhimpund@gmail.com

Abstract— In today's computing world, migration has become a necessary aspect in every field from humans to data. Migration is a very crucial process which requires lot of configurations, in-depth knowledge of the underlying functions, commands and systems which makes the migration process complex and time consuming. It requires skilled human resource, infrastructure with the best error handling capacities.

An automated system provides the best experience of migration to all the database administrators by providing an automated developed platform to extract, pump and replicate efficiently by just a click.

Index Terms—*Database Migration, Goldengate, Oracle Database, Scripting, Zero Downtime*.

I. INTRODUCTION

To perform database migration from source DB to Destination DB without manual intervention by providing all the necessary functionalities through a dashboard by using *oracle GoldenGate as our middleware software*.

There are three main components of the project:

1. Oracle Database
2. Golden Gate
3. Scripting

Oracle Database:
It is a multi model database management system produced and marked by oracle corporation. The data is stored logically in the form of table spaces and physically in the form of data files.

Golden Gate (Middleware):
Oracle Golden Gate is a software product used for data replication and integration in heterogeneous IT environment. By using oracle Golden Gate we can move committed transaction across multiple heterogeneous environment.

Scripting:
All the procedures of installations and migration that is performed manually consumes a lot of time from days to week, along with that specialized knowledge of oracle is required for installations and migration, with error handling capability.
As developers, these manually done processes of configuration and error handling will be done fully automated through scripting, which will save a lot of time of database administrators and developers avoiding manual errors and no human intervention in migrating database.

II. (A) LITERATURE SURVEY

Sr No	Published Year	Published By	Research Topic	Outcomes
1.	2012	A. Krizhevsky, I. Sutskever, G. E. Hinton	Imagenet classification with deep convolutional neural networks	Able to map out how CNN algorithm can be extended to various other techniques.
2.	2015	S. V. Radhakrishnan, A. S. Uluagac, R. Beyah	GTID: A technique for physical device and device type fingerprinting	GTID exploits the heterogeneity of devices, which is a function of the different device hardware compositions and variation in device clock skew.
3.	2016	Q. Xu, R. Zheng, W. Saad, Z. Han	Device fingerprinting in wireless networks: Challenges and opportunities	Reviewed the idea to extract characteristics from transmitted signal or frames from the wireless devices and their environments to generate non-forgeable signatures
4.	2017	A. Selim, F. Paisana, J. A. Arokkiam, Y.	Spectrum monitoring for radar bands using deep	Tested the performance for different data representations and concluded that the

		Zhang, L. Doyle, L. A.DaSilva	convolutional neural networks	proposed Amplitude+Phase Difference representation enables CNNs models to obtain high classification accuracy and it is more robust to noise.
5.	2017	T. J. O'Shea J. Hoydis	An Introduction to Deep Learning for the Physical Layer	Introduced a new way of thinking about communications as an end-to-end reconstruction optimization task using autoencoders to jointly learn transmitter and receiver implementations as well as signal encodings without any prior knowledge.
6.	2018	S. Riyaz, K. Sankhe, S. Ioannidis, K. Chowdhury	Deep learning convolutional neural networks for radio identification	A radio fingerprinting approach based on deep learning CNN architecture to train using I/Q sequence examples. The design enables learning features embedded in the signal transformations of wireless transmitters, and identifies specific devices.
7.	2017	1.Boyu Hou 2.Yong Shi 3.Kai Qian	Towards Analyzing MongoDB NoSQL Security and Designing Injection Defense Solution . [13]	Demonstrated server-side JavaScript and HTTP injection attacks and propose defense measures to promote the security of MongoDB, which will help NoSQL databases programmers and designers be aware of injection mechanism and build a more secure data environment.
8.	2017	1.Li Dianwei, 2.He Mingliang, 3.Yuan Fang	Research on Insider Threat Detection Based on Role Behavior Pattern Mining[J]. [4]	Based on user log data, we constructed three types of datasets: user's daily activity summary, e-mail contents topic distribution, and user's weekly e-mail communication history. Then, we applied four anomaly detection algorithms and their combinations to detect malicious activities.
9.	2017	1.Jitender Kumar 2.Varsha Garg	Security analysis of unstructured Data in NoSQL MongoDB	Used symmetric cryptographic techniques for providing the security (confidentiality) of unstructured data in NoSQL document stored MongoDB. DES, AES, and blowfish algorithms with random key generation are used to encrypt/decrypt the document data before storing/retrieving to/from the NoSQL MongoDB database.
10.	2017	1.Xu Guangxian 2.Zhao Yue 3.Public Zhong Sheng	Design of Double Encryption security network coding scheme based on chaotic sequence.[5]	Using chaotic sequence, designed double Encryption security network coding scheme
11.	2017	1.Alfredo Cuzzocrea 2.Hossain Shahriar	Data Masking Techniques for nosql Database Security: A systematic review. [15]	In-depth study of potential security vulnerabilities in MongoDB and Cassandra, two popular NoSQL databases.
12.	2018	1.Kosovare Sahatqija 2.Jaumin Ajdari 3.Xhemal Zenuni 4.Bujar Raufi 5.Florije Ismaili .	Comparison between relational and NOSQL databases [16]	This paper is a qualitative research, based on detailed and intensive analysis of the two database types, through use and comparison of some published materials during last few years.
13.	2019	1.Md Rafid Ul Islam 2 Md. Saiful Islam 3. Zakaria Ahmed 4. Anindya Iqbal 5. Rifat Shahriyar.	Automatic Detection of NoSQL Injection Using Supervised Learning [17]	A tool for detecting NoSQL injections using supervised learning. Applied our tool to a NoSQL injection generating tool, NoSQLMap and find that our tool outperforms Sqreen, the only available NoSQL injection detection tool, by 36.25% in terms of detection rate. The proposed technique is also shown to be database-agnostic achieving similar performance with injection on MongoDB and CouchDB databases.
14.	2015	Stanislav Lange, Steffen Gebert, Thomas Zinner, Phuoc Tran-Gia, David Hock, Michael Jarschel, and Marco Hoffmann	Heuristic Approaches to the Controller Placement Problem in Large Scale SDN Networks	Panel designed to control sdn network. Algorithms developed and implemented in the Matlab based POCO framework for Pareto-based Optimal COntroller placement.

15.	2014	Richard G. Clegg, Stuart Clayman, George Pavlou, Lefteris Mamatas and Alex Galis	On the selection of management/monitoring nodes in highly dynamic networks	Investigated the selection of management and monitoring nodes in a rapidly changing network environment.
16.	2014	Manar Jammal, Taranpreet Singh, Abdallah Shami, Rasool Asal , Yiming Li	Software defined networking: State of the art and research Challenges	--
17.	2015	Dmitry Ju. Chalyy, Evgeny S. Nikitin, Ekaterina Ju. Antoshina	A Simple Information Flow Security Model for Software-De_ned Networks	An approach which is based on the formal security-type system. This system ensures that the controller application does not violate confidentiality policy.
18.	2015	M.Elamparithi , V.Anuratha	A review on Database Migration Strategies,Techniques and Tools	
19.	2018	Virender Kumar, Cherry Kosla	Data Cleaning – A thorough analysis and survey on Unstructured data	
20.	2015	Bogdan Walek, Cyril Klimes	Expert system for data migration between different database management systems	

Table 01: Literature Survey

(B) LIVE SURVEY

There are various countries around the world that uses golden gate to perform migration operations in the companies. The top countries that uses golden gate are listed below[32][33][34]:

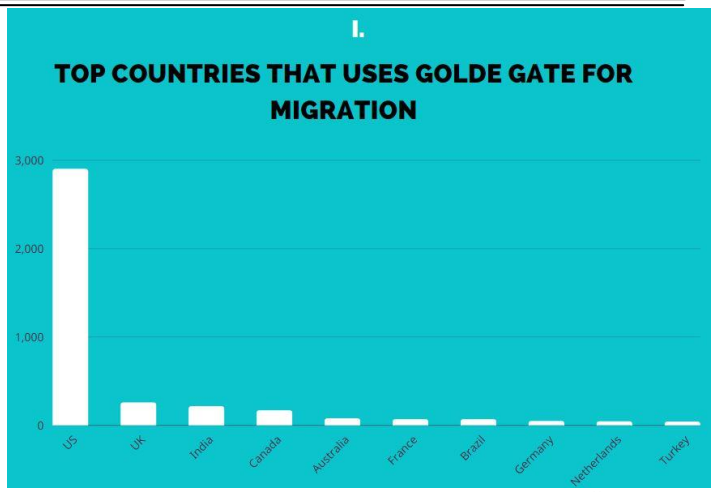


Figure 01: Graph representing the count of companies using goldengate in different countries

Distribution of the various sector companies that use golden gate based on their company size. From the graph we can see that computer software takes 25% and Information and Technology and Services takes 11% are the most largest segment.

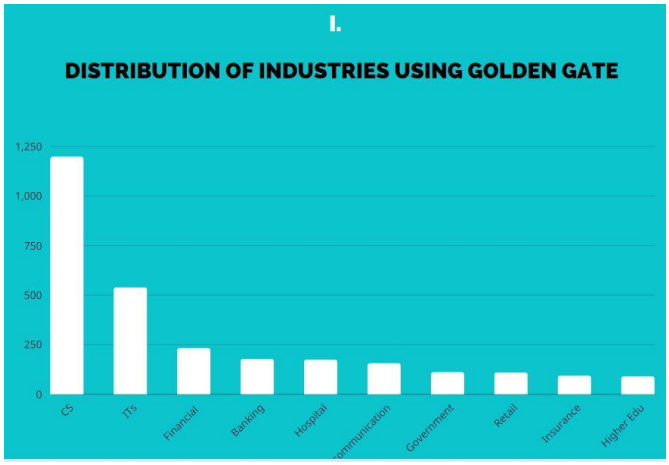


Figure 02: Graph representing distribution of industries using goldengate.

Based on II (A) Literature survey and II (B) Live Survey, we were able to identify the following research gaps:

1. There are no systems available that has all functionalities for database migration altogether.
2. Even the countries using goldengate has a wider spectrum indicating lack of knowledge about GoldenGate among the users.
3. Data Analysis for migration records is a hassle.
4. Countries not using GoldenGate face a lot of Downtime during database migration.

III. PROBLEMS IDENTIFIED IN THE EXISTING SYSTEM

1. In migration all the commands and the queries are executed on the server terminal or via PuTTy and it becomes very difficult to know whether the functionalities required for the migration are up or not (eg. If the Listener is up or not, Manager is up or not and many more).
2. It also becomes very difficult to locate errors in the error log file and to interpret the fault history.
3. It is also difficult to view the script as it requires the knowledge of all the necessary commands and queries.
4. The traditional migration requires extra tools for report generation.
5. Migration becomes a very complicated process as we are not able to view all the necessary functionalities at the same time as all the commands are executed on the server terminals one after the another as shown below:

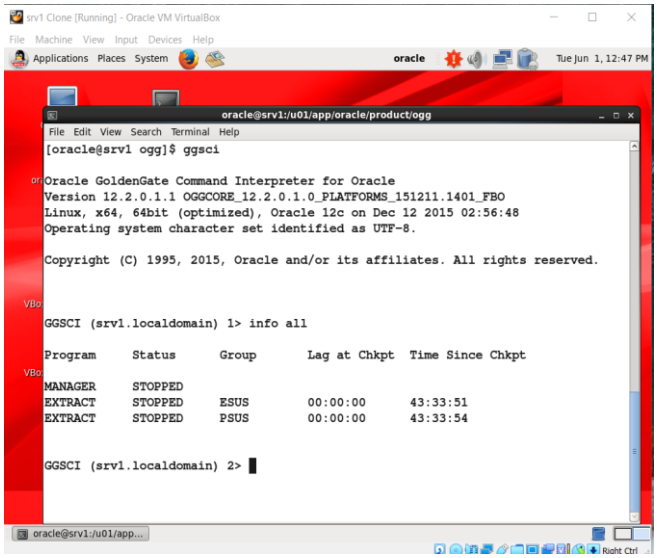


Figure 03: Server Terminal of GoldenGate- Users have to enter all the commands manually here

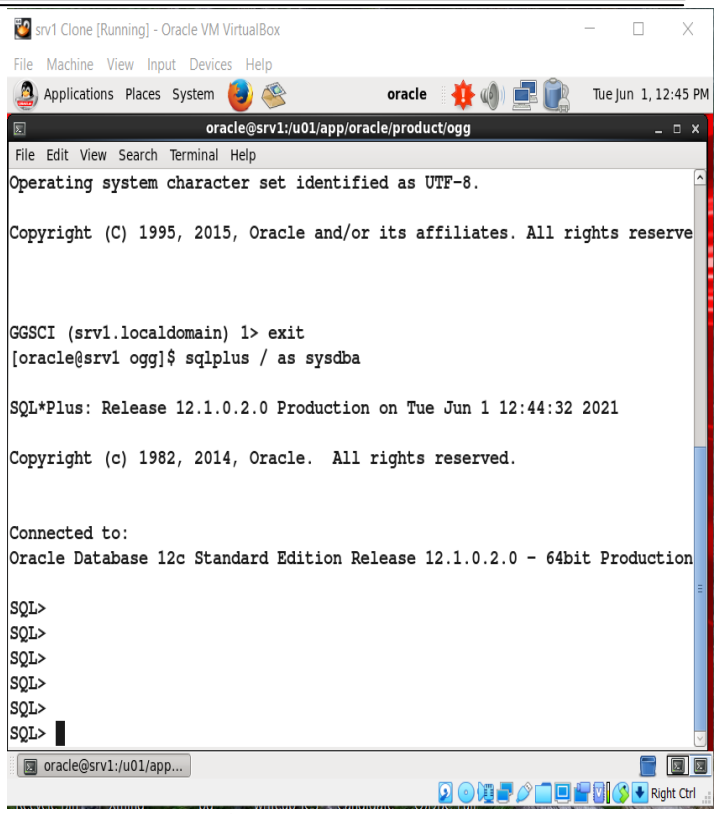


Figure 04: Server Terminal of Oracle SQLPLUS- Users have to enter all the commands manually here

IV. OBJECTIVES

1. Our system provides the best experience of migration to all the database administrators by providing an automated developed platform to extract, pump and replicate efficiently by just a click.
2. Thus making it convenient to third-party users for performing the migration as they won't need to have in-depth knowledge about oracle GoldenGate.
3. Added on security due to introduction of encryption and decryption.
4. Creating a dashboard for easy monitoring and analysis of the performance credibility of the application.

V. PROPOSED SYSTEM

1. Automated the migration process by using scripting.
2. Creation of a Dashboard for viewing all the functionalities at the same place.
3. Displaying the source and destination schema on the same page.
4. Providing extra functionalities like Performance Analysis, Fault History, View Script and Error Log.
5. Providing the report for the entire migration.

VI. SYSTEM ARCHITECTURE

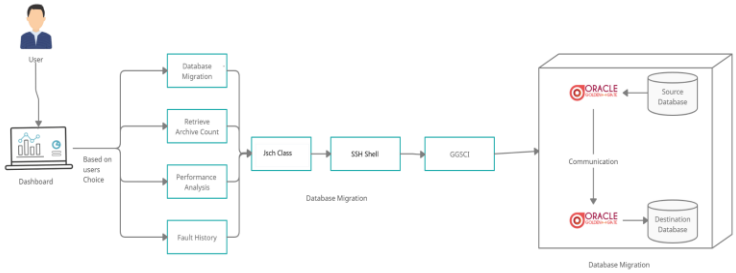


Figure 05: System Architecture

We have created one user which is the DBA who will have the complete access to the dashboard. From the dashboard they'll have the option to choose which operation to perform. Be it Checking the status of Archive log, Listener, Manager, Oracle Instance, Extract, Pump and Replicate process.
They also have the functionalities like viewing performance analysis, fault history, error logs and scripts. Depending on their choice specific commands are passed to Jsch class which relies it onto the SSH Shell.
From here the golden gate is controlled using GGSCI command which will initiate the next processed.

VII. USE CASE DIAGRAM

Figure 06: Use Case Diagram

Use cases:

1. Source Server: Source server on the Oracle linux.
2. Destination Server: Destination server on the Oracle linux.
3. Archive log, listener and other components: Components of oracle database and goldengate used for migration.
4. GUI: graphical user interface- Login page and dashboard.
5. Performance Analysis: performance analysis pie chart.
6. Fault History: Record pf transaction failures in line chart format.
7. Dashboard: One stop for all the migration functionalities.
8. Jsch Class: Java class used for connecting with SSH of the oracle linux.
9. SSH: Secure shell way of communication.
10. Shell: terminal at the server.
11. GGSCI: for controlling the goldengate.

Actors:

1. DBA: Database Administrator- has prior knowledge of all the process, can troubleshoot all queries with ease.
2. Database: Oracle Database at the virtual instances.

VIII. SEQUENCE DIAGRAM

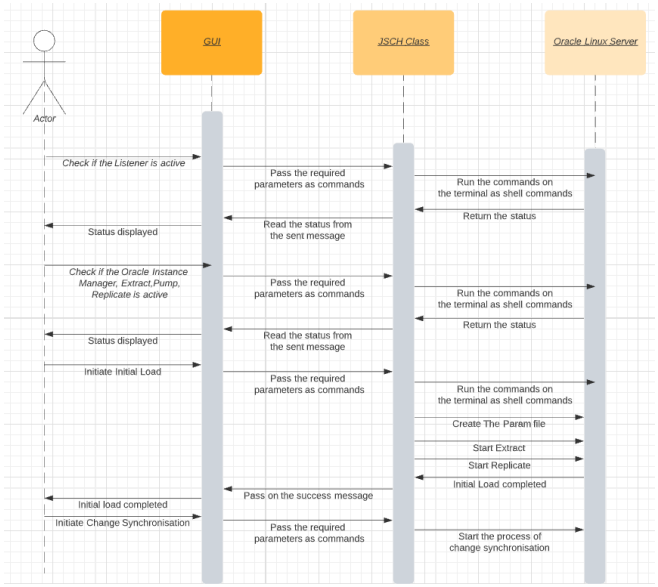


Figure 07: Sequence Diagram

The actor here is the DBA.
The DBA chooses the option from the dashboard provided and based on the option the subsequent commands are sent to the Jsch class which then posts these commands onto the shell of the server terminals.
If the actor choses for migration the component checking commands for Listener, Oracle instance, Manager, Extract, Pump, Replicate are generated by the click of their respective buttons.
Once the commands are sent to the shell, they are run on Oracle or Goldengate respectively based on the functionality to be carried out.
Then the result is sent back to the GUI via Jsch class and handled accordingly.
After initial load change synchronization is carried out.

IX. STATE DIAGRAM

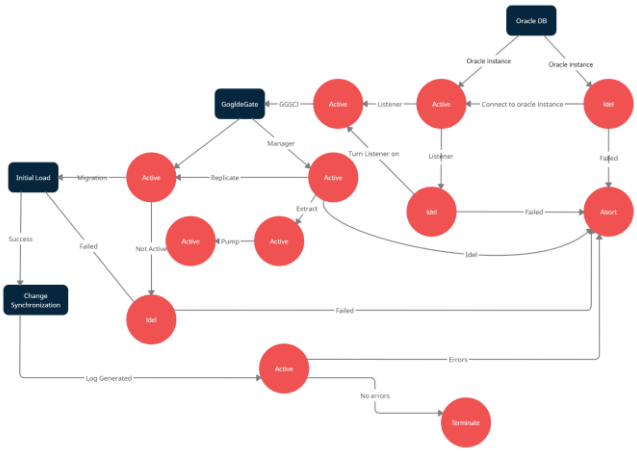


Figure 06: State Diagram

X. MATHEMATICAL MODEL

Interaction Models for migration: Most commonly the models of this type can be represented as follows: [35]

$$M_{ij} = k\frac{f(R_i,A_j)}{\phi(d_{ij})^*} \quad \text{.....[1]}$$

where Mij is the amount of interaction of two "Databases" (here the number of migrants from region i in region j); f (Ri , Aj) is a function of repulsive forces (Ri) at region of origin i and attractive forces (Aj) at region of destination j, Ri is a parameter representing repulsive factors which are associated with "leaving" region i, Aj is a parameter representing attractive factors related to going to region j, ϕ (dij) is a function of the time latency (dij) between regions i and j, k is a proportionality coefficient.

The followers of the "social physics" concept consider that *f* (*Ri* , *Aj*)= *WiWj* , where *Wi* and *Wj* are the weights of "bodies" *i* and *j*, correspondingly. Therefore the interaction model can be represented as follows:

$$M_{ij} = k\frac{W_iW_j}{\phi(d_{ij})} \quad \text{.....[2]}$$

XI. FEASIBILITY ANALYSIS

Project Description:

1. This project aims to use Golden gate features to ensure the successful migration of data from source database to destination database.
2. To provide one click migration without having to write all big and complicated commands with zero downtime.

Outlining the potential solutions:

1. Quick and easy handling to database migration configurations and processes.
2. Migrating database from source server to destination server with zero downtime.
3. Get performance analysis and reports after migration through GUI.
4. Making migration commands less prone to errors by automating configuration through scripting.

Most Feasible Solution:

1. Based on the potential solutions offered by this project, we have decided that the most feasible solution for both the project and the future scope is creating a dashboard which provides all the functionalities such as one click migration, performance analytics ,performing various checks all at one place.
2. The goal of this project is to give DBA the resources they need by making migration commands less prone to errors and perform migration task speedly.

XII. PROPOSED ALGORITHM

Proposed Algorithm for database Migration:

Migration():
Accept the SRV1 and SRV2 addresses and password
Login to Oracle Linux instances using SSH communication.
Load up the terminal
Check for Listener if off: Turn on
else: Check for Oracle instance if not connected:
Connect using STARTUP
else:
Check for manager, extract, pump and replicate process in goldengate are started.
If not connected:
Start them using START <process name>
Else:
Start with the migration.
RUN GGSCI from $GG_HOME and start the Manager, Pump process at SRV1.
RUN GGSCI from $GG_HOME and start the Replicate process at SRV2.
Check for the schema HR in Oracle database using SQLPLUS / as sysdba
Check for the tables in HR schema.
Check for records at SRV2 after Initial Load
Initiate Change Synchronisation.
Generate the Log File.

and frontend development language- JAVA)

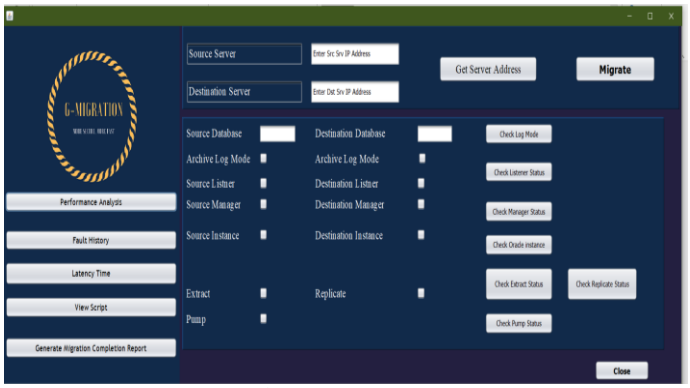


Figure 07: Dashboard- All the functionalities to be checked are easily displayed

2. Migration is performed by just the click of a button.

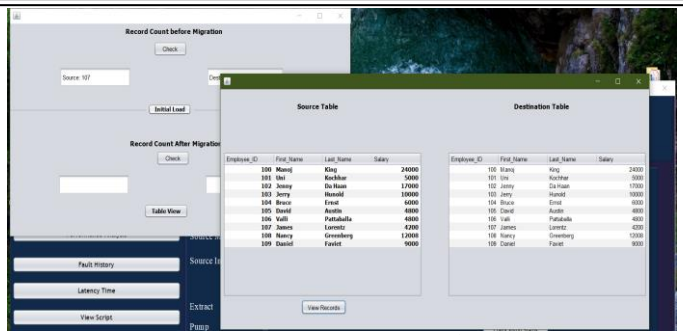


Figure 08: Migration Schema- For easy viewing of the server tables

3. Performance analysis chart.

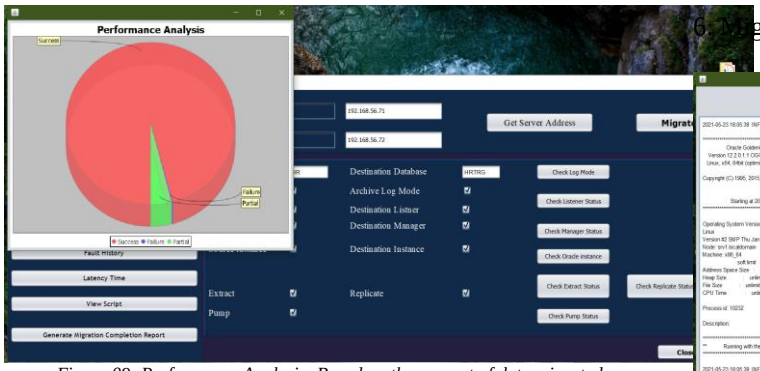


Figure 09: Performance Analysis- Based on the amount of data migrated

4. Fault History Graph

Figure 10: Fault History -Timely analysis of data migrated

5. Viewing the script.

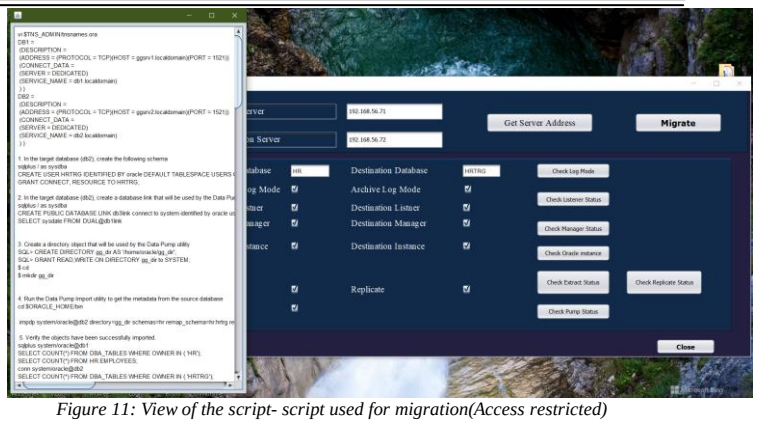


Figure 11: View of the script- script used for migration(Access restricted)

6. Migration Report Generation.

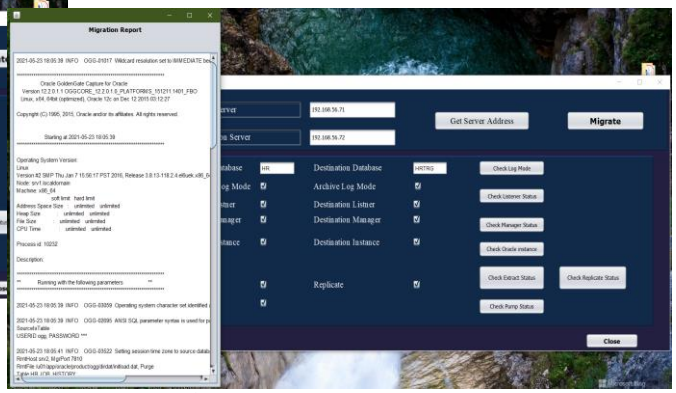


Figure 12: View of Migration Report- Migration completion record count.

7. Error Log generation.

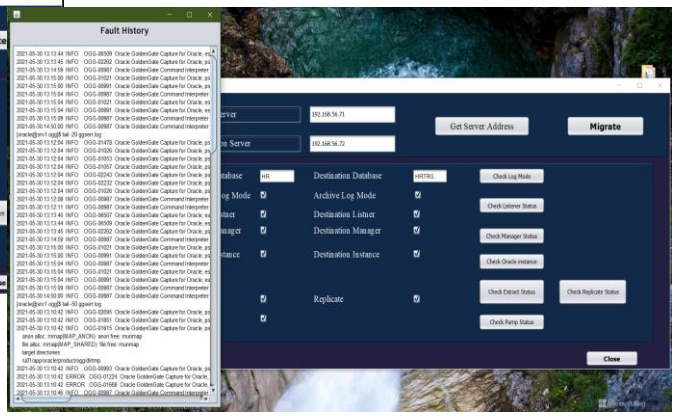


Figure 13: View of Error Log- Record of errored out transactions.

XIV. RESULTS

The result for the database migration can be expressed in two formats:

1. Table Record Count.
2. Schema Validation
3. Reconciliation Check

Table Record Count: This returns the count of the records of the specified table from the schema from both the source and destination servers.

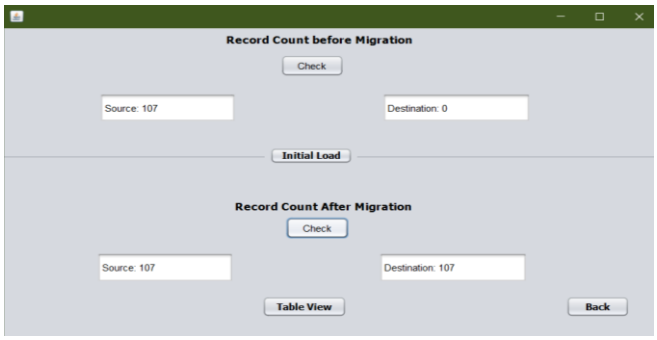


Figure 14: Table Record Count, Schema-HR, Table- Employees

Schema Validation : During data migration it is necessary to validate the schema, such as stored procedures, views, or user preferences as part of the data migration.
Here we have represented the schematic view of the Table Employees from HR schema from both source and destination servers for validating if the migration has been carried out successfully.

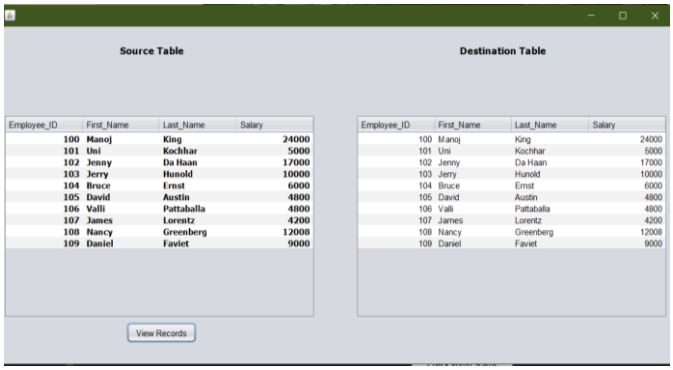


Figure 15: Schema Validation, Schema-HR, Table- Employees

Reconciliation Check: One of the most important aspects of validation is performing reconciliation checks on the source and target databases for all columns. This ensures that the data is not corrupted, date formats are maintained, and that the data is completely loaded.

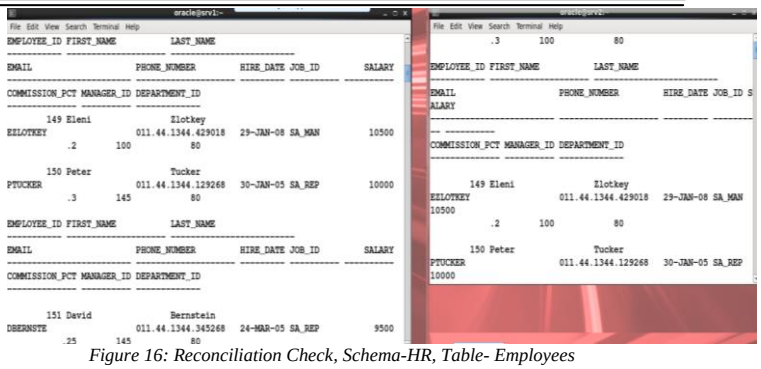


Figure 16: Reconciliation Check, Schema-HR, Table- Employees

FUTURE SCOPE

Along with the view of the functionalities on the dashboard and the source and destination servers we can also automate the entire installation process for Oracle Database, GoldenGate and all the other depending softwares and middlewares. This will also help in minimizing the time required to solve the upcoming errors.

The users with absolute no knowledge will find it easy to configure and install the necessary softwares along with the easy view of the functionality provided by us.

There can also be features added to show running processes using graphic libraries and suitable animation tools.

So with the automation of the entire configuration and installation processes with the automated system of migration, will prove to be the most efficient, effective and least time consuming along with easy handling of the entire system for not only the Database Administrators but also for Developers and users.

CONCLUSION

Thus we have **Automated the migration process by scripting** which makes migration easier with least manual intervention and with a user interface which provides the performance analysis and all the necessary information before and after the replication.

ACKNOWLEDGEMENT

We would like to thank our guide and co-author Prof. V.M.Lomte, for the undulating guidance and knowledge provided to us which was very much beneficial for the successful completion of this research topic. We would also like to take this opportunity to thank all our members for their

consistent hardwork and determination showcased during this entire process which led to the successful termination of our research work.
We would be obliged to extend our special thanks to Mr. Manoj Pund for nurturing us, guiding us, teaching us, throughout this time both in terms of time as well as resources. Without all the generous help received from sir, we wouldn't have been able to reach to this platform where we are able to elaborate about our completed research work.

Finally we would like to thank all the reviewers and peer reviewers who took out time from their schedule and helped us identify the nature of our paper, the contents of our paper that we were missing out and the quality of our paper. Without their help it would have been impossible for us to produce an end result even close to this. Thank you.

REFERENCES

[1]. *T. Jia, X. Zhao, Z. Wang, D. Gong, and G. Ding, ''Model transformation and data migration from relational database to MongoDB,'' in Proc. IEEE Int. Congr. Big Data (BigData Congr.), Jun./Jul. 2016, pp. 60–67.*

[2]. *C.-H. Lee and Y.-L. Zheng, ''Automatic SQL-to-NoSQL schema transformation over the MySQL and HBase databases,'' in Proc. IEEE Int. Conf. Consum. Electron.-Taiwan, Jun. 2015, pp. 426–427.*

[3]. *D. Serrano, D. Han, and E. Stroulia, ''From relations to multi-dimensional maps: Towards an SQL-to. -HBase transformation methodology,'' in Proc. IEEE 8th Int. Conf. Cloud Comput., Jun./Jul. 2015, pp. 81–89.*

[4]. *Virender Kumar, Cherry Kosla, "Data Cleaning – A thorough analysis and survey on Unstructured data," in Proc. IEEE 8th International Conference on Cloud computing , Data Science & Engineering, 2018*

[5]. *Zhibin Guan Tongkai Ji, Xu Qian, Yan Ma ,Xuehai Hong," A Survey on Big Data Pre-processing," in Proc. IEEE 2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD).*

[6]. *H. Gonzalez, J. Han, and X. Shen, ''Cost-conscious cleaning of massive RFID data sets,'' in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007*

[7]. *Arati Koli, Swati Shinde," Approaches used in efficient migration from relational database to NoSQL database," Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*

[8]. *Leonardo Rocha, Fernando Vale, Elder Cirilo, Darlinton Barbosa,and Fernando Mourao," A Framework for Migrating Relational Datasets to NoSQL," 2015 ICCS International Conference on Computational Science.*

[9]. *Bogdan Walek, Cyril Klimes," Expert system for data migration between different database management systems," Advances in Data Network, Communication, Computers and Materials.*

[10]. *M.Elamparithi, V.Anuratha," A review on Database Migration Strategies, Techniques and Tools," 2015 World Journal of Computer Application and Technonlogy.*

[11]. *Q. Xu, R. Zheng, W. Saad, and Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," IEEE Communications Surveys Tutorials, vol. 18, no. 1, pp. 94–104, Firstquarter 2016.*

[12]. *T. J. O'Shea and J. Corgan, "Convolutional radio modulation recognition networks," 2016. [Online].*

[13]. *A. Selim, F. Paisana, J. A. Arokkiam, Y. Zhang, L. Doyle, and L. A. DaSilva, "Spectrum monitoring for radar bands using deep convolutional neural networks," in IEEE GLOBECOM 2017.*

[14]. *J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Van Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," in ACM USENIX Security Symposium - Volume 15, 2006.*

[15]. *K. Gao, C. Corbett, and R. Beyah, "A passive approach to wireless device fingerprinting," in IEEE DSN 2010, June 2010, pp. 383–392.*

[16]. *I. O. Kennedy, P. Scanlon, F. J. Mullany, M. M. Buddhikot, K. E. Nolan, and T. W. Rondeau, "Radio transmitter fingerprinting: A steady state frequency domain approach," in IEEE VTC, Sept 2008, pp. 1–5.*

[17]. *V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in ACM MOBICOM 2008.*

[18]. *S. V. Radhakrishnan, A. S. Uluagac, and R. Beyah, "Gtid: A technique for physical device and device type fingerprinting," IEEE Transactions on Dependable and Secure Computing, Sept 2015.*

[19]. *F. Chen, Q. Yan, C. Shahriar, C. Lu, W. Lou, and T. C. Clancy, "On passive wireless device fingerprinting using infinite hidden markov random field," submitted for publication.*

[20]. *N. T. Nguyen, G. Zheng, Z. Han, and R. Zheng, "Device fingerprinting to enhance wireless security using nonparametric bayesian method," in IEEE INFOCOM, April 2011, pp. 1404–1412.*

[21]. *T. J. O'Shea and J. Hoydis, "An introduction to machine learning communications systems," 2017. [Online].*

[22]. *S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," IEEE Communications Magazine, vol. 56, no. 9, pp. 146–152, 2018.*

[23]. *A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS 2012.*

[24]. *Anduo Wang, Wenchao Zhouz, Brighten Godfrey_ Matthew Caesar, "Software-Defined Networks as Databases", University of Illinois at Urbana-Champaign, Georgetown University*

[25]. *Anduo Wang, "Database Criteria for Network Policy Chain", Temple University, SDN-NFVSec'18, March 21, 2018, Tempe, AZ, USA, adw@temple.edu*

[26]. *Chung-Ming Huang,Shu Chiang, Duy-Tuan Dao, Wei-Long Su, Shouzhi Xu, Huan Zhou, "V2V Data Offloading for Cellular Network based on the Software Defined Network (SDN) inside Mobile Edge Computing (MEC) Architecture*

[27]. *Ying-Guang Sun, "Access Control Method Based on Multi-level Security Tag for Distributed Database System", Computer Center ,Liaoning University of Technology, Jinzhou 121001, P.R.China*
[28]. *Anduo Wangy, Xueyuan Mei, Jason Croft, Matthew Caesar, Brighten Godfrey ,"Ravel: A Database-Defined Network", Temple University, University of Illinois at Urbana-Champaign*
[29]. *Ali Al-Haj, Benjamin Aziz, "Enforcing Multilevel Security Policies in Database-Defined Networks using Row-Level Security", School of Computing, University of Portsmouth Portsmouth, PO1 3HE, United Kingdom*
[30]. *Noemi Glaeser, Anduo Wang, "Access Control for a Database- Defined Network", University of South Carolina,Temple University, adw@temple.edu*
[31]. *Xu Guangxian, Zhao Yue, Gong Zhongsheng. "Design of secure network coding scheme by double encryption based on chaotic sequences[J]". Journal of Computer Applications, 2017, 37(12):3412-3416.*
[32]. https://enlyft.com/tech/products/goldengate
[33].https://www.infoclutch.com/installed-base/data-integration/oracle-goldengate/
[34].https://flywaydb.org/documentation/concepts/migrations#script-migrations
[35]. *E. Anderson · J. Hall · J. Hartline · M. Hobbes · A. Karlin · J. Saia · R. Swaminathan · J. Wilkes "Algorithms for Data Migration" , 16 July 2008, DOI 10.1007/s00453-008-9214-y*
[36].https://docs.oracle.com/en/database/oracle/oracle-database/12.2/upgrd/major-steps-in-the-upgrade-process-for-oracle-database.html#GUID-EE26CF0A-1B77-43D7-A613-9C6D0BF42DBB

Brightness Preserving Low Contrast Medical Image Enhancement Based on Local Contrast Stretching and Global Dynamic Fuzzy Histogram Equalization

[1]Mr. Vijay Panse, [2] Dr. Rajendra Gupta
[1,2] Department of Computer Science, Rabindranath Tagore University (AISECT) , Raisen, M. P., India

***Abstract*—Medical photographs such as MRIs, mammograms, X-rays, and CT scans offer pertinent detail on the soft tissue of the human brain, cancer, stroke, and a variety of other diseases. These photographs assist physicians in quickly identifying diseases. However, these images can have a poor contrast ratio, and to remedy this, we used a hybrid strategy that combines contrast stretching (CS) and Brightness Preserving Dynamic Fuzzy Histogram Equalization (BPDFHE).The below is how our suggested strategy will work: We began by doing pre-processing on the images to eliminate noise, and then calculated an accurate background surface and created a new foreground image by subtracting the estimated background from the original image. On a new image containing only foreground objects, we used the contrast stretching technique. Finally, after merging the background with enhanced foreground images, we improve the picture using the global BPDFHE process. Experiments were conducted using a variety of low-contrast medical images. Experiments demonstrate that this approach produces more reliable results than other types of contrast enhancement.**

Keywords- AMBE, BPDFHE, Contrast Stretching, Medical Image, PSNR.

I. INTRODUCTION

Improved medical imaging revolutionized the medical sector with the improvement in image accuracy, allowing doctors to detect illness in patients[1,2]. Indeed, image enhancement enables doctors to see fine distinctions of photographs that are impossible to identify or discern through the naked eye. Medical imaging techniques include computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), X-ray, PET-CT, and biomarkers. These images have been extensively used in the field of medical imaging for image acquisition due to their reliability and affordability [3].

Limitation of acquisition equipment, noisy channel transmission and faulty memory positions on equipment and inadequate illumination during signal capture or unfavourable weather environments during imaging are key causes of the low output of such pictures [4]. Thus, contrast, noise, information loss, poor lighting, blur and inappropriate color balance may be impaired in the image [5].

Improving contrast of images and sharpness of details while suppressing noise is therefore considered a necessity in the medical field [6]. Image enhancement involves manipulating an image by modifying its attributes to produce more appropriate and meaningful result for a specific task and human viewer.

II. LITERATURE SURVEY

Several enhancement techniques are available to enhance the image, some of the contrast enhancement techniques for medical images are disused in this section.

To get better contrast in digital images, a well-known method used, is Histogram Equalization (HE) [7]. This is very popular because it is very easy to implement and less expensive in the case of computation when we compare it to other existing methods. But it suffers with few limitations like unnecessary visual distortion may be introduced.

Global histogram equalization is not suitable for the situation where image contains low contrast in a specific local region, so a modified version called Adaptive Histogram Equalization [8], can be adopted, instead of global histogram equalization. It considers only those small

regions that contain low contrast based on their local CDF and perform the enhancement of contrast on those regions. However, the problem with AHE is that it does not prevent the over amplification of noise, and it is unable to increase contrast. To overcome the problem associated with AHE, improved version of AHE called CLAHE [9] can be used where noise can be reduced whereas it maintains the high spatial frequency regions of the image.

But all previous discussed techniques fail to prevent the image brightness at the same time as improving the contrast of image. To maintain the mean brightness of input image at the time of improving contrast of a given image, an updated method called BPDHE [10] was suggested. But all classical HE methods fail to prevent the excessive saturation of intensity bins causing the visual artifacts and loss of natural appearance in the processed images.

To solve the restriction of un-even expansion of intensity bins in classical multi-HE processes, many crisp and soft computing dependent algorithms have been implemented which either specifies the aim metric to monitor the enhancement process or pre-process the initial histogram before equalization process [11]. Along with crisp HE approaches, the fuzzy based HE algorithms have also been utilized to maintain the brightness and natural appearance of the pictures. Most of the fuzzy dependent HE methods either convert the crisp histogram into fuzzy histogram before segmentation phase or utilizes fuzzy logic for segmenting the crisp histogram into sub-histograms before enhancement process [12], [13].

J. Kinani et. al. [14] proposed a method based on Fuzzy system on intensity transformation through the implication method. Authors provided a better segmentation in image and achieve fast computation.

The authors of [15] concentrated on preserving picture brightness while enhancing local contrast in the original image. Their fuzzy logic–based histogram equalization (FHE) approach computes fuzzy histograms using fuzzy set theory. The fuzzy histogram is then divided in half depending on the initial image's median meaning.

[16] Proposes a novel approach for contrast enhancement based on fuzzy logic interpolation. This approach specifies a feature for transforming the degree of pixel intensity from a series of locally stretched pixel intensities.

The author of [17] proposed a fuzzy mapped HE approach in which they divided the histogram into several segments, then each segment to its maximum dynamic range using the fuzzy mapping algorithm, equalizing each segment separately, and finally normalizing the mixture of equalized segments.

In [18], the author suggested a novel method for enhancing low-contrast X-ray images dependent on brightness modification using a fuzzy gamma reasoning model. Their approach employs foreground and background intensity measurements to estimate G values using a fuzzy logic model, which will be used to adaptively operate the gamma function.

A procedure called brightness preserving dynamic fuzzy histogram equalization (BPDFHE) [19] was developed as an improved variant of brightness preserving dynamic histogram equalization (BPDHE) [10] in order to boost the method's brightness preservation and contrast enhancement capabilities while reducing its numerical difficulty. Fuzzy numbers are used to reflect and process visual images in order for the procedure to properly accommodate the imprecision of gray level values.

After analysis of the survey we identified that some images are poor in global contrast only and some are poor in local contrast only. However, for some images, besides local contrast, global contrast is also poor. For these types of images, we cannot opt only the global contrast or local contrast enhancement technique and to get improved output images, we can use the combination of global as well as local contrast enhancement.

We have proposed hybrid technique which takes the advantages of both the technique (BPDFHE and CS) with the concept of local and global contrast enhancement. Our proposed method work as follows: In the first step, we apply pre-processing to remove noises present in images, and then we estimated accurate background surface and obtained new foreground image by subtracting this estimated background from the original image. We performed local contrast stretching method on new image which contains only foreground object. Finally, after combining the background and enhanced foreground image we apply globally BPDFHE method to enhance image. The idea behind applying BPDFHE globally is to maintaining the mean brightness of the image.

The paper consists of Section III which addresses general Material and Methods used in proposed methodology. Experiments are detailed in Section IV. Conclusion is described in Section V.

III. MATERIAL & METHODS

A. Median Filter

Medical images [20] suffer from various noisy data, so the de-noising is required to get high MR image quality. In this paper, we have adopted median filter to de-noise the speckle noise from the image. The reason behind using the median filter is to its robustness against the impulsive noise and to eliminate spurious data from MR images. In this type of filter, current pixel value is replaced by mid value of the pixels that is selected surrounding the pixel under consideration.

B. Background Approximation

By using background approximation, we are capable of accurately extracting non-uniform background from the original image. The performance of non-uniform lighting varies according to the factor s(x, y) to be extracted from the Image, I(x, y). i.e.

$$f(x; y) = s(x; y)I(x; y) + n(x; y) \quad (1)$$

Where f is the observed image, s is the true signal, I is the non-uniform illumination field and n is additive noise.

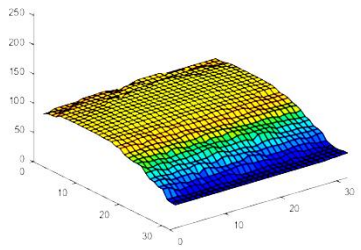


Figure 1. Image Background Estimation

Here, S(x, y) is a multiplicative quantity with the image and is difficult to remove as it is variable [21].

C. Contrast Stretching

One of the simplest point processing approaches is contrast stretching [22] which allows high dynamic range images. In this method contrast is increased by making the dark areas darker, and the lighter areas lighter.

Upon the reception of the Improved I component, Contrast Stretching is done. That will increase the overall intensity globally.

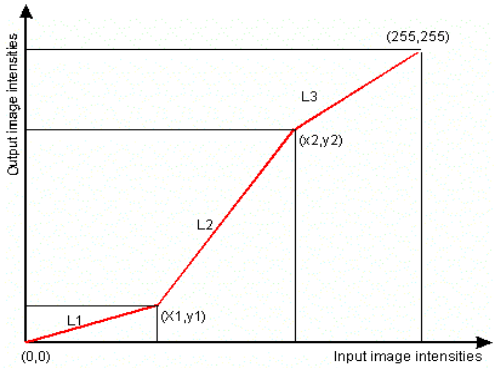


Figure 2. Contrast Stretching Process

Contrast stretching can be calculated as given below.

$$I' = \begin{cases} L1.I & 0 \le I < x1 \\ L2.(I - x1) + y1 & x1 \le I < x2 \\ L3.(I - x2) + y2 & x2 \le I < 255 \end{cases} \quad (2)$$

Where, L1, L2 and L3 are the slopes. It is clear from the figure that L1 and L3 are less than one while L2 is greater than one.

D. Brightness Preserving Using DFHE

This section focuses with BPDHE assessment processes. The BPDHE is an upgrade of DHE in that it uses almost the same contrast for the output and input images. It has many phases.

1) Fuzzy Histogram Computation

A fuzzy histogram is a sequence of real numbers $h(i)$ $i \in \{0,1,\ldots,L\text{-}1\}$, where $h(i)$ is the frequency of occurrence of gray levels that are "around i ". By considering the gray value $I(x, y)$ for the image I as a fuzzy number $F(x,y)$, the fuzzy histogram is computed as:

$$h(k) \leftarrow h(k) + \sum_i \sum_j \mu_{F(x,y),k} \qquad \forall k : 0 \le k \le L-1 \quad (3)$$

where $\mu_{F(x,y),k}$ s the fuzzy membership function defined as in eq.4 with constant $\delta = 4$

$$\mu_{F(x,y),k} = \max\left(0, 1 - \frac{|I(i,j) - k|}{\delta}\right) \quad (4)$$

2) Local Maximum

We first identified the position of the highest points in the measured histogram using first and second derivative of fuzzy histogram, and then split the fuzzy histogram into n areas denoted by $\{m_1, m_2, ..., m_n\}$. The region for each partition is determined using the formula.

$$span_i = high_i - low_i \quad (5)$$

$$factor_i = span_i \times \log_{10} M \quad (6)$$

$$range_i = (L-1) \times factor / \sum_{k=1}^{n+1} factor_k \quad (7)$$

Where $high_i$ is the maximum intensity value contained in the sub histogram; low_i is the minimum intensity value; M is total pixels.

3) Histogram Equalization

In this step fuzzy histogram of each interval is equalized separately. The transformation function is:

$$y(x) = start_i + (end_i - start_i) \sum_{k=start_i}^{x} \frac{h(k)}{M_i} \quad (8)$$

4) Image Normalization

The last step involves normalizing the intensity of the output image by approximates the mean intensity of the input image. Therefore, the brightness of the resulting image would be the same as the brightness of the input. The Normalization function has the following form:

$$g(x,y) = (M_i / M_O) f(x,y) \quad (9)$$

Where $g(x,y)$ the normalized output is image and $f(x,y)$ is the output just after the equalization.

After applying BPDHE to the image, the contrast is enhanced both subjectively and objectively compared to other enhancement methods.

Algorithm 1 shows the step by step procedure of proposed methodology for enhancing the quality of low contrast medical image.

Algorithm1: Brightness Preseving Medical Image Enhancement Using Contrast Stretching and BPDHE.
Y=MIE_CS_PLUS_BPDFHE(X)

Input: Low contrast Medical Image X.
Output: Contrast Enhanced Medical Image Y
Procedure:

1. Apply pre-processing using median filter to remove impulse noises present in images.

$X' \leftarrow MedianFilter(X)$

2.Estimate accurate background approximation as a surface from the image received from the previous step.

$B \leftarrow BGApprox(X')$

3.Extract the non-uniform background from the input image and making a new foreground image by subtracting this estimated background from the image received form step one.

$F \leftarrow X' - B$

4. Perform contrast stretching method on foreground objects which containing only Region of interest (ROI).

$\hat{F} \leftarrow ContrastStretch(F)$

5. Combining the background and enhanced foreground image.

$\hat{X} \leftarrow \hat{F} \bigcup B$

Apply global BPDHE method to enhance image to preserve the brighness and enhacement globally.

$Y = BPDFHE(\hat{X})$

End

The experiments were carried out on various low contrast medical images. Proposed Flow chart of given image is shown in following figure 3. In figure we can see that after applying the preprocessing on input image we get the noise free image from which background and foreground are separated. In the foreground image, contrast stretching is applied that give us enhanced foreground image which is merged with background image. Finally BPDFHE is applied on the merged image globally that not only improved the contrast of image but also preserve the brightness.

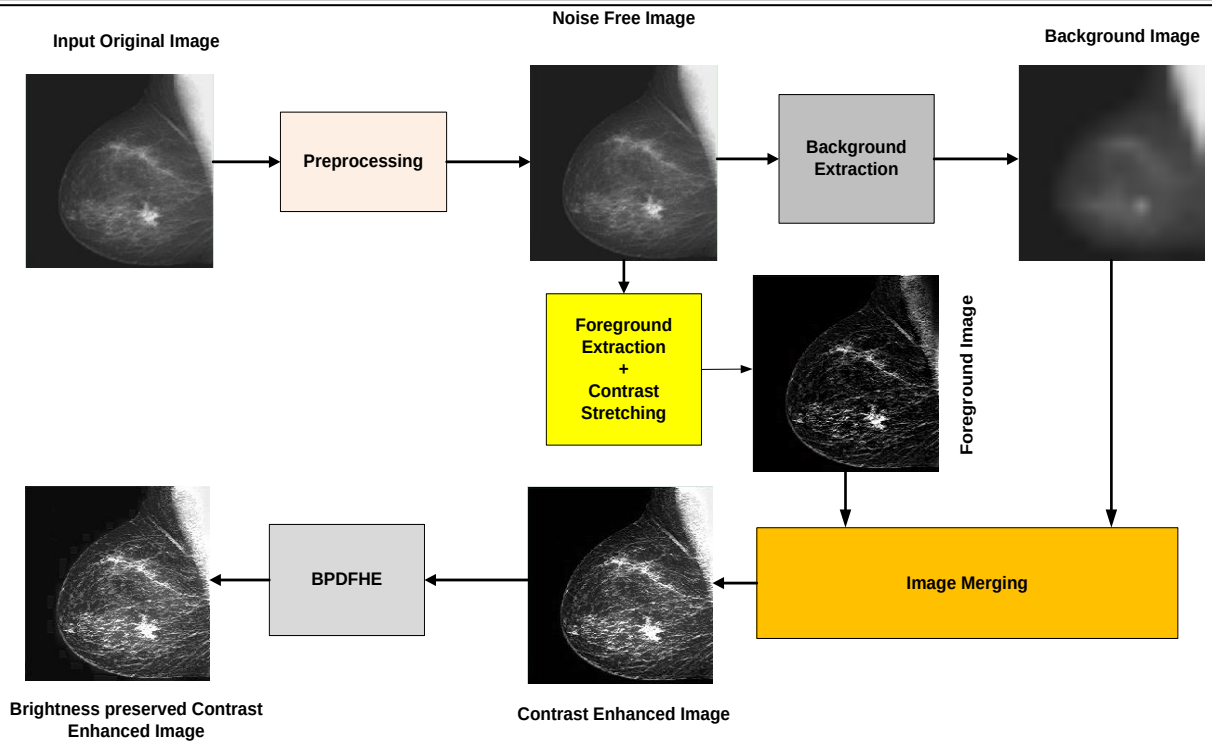


Figure 3. Proposed Block Diagram

IV. Result Analysis

In this section, we conducted experiment to analyze the performance of the proposed method in comparison with some existing HE based contrast enhancement methods, like HE, CLAHE and BPDHE. The performance is calculated on AMBE, PSNR, entropy and contrast (C) [23].

We have conducted the results on 10 medical images, these images are provided online at [24].

The enhanced images produced by the proposed methods are presented in Figures4 to 8 with their histogram.

For the exhaustive analysis of image quality, images have poor visibility in the underexposed regions and visibility is high in the overexposed regions.

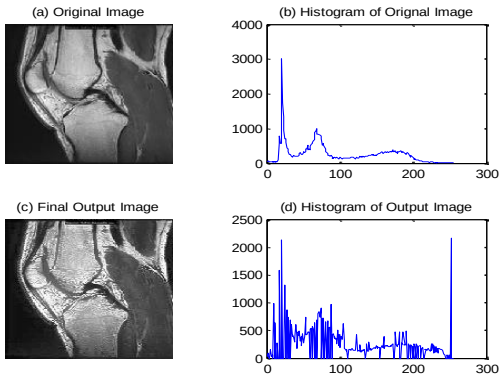


Figure 4. (a) Image MRI-of-knee-Univ-Mich, (b) Histogram of Original Image (c) Enhanced Image with Proposed Method (d) Histogram of Output Image

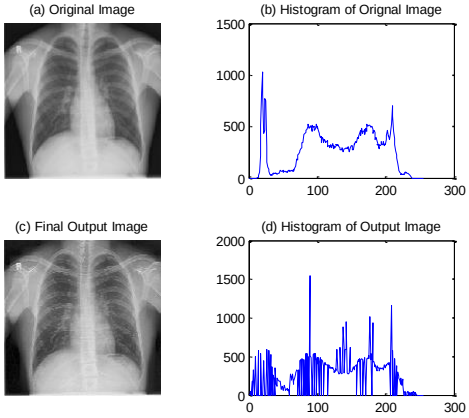


Figure 5. (a) Image chest-xray-vandy, (b) Histogram of Original Image (c) Enhanced Image with Proposed Method (d) Histogram of Output Image

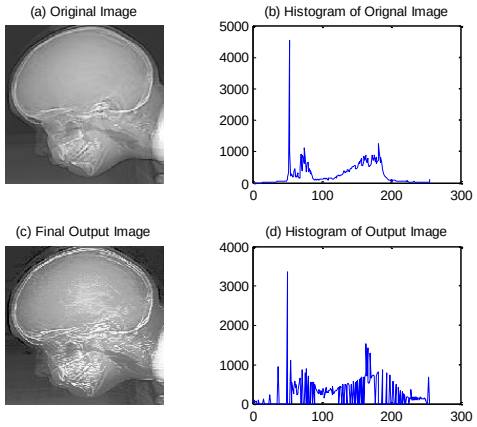


Figure 6. (a) Image ctskull-256, (b) Histogram of Original Image (c) Enhanced Image with Proposed Method (d) Histogram of Output Image

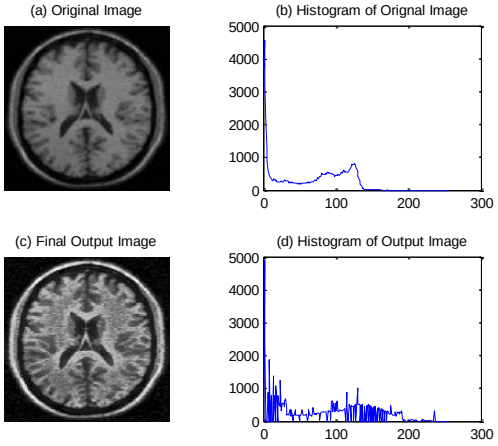


Figure 7. (a) Image headCT_Vandy, (b) Histogram of Original Image (c) Enhanced Image with Proposed Method (d) Histogram of Output Image

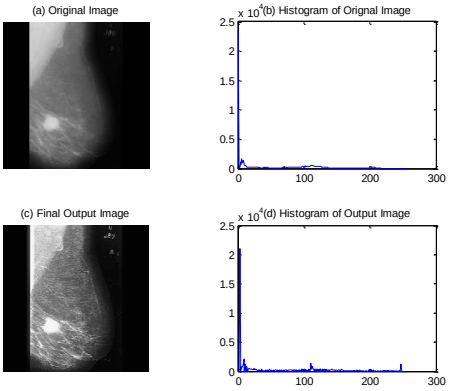


Figure 8. (a) Image Image mdb028, (b) Histogram of Original Image (c) Enhanced Image with Proposed Method (d) Histogram of Output Image

Table 1 shows the experimental result on various medical images conducted by the proposed method.

TABLE I. EXPERIMENTAL RESULT ON VARIOUS MEDICAL IMAGES.

Image Name	AMBE	PSNR	Entropy	Contrast
angiography_live_img	3.7416	24.7438	0.2252	0.2935
breast_digital_Xray	0.7378	20.0471	0.1388	0.4198
chest-xray-vandy	1.7925	32.1827	0.1854	0.3275
ctskull-256	4.1692	24.2475	0.1481	0.3590

headCT_Vandy	0.4255	18.4002	0.2919	0.7123
kidney_original	1.9460	18.6044	0.2516	0.5596
MRI-of-knee-Univ-Mich	1.5277	20.904	0.1134	0.5249
MRI-spine1-Vandy	6.3828	22.7588	0.2768	0.6869
mdb28	0.4821	22.7268	0.1310	0.3525
Meningioma_general	1.8135	20.6117	0.15309	0.70738
Average	**2.3019**	**22.5227**	**0.1915**	**0.4943**

Performance of proposed method is compared with other existing method and it is given in Table II.

TABLE II. PSNR COMPARISON OF PROPOSED METHODOLOGY ON VARIOUS LOW CONTRAST MEDICAL IMAGES AGAINST OTHER METHODS

Image Name	HE	CLAHE	CS+BP DHE	CS+BPD FHE
angiography_live_img	14.6997	19.9763	21.0623	24.7438
breast_digital_Xray	10.7449	21.8992	18.8272	20.0471
chest-xray-vandy	21.9830	20.5826	22.1768	32.1827
ctskull-256	18.9265	21.6344	17.2943	24.2475
headCT_Vandy	10.9988	14.4415	18.4095	18.4002
kidney_original	15.0352	17.4653	17.3058	18.6044
MRI-of-knee-Univ-Mich	16.1980	20.1443	22.2014	20.9040
MRI-spine1-Vandy	11.6973	18.1374	22.3530	22.7588
mdb28	8.8365	20.6631	25.9722	22.7268
Meningioma_general	13.4786	19.5030	20.6361	20.6117
Average	14.2599	19.4447	20.6239	**22.5227**

The Table 3 shows results of AMBE. We can see from table that our method has least values as compared to other methods. This shows that our proposed method is able to maintain mean brightness in the processed image.

TABLE III. AMBE COMPARISON OF PROPOSED METHODOLOGY ON VARIOUS LOW CONTRAST MEDICAL IMAGES AGAINST OTHER METHODS

Image Name	HE	CLAHE	CS+BP DHE	CS+BPD FHE
angiography_live_img	2.5059	9.6007	0.6853	3.7416
breast_digital_Xray	61.7233	10.7336	9.5313	0.7378
chest-xray-vandy	1.2835	0.6893	0.2981	1.7925
ctskull-256	1.8533	1.0599	0.1675	4.1692
headCT_Vandy	65.8141	37.5725	15.8668	0.4255
kidney_original	24.2129	19.4468	12.6998	1.9460
MRI-of-knee-Univ-Mich	34.1720	13.0401	10.8841	1.5277
MRI-spine1-Vandy	61.7342	17.6749	1.0955	6.3828
mdb28	91.3363	11.9717	4.5378	0.4821
Meningioma_general	47.0833	18.3250	5.8449	1.8135
Average	39.1719	14.0115	6.1611	**2.3019**

Figure 9 and Figure 10 shows the PSNR and AMBE comparison graphs of proposed methodology against the other existing methods respectively.

V. CONCLUSION

Medical Image enhancement offers extensive range of approaches for enhancing medical images to achieve visually acceptable images. In this paper we applied pre-processing to remove noises present in images then we estimated accurate background surface and obtained new foreground image by subtracting this estimated background from the original image. We perform contrast stretching method on foreground image which containing only required object. Finally, after combining the background and enhanced foreground image we apply global BPDFHE method to enhance image. The experiments were carried out on various low contrast medical images. Experimental results indicate that the proposed algorithm achieves good quality output image in term of PSNR, AMBE etc. parameters.

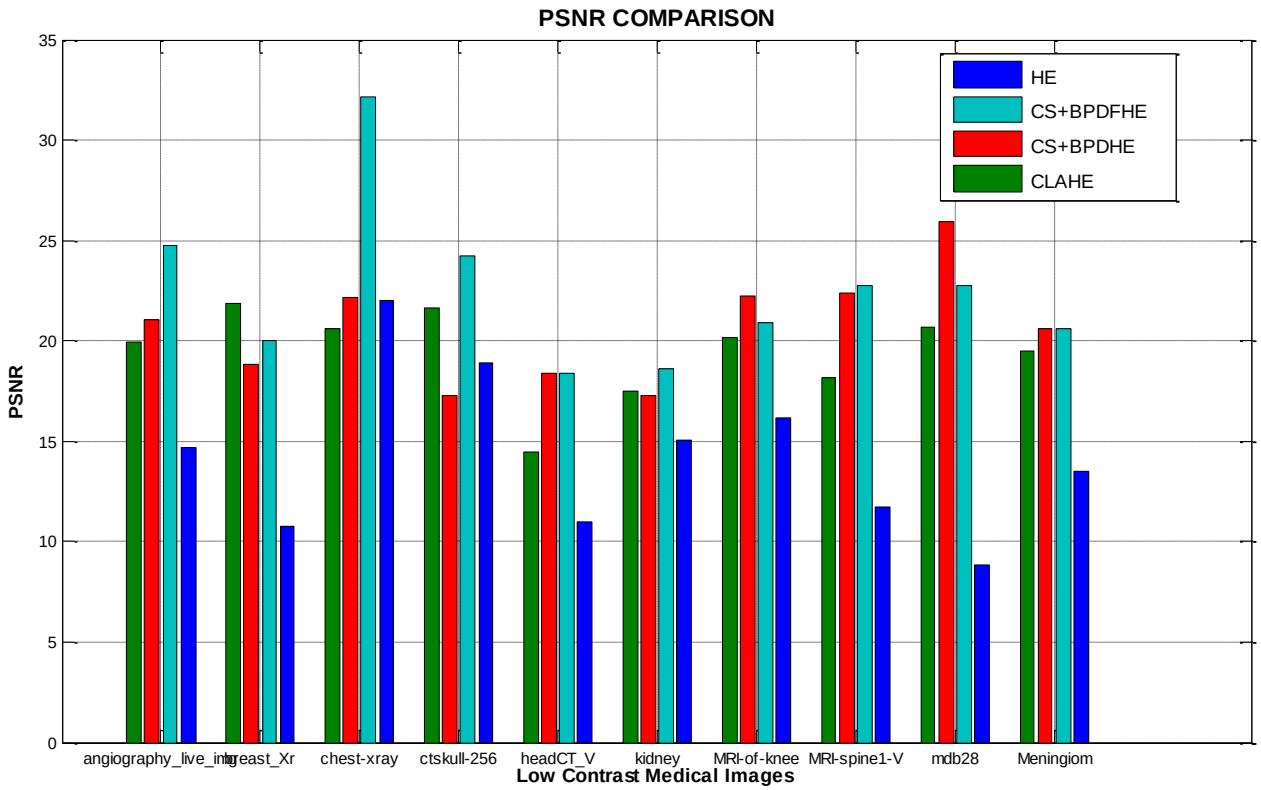


Figure 9. PSNR comparison graph of proposed methodology on various low contrast medical images.

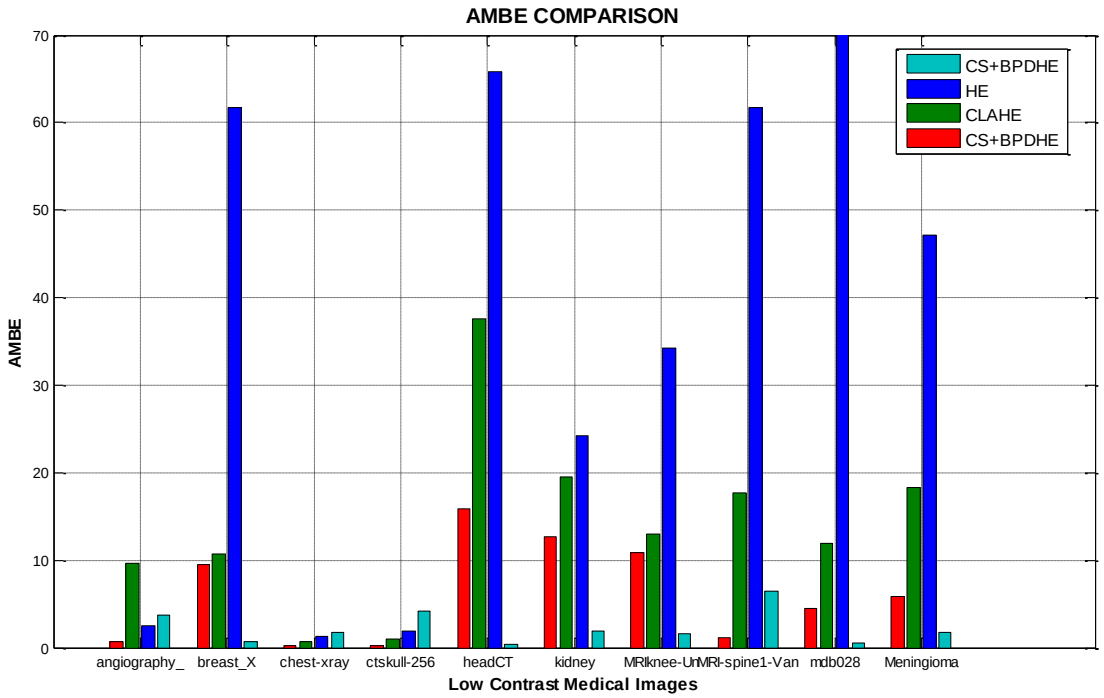


Figure 10. AMBE comparison graph of proposed methodology on various low contrast medical images

REFERENCES

[1] Huang RY, Dung LR, Chu CF, Wu YY (2016) Noise removaland co ntrast enhancement for x-ray images. J Biomed Eng MedImaging 3(1):56

[2] Menon HP, Rajeshwari B (2016) Enhancement of dental digitalx-ray images based on the image quality. In: The internationalsymposi um on intelligent systems technologies and applications.Springer, pp 33–45.

[3] T. Gong, T. Fan, L. Pei and Z. Cai, "Improved immune algorithm for medical image enhancement," 2015 International Workshop on Artif icial Immune Systems (AIS), Taormina, Italy, 2015, pp. 1-7, doi: 10.1109/AISW.2015.7469241.

[4] Voronin V, Semenishchev E, Ponomarenko M, Agaian S (2018)Com bined local and global image enhancement algorithm. ElectrImaging 2018(13):1–5

[5] alih AAM, Hasikin K, Isa NAM (2018) Adaptive fuzzy exposure local contrast enhancement. IEEE Access 6:58,794–58,806.

[6] Mustapha A, Hussain A, Ahmad WSHMW, Zaki WMDW,Hamid HBA (2019) Cbir-dsn: integrating clustering and retrievalplatforms for disk space narrowing degradation assessment.Multimed Tools Appl 78(13):18887–18919.

[7] E. Davies Machine Vision: Theory, Algorithms and Practicalities, Academic Press, 1990, pp 26-27, 79-99.

[8] S. M. Pizer, E. P. Amburn, J. D. Austin, et al.: Adaptive Histogram Equalization and Its Variations. Computer Vision, Graphics, and Image Processing 39 (1987) pp. 355-368.

[9] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," Chapter VIII, Graphics Gems IV, P.S. Heckbert (Eds.), Cambridge, MA, Academic Press, pp. 474-485, 1994.

[10] H. Ibrahim, N. S. P. Kong, "Brightness Preserving Dynamic Histogram Equalization for Image Contrast Enhancement" , IEEE Transactions on Consumer Electronics, vol. 53, Issue: 4, Nov. 2007.

[11] F. Khan, E. Khan, and Z. A. Abbasi, ''Segment dependent dynamicmulti-histogram equalization for image contrast enhancement,''Digit. Sig-nal Process., vol. 25, pp. 198–223, Feb. 2014.

[12] D. Sheet, H. Garud, A. Suveer, M. Mahadevappa, and J. Chatterjee,''Brightness preserving dynamic fuzzy histogram equalization,''IEEETrans. Consum. Electron., vol. 56, no. 4, pp. 2475–2480, Nov. 2010.

[13] M. F. Khan, X. Ren, and E. Khan, ''Semi dynamic fuzzy histogramequalization,''Optik, vol. 126, no. 21, pp. 2848–2853, Nov. 2015

[14] J. M. V. Kinani, A. J. R. Silva and F. J. Gallegos, "Fuzzy C-means applied to MRI images for an automatic lesion detection using image enhancement and constrained clustering", 4th International Conference on IPTA, Oct. 2014.

[15] Magudeeswaran V, Ravichandran C (2013) Fuzzy logic-basedhistogram equalization for image contrast enhancement. Mathe-matical Problems in Engineering 2013

[16] Panda SP (2016) Image contrast enhancement in spatial domainusing fuzzy logic based interpolation method. In: 2016 IEEEstudents' conference on Electrical, electronics and computerscience (SCEECS). IEEE, pp 1–4.

[17] M. F. Khan, E. Khan, M. M. Nofal and M. Mursaleen, "Fuzzy Mapped Histogram Equalization Method for Contrast Enhancement of Remotely Sensed Images," in IEEE Access, vol. 8, pp. 112454-112461, 2020, doi: 10.1109/ACCESS.2020.3001658.

[18] Mouzai, Meriem & Tarabet, Chahrazed & Mustapha, Aouache. (2020). Low-contrast X-ray enhancement using a fuzzy gamma reasoning model. Medical & Biological Engineering & Computing. 58. 10.1007/s11517-020-02122-y.

[19] heet D, Garud H, Suveer A, Mahadevappa M, Chatterjee J (2010)Brightness preserving dynamic fuzzy histogram equalization.IEEE Transactions on Consumer Electronics 56(4).

[20] R. C. Gonzalez and R. E. Woods, 2008, Digital Image Processing, 3r d Edition, Prentice Hall.

[21] P. Singh, A. K. Garg, "Morphology Based Non Uniform Background Removal for Particle Analysis: A Comparative Study", IJCCR, Nov. 2011.

[22] F. G. Mohammed, H. M. Rada, S. G. Mohammed, "Contrast and Brightness Enhancement for Low Medical X-Ray Images", IJSER, vol. 4, Issue 5, 2013.

[23] Shweta, K. Viswanath, "A Review of Enhancement Techniques on Medical Images", IJACET, vol. 4, Issue-3, 2017.

[24] Medical Images Database, "http://www.imageprocessingplace. com/DIP-3E/dip3e_book_images_downloads.htm".

Pandemic Crisis Fraternity

[1] Akshat sharma, [2] Mayank singh, [3]Sourish Keshav
[1,2,3] Department of Computer Science Galgotias University
Greater Noida, India

***Abstract*— As covid is on the apex treat the world is facing right now with almost 17.5Cr cases and still rising with an indomitable pace. All though some countries claim to have successfully prepared covid vaccine and India is one of those countries but the availability of the vaccine are not so good and if we were to import vaccine from one of those countries we don't know if they will be willing to share it free of cost or not and if they do charge there has not been any globally fixed price. Keeping this in mind me and my teammates have decided to build an app/website or both as soon as possible in which we will be helping people to be able to wield the second best option to fight against covid being plasma treatment we will be helping people to donate plasma and buy plasma by creating a link between the donor and the buyer, between the needy and the helper. And not only that this app/website will help the covid patients to seek the nearest hospital where they can get treatment. There will also be a charity box option where some people can donate money as well to help the people who can't get treatment because of their economical status. In future the app can be very helpful to fight against the unpredictable pandemics, the boundaries are not restricted for only covid treatment. Our aim is to make India as safe as possible because with the population so high controlling covid is very hard and relying on the vaccine only is too risky. We need to take a step now before it's too late.**

Keywords- Covid-19, Smart technologies, Smart cities, Techno-driven, Human-driven, Pandemic, Privacy

INTRODUCTION

whole world is in crisis and is facing many problems because of this disease. Cases and deaths due to corona infection increase day by day. India currently has the highest number of confirmed cases in Asia, and has the third highest number of certified cases in the world after the U.S. and Brazil. Well it is true that countries say they have developed vaccines but also just say they have developed vaccines. There is no prescriptive drug that will stop the flow of emotions, though their effects can be curtailed. We can't just rely on claims we need to find another way to ensure our safety first rather than waiting for something that hasn't been properly disclosed. And even if we think once that a failed vaccine was developed but the time required for the successful distribution of the vaccine will not be expected because looking at the people of the world this task is not easy especially in some remote areas it may take longer than expected the most you can do.

Now we are moving one step further into the speculation of a failed vaccine and now we are facing another barrier to the cost of the vaccine, whether it will cost the vaccine, whether it will be free or not, and if it does not cost how much it will cost people will be able to buy the vaccine. So we see that we have a lot of problems from now on in the course of covid therapy which is why we decided to go with what we currently have which is plasma therapy which will help the people to save their life from this pandemic

LITERATURE SURVEY

First, internet and devices availability has grown immensely so the application will be available easily.

Plasma is a component of blood. The simple process involves extracting the blood containing the antibodies of the infection the donor has recovered from. Once the plasma is separated from the blood, doctors administer it to people infected with the same virus, in this case, COVID-19.Plasma — the liquid component of blood — contains antibodies. Extracting plasma from someone that has "convalesced," or recovered, from an illness might provide a much-needed boost to the immune system of someone grappling with coronavirus.Covid-19 will probably be a part of our lives for the foreseeable future.

A vaccine has yet to materialize. As the pandemic continues to take a crushing toll, doctors are resorting to a century-old treatment that has been helpful in managing previous pandemics: taking antibodies from those who have recovered and giving it to the sick. It's known as

convalescent plasma therapy, or "survivors' blood."A plasma transfusion involves removing some antibodies from one person and infusing them into someone who is sick, providing an immediate jolt to their immune system. A dose of antibodies doesn't directly stimulate a person's immune system to start creating their own antibodies, but it does offer some protection until their own immune system ramps up.

Mounting an antibody response isn't exactly a speedy process. It generally takes one to three weeks for the immune system to produce antibodies against COVID- 19.

PURPOSE AND SCOPE

The aim of this project is to create an app that will help people fight covid 19 using plasma treatment. We will help people donate and buy plasma by building a connection between the donor and the buyer and we will also help patients with specific colors to go to the nearest hospital for plasma treatment. The vision of this project is to provide an effective and easy way for people to receive covid treatment. While the COVID vaccine is still in its infancy and the proven COVID solution remains unclear, plasma recovery treatment has yielded positive results in cases, although similar clinical trials are still ongoing. Building an app can help many people in need to recover. Going to the hospital, registering, waiting hours can put your own life and the lives of others in danger. So to alleviate the problem, we discussed building an app so you don't get stuck in the queues and make your life easier just by staying home. Here we will help those in need to communicate directly with those who wish to serve only. In all three groups of COVID patients, plasma treatment is recommended for patients moderately infected. In this case you just need to determine if you have a low, moderate or mild infection and the time of the virus first attacked in your body with more details and you will be given steps to proceed

REQUIREMENTS SPECIFICATIONS

Tools requirement :-

Android studio:- android studio is the most powerful software for developing any android application . it provide the user a interface to design its application as well as to write code for it.

Eclipse:- Eclipse is IDE for developing application using the java programming language and many more programming languages .

Android emulator:- The emulator lets you prototype , develop and test the android application without using a physical device .

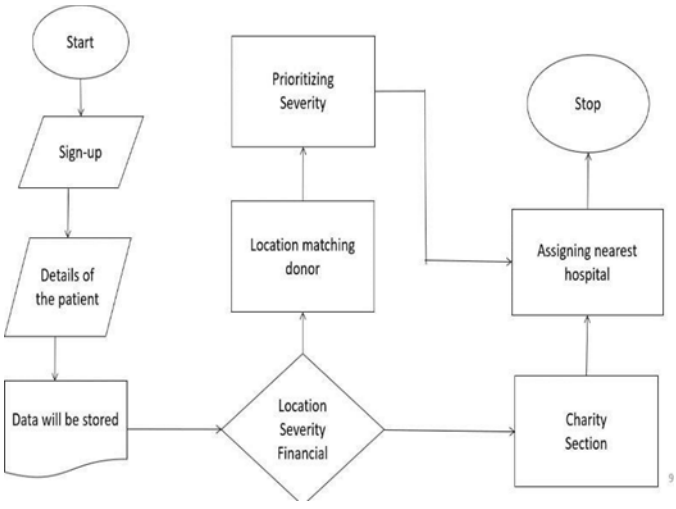
Live device:- Any live device to see how the application is running in live device

ARCHITECTURE

The new user (who needs plasma for the treatment) should sign in first by filling his basic details like name, address, phone number etc.

The application will store the user information safely in database, than the application will check all the details of the patient and will provide him location matching donor according to his location and severity filled by the user and will suggest him the nearest hospital where the patient can get the treatment .

There will be charity section in the application for the people who are not able to get the treatment because of their economical status, are that the charity section will work for only those people whose income will be less than a specific amount. People can donate as much as they want in the charity section so that everybody can get the treatment.



IMPLEMENTATION

The following features are going to be implemented in the application which will help the people a lot:-

Quick donor check button:-

There will be a Quick donor check button which will show the list of donor available, near the patient (according to the location filled by the patient)so that the patient can find the donor easily .

Charity section:-

There will be a charity section where people can donate for the people who are not financially stable.

Home profile:-

There will be a profile of each patient who has signed up for the application where he can see and edit all his details filled by him at the time of signing up.

Hospitals availability:-

The application will also going to show the list of all the hospitals where the patient can get the plasma treatment.

Prioritizing severity:-

The application will also look up for the patient first who need more help i.e. giving priority to the person who have more symptoms of covid like breathlessness rather than going for the person with less symptoms and less chances of loosing life.

FUTURE ENHANCEMENTS

This application can further be used by government to helpthe people in need and also after some time by doing some modification the app will not be restricted only for the patient who are fighting with covid, the app can also be used for fighting many other diseases and many other languages other than English can also be added to the application so it can be used globally.

CONCLUSION

We want to establish a link between the donor and the patient who is the need for plasma donation to fight againstnot only covid and help them attain a fighting chance so they don't lose their lives accidentally. We will not only create link but also guide them to the nearest hospital to get the best possible treatment. If possible we also would like to help people economically as well especially who are poor and can't afford the treatment on their own. By doing this we will be helping people to attain a fighting chance against covid pandemic and helping India to achieve a better and stable health status among other countries.

The application will be as simple as possible to use so people who don't have a good knowledge to operate any phone can also use it to the utmost potential

REFERENCES

[1] Casadevall, A. & Pirofski, L. A. The convalescent sera technique for including COVID-19. J. Clin. Invest. 130, 1545– 1548 (2020).

[2] Rojas, M., et al. Convalescent plasma in COVID-19 and the possible way of action. The treatment of autoimmune diseases. Rev 19, 102554 (2020).

[3] Sharun, K., et al. safety-tested antibody, and the use of convalescent plasma to the fight against COVID-19: the progress and potential. Opin expert. Was. Ther. 20, 1033-1046 (2020).

[4] Richardson, S., et al. showing the attribution, disease-related outcomes between 5,700 patients at the hospital with COVID- 19 in New York city. JAMA 323, 2052-2059 (2020).

[5] Duan, K., et al. The effect or effectiveness of plasma therapy in severe patients of COVID-19. Proc. Natl Make. sciences". The US-117, 9490-9496 (2020).

[6] Zhou, F. et al. Clinical course, and risk factors for death in adult patients in the inpatient treatment of COVID-19 in Wuhan, China: a retrospective cohort study. Utility knife, 395, 1054-1062 (2020).

[7] Zhao, J. et al. Antibody reaction to SARS-CoV-2 in people of novel coronavirus disease 2019. Clin. Infect.Dis. https://doi.org/10.1093/cid/ciaa344 (2020).

[8] Chen, L., Xiong, J., Bao, L., & Shi, Y. Convalescent plasma as a potential treatment for COVID-19. The scalpel is contagious. Dis. 20, 398-400 (2020).

[9] Roback, J. D. & Guarner, J. Convalescent plasma for the treatment of COVID-19: challenges and opportunities. JAMA 323, 1561-1562 (2020).

[10] Luke, T. C., Kilbane, E. M., Jackson, J. L. & Hoffman, S. L. Meta-analysis: convalescent blood products for the treatment of inflammation of the lungs of the Spanish flu pandemic, the future of the treatment of H5N1? Ann. Intern. By. 145, 599– 609 (2006).

[11] Mair-Jenkins, J. et al. The efficacy of convalescent plasma and hyperimmune immunoglobulin for the treatment of severe, acute respiratory infections, and the virus's origin: a systematic review and meta-analysis of the studies. Dis. 211, 80-90 (2015).

[12] Enria, D. A., Briggiler, A. M., Fernandez, N. J., Levis, S. C. & Maiztegui, J. I. the Importance of neutralizing antibodies, the dosage in the treatment of Argentine haemorrhagic fever and the immune system of the plasma. Utility knife, 2 in, 255-256 (1984).

[13] Lee, J. S. et al. The Anti-ebola therapies for patients with Ebola virus disease: a systematic review. To infect the COMPANY. Dis. 19, 376 (2019).

[14] That is, M., et al. Convalescent plasma therapy for COVID - 19 patients, in Wuhan, China. Virol. https://doi.org/10.1002/jmv.25882 (2020 г.).

[15] Salazar, E. et al. Treatment of COVID-19 patients with convalescent plasma. Am. J. Pathol. 190, 1680–1690 (2020).

[16] Perotti, C. et al. The reduction of the mortality rate, there are 46 of severe COVID-19 patients who were treated with hyperimmune plasma. For example, the concept of the one- handed multi-center intervention study. Haematologica https://www.haematologica.org/content/early/ 2020/07/20/haematol.2020.261784.long (2020).

[17] Gharbharan, A. et al. Convalescent plasma for COVID-19. Randomized clinical trial. Preprint at https://www.medrxiv.org/content/10.1101/2020 .07.01.20139857 v1 (2020).

[18] Li, L. et al. The effect of the convalescent-plasma therapy, the time to clinical improvement in patients with severe and life-threatening COVID-19: a randomized clinical trial. JAMA 324, 460 to 470 (2020).

[19] Finlandia, M. Serumotochnoe lechenie lobarnoi pnevmonii [Serum in the treatment of lobar pneumonia]. N. Engl. J. Med. 202, 1244-1247 (1930).

[20] Cecil, R. L Remarks on the TREATMENT of inflammation of the lungs, in the BLOOD SERUM. Br. Med. J., 2, 657-662 (1932).

[21] Hung, I. F. N. et al. Hyperimmune IV immunoglobulin treatment: a multi-center, double-blind, randomized, controlled study of patients with severe 2009 influenza A (H1N1)virusnfection. Chest, 144, 464-473 (2013)

[22] Austin, P. C. a Comparison of 12 algorithms at scale, according to the tolerance. State to state. Med. 33, 1057-1069 (2014).

[23] Amanat, F. et al. Serological assay for the detection of seroconversion of the SARS-CoV-2 in humans. Nat. Med. 26, 1033-1036 (2020).

[24] Stadlbauer, D. et al. The human SARS-CoV-2 seroconversion, what it is: a detailed description of the protocol for the serological analysis of antigens and production of test setup. Carr. "No. Microbiol. 57, e100 (2020).

[25] Wajnberg, A. et al. The SARS-CoV-2 infection as the cause of persistent neutralizing antibodies are stable for at least three months of the year. Preprintathttp://medrxiv.org/content/early/2020 /07/17/2020.07.14

[26] .20 151126.abstract (2020).

[27] Choe, P. G. et al. Antibody response to the SARS- CoV-2 and 8 weeks post-infection in the asymptomatic patient. Emerg. Infect. Dis. 26 (2020).

[28] Klein, S., et al. The gender, the age, and to determine the hospital in response to the antibody, as reconvalescent donor, the blood of COVID-19 in the plasma . To put it down. https://doi.org/10.1172/jc1142004 (2020).

[29] Okba, G., M. A., et al. The Severe Acute Respiratory Syndrome Coronavirus 2-specific antibody responses
[30] of the patients with the coronavirus disease. Emerg. Infect. Dis. 26, 1478-1488 (2020).

[31] Joyner, M. J., et al. Safety Update: COVID-19 convalescent plasma is 20 000 in-patients. Klin, swimsuit, bathing suit.
[32] Proc. https://www.mayoclinicproceedings.org/article/ S0025 - 6196(20) 30651-0 / full text (in 2020).

[33] Joyner, M. J., et al. Early on, the safety indicators for the convalescent COVID-19 plasma

Large Dataset Clustering Using K-Means with Hadoop Mapreduce

[1]Meenakshi Dayal , [2]Dr. Rajendra Gupta

[1,2] Rabindranath Tagore University, Bhopal

Abstract: Large dataset has become more popular for process data, store data and manage data enormous data. The clustering of datasets has become a challenging issue in the field of large dataset analyzer. The K-means algorithm is perfect to find the relation between objects which is created on distance measures with small datasets. Clustering algorithms involve accessible solutions to manage large datasets. This study presents two approaches to the clustering of large datasets using MapReduce. The first method, K-Means Hadoop MapReduce (KM-HMR), emphases on the MapReduce execution of standard K-means. The second method develop the quality of clusters to produce clusters distances for large datasets. The results of the proposed methods show significant improvements in the efficiency of clustering in terms of execution times. This Research conducted on standard K-means and proposed solutions show that the KM-I2C approach is well- organized.

Keyword- Large Dataset, Clustering, K-Means, Hadoop, Mapreduce, KM-12C.

1. Clustering Large Dataset With Km-Hmr

Dataset clustering is a single technique for dividing the articles that have a place with the data with some equivalent attributes. In distributing technique, the objective of clustering is to isolate 'n' objects into 'k' gatherings, where k<=n and the individuals from a cluster are nearer to one another when conflict from different groups. Dataset clustering has picked up its frame in the field of large data investigation. Clustering issue can be characterized as distributing of given datasets that are requested dependent on specific rules, for example, coterminous, ideal, likeness measure and cluster that are non-covering. Clustering technique focuses on gathering of dataset that is dependent on certain proportion of equivalences. With developing size of datasets in associations, for example, broadcast communications, drug store, bioinformatics and online media, there is a need to separate important experiences from the data put away in the workers.

1.1. Issues identified with clustering large datasets

K-implies clustering technique functions admirably with datasets that are with low size in volume and manages data that has predefined data design. There emerge issues with large dataset, for example, stockpiling, handling, dissecting and picturing with data organizes that either doesn't have determined organization or with certain structured configuration. Storing colossal volumes of data requires various hard

drives. For instance, to store 100 Terabyte of data, around 250 hard drives each with a limit of 400 GB is needed to store the data. Data preparing is identified with the handling of gigantic volumes of data. Utilizing a solitary PC to peruse 100 Terabyte of data with around 35 MB/sec from the circle, it needs roughly 40 days to measure and peruse the data. Additionally investigating and envisioning the acquired outcomes takes a lot of time if either independent or multinode frameworks are utilized. Clustering arrangement in large dataset needs in preparing of large quantities of data with disseminated processing highlight that needs calculation errands to be executed at a quicker rate. Hadoop is versatile and conveyed structure, which takes into consideration circulated handling of gigantic measures of data utilizing MapReduce programming model. The upside of utilizing MapReduce likewise incorporates recognizing, dealing with disappointments and giving high-solid assistance.

Clustering can be applied as the establishment for contravention datasets, for example, exchanging, wrap up the news and finding the shrouded information in the data. Finding definite examples in client databases, recognizing assemblies, medical coverage with a high normal case rate, ID of geographical territories of comparative land use in the geographical database are a portion of the continuous applications of clustering.

The target capacity of clustering can be characterized regarding likeness or separation between the items. In the proposed work, K-Means clustering is utilized which utilizes unaided learning strategy to unravel realized clustering issues. K-Means isolates the whole dataset into 'k' groups, where 'k' is predefined. K-Means clustering calculation suits well for little datasets with determined data designs. As the size of the dataset builds, the time taken for clustering monstrous measures of data takes a lot of time. Unstructured data is the data that either doesn't have a pre-characterized data design or isn't composed in pre-characterized way. Clustering joined with Large dataset innovation can shape an answer in gathering the unstructured data to get important experiences from monstrous volumes of data. In this work, we propose equal and dispersed registering rendition of K-implies clustering.

Dataset clustering is a mind boggling issue and there exist a few clustering calculations. Because of the expanded size of the data in different associations, there is a requirement for effective clustering technique which can measure the datasets in a conveyed way. large dataset manages different data configurations, for example, structured, unstructured and semi-structured arrangement.

Structured data manages the data that are put away in a database containing lines and segments. To cluster the structured data, customary clustering calculations can be utilized. Unstructured data speak to a significant bit of data created in the current

situation. Text, eBooks, records, interactive media content, sound, atmosphere data and site pages are a portion of the instances of unstructured data. Clustering unstructured data needs a proficient clustering calculation which meets adaptability issue and with better execution times. Semi-structured data, for example, data in CSV (Comma Separated Values) structure and JSON (Java Script Object Notification) archives speaks to an exceptionally less bit of the data created, contrasted and unstructured data.

1.2. Data analysis on Clustering Dataset utilizing KM-HMR

The proposed arrangement is a MapReduce adaptation of K-implies clustering technique. MapReduce is one of the most effective clustering answers for Large data issues. KM-HMR works on the rule that every data can be isolated into definite groups. The proposed arrangement functions admirably with large datasets which can bring about better execution times as conflicted and the traditional K-implies clustering technique. The KM-HMR depends on MapReduce programming model and is included mappers and reducers. Each guide cycle is liable for allotting the items or data focuses closest to the cluster. The reducer will peruse each file created by the mapper and travel through the rundown of cluster appointing each object to it.

Figure -1 shows KM-HMR point by point cascade graph. The objective of the proposed work is to do the calculation task of clustering utilizing disseminated and equal preparing of guide and lessen errands with the assistance of Hadoop stage. Introductory cluster centroids file, number of groups (k) and most extreme number of emphases are given as input to the proposed algorithm. MapReduce work contains map errands and decrease undertakings which are executed by Mapper class and Reducer class individually. In KM-HMR, the initial step is to part a large dataset input file into blocks. The default size for the square can be set consequently by the MapReduce programming.

In KM-HMR, a right-angled of size 128 MB is thought of. The guide stage takes the underlying group centroid file, where the cluster places are chosen indiscriminately. The quantity of decrease assignments can be predefined by the program. The more the quantity of reduce assignments are utilized; it can expand time taken to execute the errands put together by an application. In the proposed work, default setting for number of reduce assignments is thought of, which is set by Hadoop stage.

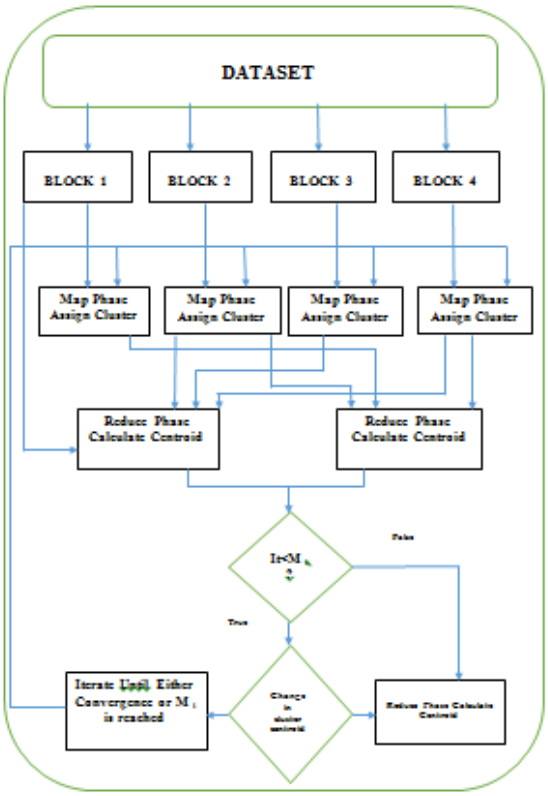


Figure 1: Clustering Dataset Utilizing KM-HMR

The guide stage ascertains the separation between objects utilizing Euclidean separation and the new values of separations between objects are refreshed at each cycle. As a rule, for each Input Split one guide task is made and is executed by mapper class.

Toward the finish of every cycle, the cluster framed beforehand, which is put away in old group file is contrasted and the new cluster centroids shaped, which is put away in new file. In the event that there is an adjustment in the group centroid values, at that point the old cluster file is refreshed with the values of new file. This cycle of refreshing the cluster centroid and gathering it to the separate group proceeds till either the contrast between new cluster centroid values and old cluster centroid values isn't more than the reach from 0 to 1 or it has arrived at the most extreme number of emphases, alluded as combination. The decrease stage summarizes the cluster centroids and composes the new groups in a file.

In our analysis, different parts are utilized for the given dataset with the end goal that the time taken for preparing the whole dataset is less when conflict and the handling of whole dataset. As the split sizes are too little the time taken for in general occupation execution time increments relatively in single hub mode. Of course, the suggested split size of HDFS block is 64

MB.

The split size can likewise be expanded by making fundamental alterations to the setup files. The guide work circulates the errands over a few DataNodes and is equipped for recuperating the assignments and data if there should be an occurrence of breakdown of the framework. Execution of guide assignments brings about composing output files into a neighborhood circle on the separate data hub. Of course, the Hadoop replication factor is set to three. So as to dodge replication which when all is said in done happens at HDFS store activity, nearby circle is picked over HDFS. The halfway output is the Map output which is handled by Reduce errands which produce last output file. The output of each guide task is taken care of to the reduce task. Guide output is moved to the machine where reduce task is running. The output is converged on this machine and later passed to the client characterized diminish work. The diminish output is put away in HDFS. Hadoop multinode climate can handle the positions at a quicker rate than contrasted and the single hub climate. MapReduce is a doable answer for the issues, for example, clustering which includes enormous measures of data to be handled.

2. Km-Hmr Algorithm Used In Clustering Dataset

With expand in size of data in the workers, the issue gathering of data dependent on a specific similitude likewise increments. There is a requirement for proficient clustering algorithm, which can meet the prerequisites, for example, adaptability and better execution times.

Table 1: Notations used in Algorithm 1

Notation	Description
It	Number of iterations
Ic	Initial centroid
D	Dataset
K	Number of clusters
oc	Previous centroid values

nc	New cluster centroid values
Result	Final result
select( )	Function to select data based on k value
input( )	Function to upload the data file
job.mapper( )	Map function
job.reducer( )	Reduce function
write( )	function to write centroid values to a file
read( )	function to read centroid values to a file
update( )	function to check for updated centroid values

Algorithm-1 portrays the proposed work, KM-HMR, which meets the necessity of better execution time during the time spent clustering datasets. The documentations utilized in the algorithm is depicted in Table 1.

Algorithm 1: KM-HMR ALGORITHM USED IN CLUSTERING DATASET

Input:

O : { o1,o2,o3,…..on}; where O represents data objects and

oi represents entities in the data object Ox

K : predefined number of clusters

Mi: Maximum number of iterations

Output:

Final output clusters

KM-HMR(data)

it ← 0

for each datapoint d ϵ D do

ic ← select(k, d)

input(d)

write(ic)

oc ← ic

while (true)

call to job.mapper( )

call to job.reducer( )

nc = read()

repeat until convergence

```
        if update((nc, oc) > 1 )

        oc = nc

        else update nc to result

        it++

    result = read( )
```

MapReduce programming can be utilized to deal with large-scale calculations with adaptability and dependability highlights. With plan and diminish stages, equal execution of the errands is accomplished. Mapper task measures each InputSplit and creates a key-esteem pair. The key-esteem sets produced is not quite the same as the input, which is given to the mapper function. The key-esteem sets got from each guide task are arranged by the key. The diminish task handles and cycles each key and consolidates all the values related with that key.

2.1. Dataset Description for Mapreduce

The US Climate Reference Network (USCRN) is a continuous venture of a Climate Reference Network (CRN), which has been embraced to screen present and future climatic changes. The undertaking comprises of an organization of around 250 stations used to catch normal climatic boundaries. The climatic boundaries incorporate, for example, temperature, wind speed, precipitation, surface temperature and precipitation. All the stations are furnished with great sensors to catch the data precisely. Consistently the subtleties caught are put away in the worker. The requirement for storing these values is that to watch the climatic conditions, examine and anticipate climate gauge. The subset of the whole dataset is considered to test the proposed work. MapReduce programming causes the calculation undertakings to be acted in corresponding with the squares of data put away in the SlaveNodes.

The dataset for MapReduce task is put away as input files. Hadoop Input Format strategy checks the input detail of the occupation put

together by the client. Input Format parts the dataset file into Input Split and appoints the parts to every mapper.

Million Song Dataset (MSD) is an assortment of sound highlights and metadata accessible in Amazon Public Dataset which can be appended to an Amazon EC2 virtual machine to run the investigations. MSD is an unstructured data that contains data of around 1 million tunes including 53 highlights and is around 300 GB in size. A subset of the MSD dataset is considered as an aspect of the investigation work. MSD dataset comprises of the fields, for example, songno, songid, albumid, albumname, artistid, artistname, artistloc, length, type, rhythm, mark, title and year.

Venture Gutenberg (PG) is an assortment of around 3,000 English records composed by more than 140 creators. The dataset considered in the proposed work is identified with reports as eBooks put away in the Project Gutenberg (PG) worker and downloaded from. PG records put forth an attempt to utilize and circulate reports for research. Reports are put away in unstructured organization, where there is no predefined data model and isn't composed into lines and segments as found in structured configuration. The explanation behind picking this dataset is that the data is in unstructured configuration, which meets the qualities of Large dataset, for example, Variety (unstructured), Veracity (confirmed data), Volume (contains a few great many archives, where each record contains a few several MB file size) and Value (Gaining information through clustering, which can extricate significant bits of knowledge from the data put away).

Clustering and breaking down these records is exceptionally intricate to be explained physically. The proposed arrangement, KM-HMR utilizes circulated and equal handling of calculation undertakings utilizing HDFS, plan and reduce errands. There is a requirement for the issues, for example, clustering to utilize equal preparing of calculation assignments to get effective and with better execution times as contrasted and

the traditional clustering algorithms. The distinction between the proposed work and with the current K-implies clustering is that KM-HMR fulfills the significant qualities of Large data and furthermore addresses the issue of adaptability.

The namespace structures the metadata, which is kept up by NameNode and the squares of the data dwell in the DataNodes. NameNode contains the subtleties of data area and afterward moves the data to or from the predetermined DataNodes. Dependability in Hadoop is accomplished through square replication. Each square is repeated by the customer to three DataNodes utilizing HDFS. The NameNode distinguishes and recognizes consequently the undermined and malfunctioned DataNodes and because of copies, it can reestablish the data from the duplicated block. The InputFormat class of MapReduce characterizes how the input files are part and how the data is perused from the InputSplit.

```
map(DocID, DocData)

for each word W in DocData

        emit(W, DocID);

Done
```

Algorithm 3. Pseudo-code for Red

```
reduce(word, values)

for each DocID W in values

     AppendtoOutput(DocID)

Done

emit_Final(FormtDoctIDListForWord);
```

The split is intelligent and input files are not truly part into pieces. Split can be of the structure <input-file-way, start, offset> sets. The input files are spilt utilizing InputFormat class and key,value sets are created from each InputSplits. The whole

Algorithm 2. Pseudo-code for Mapper phase

key,value sets from the InputSplits are shipped off a similar guide function.

The input to the Map function contains the documentid (DocID) and part of the split of the data (DocData). The guide function for each input split radiates the word in the DocData showed by 'W' and its separate DocID. Algorithm 2 and 3 portrays the pseudo-code for the guide stage and Reducer stage individually.

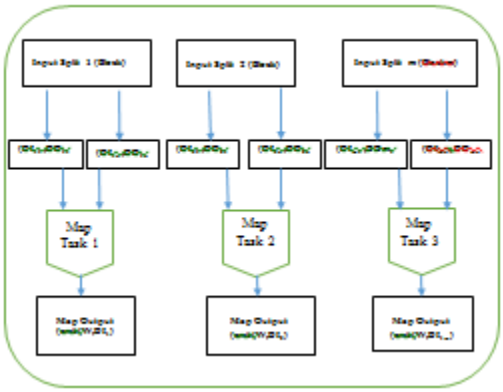


Figure 2: Map Phase in KM-HMR

Clustering the records could limit the area of sorting. The primary inspiration of this work is to improve the effectiveness and execution of the proposed work. Productivity is accomplished through parallelization of clustering issue utilizing MapReduce and execution can be accomplished with better execution times as contrasted and traditional K-implies clustering algorithm.

Figure -2 portrays the guide stage in the proposed arrangement. All the inputs to the guide stage and diminish stage are as key-esteem pair structure. Guide function is composed by the client, which changes the input parts over to (k,v) sets. A few (k,v) sets can be delivered with a similar key by the guide function. The guide assignments read a report Di and parts into words w1,w2,.....wn. Here k contains w1,w2,w3, wn what's more, v contains at first the worth 1. The output of the diminish task contains sets of key-values, where each input key k is matched with the joined worth built from a rundown of values.

2.2. Results and Discussion

Report clustering is the way toward breaking down the records identified with the content.

Reports don't have an unmistakable structure and can be sorted into a semi-structured data design by and large. Archives can be spoken to as an arrangement of words. On the off chance that there exists an archive, at that point there is a more possibility of looking for the report which is identified with each other. Gathering of such archives is identified with clustering of records. Given a lot of reports D, signified by

$$Di = [w1, w2, \ldots . wn] \qquad (2)$$

The proposed algorithm is made to run on the subset of PG dataset. Changed number of archives alongside various sizes was considered as the aspect of the analysis.

Clustering records identified with PG dataset is a difficult issue. Archives are isolated into a few sections containing, the title of the book, writer and are comprised as a semi-structured data design. In clustering the records in PG dataset, tokenization is actualized on the dataset. A large portion of the reports speak to of a grouping of words. The target of tokenization is to separate into words called as tokens. The input dataset is pre-prepared so as to get the ideal highlights which are considered in our clustering cycle.

The pre-handling stage includes eliminating stop words, for example, the, is, at, an, which, these, and to and in. The expulsion of stop words is the most well-known pre-handling technique in archive clustering. This stage likewise incorporates evacuation of words with low word frequencies in a record. The following stage is to eliminate accentuation in the given dataset with the end goal that to the images has less or no significance in the report clustering. Stemming is remembered for record clustering because of the explanation that expressions of comparable importance are to be killed for decreasing the complexities and redundancy of comparatively related words. Stemming words, for example, separation, separations; group, clustering; cook, cooked; walk, strolling; serve, serving; large, large, tremendous; same, comparative.

The square of the data after the parting of the whole archive is known as InputSplit. A

few number of InputSplits (InputSplit1, InputSplit2, ...InputSplitn) are framed. HDFS is intended for adaptability and unwavering quality of data which depends on decoupling namespace from the data.

The execution seasons of standard K-implies, proposed arrangement KM-HMR on PG Dataset were considered in the test work and it is obviously seen that as the size of the archives builds, the execution time for standard K-implies is likewise expanded because of independent PCs utilized in the clustering cycle.

Table 2: Standard K-Means Vs. KM-HMR Execution time

Number of clusters	Execution times	
	Standard K-Means	KM-HMR
5	3.43	2.86
10	3.32	2.72
15	3.24	2.56
20	2.94	2.33
25	2.63	2.21
30	2.28	1.88

With the impact of equal registering, KM-HMR functioned admirably with the expanded number of archives and the execution time is better when contrasted and standard K-implies clustering algorithm.

Table- 2 shows Standard K-Means Vs KM-HMR Execution time. As the quantity of cluster expands, the proposed arrangement accomplishes better execution time when contrasted with standard K-Means due with equal calculation of undertakings.

Table- 3 depicts the consequences of the Standard K-Means Execution time in Multinode climate.

Table 3: Standard K-Means Execution time in Multinode Environment

Run number	Standard K-Means execution time in Multinode environment (in Minutes)		
	1	**2**	**3**
1	10.03	11.06	9.28
2	10.05	9.09	9.53
3	10.01	11.14	11.46
4	9.02	11.25	11.41
5	9.04	9.08	12.49
6	9.07	11.34	9.57
7	10.01	10.38	12.44
8	9.03	11.01	11.49
9	9.07	11.05	12.34
10	9.08	10.48	12.16
Average	9.44	10.59	11.22

Jobs are presented by the MasterNode where each job speaks with the Hadoop datanodes through the JobClient. Mapper class, WordCount job, Reduce class and different classes are added while executing the job to make the job run two guide assignments in equal. The dataset utilized in this analysis is PG dataset, which contains the reports.

Table- 4 depicts the consequences of the Standard K-Means Execution time in Multinode Environment. The outcomes are acquired and are contrasted and the proposed arrangement, KM-HMR and standard K-implies clustering algorithms in Multinode climate. At each run, execution times were noted and introduced in the table.

Figure -3 shows the execution seasons of K-Means Vs. KM-HMR in Multinode climate.

Table 4: KM-HMR Execution time in Multinode Environment

Run number	KM-HMR execution time in Slave Node (Minutes) in Multi node environment		
	1	**2**	**3**
1	4.02	4.46	2.8
2	5.29	5.02	4.55
3	4.52	4.04	4.46
4	4.06	4.05	5.46
5	4.03	4.08	2.59

6	5.39	5.34	4.07
7	4.01	5.36	4.44
8	4.23	3.31	5.49
9		5.07	4.04	5.44
10		5.02	4.54	5.41
Average		4.56	4.42	4.47

An aggregate of twelve cluster were framed for the subset of PG Dataset. The clustering results are put away in output index.

Figure- 4 depicts the adaptability of the proposed work, where numbers of archives were expanded and execution times were looked at.

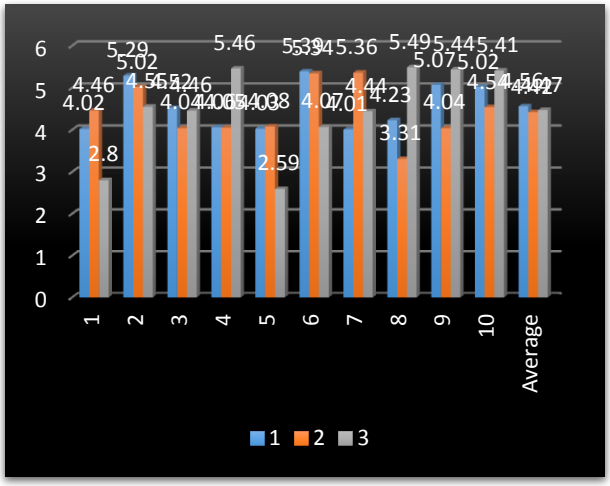


Figure 3: K-Means vs. KM-HMR

comparison in multinode environment

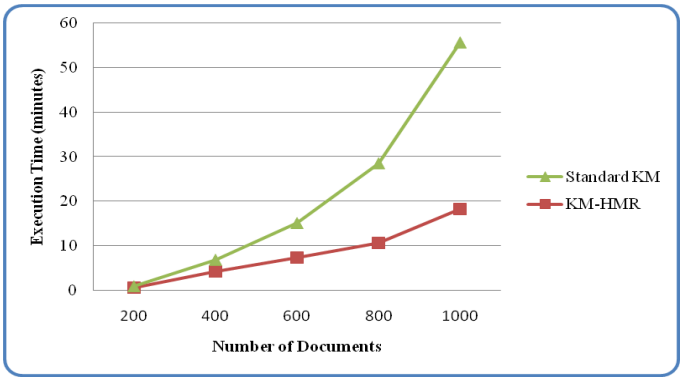


Figure -4: Scalability of the proposed work: number of documents vs execution time

With the expanded number of records presented by the client, it is seen that because of single hub calculation, time taken by the standard K-implies clustering algorithm sets aside more effort to frame groups than KM-HMR.

Table 5: KM-HMR results with five clusters

	MSD	USCRN	PG dataset
Single node	7200	5500	4500
VM=2	7000	5500	4500
VM=4	6800	5300	4000
VM=5	6500	5000	3900
VM=8	5500	4800	3500

VM=10	5000	4800	3300

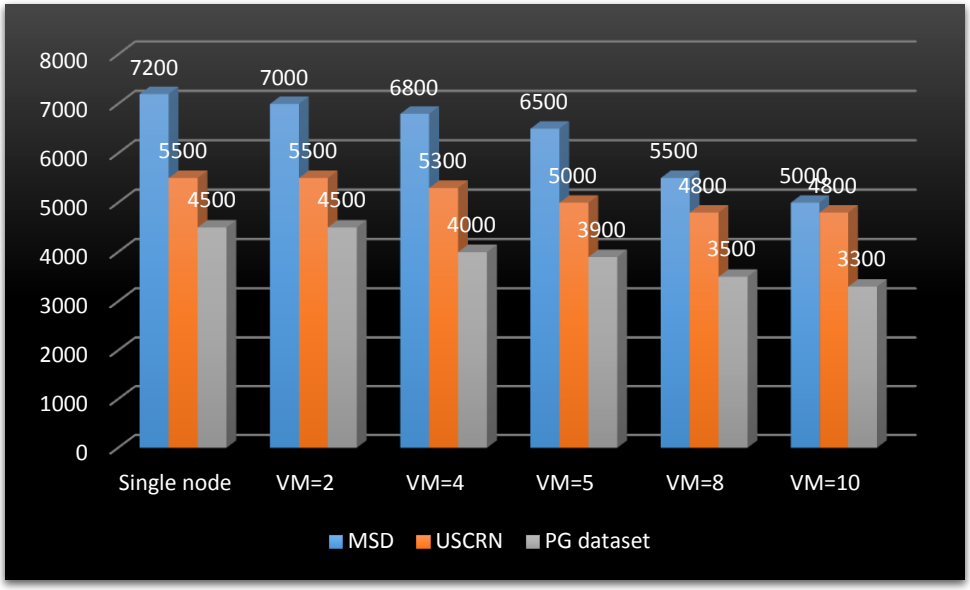


Figure 5: KM-HMR results with five clusters

Table 6: KM-HMR results with ten clusters

	MSD	USCRN	PG dataset
Single node	6000	4300	2900
VM=2	5900	3200	2800
VM=4	5500	3000	2800
VM=5	5500	2900	2500
VM=8	4000	2500	2500
VM=10	3500	2500	2100

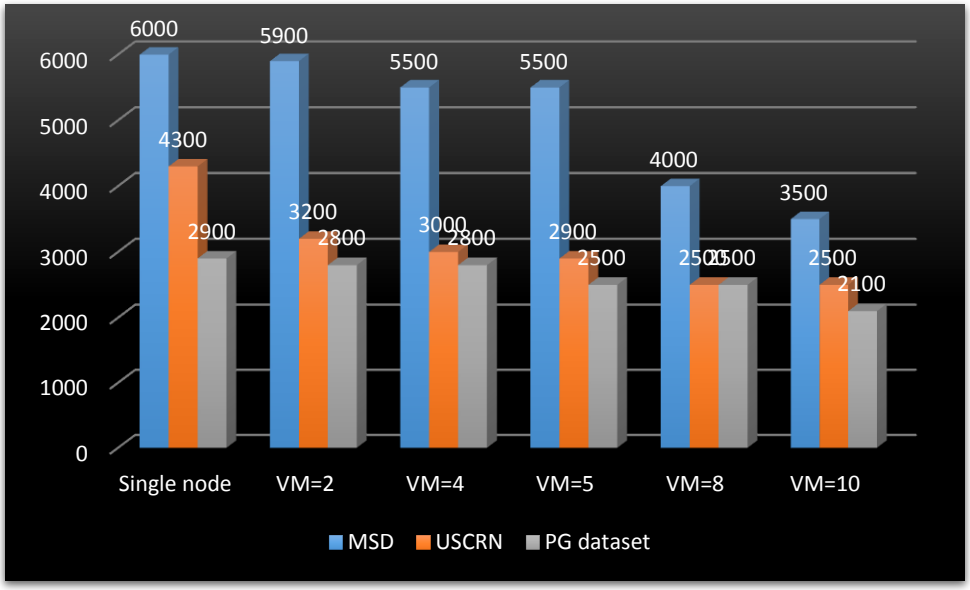


Figure 6: KM-HMR results with ten clusters

Table 7: KM-HMR results with fifteen clusters

	MSD	USCRN	PG dataset
Single node	5500	4000	2700
VM=2	5500	3800	2700
VM=4	4800	3500	2300
VM=5	3500	3400	2000
VM=8	3400	2800	2200
VM=10	3000	2500	1900

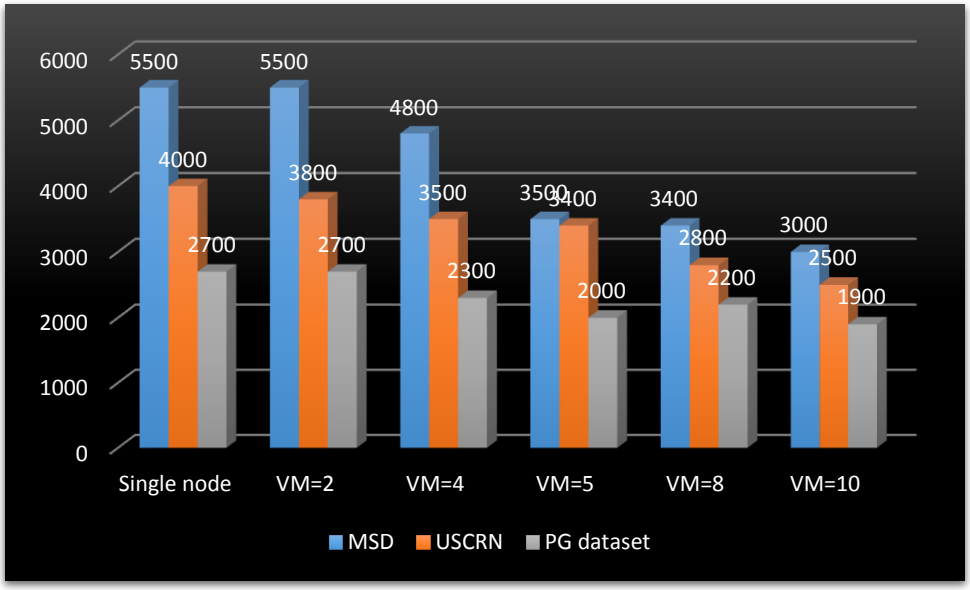


Figure 7: KM-HMR results with fifteen clusters

Figure- 5, 6 and 7 shows the KM-HMR results for the different datasets (Million Song Dataset, US Climate Reference Network Dataset and Project Gutenberg Dataset) with shifted number of machines. The outcomes likewise show that execution time is noted for fluctuated number of machines. X-pivot speaks to number of machines and the y-hub shows time taken to group the datasets like a flash.

As the size of datasets expands, time taken by the traditional K-Means clustering expanded continuously and with the proposed arrangement, KM-HMR better execution times were accomplished because of parallelization of the undertakings. In clustering, adaptability is the significant issue to be unraveled. The proposed KM-HMR functions admirably with expanded size of the datasets because of equal calculation undertakings of guide and diminish assignments in a few DataNodes.

Table 8: KM-HMR vs. K-Means Single and MultiNode Environment

KM-HMR Vs. Traditional K-Means on USCRN Dataset								
Run Number	Single Node		Multi Node					
			2		3		4	
	K-Means	KM-HMR	K-Means	KM-HMR	K-Means	KM-HMR	K-Means	KM-HMR
1	18.23	12.11	17.32	12.08	17.02	12.93	17.22	12.01
2	17.67	12.43	17.07	12.13	17.77	14.68	16.47	13.29
3	16.23	14.21	17.46	14.12	17.26	14.12	16.86	13.98
4	16.78	13.87	17.18	14.57	17.08	14.57	15.18	13.87
5	18.23	16.01	18.12	15.12	18.82	15.62	16.01	15.02
6	20.54	14.48	17.12	14.14	17.01	14.02	18.10	13.77
7	18.32	16.11	16.22	15.49	17.06	14.09	17.22	13.94
8	20.28	14.01	17.87	11.28	18.19	13.28	17.87	13.02
Average	18.29	14.15	17.30	13.62	17.53	14.16	16.87	13.61

The outcomes show that the KM-HMR when applied with various boundaries takes lesser time when contrasted with the traditional K-implies clustering algorithm. KM-HMR utilizes equal calculation of errands in a few DataNodes and thus the assignment of clustering is part into a few hubs and the outcomes are gotten at a quicker rate as contrasted and the traditional K-Means clustering algorithm.

Table- 8 portrays the execution seasons of the proposed arrangement, KM-HMR Vs K-Means clustering algorithms in single and MultiNode conditions. In USCRN dataset, term and rhythm were considered as the piece of testing the proposed arrangement. In spite of the fact that the MultiNode idea is utilized in traditional K-implies clustering algorithm, the execution times for the proposed arrangement, KM-HMR is better because of the explanation that the calculation errands were actualized utilizing equal and disseminated registering hubs utilizing Hadoop stage.

Table 9: Clustering for USCRN dataset

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	20.41	95.18	121.02	14.57	102.04	70.64	63.21	114.58	161.89	92.37	55.28
B	20.41	0	74.87	101.80	8.44	81.63	52.22	67.71	94.39	141.93	73.14	34.89
C	95.18	74.87	0	35.44	82.26	9.93	45.12	113.17	19.99	68.08	36.38	40.13
D	121.02	101.80	35.44	0	109.97	38.19	80.47	122.07	22.14	44.02	70.19	69.75
E	14.57	8.45	82.26	109.97	0	88.61	56.07	70.46	101.95	149.69	77.83	42.13
F	102.04	81.63	9.93	38.19	88.61	0	45.67	122.81	17.98	64.26	32.08	46.76
G	70.64	52.22	45.12	80.47	56.07	45.67	0	113.28	63.31	109.59	22.60	29.52
H	63.21	67.71	113.17	122.07	70.46	122.81	113.28	0	127.10	165.98	129.37	85.59
I	114.58	94.39	19.99	22.14	101.95	17.98	63.31	127.10	0	48.09	49.70	59.89
J	161.89	141.93	68.08	44.02	149.69	64.26	109.59	165.98	48.09	0	92.06	107.80
K	92.37	73.14	36.38	70.19	77.83	32.08	22.60	129.37	49.70	92.06	0	43.88
L	55.28	34.89	40.13	69.75	42.13	46.76	29.52	85.59	59.89	107.80	43.88	0

Table- 9 depicts clustering values that are gotten for USCRN dataset, where term and rhythm ascribe in dataset were utilized to get last clustering outcomes. Just the pieces of the values are considered to show execution of KM-HMR for USCRN dataset.

3. Cluster Enhancement Using Km-I2c

Clustering alludes to sorting out the data into clusters with the end goal that there is a high intra-cluster comparability and low between cluster similitudes. The center of clustering is to clusters with a characteristic gathering of articles with recognizable gatherings among the clusters framed. K-implies clustering is a classical clustering algorithm that utilizes the square of the Euclidean separation to parcel various data focuses into 'k' clusters. Productive clustering technique expands the intra-cluster likenesses and limits the between cluster similitudes. Hadoop actualizes equal calculation worldview where the application is partitioned into a few sections of work,

every one of which is executed and additionally re-executed on any hub in the cluster of hubs.

Closeness proportion of a clustering algorithm ought to likewise think about the separation among clusters and separation among centroid and different items inside a similar cluster. Productive clustering relies upon likeness measure utilized, which speaks to how close the articles are inside a similar cluster. The entomb and intra clustering condition utilized in this work speaks to an estimation of a separation between two vectors to be specific vector file and centroids file.

3.1. Clustering Process in KM-I2C

Cluster algorithm must keep up the nature of clusters shaped. The two significant variables that decide the compelling clustering are adaptability, productivity and execution time. The goal of the clustering is to create clusters that are reduced and very much isolated from each other. Successful clustering algorithm must contain clusters with least intra-cluster separation and most extreme between cluster separations. A cluster expanding an intra-cluster separation measure is picked for parting the given dataset until assembly is met or when the algorithm arrives at the greatest number of cycles. Dataset is spoken to as the vector which is given as input to the MapReduce. In the underlying stage, the guide step peruses the cluster communities into the memory from the input file.

KM-I2C acquires quality clusters by figuring bury and intra clustering separations among the clusters framed. In KM-HMR, Euclidean separation metric is utilized to cluster the items, which are more like each other. The last clusters got needs a proficient metric to locate the best prospects of collection comparable articles near one another, when new items are added to the dataset. Subsequent to finding the centroids utilizing Euclidean separation metric, another item is put to a specific cluster dependent on the Inter cluster separation. Entomb cluster separation demonstrates the

separation between any two clusters concerning cluster centroids.

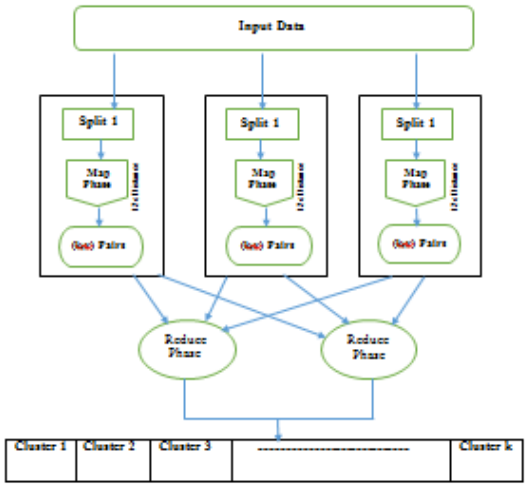


Figure 8: Clustering Process in KM-I2C

In every emphasis, separation between objects in a similar cluster is recalculated utilizing intra cluster separation. Intra cluster separation gauges the good ways from centroid to all different items inside a similar cluster. Cluster proficiency can be improved by augmenting the intra cluster likenesses and limiting the bury cluster similitudes. KM-I2C utilizes Hadoop stage to discover proficient clusters. HDFS is intended to store tremendous volumes of files in a cluster of product equipment. In MapReduce execution of KM-I2C, the job is separated among a few data hubs dependent on Divide and Conquer worldview. As the data is handled by different data hubs, the time taken to deal with the data is diminished as contrasted and independent frameworks for clustering datasets. In our methodology, rather than moving the data to the handling unit, the algorithm is moved to the data in the MapReduce system. The significant bit of leeway of executing clustering technique in MapReduce is that of versatility of data handling over a cluster of hubs.

Figure- 1 shows the outline of the proposed cluster upgrade utilizing entomb and intra clustering separation for the K-implies clustering algorithm. The input dataset is part into a few InputSplits relying on the size of the data given as input to the proposed arrangement. Traditional K-implies clustering algorithm utilizes Euclidean separation. In KM-HMR a clustering algorithm, plan and decrease stage

are utilized to take the underlying clusteroid focuses and ascertains the separation between each item and each cluster utilizing Euclidean separation over a circulated climate in Hadoop.

The second methodology KM-I2C improves the clustering component with better execution times as contrasted and the traditional K-means and KM-HMR clustering techniques.

Conclusion

Clustering is a very difficult problem that is heavily shaped by large dataset used and problems considered the problem. It show the improvement in execution time through proposed algorithm. Hadoop can perform operation on map and reduce jobs in simultaneously to cluster large datasets well organized. The main objective of this work was to increase speed and balance large datasets to obtain resourceful good-quality clusters. The standard K-means method is the most popular clustering method due to its simplicity and reasonable execution efficiency when applied to small datasets. We have introduced a KH-HMR algorithm to make use of simultaneously tools through Hadoop and to obtain better execution times than those of the standard K-means approach. We have developed a unique KM-I2C algorithm by making amendments to the clustering distance metric. We, convert, have developed well organised method relative to other clustering techniques. Future work should enhance the performance of map and reduce jobs to suit large datasets. The performance of Hadoop can be enhanced by using multilevel queues for the efficient scheduling of jobs suitable for large datasets.

Reference

[1] Tang, Tinglong & Chen, Shengyong & Zhao, Meng & Huang, Wei & Luo, Jake. (2018). Very large-scale data classification based on K-means clustering and multi-kernel SVM.

Soft Computing. 10.1007/s00500-018-3041-0.

[2] Abualkishik, Abedallah. (2019). Hadoop And Big Data Challenges. Journal of Theoretical and Applied Information Technology. 97. 3488.

[3] Al-Sharo, Yasser & Shakah, Ghazi & Mutasem, Sh & Alju-Naeidi, Bajes & Alazzam, Malik. (2018). Classification of Big Data: Machine Learning Problems and Challenges in Network Intrusion Prediction. International Journal of Engineering and Technology(UAE). 7. 10.14419/ijet.v7i4.36.25381.

[4] Chen M. Soft clustering for very large data sets. Comput Sci Netw Secur J. 2017;17(11):102–8.

[5] Tsai CW, Hsieh CH, Chiang MC. Parallel black hole clustering based on MapReduce. In: Proceedings of IEEE international conference on systems, man and cybernetics. 2015.

[6] Fadnavis, R. A. and Tabhane, S. (2015) "Big Data Processing Using Hadoop", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (1), 443-445.

[7] Balasubramanian, K. S. and Esmailpour, A. (2014) "Hadoop Framework to Provide Fault Tolerance in the Cluster", ASEE 2014 Zone I Conference, University of Bridgeport, Bridgeport, CT, USA.

[8] Ferreira Cordeiro RL, Traina Junior C, Machado Traina AJ, López J, Kang U, Faloutsos C. Clustering very large multidimensional datasets with MapReduce. In: Proceedings of KDD'11, ACM, California, August 21–24. 2011.

[9] Chao L, Yan Y, Tonny R. A parallel Cop-K means clustering algorithm based on MapReduce framework. Adv Intell Soft Comput J. 2011;123:93–102.

[10] Wang C, Guo M, Liu Y. EST clustering in large dataset with MapReduce. In: Proceedings of pervasive computing, signal processing and applications, Sept 2010.

[11] Zhao W, Ma H, He Q. Parallel K-means clustering based on MapReduce. In: CloudCom 2009, LNCS 5931. Berlin: Springer; 2009. pp. 674–9.

Multilevel Streaming Clustering Algorithm for High Dimensional Data Sets

Ankit Kumar Dubey[1] , Rajendra Gupta[2], Satanand Mishra[3]
[1] Department of Computer Science, St. Aloysius College (Autonomous), Jabalpur, MP, India.
[2] Department of Computer Science, Rabindranath Tagore University, Raisen, MP, India.
[3] CSIR-Advanced Materials and Processes Research Institute (AMPRI), Bhopal, MP, India.

Abstract: The Data stream clustering is an active area of research that has recently emerged with the goal of discovering new knowledge from a large amount and variability of constantly generated data. In this context, different-different algorithm for unsupervised learning that clusters multiple data streams has been proposed by many researchers. There is a need for a more efficient and efficient data analysis method. This paper introduces a multi-level K-Means Density-based flow Clustering Algorithm (MKDCSTREAM) for clustering problems. This approach proposes to view the problem of clustering is a optimization process hierarchy that follow different levels, from unrefined to subtle. In the clustering problem, for the solution first divide the problem in parts, by following different levels to make the first clustering a coarser problem than calculated. Coarse problem clustering is mapped level by level and improves the Clustering the original problem by improving intermediate clustering using the general K-means algorithm. Compare the performance of the hierarchical approach with its single-tier approach using tests with a set of datasets collected from different areas.

Keywords: multi-level K-Means, MKDCSTREAM, Clustering, unsupervised learning.

1. Introduction

The amount of data stored on your computer is growing at an alarming rate. However, it turned out to be very difficult to obtain useful information. In many cases, regular data research techniques and analysis tools are simply not suitable for meeting these growing demands for information [1]. Data clustering is one of the main tasks of data analysis, Different pattern identification, data densification, analysis of image, and machine learning. [2]. Our method first estimates multiple sub-clusters from the stream the data and then apply task-specific clustering algorithms. to these sub-clusters. Each subcluster is represented by a set of centroids that are one by one assessed using

various parameters. Depending on the arrival of new data object in the streaming input, centroid updates continuously. In the next step clustering all fit points is for the centroid to coincide with the cluster. This is based on the attribute of the data-set Opposite to kmeans clustering, the method which we are proposing a cluster can have multiple center points, one typical point per cluster[3]. Major contribution of this research is:- A Multi-level stream clustering algorithm capable of processing large multidimensional datasets.

2. Literature review

Researchers have made several attempts to improve efficiency and effectiveness of data stream clustering. In this[4] the author proposes a density-based clustering algorithm for IoT streams. This method is fast and very useful for real-time IoT applications. In the Experimental results author finds that this method can provide high quality results for real data sets and synthetic data sets in short computation time. In this[5] paper, author's have developed an efficient and effective client method called CluStream for large-scale clustering. Evolving data streams. This method has obvious advantages. Discourse about the latest technologies to combine Flow as a process that changes over time rather than looking at the entire stream at once. The stream CluS model provides a wide range of capabilities for describing clusters of data streams over different time periods in an evolving environment. In [6] author's proposed a hybrid K-Means algorithm that joint the steps of reduction of dimensionality with PCA, a new approach for initializing cluster centers, and the steps for assigning data points to the appropriate clusters. Use The suggested algorithm divides the specified dataset into k clusters so that reducing the total clustering errors for all available clusters as much as possible while keeping the distance between the clusters as large as possible. Authors of [7], proposed UMicro's algorithm for clustering undefined data streams. Undefined data streams can be present in many real-world applications due to inaccurate write mechanisms. Data flow uncertainties have a significant impact on the clustering process, as different attributes behave differently relative to each other and affect distance calculations. This paper[8] gives an overview of the data A stream clustering algorithm applied on stream of big data and large datasets. The documentation shows a comparative deep analysis of all the methods reviewed and an overview of progression and progress Dataflow clustering algorithms for large datasets. The article is also reviewed a proposed and recently implemented algorithm. This article[9] has summarized a simple set of conditions for a data stream clustering algorithm. An important requirement that algorithm designers often ignore is the need for clear isolation of outliers in the data stream. This is because a sufficient number of outliers may indicate that the clustering model needs to be modified. Authors provide analytical tools to effectively track these changes. This[10] study presents a new discovery method based on a scalable framework for identifying relevant ones in a community on the web. We propose a multilevel clustering

(MCT) method that uses a textual information structure to identify a local community, called a microcosm. Experimental evaluations of reference models and datasets show the effectiveness of the approach. This research contributes to a new dimension for identifying cohesive communities on social media. This approach provides better understanding and clarity to explain how low-level communities develop and operate on Twitter.

3. *Multilevel Kmeans_clustering Algorithm*

This clustering technique creates diverse views for each and every cluster. The algorithm is divided mainly in two stages. We call this the Subset Generator. The first step in sub-clustering phase clusters x injects data into y sub-clusters (k <x <y) similar to the algorithm_ kmeans. In the next step , few of these centers of gravity are clubbed together into one large group. This is called the recovery stage. Using different parameters, each cluster is constructed from the same set of input. The multi-level K-Means algorithm is a combination of the popular greedy K-means algorithm and the multi-level paradigm of clustering problems. The layered paradigm is a simple problem-solving technique with a recursive coarse structure that easily solves both simple and minor problems. The layered paradigm consists of four stages: broad outline, initial decision, forecasting, and correction. The broad outline phase aims to combine the volatile data related to the problem to form a cluster [1]. Clusters are used recursively, each cluster representing the original problem, it builds a hierarchy of problems with less freedom. This phase is continuing till the minimum problem size reached the specified reduction limit. Then the solution to the problem is generated at unrefined level is projected in reverse order to each intermediate level. "The solution at each child level is improved before moving to the parent level. A common feature that characterizes multilevel algorithms is that the solution to one of the critical problems is a legitimate solution to the original graph".

Algorithm 1: The Multilevel Algorithm

$Input: Problem\ x_0$
$Output: Solution\ y_{final}(x_0)$

1. $Begin$
2. $lvl := 0;$
3. **/* Proceed with Coarsening */**
4. ***while*** *not stop* ***do***
5. $d_{lvl+1} := lvl + 1;$
6. $lvl := lvl + 1;$
7. $QC_{start}(d_{lvl}) = initial\_clustering\ (d_{level});$
8. **/*Proceed with Uncoarsening and Refinement */**
9. $\textbf{while}(lvl > 0)\textbf{do}$
10. $Qy_{start}\ (d_{lvl-1}) := Extend\ \left(Qy_{final}\ (d_{lvl})\right);$
11. $Qy_{final}\ (d_{lvl-1}) := K - means\ (Qy_{start}\ (d_{lvl-1}));$
12. $lvl := lvl - 1;$

end

The first phase of multilevel clustering is the pruning phase, where d_0 is the set of data objects to cluster. From d_0 with two different algorithm, d_1 the next level of coarser is constructed.Random coarsening scheme is the first level algorithm. In the random order data objects are accessed. If the 0_i data object is not yet mapped, a mismatched 0_j data object is randomly selected consisting of the two data object 0_i and the 0_j, a new data object 0_k randomly selected. A new set of data object attributes $0_k$is calculated by averaging each attribute from $A0_i$and the corresponding attribute from $A0_j$. Just copy the data object to the next level without merging. The second algorithm is coarsening algorithm, based on the concept of distance this algorithm uses a measure of the strength of the connection between data objects. However, instead of adding the 0_i object with a random 0_j object, the 0_i data object is combined with 0_m so that (2) is minimized. Use the newly created data object to identify a new small problem and repeat the clipping process recursively until the size of the problem reaches the desired threshold. (1, 2, 3, 4, 5,6 line).

If we see the line 7, initialization is easy and consists of using a random procedure to generate an initial clustering of problems (d_m). All individual clusters in the population are assigned random labels from a set of cluster labels. Due to the improved quality of clustering at the d_{m+1} level, it is necessary to expand the parent level of d_m.In the line 10, If the $0_i \in d_{m+1}$ assigned the cluster y_1 , The combined pair of data objects that it represents, $0_1, 0_m \in d_m$ is also assigned the cluster label y_1 . Line 11, The clustering found at level $m + 1$ is minimal, with respect to m, predicted clustering may not be optimal. Predictive clustering is already fine and the K-mean converges faster and clusters better over several iterations. Premature convergence occurs if the cluster does not change over the number of iterations. At each level, the k-means are expected to converge if all clusters do not change within 5 consecutive iterations.

4. **Streaming sub-clustering and Proposed MKDCSTREAM Algorithm**

Our sub-clustering module is working on this streaming vector mapping method. The input data is randomly ordered. When a fresh data point is displayed, the center which is closest to the fresh data shifts gradually in that direction. It holds this center point updated and moved.

Figure 1. This process continues to update and move this center point as new data is displayed.

Algorithm 2: Streaming sub clustering

Input : *streaming input(s) in dimension(m), a current set of centroids(x)*
Output: *X is the new centroids set*

1. *Take a new input s*
   Calculation of the each center point Distance x ∈
2. *X and the new s*
3. *find a center point nearest cluster, x_m*
4. *shift x_m nearest to s*
5. *Return the updated X*

Algorithm 2 shows the stream sub clustering algorithm. The input s is the m-dimensional point of the streaming data and X is the current set of centroids for the y-cluster. The output :- fresh set of X centroids. First, the processing engine receives the s and from this latest s each centroid of the y cluster. This point will be attached to the nearest one.

The steps of the MKDCSTREAM algorithm are as follows.

Algorithm 3: MKDCSTREAM Algorithm

Input: *−h dimension data stream*
Output:- *the latest set of centroids* with Cluster Id
** ***Step 1*** **
*Use the*Multilevel K − Means algorithm for Coarsening and Refinement.
** ***Step 2*** **
For merging the micro cluster use merge and sort algorithm.
** ***Step 3*** **
For the stream cluster use the stream subclustering algorithm
End.

The sub clustering module using different parameters, each cluster is constructed from the same set of input sets. Each process is completely independent, which makes it highly extensible. Our method of streaming minimizes complexity of computation and resources and parallelizes these independent operations.

5. Experimental Activities and Result Discussion

All performance tests were performed on a 2.40GHz Intel® Core ™ i3-3110M CPU. Windows 8 (64-bit) operating system with 3.25GB of RAM. The implementation was done in Pycharm 2020 3.5 for MKDCSTREAM encoding.

We test our approach using several different sets of data parameters. Table 1 shows the data set which we have used for experimental purpose. The dataset which we have used here is datasets of 2D and 3D dimensions –moons and 3D clouds, from UCI Machine Learning Repository (spambase and census 1990)[12]. We compared algorithm MKDCSTREA with three frequently used clustering algorithms for processing data streams, namely with algorithm MiniBatchKmean and with Birch algorithm.

Table 1. Multidimensional DataSets for Test

Data_set	Dimension (d)	Size of the data	Clusters (k)
moons	2	1500	2
3D clouds	2	16384	128
Spambase	57	4601	10
census 1990	68	2458285	10

$$\operatorname{argmin}(k) = \sum_{x=1}^{c} \sum_{k \in K_1} ||k - a_i||Y \quad (3)$$

In the above function, k are n input data, $\{K_1, K_2, \ldots K_c\}$ present c which denotes clusters, and every one of them is presented using one centre point. a_i, is a minimum cost method, cost value. Our cost result comparison is with MiniBatchKmean and Birch in Table 2.

6.Units compare clustering costs

Average distance between each centroid in the cluster and data points. The cost is estimated by the value of the objective function, similar to the k-means algorithm.

Table 2. Cost Result Comparison

Data Set	MiniBatchKmean	Birch	MKDCSTREAM
3D clouds	158.95	157.18	152.12
Spambase	96.96	111.92	102.19
census 1990	35.36	35.47	35.31

CONCLUSIONS

In case of a multilevel paradigm the strength to improve the merging behavior of the K-Means algorithm is expected to cover all cases. Although the reason for this merging behavior, which is detected in the multilevel paradigm, is not clear, it can be concluded. As mentioned before, in a layered paradigm, one solution to a gross problem must provoke a legitimate solution to the original problem. In order to obtain the final solution to the problem, after initialization at any step, the latest solution of the problem can be used to all levels of the task. In our case, we are violating this requirement. Derived level element of each object, are calculated at base level by calculating the average of the elements at two different objects. We designed a stream clustering algorithm for a multilevel streaming high dimensional data sets. Our clustering algorithm can handle an unlimited amount of large streaming data. It presents clustering results comparable to available clustering algorithms. The planned method evaluates subclusters based on large amounts of data and applies task-specific clustering algorithms to these sub-clusters.

REFERENCES

[1] Bouhmala, N., Viken, A., & Lonnum, J. B. (2016). A multilevel K-Means

algorithm for the clustering problem. 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). doi:10.1109/icccbda.2016.7529544

[2] Ng, H.P., Ong, S.H., Foong, K.W.C., Goh, P.S., Nowinski, W.L.: Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm. 2006 IEEE Southwest Symposium on Image Analysis and Interpretation.

[3] Lee, D., Althoff, A., Richmond, D., Kastner, R.: A streaming clustering approach using a heterogeneous system for big data analysis. 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). (2017).

[4] Amini, A., Saboohi, H., Ying Wah, T., Herawan, T.: A Fast Density-Based Clustering Algorithm for Real-Time Internet of Things Stream. The Scientific World Journal. 2014, 1–11 (2014).

[5] Aggarwal, C.C., Yu, P.S., Han, J., Wang, J.: A Framework for Clustering Evolving Data Streams. Proceedings 2003 VLDB Conference. 81–92 (2003).

[6] Dash, B., Mishra, D., Rath, A., Acharya, M.: A hybridized K-means clustering approach for high dimensional dataset. International Journal of Engineering, Science and Technology. 2, (2010).

[7] Aggarwal, C.C., Yu, P.S.: A Framework for Clustering Uncertain Data Streams. 2008 IEEE 24th International Conference on Data Engineering. (2008).

[8] Dubey, A.K., Gupta, R., Mishra, S.: Data Stream Clustering for Big Data Sets: A comparative Analysis. IOP Conference Series: Materials Science and Engineering. 1099, 012030 (2021).

[9] Barbará, D.: Requirements for clustering data streams. ACM SIGKDD Explorations Newsletter. 3, 23–27 (2002).

[10] Inuwa-Dutse, I., Liptrott, M., Korkontzelos, I.: A multilevel clustering technique for community detection. Neurocomputing. 441, 64–78 (2021).

[11] Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering. 63, 503–527 (2007).

[12] Khalilian, M., Mustapha, N., Sulaiman, N.: Data stream clustering by divide and conquer approach based on vector model. Journal of Big Data. 3, (2016).

[13] Lichman, M.: UCI machine learning repository, http://archive.ics.uci.ed/ml.

IOT and ML architecture for predictive maintenance in industry 4.0

[1]Madhurima Sharma, [2]Mohnish Sharma, [3] Dr. Shailja Shukla
[1] Research Scholar ,[2] Research Scholar, [3] Associate Professor, RNTU, Bhopal
[1] madhurimasharma7@gmail.com,[2] mohnish.sharma272@gmail.com, [3] shailja.sharma@aisectuniversity.ac.in

Abstract
To satisfy the demands of a more difficult and quickly changing industry, future manufacturing processes will need to be more adaptable. They must allow greater use of info, ideally all of it. Low-level details can be refined to actual knowledge for decision-making to encourage competition viatimely decisions and informed. The automotive sector has a significant impact on economic and social growth. Since it is a common idea for colleges and research centers, the Industry 4.0 program has attracted a lot of support from the market and academic communities. Industry 4.0, as well as its synonyms such as Smart Production, Smart Manufacturing, and the (Internet of Things) IOT, have been recognized as significant suppliers to the digital manufacturing and automated climate. Industry 4.0 (I4.0), smart networks, (ML) machine learning, a branch of (AI) artificial intelligence, and PdM (predictive maintenance)methods are now frequently employed in factories to control the strength of manufacturingtools. Digital convergence forI4.0, computerized management,communication networks and information techniques, it is quite easy to gather vast quantities of process and functionalsituationsinformationproduced by various parts of tools and produce data for diagnostic and automatic fault detection with the intention of reducing downtime and increasing component utilization rate. This paper goals to offer a completeevaluation of currentdevelopments in machine learning methodsbroadlyuseful to PdM for smart manufacturing in I4.0 throughcategorizing the studybasedon the ML algorithms. In this paperfuture prediction of temperature is done using the time series analysis and multivariate analysis.

Keywords: Industry 4.0, IOT, Lean manufacturing, Predictive maintenance, CPA

1. Introduction

Production systemsof the future industries needto make better use of the information as well as the raw data[1][2]. To help decision-making, low-level data must be converted intosmart servicesand usable data, must be incorporated. In recent years, the task of managing data, transforming it into information, and making wise computerizedevaluations has gotten a bunch of interest. In partnerships including Smart Manufacturing Leadership Coalition 2016, Industry 4.0 (Industry 4.0 Working Group 2013),Internet of Things (Atzori, Iera, and Morabito 2010)[3], the Industrial Internet (Evans and Annunziata 2012), and automationand cloud robotics, the emphasis has been on the overall architecture (Kehoe et al. 2015)[4].While the concept is not innovative and has been on the program of theoreticalstudy for several centuries with various interpretations, the word "Industry 4.0" has only

recently been coined and is widely embraced not only in academia but also in industry. Professionals have coined the phrase "Industry 4.0" to describe how industries are now undergoing "The 4thIndustrial Revolution." (I4.0). Industry 4.0 is a technique for transforming manufacturing from a machine-dominated to a digital-dominated state. Industry 4.0 should be well known to achieve an effective transition, and a simple path map should be created and enforced.

With the introduction of I4.0, the idea of (PHM) prognostics and health management has developed an inevitable trend in the context of smart manufacturing and industrial big data; it also offers a dependableexplanation for handling the health status of industrial tool. I4.0 and its main innovations are important for making industrial systems autonomous [5,6], allowing for automated data gathering from components and industrial machines. ML algorithms can be used to automate diagnosis and fault detection founded on the gathered data. However, selecting suitable (ML) machine learning methods, data types, data sizes, and tool to implement Machine Learning in manufacturingmethods is extremely hard. Infeasible maintenance scheduling and Time loss may result from choosing the wrong (PdM) predictive maintenance methodology, data size and dataset. As a result, the aim of this research is to introduce a systematic review ofliterature to discover current research and Machine Learning applications, thus assisting practitioners and researchers in selecting suitable Machine Learningmethods, data type, and data size to obtain a feasible ML application. Since it was created to accomplish near-zero; hidden risks, faults, emissions, and near-zero accidents in the completeatmosphere of industrialmethods, industrial equipment(PdM) predictive maintenance can detect deterioration results.[7].

These massive volumes of data collected for ML provide a wealth of useful information and expertise that can help increase the overall competitiveness of manufacturing processes and system dynamics, as well as decision support in a variety of areas, most notably maintenance based on conditions, and health checking[8]. Now it is possible to gatherhuge quantities of process conditions and operational data produced from numerousparts of tools to be gathered in creating an automated (FDD) Fault Detection and Diagnosis [9]acknowledgements to recent developments in technology, communication networks,information techniques, and computerized control. The collected data could be used to establish further client-specific methodologies forintelligentprecautionary maintenance practices, also recognized as PdM [10].

Recently, machine learning (ML) has emerged as one of the most important methods for designing intelligent predictive algorithms in the sense of AI (Figure 1, copyright permission of Figure 1 has been taken on 20 September 2020). Over the last few decades, it has developed into a large area of study.

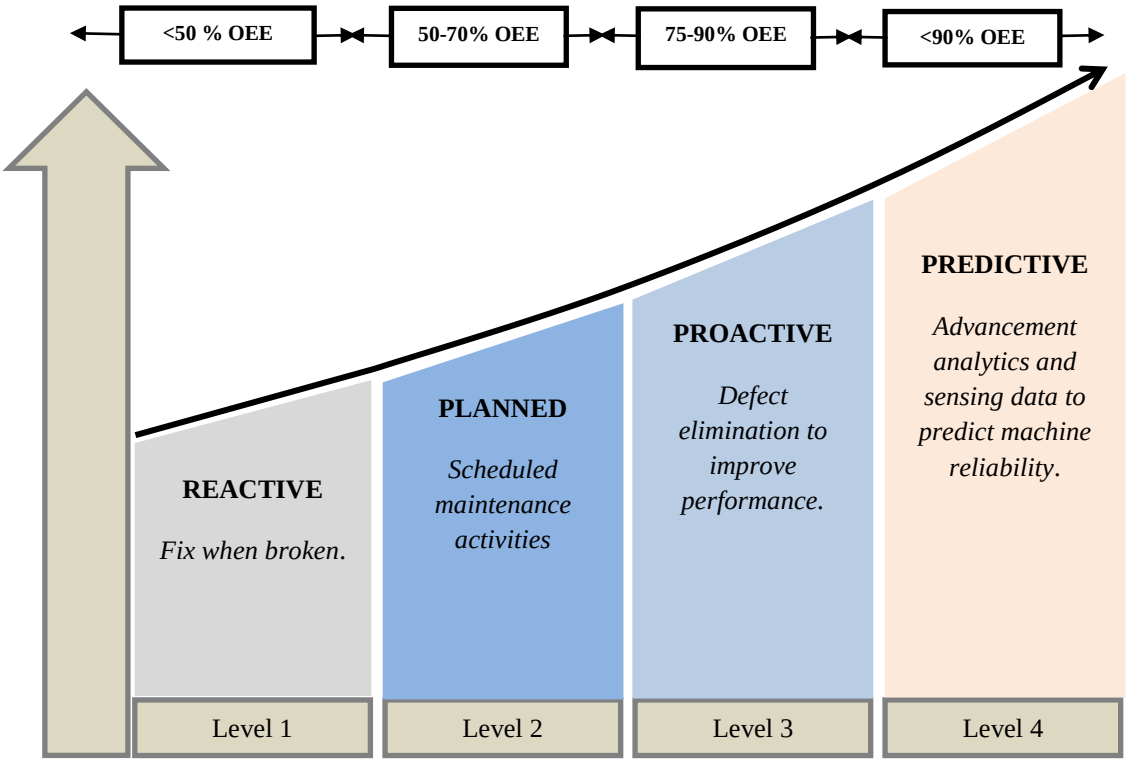


Figure 1: Maintenance type[1].

PDM is constructed on the premise that specificattributes of machinery could be tracked, and the data collected used to measure the equipment's remaining useful life. As a result, this form of maintenance strategy requires a range of major changes in the manufacturing and maintenance processes, all of which will substantially lower production costs [11]. First, since predictive maintenance is not dependent on periodic maintenance cycles tied to average lifespan, it can minimize the amount of excessive maintenance activities. As a result, the total number of maintenance tasks performed over the life of a system can be decreased. Because of the configuration in which some part is used in biggerequipment, for example. Both the reduction of fatal breakdowns and the reduction of excessive maintenance result in increased efficiency and less downtime in the manufacturing process. In comparison to traditional maintenance policies, Pdm can be considered an overall rise in productivity based on the accuracy of the prognostic approach used [12,13].As a result, the aim of this paper is to deliver a complete overview of recent advances in machine learning techniques applied to PdM. This paper aims to pinpoint and categorize based on the ML technique considered, ML category, equipment used, system used in data acquisition, applied data definition, data size, and data form from a detailed perspective. This paper's goal is to give an overview of these

initiatives, with an emphasis on I4.0 and Smart Manufacturing, as well as some application examples. Present and potential research problems for Smart Manufacturing will be recognized based on the findings.

Temperature readings from IoT sensors deployedoutdoorand inside an anonymous room have been collected. Because the device was still being tested, it was uninstalled or turned off numerous times during the reading period, resulting in some outliers and missing values.

1.1 Industry 4.0

The word Industry 4.0 refers to the fourth Industrial Revolution, a period in the growth of humanity's production systems. Industry 4.0's main goal is to make manufacturing – and allied industries like logistics – faster, more efficient, and more customer-centric, while also going beyond automation and optimization to discover new business prospects and models. The first three technological revolutions brought mechanization, energy, and IT to human development. Germany is one of the top technology production countries and has many of the most specialized suppliers and factories. In addition, two of three research and development funds are supported by the German government for industrial development, allowing industrial technology to rapidly expand. The passive machines and robotics have replaced the labor powers, which means that they are operated by an unconscious human. In 2012, the number of industrial robots in Germany was around 273 per 1000 employees, (K. Sector).

However, the use of personnel and supplementary services for monitoring, testing or effective servicing is indeed costly. Recently, the Internet of Things (IOT) and cyber physical system (CPS) have been used to link industry-related items such as content, sensors, equipment, goods, supply chain and consumers, which means that these required objects share information and control measures independently and autonomously with each other. German engineers understand that development has been a modern paradigm change, the so-called 'Industry 4.0,' in which consumers manage their own manufacturing processing.

Since then, Industry 4.0 is one of the world's most common manufacturing issues and has also been the fourth industrial revolution with severe potential effects on manufacturing. Many other developed countries are almost simultaneously conscious of this modern technological age. In China, the industrial growth strategy 'Made in China 2025' was issued in 2015. A business growth framework for the same reasons as Industry 4.0 was also developed [14] According to several analysts' studies and views, the future view of production, the latest business models and the framework architecture are discussed here to suggest the core ideas of Industry 4.0.

Since the late 1700s, when the first industrial revolution began, industrial maintenance and reliability techniques have evolved. We are well into the fourth industrial revolution (Industry 4.0), and PdM (predictive maintenance) solution suppliers promise a reliability panacea.

1.2. IOT and ML

Industry, infrastructures are all undergoing digital transformations. Whether it's referred to as the Industrial Internet of Things (IoT), Industrie 4.0. discrete and process manufacturing organizations have begun to reinvent their business models using accessible technology.Since IoT devices produce enormous amount data, conventional storage,data collection, and processing techniques may not be adequate. Patterns, habits, forecasts, and evaluation can all be aided by the massive amount of data. Furthermore, the heterogeneity of IoT data presents a new front for

current data processing frameworks. As a result, new mechanisms are needed to unlock the value of IoT-generated data. ML is one of the best computational models for embedding knowledge in IoT devices. [15]

Machine learning can assist machines and smart devices in deducing useful information from data created by devices or humans. It can also be described as a smart device's ability to change or automate a situation or behavior based on experience, which is an important component of an IoT solution.In tasks like classification, regression, and density estimation, machine learning techniques have been used. ML algorithms and techniques are used in several applications, including computer vision, bioinformatics, malware detection, authentication, and speech recognition. In the same way, ML can be used in IoT to provide intelligent services.However, this paper, concentrate on the use of machine learning to architecture for predictive maintenance. [15]

1.3.The Vision and concept of Industry 4.0

Many scholars believe that industrial developments entail a lengthy gestation cycle and address the following four elements, known as potential output visions:

- **Factory.** As one of Industry 4.0's key components, the future factory will include a new integrative facility, which can not only link and share information on all the manufacturing tools (sensors, actuators, motors, robotics, conveyors, etc.) but also make the factories responsive and intelligent enough to anticipate and manage their devices. Many manufacturing processes, such as product design, production planning, development engineering and development and services, can also be represented as hierarchical and linked, which means that these processes are not only managed by a decentralized structure, but are interdependently managed. This sort of future plant is called the Smart Factory.
- **Business.** Industry 4.0 means the presence of a full communication network among different businesses, manufacturers, vendors, logistics, tools, customers and so on. Each section optimizes the setup in real time based on the criteria and status of related network segments, allowing optimum benefit for all cooperatives with minimal capital for sharing. Cost and waste, raw materials, CO2 emissions and so on can also be decreased. In other words, each co-operating segment affects the future business network which can attain a self-organizing status and relay real-time replies.
- **Products.** The gains of Industry 4.0 would be a different form of industrial tool, that of smart devices. These goods are integrated into sensors, recognizable modules, and processors, which hold information and expertise to provide consumers with practical guidance and transmits feedback on the application to the production system. Many features may be applied to goods with these components, for example calculating product status or customers, distributing this information, monitoring items, and evaluating the results based on it. In addition, a total development details log with a product support developer can be implemented to improve the process, forecast and maintenance.
- **Customers.** In Business 4.0, consumers would still have certain benefits. A new form of payment would be introduced for consumers. It enables customers to order any product feature with any number, even though there is only one. In addition, also at the last minute, customers could without charge adjust their order and ideas at any time during production. In the other hand, the advantage of smart goods helps consumers not only to know the product 's development details, but also to get guidance on the use according to their own conduct[16].

2. Lean Manufacturing

Lean Manufacturing can be better stated as a multi-faceted development strategy having a wide range of organizational processes aimed for the detection of customer-scope value-adding processes and enabling these processes to flow through the enterprise at the pull of the customer[17]. It originated from the conceptualization of the Toyota Production System (TPS) at Toyota Motor Company by Taichii Ohno 's initiatives (Ohno, 1988). The primary goal of lean manufacturing is to create a streamlined flow of processes to produce the finished products with little or no waste at the required pace of customers. To define the dimensional structure of lean manufacturing, [17] conducted a systematic, multi-step approach-based analysis and built accurate scales to describe them. As described below, they quantified in ten variables the conceptual description and measurement of lean manufacturing.

1. ***Feedback ofSupplier:***Criticism and output of goods and services purchased from consumers to be conveyed regularly to suppliers to transmit knowledge effectively.
2. ***Just-In Time (JIT) suppliers'delivery:***Just the quantity of goods needed to be supplied by suppliers at a given time when they are required by customers.
3. ***Supplier development:***Suppliers would be produced in collaboration with the producer to avoid confusion or a discrepancy in competence levels.
4. ***Customer involvement:***Customers are the key drivers of a business, and high priority should be given to their needs and expectations.
5. ***Pull production:***An initiation of the need from the successor through Kanban should allow the predecessor's production flow, signified as production of JIT.
6. ***Continuous flow:***A efficient flow of goods should be formed through the factory without wide stops.
7. ***Setup time reduction:***The periodcompulsory to adapt resources for product variations should be maintained to the minimum possible extent.
8. ***Total productive:***Successful periodic maintenance procedures should prevent the breakdown of machines and equipment. In the event of failure, it is important to maintain a low rectification period.
9. ***Statistical process control:***Product quality is of prime importance, from a system to a subsequent one, no defect should be percolated.
10. ***Employee involvement:***Employees are to be motivated with enough encouragement and entitlement to make an overall contribution to the business.

3. Predictive Maintenance

Predictive maintenance (PdM) has seemed like the ideal application for the Internet of Things (IoT), particularly for Industrial IoT (IIoT) and contexts where asset uptime is vital, and breakdowns can have serious consequences for a variety of reasons.It's no surprise that predictive maintenance is one of Industry 4.0's most often discussed use cases.PdM, on the other hand, is not only about smart manufacturing. Transportation, oil and gas, process industries in general, and numerous segments with critical power settings are among the industries/segments that use predictive maintenance as a use case.Industry 4.0, the arrival of modern numericalindustrial technologies, seeks to make it possible for factories to produce higher-quality products at reduced costs more easily, more flexibly and

more effectively (Industry 4.0 Working Group 2013). As in 'Predictive Maintenance 4.0' of Industry 4.0[18] PdM has been featured as a key theme. PdM tracks the health of equipment and indicates when a maintenance event will be required in the future, when the primary enabler for optimizing the availability of tools has been elevated to the highest priority. The factory wide PdM specifications are:

a. Strong infrastructure and fast platform for communication and processing of data.
b. Efficient diagnostic and prognostic engine for faults.
c. Manageable health index hierarchy from a factory-wide view.

Industry 4.0 is a collective concept for technology, paving the way for a smart factory and manufacturing that can be accomplished by both IoT and CPS integration. A smart factory has smart-manufacturing scenarios; and by incorporating IoT, CPS, cloud-based techniques, and big-data technologies, smart manufacturing emphasizes man-machine collaboration and production logistics management. In order not only to achieve the goals of Industry 4.0, but also to achieve the target of NoDefects, the authors suggested a platform called AMCoT. To act as the IoT agent, the AMCoT platform adopts the so-called CPA. The predictive maintenance system's major integrated components are as follows:

- **Cyber-physical agent (CPA*):*** CPA plays a significant role in the AMCoT platform by
  a) Collecting data and dealing with physical objects, cyber networks, and human operators,
  b) identifying all the physical objects, and
  c) supporting intelligent applications. CPA is composed of CPA control kernel, communication service, data collection manager (DCM), data collection plan (DCP), data collection report (DCR), equipment driver (ED), application interface (AI), and database.
- **Advanced manufacturing cloud of things (AMCoT):** AMCoT offers a cloud-based network between the seller and its customers to communicate and exchange all knowledge about items. In this way, the seller will create attractive after-sales services, such as building AMCoT on-demand services / models directly, fanning out AMCoT real-time models, and tracking all machine tools through AMCoT to minimize maintenance costs. AMCoT offers a forum for bridging suppliers, consumers, and manufacturing tools with technology support.

4. Review of Literature

A literature review is an important component of every research project. The writers used a similar analysis technique in this article. First, related sources of publication about innovations in the fields of Industry 4.0 and smart manufacturing were established for this article. The authors cited articles from the Web of Science (WoS) website, which features many prestigious publications such as Emerald, Taylor and Francis, Springer, IEEE, and Elsevier. The systematic analysis approach followed a six-step procedure, as seen in Fig. 2.[19,20]

O'Donovan et al. (2015)[21] published a survey that analyzed and rated the papers, indicating research progress relevant to Big Data and Industry 4.0. They start with a thorough mapping study of Big Data in manufacturing. The mapping's findings for two study questions piqued our interest: "What kind of analytics are used

in big data in manufacturing?" and "What kind of analysis is done in big data in manufacturing?" Moreover, the agitation in coming future works mentioning the necessity for repair and diagnostic studies aided in the continuation of this article

In terms of maintenance,[22] presented a technique and framework that described a CPS template for Big Data for PdM and demonstrated that it is possible to render the degradation of an asset visible to human users by leveraging technology. The same can be said for report[23], which included concepts, implementations, and challenges in CPS domain forms. Current improvements in industrial information science surrounding Big Data environment, CPS, and Industry 4.0,' according to [24]. Similarly, [25] present Big Data's effect as a knowledge domain, categorizing works into twelve technology fields and six Big Data challenges

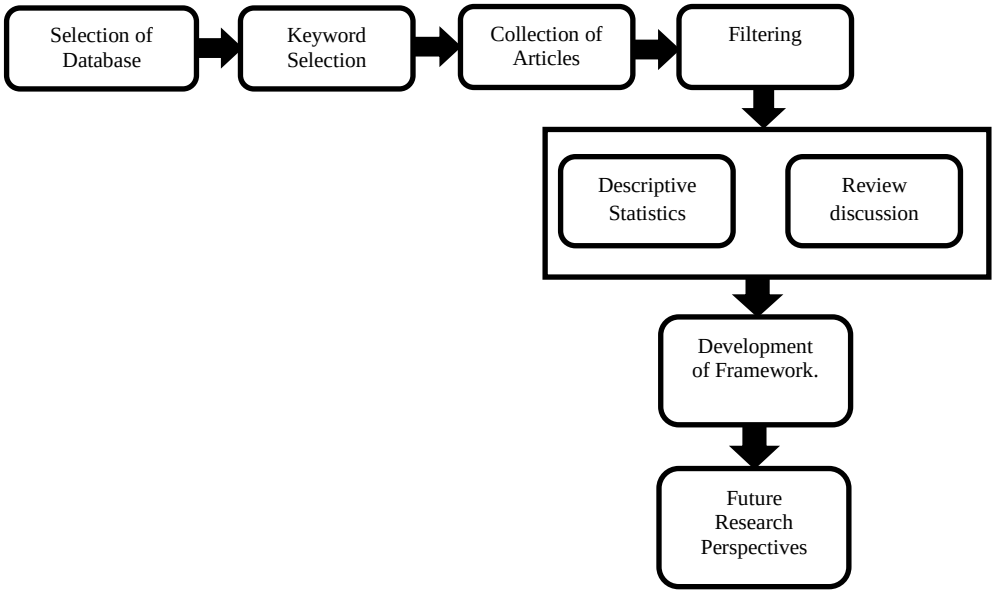


Figure 2: Research process adopted for the structured literature review.

In terms of maintenance,[22] presented a technique and framework that described a CPS template for Big Data for PdM and demonstrated that it is possible to render the degradation of an asset visible to human users by leveraging technology. The same can be said for [23]. report, which included concepts, implementations, and challenges in CPS domain forms. Current improvements in industrial information science surrounding Big Data environment, CPS, and Industry 4.0,' according to [24]. Similarly, [25] present Big Data's effect as a knowledge domain, categorizing works into twelve technology fields and six Big Data challenges.

According to recent studies, the number of works implementing Industry 4.0 models, platforms, technologies, usage cases, and other facets is growing [26]. As addressed in this article, the techniques provide a wide variety of applications in fields such as process and preparation (value increment and waste reduction), supply chain, transportation and logistics, health and safety, product design, and, most notably, maintenance and diagnosis. For example, in their guidelines,)[27] discuss their use for energy utilization reduction, endwise product engineering

through the complete supply chain, custom manufacturing support, telepresence, and unexpected supplier adjustments during growth. Energy conservation was also discussed in the work of [28]. [29]studied the trends of process improvement that developed because of the industrial revolution.

Another interesting application, as described by[30] is the development of simulated worlds utilizingenhancedexperience, which may be used to assist operators in a dynamic manufacturing setting. [31]. built on this notion in their study. This project investigates the usage of Industry 4.0 concepts for asset predictive management, a crucialelement of a company's effectiveness and product quality [32,33]

The number of case studies in the literature is increasing as PdM becomes more commonly used. Such as,) [34]identified a cloud-based tool for condition tracking of a cutting system, through which the machine's health status is communicated to the operators through a web interface. [35], which is more relevant to our function, employs a Kalman filter to approximate the state of a DC motor for PdM purposes. As mentioned by Prajapati et al. (2012)[36], this approach has the downside of being complex and computationally expensive, rendering it unsuitable for sensitive systems. Papers by [37], [38], and [39] provide additional case studies.

They discuss the need for predictive maintenance in the field of Industry 4.0, but they don't go into detail about it. They explain CPS, Big Data, and problems associated with large volumes of data, but they don't go into detail about it[40,41,42,43]. Paper discovered a few technical papers that provide a more detailed description of PdM solutions and services[44]. Except for two recent systematic reviews that explicitly cover ML and implementations of data-driven approaches in PdM, none of them specifically concentrate on PdM implementation[45,46]. Despite the fact that the subject is not as broad as suggest in this article. paper looked at and identified a few key topics for PdM researchers.

Data collection, pre-processing, wear detection, and indicating out the probability of collapse are all common themes in works that approach PdM. [47,48]. This definition is remarkably similar in works about PHM and the period of study, observation, and action that it entails.[49], PHM has been examined in the literature by investigators from various engineering fields to improve the dependability, accessibility, protection, and cost-effectiveness of engineering properties. Authors such as [34]see advanced forecasting methods as an integral part of their study. The explanation for this review's focus on PHM content is that.

These works focus on various strategies for deciding the condition of the machinery, the majority of which come from the (AI) Artificial Intelligence or(ML) Machine Learning sectors [50]. Machine Learningmethods are data-focused methods capable of detecting complicated and non-linear trends in data and constructing models from them for regression, classification, detection, or estimation[51,52,53]. Among these methods are Support Vector Machines, Decision Trees, Neural Networks, and so on, with the model selected depending on the following criteria: the kind of data it must operate with, the operating conditions, and the type of outcomes it must [54].

In this paper, itrecommends a PdM model that uses an ML technique called a Discrete Bayes (DBF) to transform data from sensors, systems, and domain experts into details about the machinery's deterioration condition and potential actions. The filter naturally models the knowledge latent in these types of processes, offering a valuable measure of confidence in its result. Another significant advantage of DBFs over other ML alternatives is their resistance to fluctuating and noisy data. The suggested filter will also benefit from practice, addressing the time constraints imposed by industrial environments. Other filters, such as the Kalman filter its extensions may be

considered as well, but they either depend on the underlying system's linearity assumptions or involve the calculus of complex Jacobians to linearly approximate its dynamics and propagate uncertainty. This study is being done as part of the SiMoDiM project, with an emphasis on the Steckel mills used in the Hot Rolling method to manufacture stainless steel. To the best of our understanding, this is the first paper discussing the predictive management of Steckel Mill components.[55]

The RF algorithm has been discovered to be the most widely used MLmethod for predictive management having been used on a wide range of components,industrial equipment, or systems, like turbofan engines,aircrafts, rotatingmachineries, rotor bar-LS-PMSM, production lines, semiconductors, industrial pumps, cutting machines, supermarket refrigeration systems, (HDD)hard drive disc, wind turbine, and vending machines. CNC machines, wind turbines, aircraft, and semiconductors seemed to be the authors' main focus.[56]

The 4th industrial revolution, also recognized as Industry 4.0 (Germany/EU) and Smart Manufacturing, has received a lot of attention (USA). The traction and prominence that both programs (and related ones in many other countries) have achieved in recent years demonstrates the dramatic, paradigm-shifting change that the automotive industry and manufacturing science are undergoing today. I4.0 and Smart Manufacturing are terms used to describe the shift to a highly data-focused production network with improved integration and adoption of knowledge and networking technology thus holding people in the loop. Among the goals are energy conservation, sustainability (social, economic, and environmental), agility/resilience, and consistency and performance. [57]

Despite the availability of reliable testbeds, I4.0 and Smart Manufacturing are both in their infancy. Given the funding agencies' interest and accessible grants, as well as the strong interest from business (both major corporations and small and medium-sized enterprises), rapid developments in this area are anticipated in the near future. Because of their interdisciplinary nature, advances in basic research fields could make their way to commercial use more quickly than in previous years. This may be a chance for researchers who haven't had any contact with practical science in their area to collaborate with researchers from other fields to sectors to see their study come to fruition. [57]

Industry 4.0 would allow predictive and smart manufacturing in the future. An industry 4,0 factory's machines are joined as a joint group, allowing for a wide range of predictive maintenance options. The development of predictive maintenance, its technical problems, and its future in the Industry 4.0 ecosystem is discussed in this paper. In the one side, the field of smart factories, industrial big data, and cloud computing is enhanced and accepted in the Industry 4.0 period, paving the way for predictive maintenance. Predictive maintenance, on the other side, can play an important role in potential maintenance operations and will assist in meeting Industry 4.0's requirements for smart production and self-aware robots[58].

5. Research Gap

In the literature various of the forms are considered for the achievement of the goals of the industry 4.0 with the consideration of various systems and frameworks like lean management, smart data management, predictive maintenance, etc. While in none of the work the integration of the lean management with respect to the predictive management is being considered. As in predictive management the purpose of the system working is for the prediction of the coming process and requirement steps in terms of the data, tools, and other requirements. And also,

at the same lean management will help the industry to remove out the waste from the organization, which directly will affect the performance of manufacturing process.

Beyond this all the gap identified in the literature is about the consideration of the optimized techniques for information system in terms of the decision making, also for the integration of the optimization process online and information visualization. In the information system available the data first is supposed to be classified and then it to be picked for the decision-making process for which some specified machine learning techniques are supposed to be opted. In the current work the major focus is on the optimization process in terms of the performance, energy, etc. and also considered the decision-making process via data classification.

6. Objectives of the Research

i. To study and evaluate about industry 4.0.
ii. To study and evaluate the concept of the lean management and predictive maintenance.
iii. To design an integrated system as predictive maintenance module considering the principles of the lean management.
iv. To validate the research work using the comparison strategy with previous works done.

7. Research methodology

Industry 4.0 is all about the automation of the overall process and also about the inclusion of the techniques like IoT, Cloud, AI, etc. over the traditional system for manufacturing process. Complete manufacturing process is the integration of the various tools, machines, techniques working together for an integrated outcome. In the various of the frameworks towards the achievement of the industry 4.0 predictive maintenance is one which works towards the enhancement of the manufacturing process with efficient utilization of the resources used for the complete process. The maintenance process includes the consideration various processes like data collection, data storage, data processing, and checking working of the tools running for the manufacturing process.

Based on the previous recorded data the system provides a decision for the maintenance related steps, other than the maintenance related process the machine should be made to react for the lean management which works for the betterment of the organization is one of the better tools considered in the industry 4.0. In the prediction system the modules like Cyber-physical agent (CPA) and Advanced manufacturing cloud of things (AMCoT) are integrated together for better prediction process. In the process of predictive maintenance, the five principles of the lean management are to be considered for the better efficiency and ensuring the integration level of the industry 4.0 and lean.

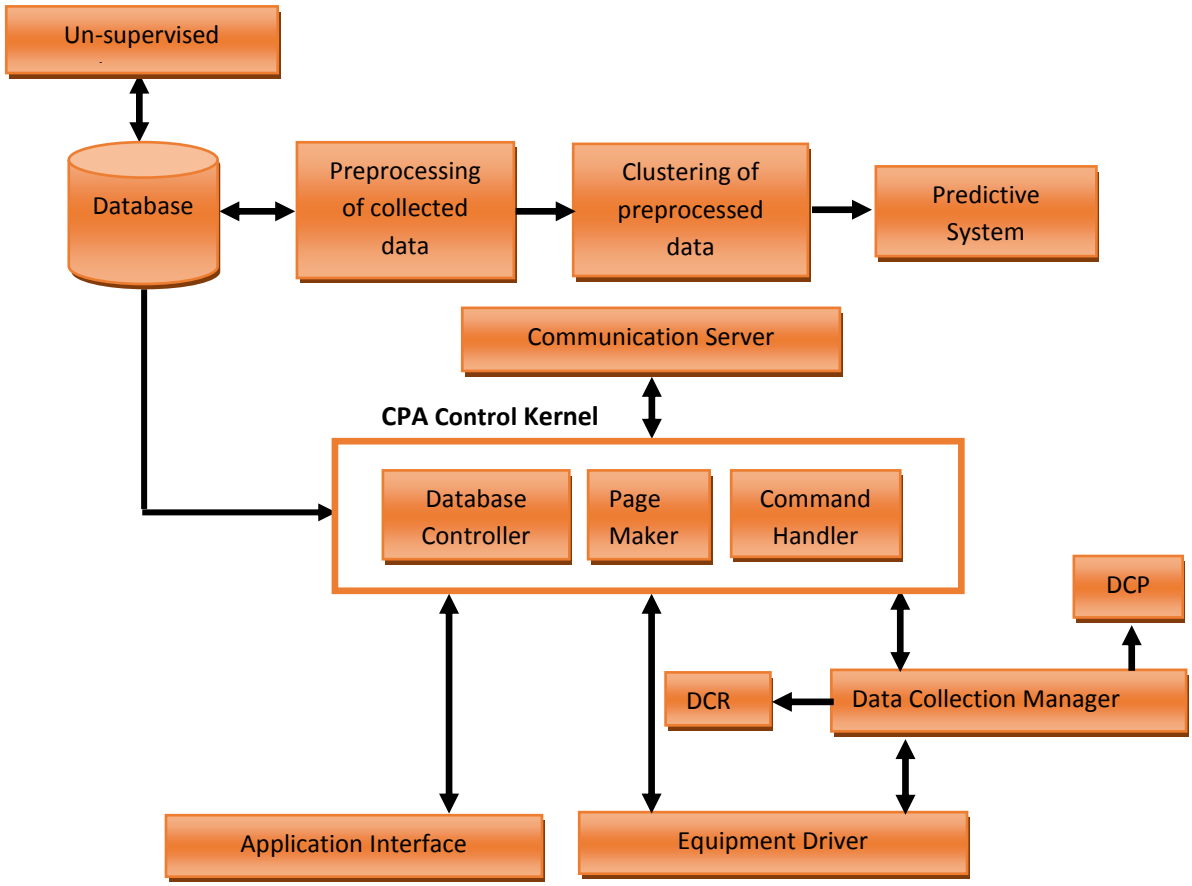


Figure 3:Working Architecture of the predictive system.

The complete research work processes in three different modules as, Data gathering/collection, Dataset training and Decision making. For the data gathering the process defined in CPA architecture (Y. Chiu *et al.*, 2017) is considered which is having various modules integrated for the data collection. CPA architecture is composed of CPA control kernel, communication service, data collection manager (DCM), data collection plan (DCP), data collection report (DCR), equipment driver (ED), application interface (AI), and database (Cheng et al. 2016).

The un-supervised learning technique is considered for data set training, where the algorithm is expected to analyse the actual data with unsupervised machine learning to identify similarities, patterns, and correlations in the data to explore and learn about relationships within the data. The training of the dataset is then accompanied by a clustering process, clustering is the method of identifying correlations in un-labeled data in order to combine related data items into a cluster together.

The work majorly focuses for the lean manufacturing which is the integrated contribution of the various modules like supply chain, process/operation, human employment, control and human factors, predictive maintenance, etc. In the present work the major consideration will be towards the smart- supply system and lean manufacturing with the help of the predictive maintenance in temperature analysis. In the proposed study the predictive system is to be considered which is optimized with the help of the learning process to present the smart supply system and work towards the lean manufacturing considering various parameters that defines the lean manufacturing.

8. Implementation Results

In the datasets, it has the temperature readings from IoT devices installedoutdoorand inside of an anonymous room.

Dataset details:

- **temp**: temperature readings of the dataset.
- **out/in** whether reading was taken from device installed inside oroutdoorofroom.

Finding of the dataset:

- the temperature difference between interior and exterior
- trend or seasonality in the data
- forecasting future temperature by using time-series modeling
- characteristic tendency through year, month, week, or day/night

Time series analysis of the temperatures.

1. Inside temperature is composed of a single distribution, whileoutdoortemperature is composed of multiple distributions.
2. The temperature indoor the room is kept constant by the air conditioner, but the outdoor temperature is easily affected by time-series factors such as seasons.
3. The outdoor temperature has a larger time series change than the indoor temperature.

According to this observation, India has four climatological seasons as below.

- Winter: December to February
- Summer: March to May
- Monsoon: June to September
- Post-monsoon: October to November

We can create seasonal variable based on month variable.

The time series analysis to predict the future temperature,inside and outdoor the industries.

Result 1: Figure 4 shows the inside temperature prediction of the industry based on the monthly basis.

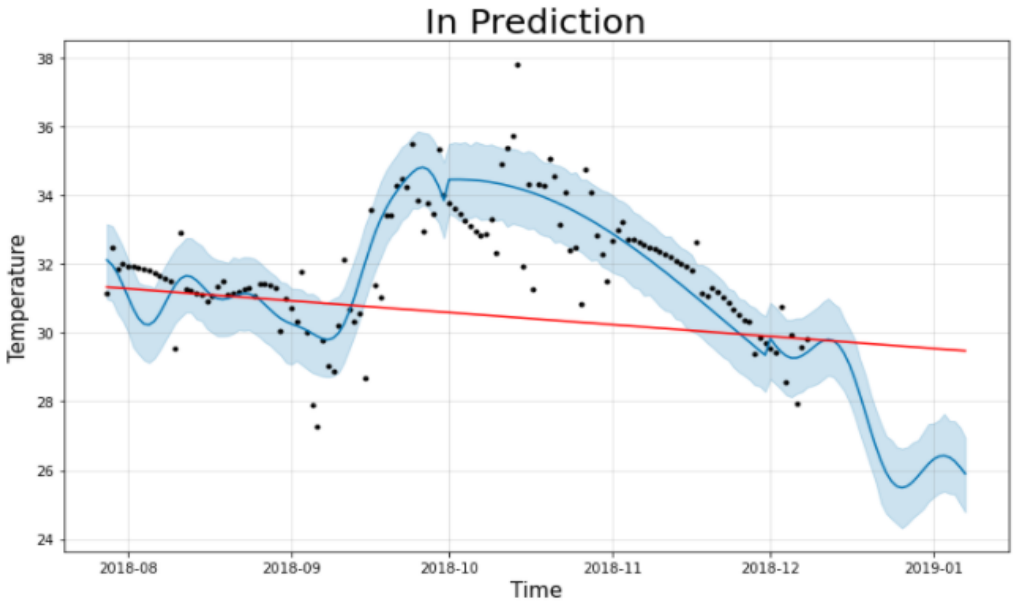
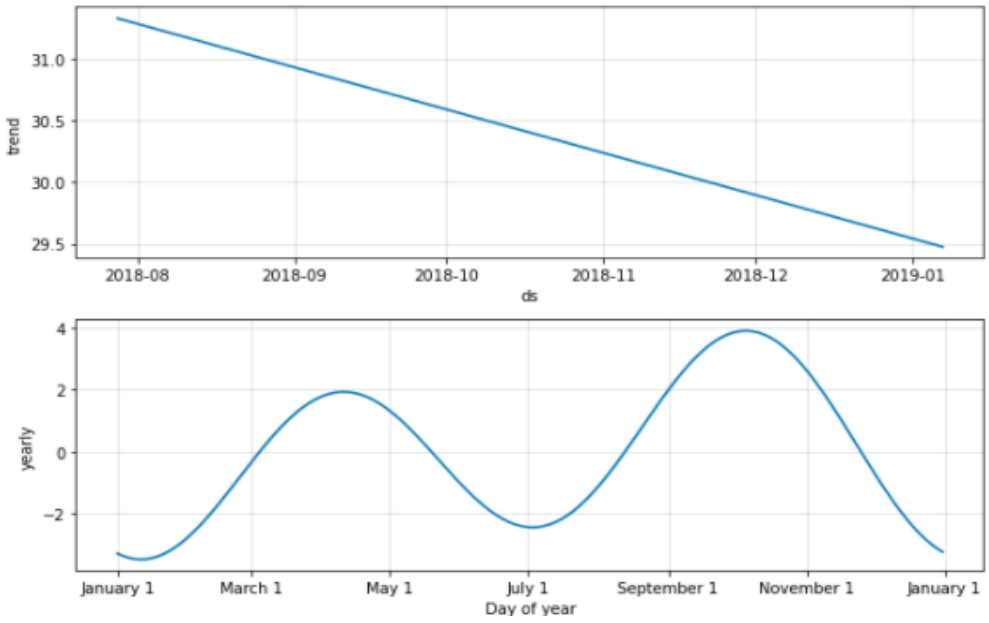


Figure 4: Prediction of inside temperature.



Result 2: Figure 5 shows trend of temperature according to seasonality.

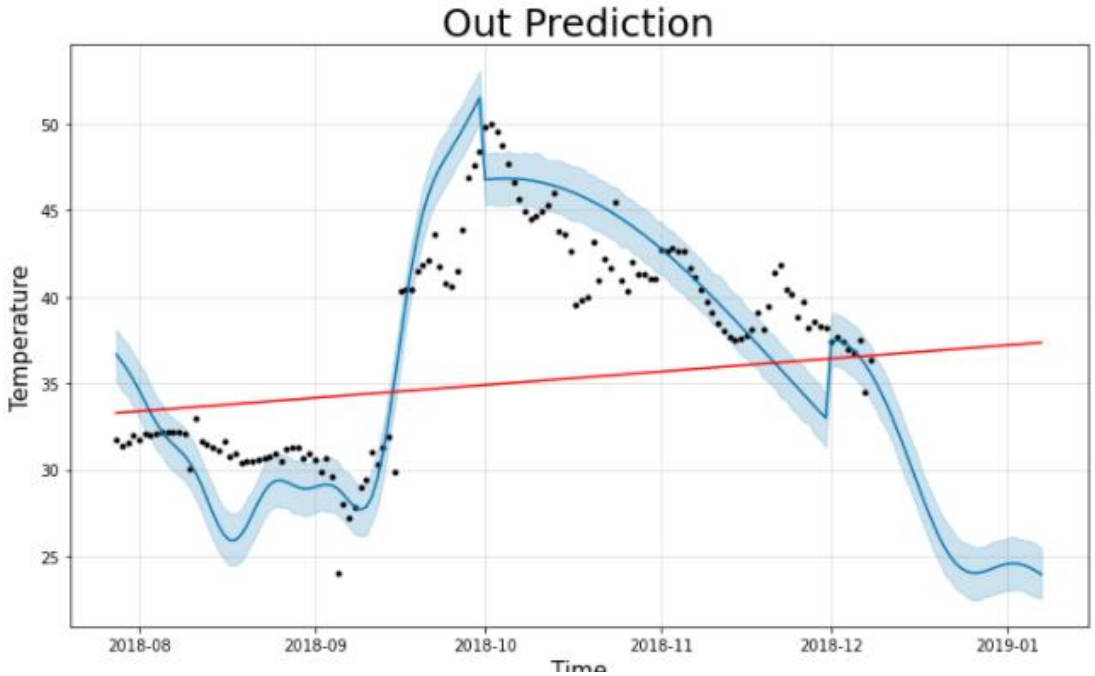


Figure 5: Prediction of inside temperature

Result 3: Figure 6 shows outdoor temperature prediction of the industry based on the monthly basis.

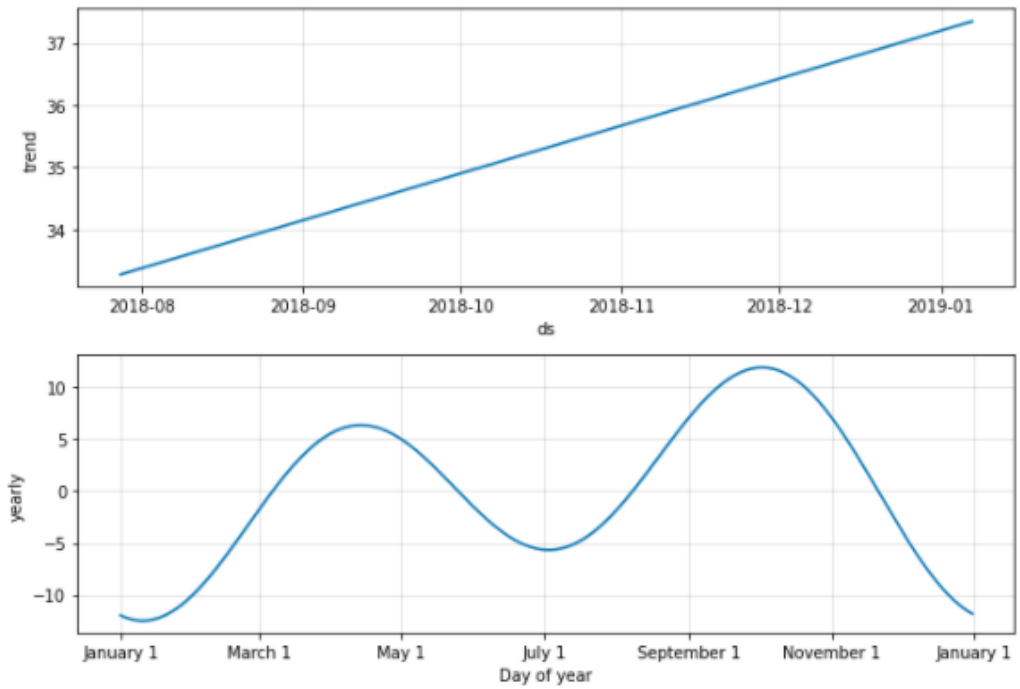


Figure 6: Trend of inside temperature

Result 4: Figure 7 shows trend of temperature according to seasonality outdoor.

9. Conclusion

The presented work focuses on the predictive system for the lean manufacturing as major goal. Lean manufacturing goes with the consideration of the customer involvement, supply system, employee involvement, continuous flow of the end results. The work considers the decision-making system where the un-supervised learning techniques is being used for training the dataset as the data generated is a un-labelled temperature data generated by different sensors, manual notations or machine generated temperature data and the outdoor fluctuated temperature data. The data after pre-processing is clustered to generate the data in different groups based on similarity, which is then is being used for the decision-making system for the future prediction of temperature. The system is also trained for making the decisions for the maintenance related decisions about the machines operating. The end outcome of the system is the seasonal information of the temperature for analysis of the future temperature prediction. The forecast model is implemented with the indoor and outdoor temperature. Annual temperature swings can have a greater impact on outside temperature than on indoor temperature.Outside temperature is made up of numerous distributions, whereas inside temperature is made up of a single distribution.

REFERENCE

1. Hill and Smith. 2009. *Performers at ISA Expo 2009*. Enterprise Integration Track, Houston, TX, USA: Reliant Center
2. Panetto, Hervé, and Arturo Molina. 2008. "Enterprise Integration and Interoperability in Manufacturing Systems: Trends and Issues." *Computers in industry* 59 (7): 641–646
3. Atzori, Luigi, Antonio Iera, and Giacomo Morabito. 2010. "The Internet of Things: A Survey." *Computer Networks* 54 (15): 2787–2805.

Figure 7: Trend of outdoor temperature

4. Kehoe, B., S. Patil, P. Abbeel, and K. Goldberg. 2015. "A Survey of Research on Cloud Robotics and Automation." *IEEE Transactions on Automation Science and Engineering* 12 (2): 398–409.
5. Cinar, Z.M.; Nuhu, A.A.; Zeeshan, Q.; Korhan, O. Digital Twins for Industry 4.0: A Review. In Industrial Engineering in the Digital Disruption Era. GJCIE 2019. Lecture Notes in Management and Industrial Engineering; Calisir, F., Korhan, O., Eds.; Springer: Cham, Switzerland, 2020.
6. Cinar, Z.M.; Zeeshan, Q.; Solyali, D.; Korhan, O. Simulation of Factory 4.0: A Review. In Industrial Engineering in the Digital Disruption Era. GJCIE 2019. Lecture Notes in Management and Industrial Engineering; Calisir, F., Korhan, O., Eds.; Springer: Cham, Switzerland, 2020.
7. Zhang, W.; Yang, D.; Wang, H. Data-Driven Methods for Predictive Maintenance of Industrial Equipment:A Survey. IEEE Syst. J. **2019**, 13, 2213–2227.
8. Borgi, T.; Hidri, A.; Neef, B.; Naceur, M.S. Data analytics for predictive maintenance of industrial robots.In Proceedings of the 2017 International Conference on Advanced Systems and Electric Technologies(IC_ASET), Hammamet, Tunisia, 14–17 January 2017; pp. 412–417.
9. Dai, X.; Gao, Z. From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis.IEEE Trans. Ind. Inform. **2013**, 9, 2226–2238.

10. Lee, J.; Lapira, E.; Bagheri, B.; Kao, H.A. Recent advances and trends in predictive manufacturing systems inbig data environment. Manuf. Lett. **2013**, 1, 38–41
11. A. Grall, L. Dieulle, C. Berenguer, M. Roussignol, Continuoustime predictive-maintenance scheduling for a deteriorating system, IEEE Transactions on Reliability 51 (2) (2002) 141– 150. arXiv:arXiv:1011.1669v3, doi:10.1109/TR.2002. 1011518.
12. K. A. Nguyen, P. Do, A. Grall, Multi-level predictive maintenance for multi-component systems, Reliability Engineering and System Safety 144 (2015) 83–94. doi:10.1016/j.ress. 2015.07.017.
13. R. C. Yam, P.W. Tse, L. Li, P. Tu, Intelligent predictive decision support system for condition-based maintenance, International Journal of Advanced Manufacturing Technology 17 (5) (2001) 383–391. doi:10.1007/s001700170173.
14. Schlechtendahl, J., Keinert, M., Kretschmer, F., Lechler, A. and Verl, A., 2015. Making existing production systems Industry 4.0-ready. *Production Engineering*, *9*(1), pp.143-148
15. Hussain, Fatima, Rasheed Hussain, Syed Ali Hassan, and Ekram Hossain. "Machine learning in IoT security: Current solutions and future challenges." *IEEE Communications Surveys & Tutorials* 22, no. 3 (2020): 1686-1721.
16. Ohno, T., 1988. *Toyota production system: beyond large-scale production*. crc Press
17. Shah, R. and Ward, P.T., 2007. Defining and developing measures of lean production. Journal of operations management, 25(4), pp.785-805
18. Hitpass, Bernhard, and HernánAstudillo. "Industry 4.0 challenges for business process management and electronic-commerce." *Journal of theoretical and applied electronic commerce research* 14, no. 1 (2019): I-III
19. Lamba, Kuldeep, and Surya Prakash Singh. "Big data in operations and supply chain management: current trends and future perspectives." *Production Planning & Control* 28, no. 11-12 (2017): 877-890.
20. Kamble, Sachin S., Angappa Gunasekaran, and Shradha A. Gawankar. "Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives." *Process Safety and Environmental Protection* 117 (2018): 408-425.
21. O'Donovan, Peter, Kevin Leahy, Ken Bruton, and Dominic TJ O'Sullivan. "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities." *Journal of Big Data* 2, no. 1 (2015): 1-26.
22. Lee, J., Jin, C., & Bagheri, B. (2017). Cyber physical systems for predictive production systems. *Production Engineering*, *11*(2), 155–165.
23. Gunes, V., Peter, S., Givargis, T., & Vahid, F. (2014). A survey on concepts, applications, and challenges in cyber-physical systems. *KSII Transactions on Internet and Information Systems*, *8*(12), 4242–4268.
24. Lee, J., Bagheri, B., & Kao, H.-A. (2014). Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. In *12th IEEE international conference on industrial informatics* (pp. 1–6). Porto Alegre, Brazil: IEEE
25. Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., &Alor-Hernández, G. (2016). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, *72*(8), 3073–3113
26. Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10
27. Kagermann, H., Wahlster, W., Helbig, J., 2013. Recommention for implementing strategics initiative industries 4.0. Industrie 4.0 working group Germany.
28. Shrouf, Fadi, Joaquin Ordieres, and Giovanni Miragliotta. "Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm." In *2014 IEEE international conference on industrial engineering and engineering management*, pp. 697-701. IEEE, 2014.

29. Tamás, Péter, and Béla Illés. "PROCESS IMPROVEMENT TRENDS FOR MANUFACTURING SYSTEMS IN INDUSTRY 4.0." *Academic Journal of Manufacturing Engineering* 14, no. 4 (2016).
30. Paelke, Volker. "Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment." In *Proceedings of the 2014 IEEE emerging technology and factory automation (ETFA)*, pp. 1-4. IEEE, 2014.
31. Ruiz, RF Garcia, M. L. Bissell, Klaus Blaum, A. Ekström, N. Frömmgen, G. Hagen, M. Hammen et al. "Unexpectedly large charge radii of neutron-rich calcium isotopes." *Nature Physics* 12, no. 6 (2016): 594-598.
32. Liu, Jie, James A. Platts-Mills, Jane Juma, Furqan Kabir, Joseph Nkeze, Catherine Okoi, Darwin J. Operario et al. "Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study." *The Lancet* 388, no. 10051 (2016): 1291-1301.
33. Medina-Oliva, Gabriela, Alexandre Voisin, Maxime Monnin, and Jean-Baptiste Leger. "Predictive diagnosis based on a fleet-wide ontology approach." *Knowledge-Based Systems* 68 (2014): 40-57.
34. Lee, Jay, Edzel Lapira, Behrad Bagheri, and Hung-an Kao. "Recent advances and trends in predictive manufacturing systems in big data environment." *Manufacturing letters* 1, no. 1 (2013): 38-41.
35. Yang, Ming-Hsuan, David J. Kriegman, and Narendra Ahuja. "Detecting faces in images: A survey." *IEEE Transactions on pattern analysis and machine intelligence* 24, no. 1 (2002): 34-58.
36. Prajapati, D. R. "Implementation of failure mode and effect analysis: a literature review." *International journal of management, IT and Engineering* 2, no. 7 (2012): 264-292.
37. Li, D. (2016). Perspective for smart factory in petrochemical industry. *Computers and Chemical Engineering*, *91*, 136–148.
38. Shin, Jong-Ho, and Hong-Bae Jun. "On condition based maintenance policy." *Journal of Computational Design and Engineering* 2, no. 2 (2015): 119-127.
39. Choudhary, Poonam, Sanjeev Kumar Prajapati, and Anushree Malik. "Screening native microalgal consortia for biomass production and nutrient removal from rural wastewaters for bioenergy applications." *Ecological Engineering* 91 (2016): 221-230.
40. Ayad, Soheyb, Labib SadekTerrissa, and Noureddine Zerhouni. "An IoT approach for a smart maintenance." In *2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, pp. 210-214. IEEE, 2018.
41. Jin, Wenjing, Zongchang Liu, Zhe Shi, Chao Jin, and Jay Lee. "CPS-enabled worry-free industrial applications." In *2017 Prognostics and System Health Management Conference (PHM-Harbin)*, pp. 1-7. IEEE, 2017.
42. Kwon, Daeil, Melinda R. Hodkiewicz, Jiajie Fan, Tadahiro Shibutani, and Michael G. Pecht. "IoT-based prognostics and systems health management for industrial applications." *IEEE Access* 4 (2016): 3659-3670.
43. Yan, Hehua, Jiafu Wan, Chunhua Zhang, Shenglong Tang, Qingsong Hua, and Zhongren Wang. "Industrial big data analytics for prediction of remaining useful life based on deep learning." *IEEE Access* 6 (2018): 17190-17197.
44. Haarman, Mark, Michel Mulders, and Costas Vassiliadis. "Predictive maintenance 4.0: predict the unpredictable." *PwC and Mainnovation* (2017).
45. Carvalho, Thyago P., Fabrízzio AAMN Soares, Roberto Vita, Roberto da P. Francisco, João P. Basto, and Symone GS Alcalá. "A systematic literature review of machine learning methods applied to predictive maintenance." *Computers & Industrial Engineering* 137 (2019): 106024.
46. Zhang, Weiting, Dong Yang, and Hongchao Wang. "Data-driven methods for predictive maintenance of industrial equipment: a survey." *IEEE Systems Journal* 13, no. 3 (2019): 2213-2227.
47. Kwon, Daeil, Melinda R. Hodkiewicz, Jiajie Fan, Tadahiro Shibutani, and Michael G. Pecht. "IoT-based prognostics and systems health management for industrial applications." *IEEE Access* 4 (2016): 3659-3670.

48. Terrissa, Labib Sadek, Safa Meraghni, Zahra Bouzidi, and Noureddine Zerhouni. "A new approach of PHM as a service in cloud computing." In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 610-614. IEEE, 2016.
49. Atamuradov, Vepa, Kamal Medjaher, Pierre Dersin, Benjamin Lamoureux, and Noureddine Zerhouni. "Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation." *International Journal of Prognostics and Health Management* 8, no. 060 (2017): 1-31.
50. Dopico, M., Gomez, A., De la Fuente, D., García, N., Rosillo, R., Puche, J., 2016. A vision of industry 40 from an artificial intelligence point of view. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), the Steering Committee of the World Congress in Computer Science. Computer Engineering and Applied Computing (WorldComp, p. 407
51. Wuest, T., D. Weimer, C. Irgens, and K. D. Thoben. "Machine learning in manufacturing: advantages, challenges, and applications. Prod. Manuf. Res. 4 (1), 23–45 (2016)."
52. Wu, Dazhong, Connor Jennings, Janis Terpenny, Robert X. Gao, and Soundar Kumara. "A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests." *Journal of Manufacturing Science and Engineering* 139, no. 7 (2017).
53. Amruthnath, Nagdev, and Tarun Gupta. "A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance." In *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 355-361. IEEE, 2018.
54. Djurdjanovic, Dragan, Jay Lee, and Jun Ni. "Watchdog Agent—an infotronics-based prognostics approach for product performance degradation assessment and prediction." *Advanced Engineering Informatics* 17, no. 3-4 (2003): 109-125.
55. Ruiz-Sarmiento, Jose-Raul, Javier Monroy, Francisco-Angel Moreno, Cipriano Galindo, Jose-Maria Bonelo, and Javier Gonzalez-Jimenez. "A predictive model for the maintenance of industrial machinery in the context of industry 4.0." *Engineering Applications of Artificial Intelligence* 87 (2020): 103289.
56. Çınar, Zeki Murat, Abubakar Abdussalam Nuhu, Qasim Zeeshan, Orhan Korhan, Mohammed Asmael, and Babak Safaei. "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0." *Sustainability* 12, no. 19 (2020): 8211.
57. Thoben, Klaus-Dieter, Stefan Wiesner, and Thorsten Wuest. "“Industrie 4.0” and smart manufacturing-a review of research issues and application examples." *International journal of automation technology* 11, no. 1 (2017): 4-16.
58. Li, Zhe, Kesheng Wang, and Yafei He. "Industry 4.0-potentials for predictive maintenance." *Advances in Economics, Business and Management Research* (2016).

IoT Based Low Cost Bridge Health Monitoring System With Future Predictive Analysis Using MATLAB and ThingSpeak

[1] Zaheen Shaikh,[2]Snehil Singh, [3] Mohammed Zaid Nidgundi, [4]Jahida Subhedar

[1,2,3] *Student, B.Tech., Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, Maharashtra, India*

[4]*Asst. Prof, Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, Maharashtra, India*

***Abstract*—As the expeditious of technological advancements emerging in civil engineering, structures like bridges are getting structurally complicated, and safety has become an issue. To overcome this issue, we have come up with a low-cost remote health monitoring system for bridges. This paper describes a proposal for an Arduino based low-cost bridge health monitoring system using IoT. This uses ThingSpeak with Arduino UNO and ESP8266 for accessing sensor readings. This system aims to record: (1) Temperature (2) Humidity (3) Water level (4) Vibration in X, Y, Z axis. (5) Forecasting has been carried out by creating regression model using MATLAB**

Keywords—Arduino UNO, Internet of Things, Bridge, Health Monitoring System, Sensors

I. INTRODUCTION

With the development of our society and improvement of the economy, various complicated Bridge systems are increasing. However, because of the complicated surroundings faced by those Bridges, it is subject to several environmental impacts which include corrosion, several coupling factors, accidental overload and impact, scour, natural calamities, and complex vibration mechanisms. Therefore, the use of science and technology to evaluate and monitor the structural state has obtained increasingly more attention. The Bridge Health Monitoring System is a vital tool to improve the safety and maintainability bridges. It provides real-time and accurate information about the bridge health condition. It is a process of continuous evaluations to detect location and extent of damage, calculate the remaining life, and predict upcoming accidents.

This model incorporates many sensors such as Temperature and Humidity sensor (DHT11), Accelerometer (ADXL345), Ultrasonic sensor(HC-SR04) etc. and Arduino UNO, ESP8266 WiFi Module. This model uses an 3-axis accelerometer sensor to measure the angle in X-Y-Z planes to measure the inclination. All raw data are processed through in-built processor and information is sent to the Arduino UNO. The novelty of this proposed model is listed as follows:

(i) Measuring the bend and inclination of column joints, bridge joints caused by malformation.

(ii) For alert system , a tweet is posted via ReactApp in ThingSpeak, Once values cross threshold limit.

(iii) Making a robust and mobile device system which can be monitored easily through user's smartphone and even can work autonomously without a user.

This designed system focuses on the health monitoring of the bridges using sensors. There had been numerous techniques to set up the Bridge Health monitoring system, Amongst these techniques, the internet of things (IoT) is an innovative and modern approach, that is a network concept for data exchange and extends the patron via the internet to any kind of information. The sensors real-time information is dispatched onto the server which may be accessed by the person using cloud, however the user need to have the login information. The most primary characteristic of ThingSpeak is the term 'channel' that have space for records, space for vicinity, space for status for numerous sensed statistics. As soon as channels are created in the 'ThingSpeak' the data can be implemented and alternately you could visualize and analyze the data the use of the MATLAB and respond to the data with tweets and other styles of alerts. ThingSpeak also provide a function to create a public based channel to research and estimate it through public. To engage the 'things' in sensing the respective information and transmitting it across the internet and one includes to head similarly just connecting data from a laptop, gadgets to gather (sensors) and to achieve this the information require to network uploaded which are inside the form of servers (that

runs programs) and such sorts are taken into consideration as cloud .

The 'cloud' utilizes the operations of graphical visualization and available within the form of digital server for the customers and the items are communicated with the cloud via possible 'wireless net connections' available to the customers and majority objects uses the sensors to inform regarding our environmental analogue statistics.

Regression Modeling is a machine learning algorithm that fall under supervised learning. Regressive models tell us the relation between an output variable and input variable. In the project, Auto Regressive Models are generated using System Identification Toolbox available in MATLAB and Forecasting (output) of data has been carried out based on measured (input) values at a given period of time. Regression model has its own advantages and disadvantages such as, Linear Regression is simple to implement and easier to interpret the output coefficients but on the other hand the outliers generated can have huge impact in decreasing the efficiency of the system. Linear Regression acts as a considerable tool for deriving relationships between variables but not always recommended for practical implementation as it simplifies real life issue to a greater extent by considering a linear relationship between the variables.

II. LITERATURE REVIEW

Fig 1*: Gujarat: Seven injured as bridge collapses on Junagadh-Sasan highway*

Seven persons were injured after the breakdown of a 60 feet long bridge over Madhuvanti river near Malanka village late Sunday evening. The bridge gave in from the middle at around 6pm. Three cars, which were passing over the bridge, also fell with the debris, injuring seven people. Those trapped in the stranded vehicles were rescued by the locals and rushed to hospitals in Junagadh. Two cars and three two-wheelers were stuck under the crashed slabs of the bridge and were later pulled out of the debris. Officers said the bridge upstream the Madhuvanti dam, was around four decades old. Stretches of the road, which leads to Sasan — a popular tourist hub in Gir forest — were widened in 2017-'18. However, the bridge was not widened nor a new one was constructed in its place. Following the collapse of the bridge near Malanka, vehicular traffic had to be diverted on alternate routes. No diversion at the place of the collapsed bridge was created till Monday evening.

Sr No.	Title	Year	Reference	Technologies used
1	Design of Bridge Monitoring System Based on IoT	2018	MVPJES	Wireless Sensor Network, Cloud Server, TCP/IP Protocol, C#, Android Smartphone
2	A Real-Time and Low-Cost Flash Flood Monitoring System to Support Transportation Infrastructure	2020	IEEE	Arduino Board, NRF24L01 radio transceiver, Adafruit SHT31-D Temperature & Humidity sensor, ultra-sonic sensors, C++, My SQL, Python
3	An Internet of Things (IOT) Based System to Analyze Real-time Collapsing Probability of Structures	2018	IEEE	Vibration sensor, Flex sensor (FS7548), Piezo Buzzer (PS1927P02), Red LED, Arduino 101, Integrate Bluetooth Low Energy (BLE), Wifi Module (ESP8266), Blynk software
4	Bridge Safety Monitoring System using IOT	2020	IJITEE	ESP8266 Nodemcu, Water level sensor, Vibration sensors, HX711 Load Cell Amplifier, servo motor
5	Design and Implementation of Real time monitoring of bridge using Wireless technology	2020	IEEE	Raspberry Pi, Load Sensors, Vibration sensor, Water level sensor, tilt sensor, Wi-Fi module

Table 1: *Literature survey*

III. PROPOSED METHODOLOGY

The framework is developed using sensors unit which consists of Temperature and Humidity sensor (DHT11), Accelerometer (ADXL345), Ultrasonic sensor(HC-SR04). Data acquisition and processing unit includes Arduino UNO, ESP8266 WiFi Module. Outputs of all the sensors go to Arduino UNO. This module constantly takes readings from the sensors, it takes about a hundred microseconds (0.0001 s) to read an analog input, therefore the most reading rate is about 10,000 times a second. This output is further given to the ESP8266 WiFi module for sending it to the ThingSpeak cloud platform over the internet for doing analysis on it using MATLAB. Further this data is analyzed in MATLAB, and future predictive analysis is done.

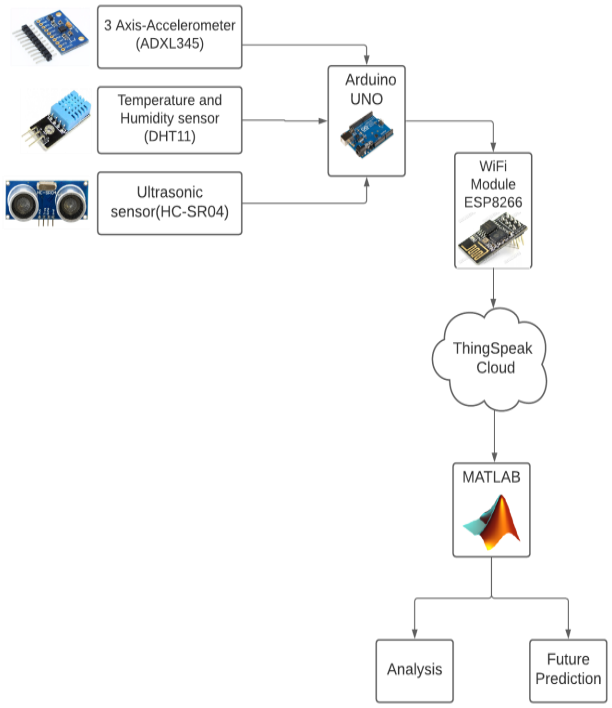


***Fig 2**: Schematic Flow Chart of Proposed framework*

A. Hardware Specifications

The proposed model's hardware unit comprises of Arduino UNO, ESP8266 WiFi Module, Temperature and Humidity sensor (DHT11), Accelerometer (ADXL345), Ultrasonic sensor(HC-SR04).

1) Arduino UNO : A Data Processing Board

Arduino UNO is an open-source microcontroller board, which is primarily based on the ATmega328P microcontroller. The board is geared up with units of digital and analog input/output pins that can be interfaced to numerous extension boards and different circuits. The board has 14 digital I/O pins (six able to PWM output), 6 analog I/O pins, and is programmable with the Arduino IDE (Integrated Development Environment), thru a A/B USB cable. It can be powered via way of means of the USB cable or via way of means of an outside 9-volt battery, although it accepts voltage range from 7 to 20 volts.

The word "UNO" means "one" in Italian and became selected to mark the preliminary launch of Arduino Software.

2) ESP8266 : A WiFi Module for IoT

ESP8266 is a Wi-Fi supported with SoC (System on Chip) board established by Espressif system. It is popularly used for the development of IoT embedded application. ESP8266 Wi-Fi module is an economical wireless transceiver that is used for end-point IoT development. For the communication of ESP8266 WiFi module, Arduino UNO needs to use a set of AT commands. Arduino UNO communicates with ESP8266 WiFi module using UART having specified Baud rate. It has 2.4 GHz WiFi (supporting WPA/WPA2). It has 16 General Purpose Input/Output. It has 10-bit ADC.

Its Pin configuration is as follows:

3V3- 3.3 Volt Power Pin
EN- Enable Pin
TX- Transmit Pin of UART
GND- Ground Pin
RST - Reset Pin
RX- Receive Pin of UART

***Fig 3**: ESP8266 WiFi Module*

3) DHT11 : A temperature and humidity sensor

DHT11 is a Temperature and Humidity Sensor. This sensor has 4 pins out of which 3 pins are used. It comes with a NTC (Negative Temperature Coefficient) for measuring temperature. It is already factory calibrated using a master; therefore, it is also easy to interface with Arduino UNO. Its output is in Serial Data form. It can measure temperature from 0°C to 50°C with an accuracy of ±1 ° C and humidity from 20% to 90% with an accuracy of ±1%. Resolution of Temperature and Humidity Output is 16-bit. It has an Operating voltage of 3.5V to 5.5V. It has an Operating Current of 0.3mA while measuring and 60 µA while it's in standby state.

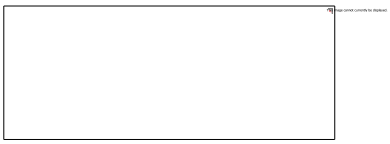


***Fig 4**: DHT11 (Temp & Humi)*

4) ADXL345 : Accelerometer Sensor

The capability of this sensor system is to analyze the property of the bridge using the data's collected from the accelerators which are fixed on bridge piers. Accelerometer is a widely used device for application such as wrecking of bridges and damage of bridges for bridge monitoring. Health monitoring systems generally include information acquisition, analysis, diagnosis, evaluation, feedback and safety. It enables the timely realization of the function of data acquisition. The most vibration will happen along the pillar which we considered as Z axis of the accelerometer.

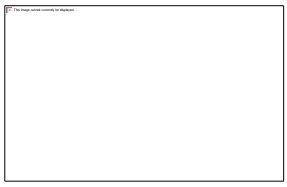


***Fig 5**: ADXL345 Accelerometer*

5) HC-SR04 : Ultrasonic level detection sensor

To perform the Structural health monitor of a structure, we have used sensors such as Ultrasonic sensor (HC-SR04) to detect the distance between water surface and bridge. If the water level crosses the threshold, it will generate the alarm to respective authority.

***Fig 6**: HC-SR04 Ultrasonic Sensor*

Sensors	Sensor Characteristics						
	Model	Type	Range	Accuracy	Current	Voltage	Frequency
Humidity	DHT11	Capacitive	20% - 90%RH	±5%RH	0.3mA	5 V	1 Hz
Temperature	DHT11	Thermistor	0°C to 50°C	±1°C and ±1%	0.3mA	5 V	1 Hz
Ultrasonics	HC-SR04	Non-contact	2 cm to 400 cm	3mm	15mA	5 V	50Hz
Accelerometer	ADXL345	3-axis acceleration measurement system	−55°C to +105°C	+/-16g	40uA	3.3 V	100 Hz
WIFI Module	ESP2866	WIFI Module	366 meters with the PCB antenna	10-bit precision SAR ADC	100mA	3.3V	80MHz

***Table 2:** Sensor characteristics*

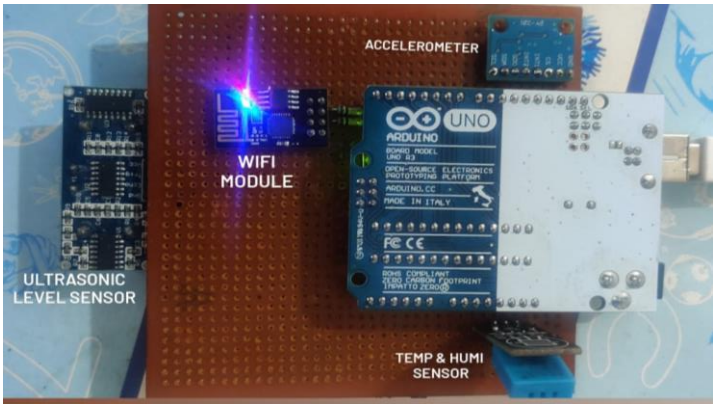


***Fig 7:** Hardware Setup*

B. Software Specifications

1. Arduino IDE
   Arduino Integrated Development Environment has been used to write program and upload to Arduino Board, integrate various sensors and WiFi module and send data to cloud for further analysis.

2. MATLAB R2021a
   MATLAB R2021 has been used to read data from ThingSpeak and data analysis has been carried out using inbuilt MATLAB functions, libraries. Also,
   The analyzed data has been represented using graphical representation in MATLAB. presently ThingSpeak is the only IoT web service that offers the data analysis on the MATLAB platform as open source with full profile access.

3. PROTEUS
Proteus is a Virtual System Modelling and circuit simulation application. Schematic circuit has been designed with the help of proteus. Schematic capture contains all the electronic design tools one's need.

Steps to create schematic diagram: -

- Open Proteus ISIS Schematic Capture
- Select the Component Mode from the left Toolbar
- Click on P (Pick from Libraries)
- Add all the required components
- Place the components on the workspace
- Wire up the circuit
- Click on Play Button on the bottom left to start simulation

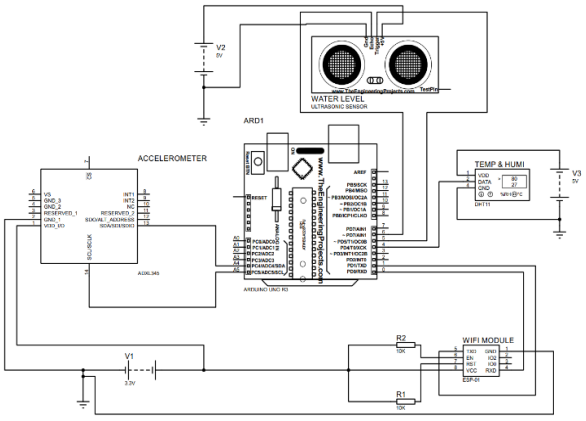
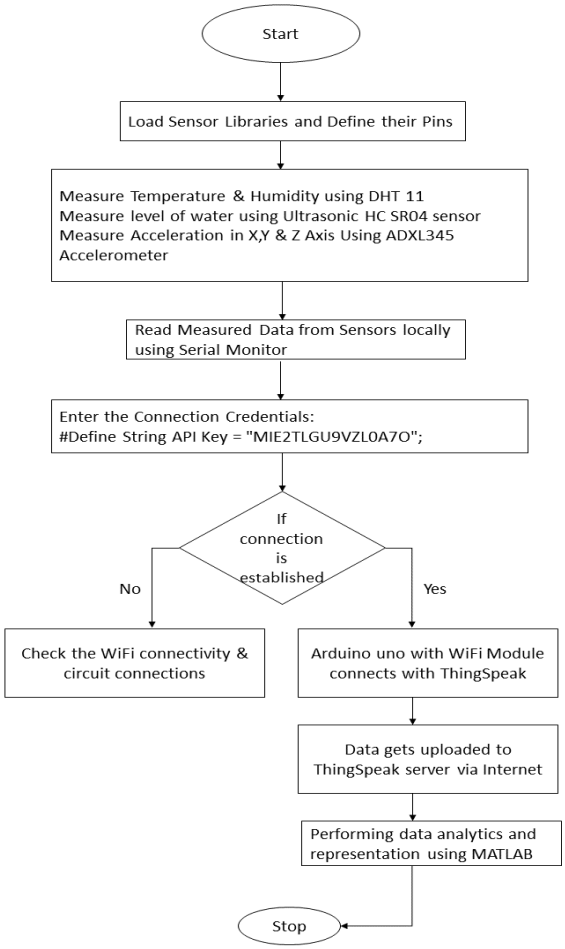
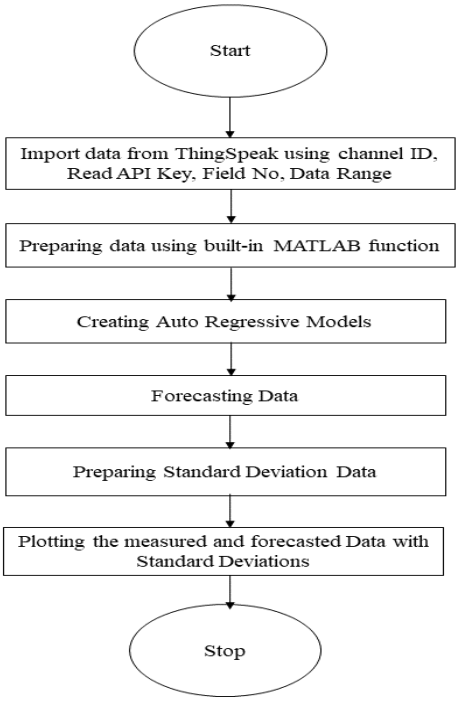


***Fig 8**: Schematic Circuit Design*

C. Process Flowchart



***Flowchart 1**: ThingSpeak Based Sensing-Monitoring System process Flowchart for IoT*



***Flowchart 2**: MATLAB process flow chart for Forecasting Data*

D. ThingSpeak Breakthrough: -

1- ThingTweet Interface

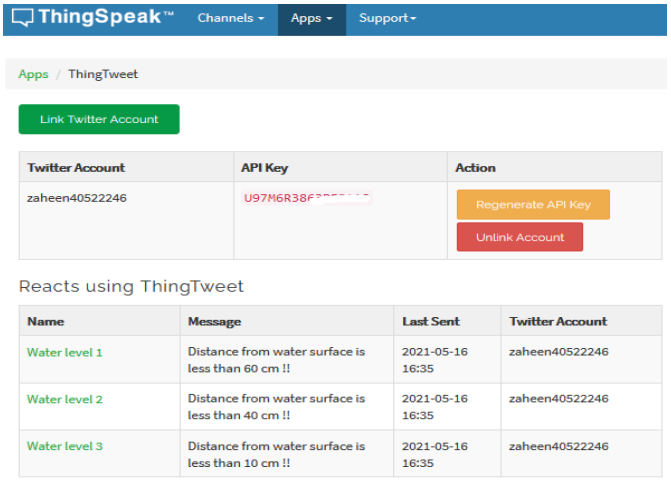


***Fig 9:** ThingTweet*

2- Steps to create React on ThingSpeak

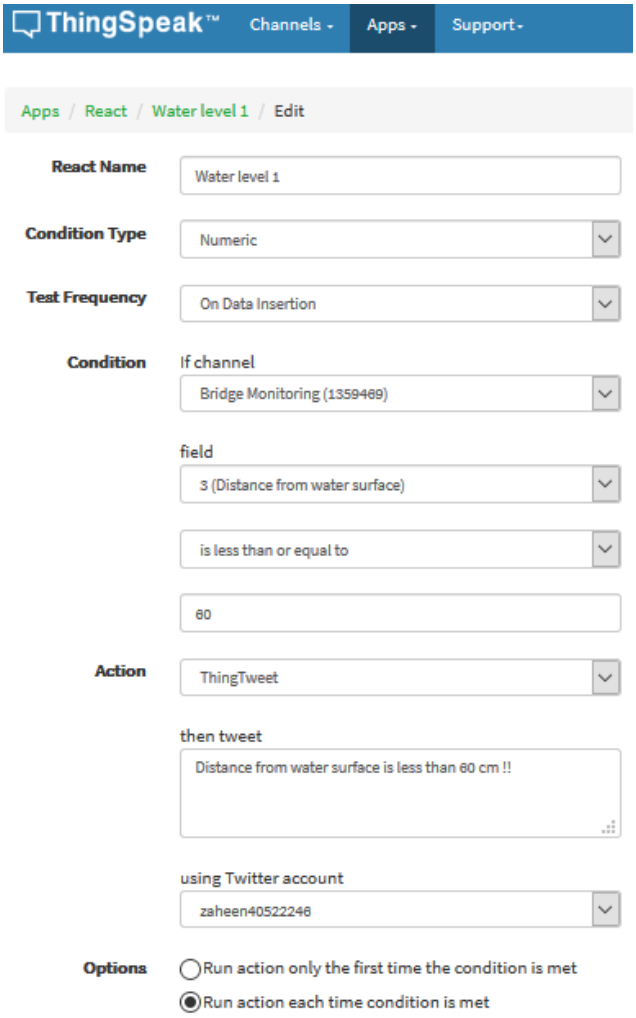


***Fig 10:** React on ThingSpeak*

3- Steps to create Time Control on ThingSpeak

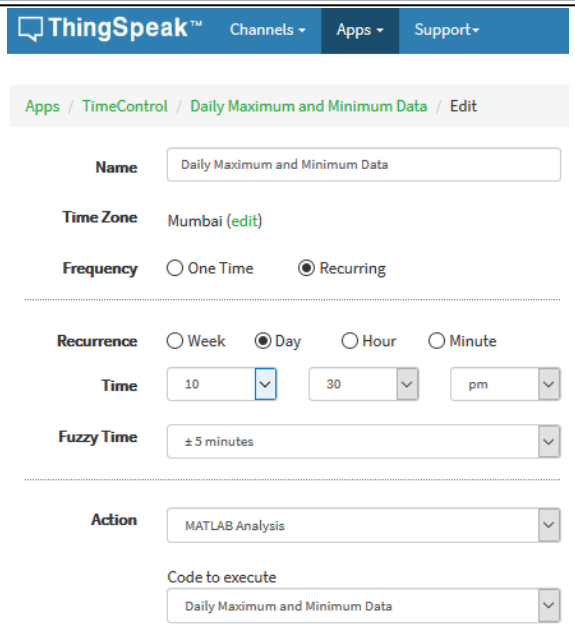


Fig 11: *Setup for Time control on ThingSpeak*

IV. RESULT

1- ThingSpeak Interface

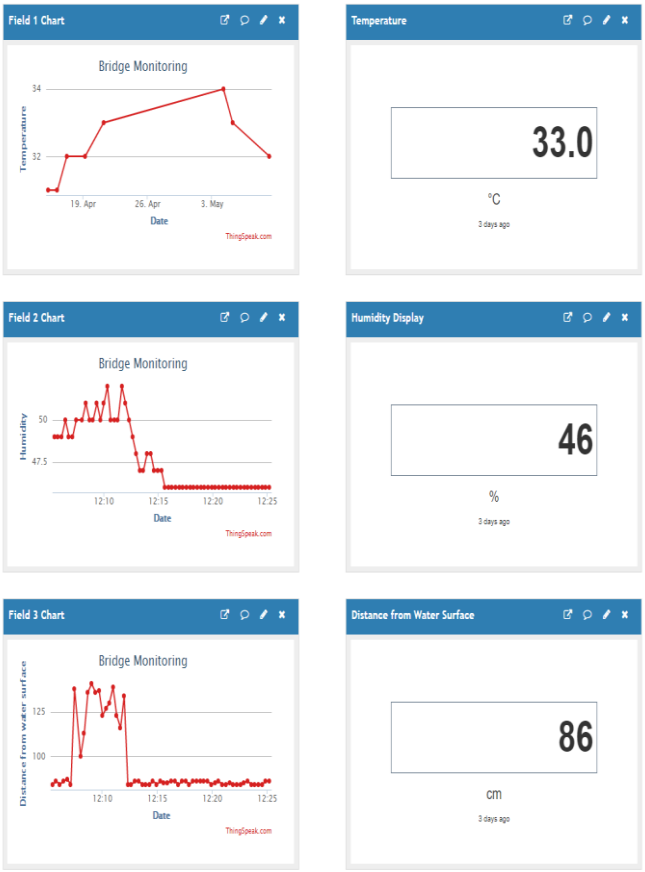


Fig 12*: Sensor Data on ThingSpeak Cloud*

***Fig 13**: Sensor Data on ThingSpeak Cloud*

2- MATLAB Result

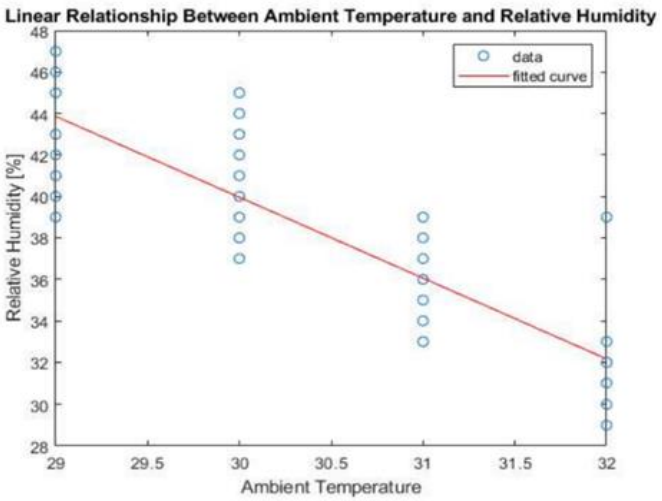


***Fig 14:** Linear relation between Temperature and Humidity*

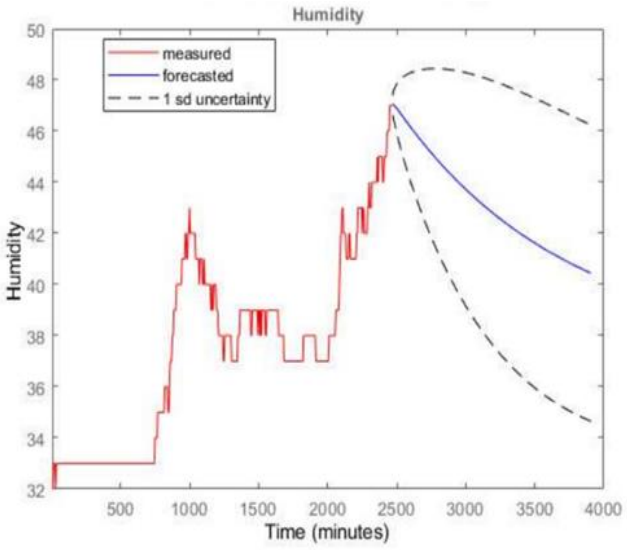


***Fig 15**: Measured and Forecasted Humidity Graph*

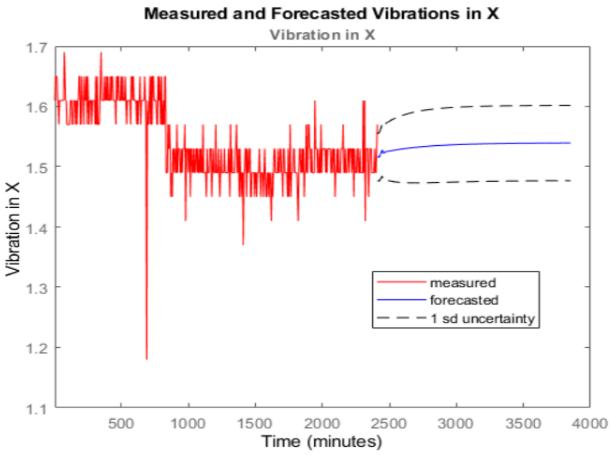


***Fig 16**: Vibrations in X*

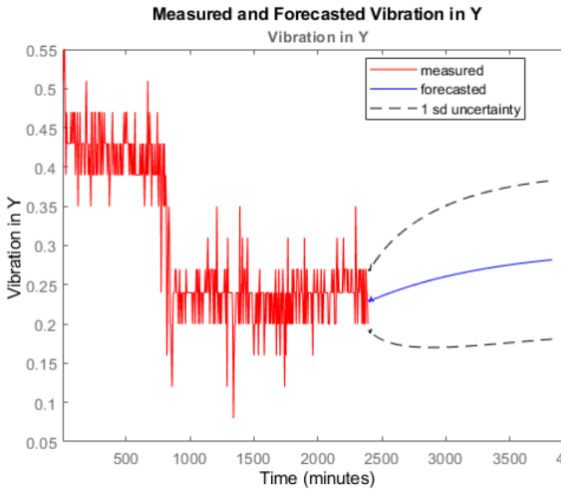


***Fig 17**: Vibrations in Y*

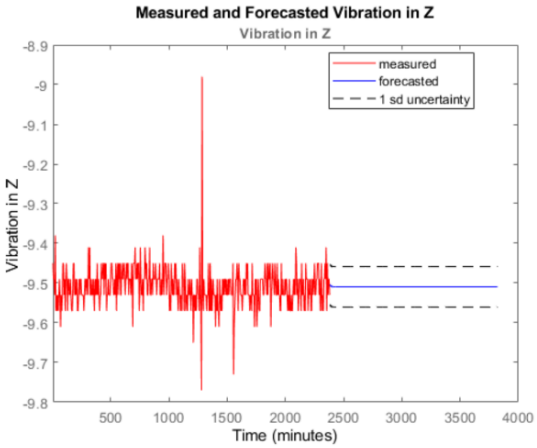


***Fig 18**: Vibrations in Z*

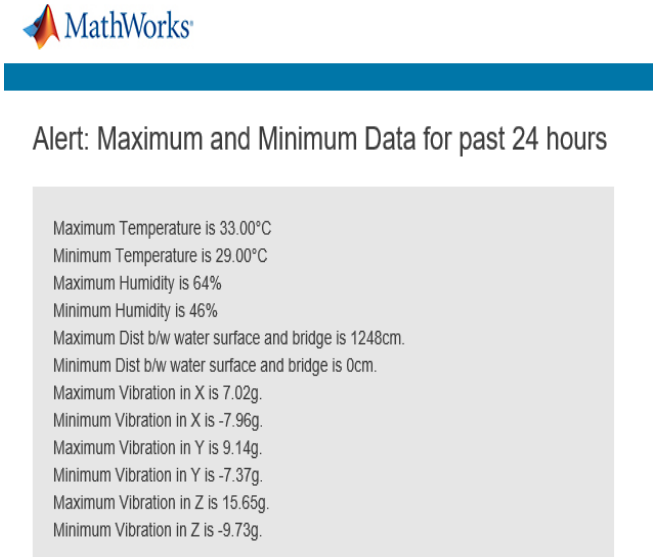


***Fig 19**: Output of MIN & MAX of sensor data*

3- Thing Tweet Output

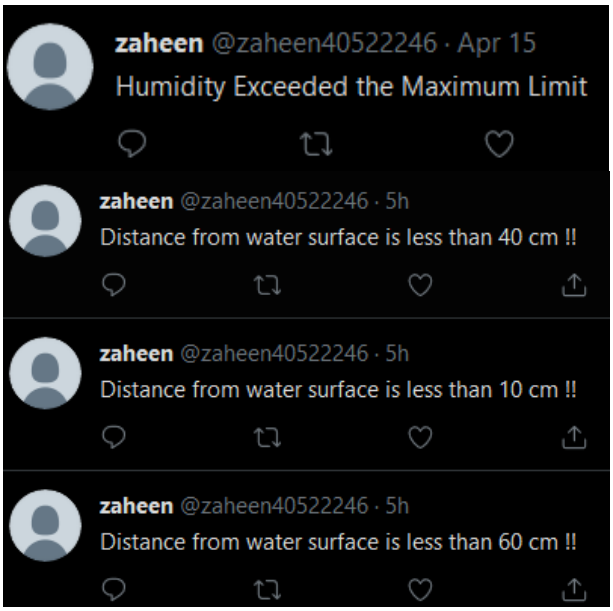


Fig 20: Output on Twitter Using ThingTweet and React

V. Conclusion

The IoT has the potential to dramatically increase the availability of information, and is likely to transform companies and organizations in virtually every industry around the world.

MATLAB is a great tool for data visualization and analysis IOT along with MATLAB gives extreme flexibility to work on different data and do analysis as per the requirement.
The results obtained from using ThingSpeak and MATLAB shows the analysis of sensor data is easy to handle and user friendly. We have analyzed sensor data for: -

A) forecasting temperature and humidity.
B) forecasting vibration in X, Y, Z
C) generating alarm using ThingTweet and React.
D) finding Max and Min of all sensor Data every 24 hours and generating email.

VI. REFERENCE

[1] IOT Based Bridge Health Monitoring System, Journal of Emerging Technologies and Innovative Research (JETIR), © 2018 JETIR October 2018, Volume 5, Issue 10
[2] Design of Bridge Monitoring System based on IoT, MVP Journal of Engineering Sciences, Vol 1(1), DOI: 10.18311/mvpjes/2018/v1i1/18258, June 2018
[3] Bridge Monitoring System Using IOT, Journal of Advances in Electrical Devices, Volume 3, Issue 2, © MAT Journals 2018.
[4] Bridge Safety Monitoring System using IOT, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume 9 Issue-6, April 2020.
[5] Low-Cost Internet of Things Platform for Structural Health Monitoring, Institute of Electrical and Electronics Engineers, 978-1-7281-4001-8/19/$31.00 ©2019 IEEE
[6] ThingSpeak Based Sensing and Monitoring System for IoT with MATLAB Analysis, International Journal of New Technology and Research (IJNTR), ISSN: 2454-4116, Volume-2, Issue-6, June 2016
[7]https://paradisetronic.com/media/image/org/ESP8266.jpg
[8]https://image.madeinchina.com/2F0j00ylEfKvtsgpku/Arduino-Uno-R3-Development-Borad-Microcontroller-for-DIY-Project.jpg
[9]https://i0.wp.com/themachineshop.uk/wp_content/uploads/2021/04/DHT11-2-750x750-1.jpg?fit=750%2C750&ssl=1
[10]https://5.imimg.com/data5/BS/PN/MY-66278010/hcsr04- ultrasonic-sensor-module-500x500.jpg
[11] https://www.labcenter.com/schematic/
[12]https://www.researchgate.net/publication/330134632_Real_Time_Weather_Analysis_Using_ThingSpeak
[13]https://indianexpress.com/article/cities/rajkot/gujarat-seven-injured-as-bridge-collapses-on-junagadh-sasan-highway-6058192/

AUTHORS PROFILE

Zaheen Shaikh, Student,
B.Tech., Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, Maharashtra, India

Snehil Singh, Student,
B.Tech., Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, Maharashtra, India

Mohammed Zaid Nidgundi, Student,
B.Tech., Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, Maharashtra, India

Performance Assessment in Precision Agriculture Using Decision Tree Approach

Shikha Ujjainia[1], Pratima Gautam[2], S. Veenadhari[3]
[1]Computer Science and Application, Rabindranath Tagore University, Bhopal, India
E-mail: shikhaujjainia90@gmail.com
[2]Dean (CSIT), Rabindranath Tagore University, Bhopal, India.
E-mail: pratima_shkl@yahoo.com
[3]Associate Professor (CSE), Rabindranath Tagore University, Bhopal, India.
E-mail: veenadhari1@gmail.com

ABSTRACT

According to statista's research analytics, in 2018 the gross domestic product (GDP) in India was around $ 2.72 trillion. This was about 6.81 percent over the previous year. Agriculture is playing a major role in the Indian economy. Rice is one of the major food sources of the modern world. Its production requires good environmental conditions, which are not always available. However, farming methods can mitigate the effects of adverse weather or poor quality soil. The impact of environmental and crop management variables on yield can only be assessed based on representative long-term data collected on farms. The most important crop management deficiencies that may cause low yields are insufficient pesticide use, non-optimal date of sowing, poor quality of seeds, etc. Environmental variables such as temperature, rainfall, groundwater level, as well as other variables such as soil pH, area have great importance for achieving high yields. Nowadays, machine learning algorithms are used to increase performance according to their data in almost every field. Machine learning is gradually spreading its wings in the agricultural sector as well. Through this we can improve food quality and yield production. This paper investigates the potential of agricultural dataset for providing better crop management by using machine learning algorithm. In this paper, the ability of the Classification and Regression Tree (CART) algorithm is examine on the different environmental as well as crop parameters like temperature, rainfall, area, soil ph etc. The CART decision tree algorithm was able to predict the yield with 85% of accuracy.

Key terms - Machine Learning, Decision Tree, Classification and Regression Tree, Precision Agriculture.

1. INTRODUCTION

Smart farming is an emerging idea that refers to dealing with farms using current information and communication technology to optimize the required human labor as well as increase the quantity and penalty of products. Predictive analytics is very successful in increasing the productivity of the agricultural system and increasing the efficiency of crop production. However, the population grows continuously, while the resources of crop production are reduced day by day. Traditionally agriculture involves sowing or harvesting the crop against a predetermined schedule. Accurate agriculture involves collecting real-time data on weather, air quality, soil, crop maturity, equipment, labor costs, and the availability of existing data [1]. This analysis will also provide benefits to the farmer and will help in increasing production. Selecting the right crop variety, the exact type and dosage of fertilizers, pesticides and herbicides, and

proper irrigation meet the demands of crops for optimal growth and development [2].

The experience of farmers in the agricultural sector is involved in crop prediction. Due to lack of reliable and timely information, farmers who were in rural areas are farming according to their personal experience and knowledge. Now-a-days the weather conditions are not the same as in previous decades. Due to day-to-day globalization, farmers have faced difficulties in assessing weather conditions. Due to prediction, better decisions are taken at the right time. In addition, agricultural tools help farmers analyze and produce accurate information to make appropriate decisions in land preparation, seed technologies, fertilizers, pesticides and herbicide applications, irrigation and drainage, and post-production activities [3].

Many experts are implementing automated farming. Since the decision tree is a well-known algorithm, which is used for prediction of supervised learning algorithm. In [4], Classification and Regression Tree (CART) algorithm is used to explain the effect of environmental and crop management variables on wheat yield based on a large and complex database. Dataset consist soil/weather and crop dataset to make yield prediction. In [5], it is shows that there are various classification methods such as Naïve Bayes, Random Forest, Artificial Neural Network, Decision Tree and Support Vector Machine (SVM) to solve the yield prediction problem. The paper describes that predicting agricultural inputs leads to improved yields and improved yields leads to improved agricultural efficiency. In [6], author analyzed various environmental and soil factors such as pH, soil salinity to determine production of crops in Bangladesh. Based on these factors they clustered the data set and then implemented Id3 and cart classification algorithms to determine the predictions for crop yield.

2. PREDICTIVE MODEL

Predictive analytics encompasses various technologies from machine learning, data mining techniques that derive from various historical information and currents factors to make a smart decision about future events [7]. This model incorporates patterns different from historical facts to analyze future predictions. It is a type of data mining technique that is derived from the data of the previous year and predicts behavior using the history of the data [8]. This involves analyzing what happened in the past and monitoring facts about current terminology and data, then making a better decision using machine learning methods. In prediction, the previous year's data is referred to as the training set and the data are classified based on the training set [9]. The machine learning model is divided into three main categories i.e. supervised learning, unsupervised learning and reinforcement learning.

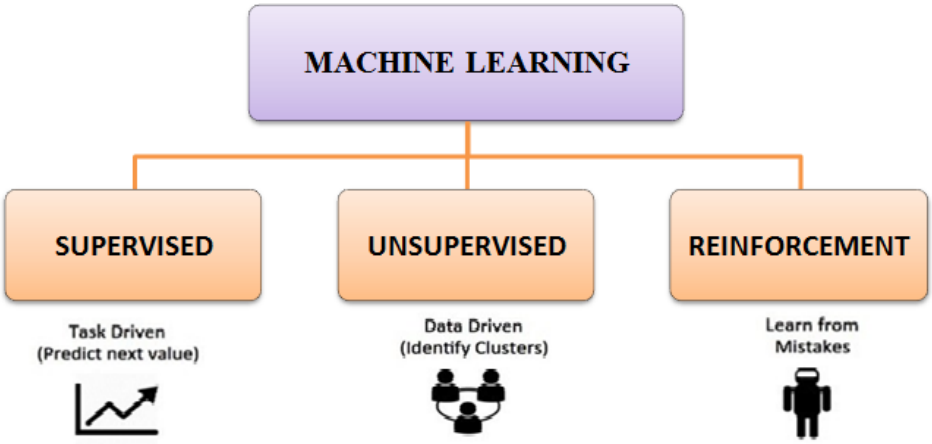


Fig. 1 Types of Machine Learning

Supervised learning is the most widely used learning algorithm for future predictions. It is trained on label data. The given dataset is further subdivided into train and test data, on which the applied algorithm is to make predictions by finding the relationship between the variables in the given dataset. Once a model is trained it can begin to predict or make decisions when given new data. The Unsupervised learning model learns through observation and finds structures in the data. Once a model is given a dataset, it automatically discovers patterns and relationships in the dataset by creating a cluster in it. Reinforcement learning improves on its own and learns from new situations using a trial-and-error method. Favorable output is encouraged or 'reinforced', and non-favorable output is discouraged or 'punished'.

3. METHODOLOGY

Machine learning is an area of research that focuses precisely on performance, and properties — based on algorithms and learning systems. Due to its implementation in a wide range of applications, machine learning has covered almost every scientific domain, which has significantly impacted science and society [10]. Machine learning approaches are applied to predict real-time and historical big data. In our work, we are using a supervised learning algorithm to predict crop yield with favorable weather conditions. This type of algorithm helps to create the most effective and accurate model because the learning data comes with a label or preferred output and is intended to find a general rule of mapping input to the output. It involves creating a machine learning model that is based on label samples. The prediction procedure of the decision tree model is depicted in the following figure:

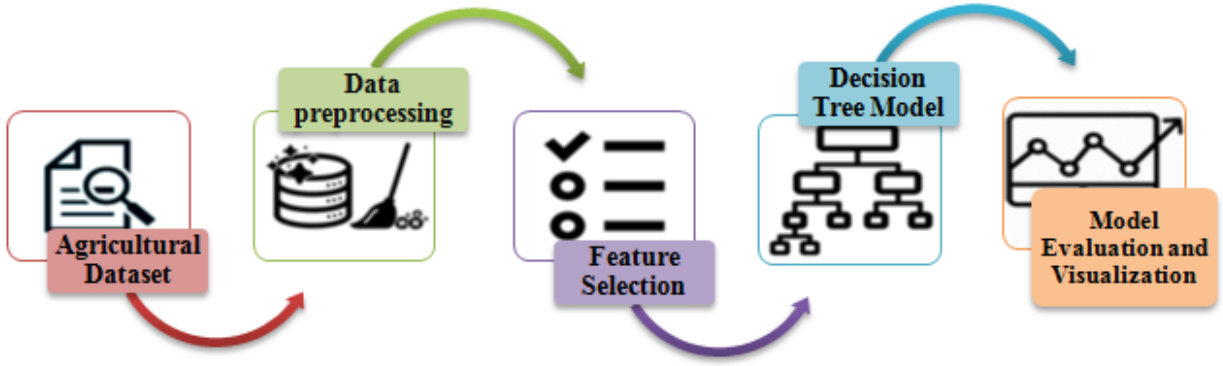


Fig. 2 Steps in machine learning

3.1 DATASET

The data is collected from various sources and prepared for the data set. And this data is used for descriptive analysis. The data is available from many online abstract sources such as data.gov.in and kaggle.com. We will use the 10-year rice crop data. The data sets used in this paper are crop yield data, temperature dataset, and rainfall dataset. Parameters used for our study are as follows:

- Area (In Hectare)
- Temperature (Degree Celsius)
- Rainfall (mm)
- Groundwater level (m)
- Soil Ph
- Potassium (kg/Hectare)
- Magnesium (kg/Hectare)
- Sodium (kg/Hectare)

3.2. DATA PREPROCESSING

Data preprocessing is an essential step in machine learning as data quality and the useful information derived from it directly affect the learning ability of our model. Therefore, we have to preprocess our data before feeding into our model. This entire process can be automated using machine learning algorithms, mathematical modeling and statistical knowledge. It can be the output of this entire process in any desired form, such as graphs, videos, charts, tables, pictures and many more, depending on what we are doing and the requirements of the machine. After collecting the data, we want the valuable information to be accurate. With some business criteria, we classify the data into 2 groups - training data and test data with 70% and 30% respectively.

3.3. FEATURE SELECTION

Feature selection is the process of selecting a subset of attributes that have a significant or equal impact on the evaluation target using all features [11]. This assumes that the

data have irrelevant or/and redundant features that may impair the performance of the model. Feature selection was initially proposed to increase the accuracy of induced classifiers in a supervised learning algorithm [12]. Here, we used a correlation method for feature selection. It selects the features which are highly correlated with the target values. This means that parameters with high correlation coefficient value are considered an important feature for predicting crop yields.

3.4. DECISION TREE REGRESSOR MODLE

The decision tree is one of the successful forms of supervised machine learning. It is a flowchart like a tree where each branch node indicates different attributes and each leaf node indicates a decision on the input set. Decision tree learning uses a decision tree as a predictive model to overview findings about the target value of an item. It is one of the predictive modeling approaches used in data mining, statistics, and machine learning. Decisions can be made relatively fast when compared to other methods of classifications. The SQL statement can be used to access a database that the tree has been created efficiently. Decision tree models achieve the same or better accuracy when compared with other classification methods.

Decision trees are constructed through an algorithmic approach that identifies ways to segment data sets based on different conditions. The classification and regression tree (CART) presented by Leo Breiman is a type of decision tree, which can be used for classification as well as regression prediction modeling problems. The decision tree which has a continuous target variable then it is called the regression tree while the decision tree that has categorical target variable is known as the classification tree. The classification and regression tree (CART) is a common term for this.

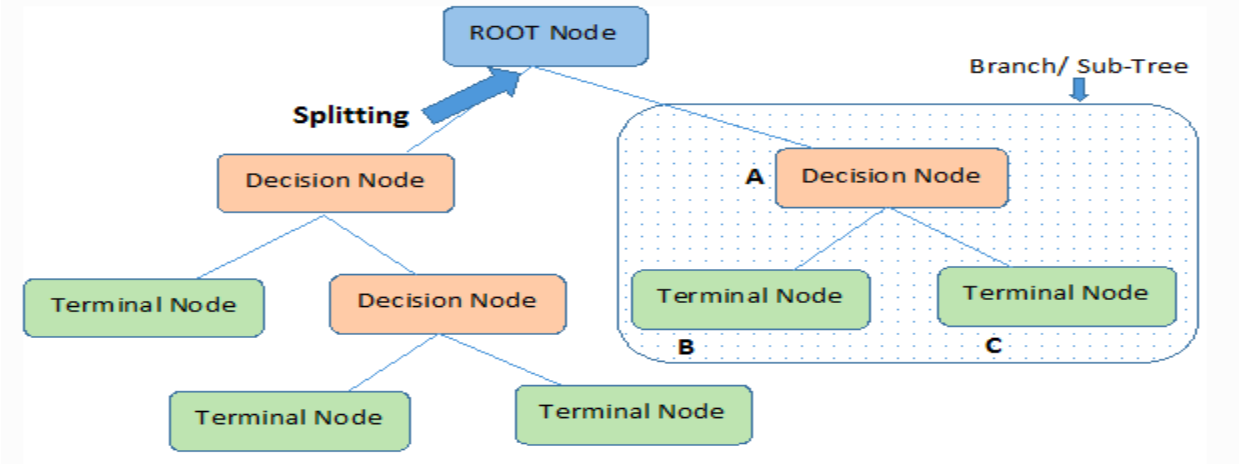


Fig. 3 Representation of decision tree

Classification and regression decision tree is used in our study. Decision trees are built by recursively splitting our training samples using the features from the data that work best for the specific task. The decision tree clarifies all possible alternatives and locates each alternative for its conclusion in a single view, to make an easy comparison between different alternatives [13]. This is done by evaluating certain metrics, like the gini index or the entropy for categorical decision trees, or the residual or mean squared error for regression trees. The process is also different if the feature that we are evaluating at the node is discrete or continuous. For categorical features, all possible values are evaluated, after that features are selected through the information gain and entropy method. These criteria will calculate values for each attribute. After that, the values of the features are sorted and placed in a tree in order to mean the attribute with the highest information gain is placed at the root. Entropy measures the randomness of the information being processed means higher the entropy, the harder it is to draw any conclusions from that information. For continuous features, for a certain node, mean squared error (MSE) values are calculated in the list of features. Then we pick the feature that gives the lowest MSE value that we are using for the resulting child node in the decision tree.

The tree will show the prediction of maximum production with the combinations of favorable weather conditions. Since we are evaluating continuous variable that is crop production, we will use regression decision tree. Features like temperature, rainfall, area, groundwater level, soil ph,

potassium, magnesium, and sodium are given as input to the model. To decide to split the node in two sub nodes we calculate the mean square value by given formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2 \quad \textbf{(1)}$$

Where,

y_i = Actual values

$\overline{y_i}$ = Predicted values

n = Number of data points in that feature

Above steps are helpful to create root node of decision tree among all the selected features from the dataset. Mean square error value is calculated for each selected features and minimum mean square value feature will be selected for the root node and this process will continue till the leaf node.

4. MODEL EVALUATION AND VISUALIZATION

Here are some results depicting the relationships of input attributes such as different weather parameters and crop conditions over the output that is crop production.

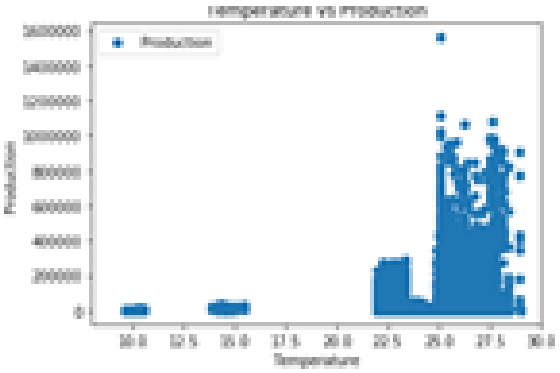


Fig. 4 Effect of temperature on yield production

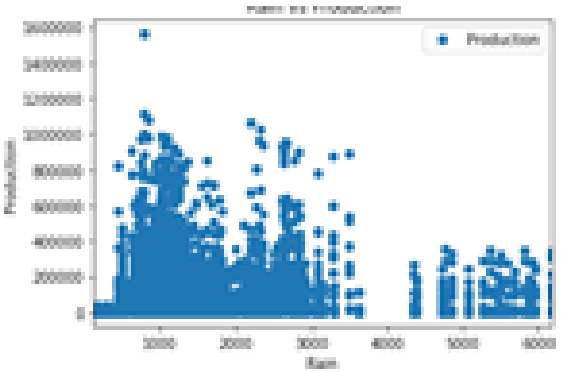


Fig. 5 Effect of rainfall on yield production

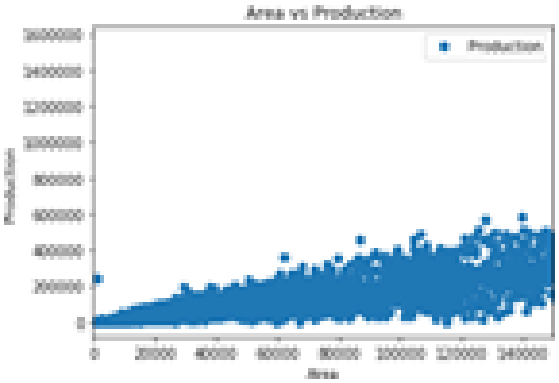


Fig. 6 Effect of area on yield production

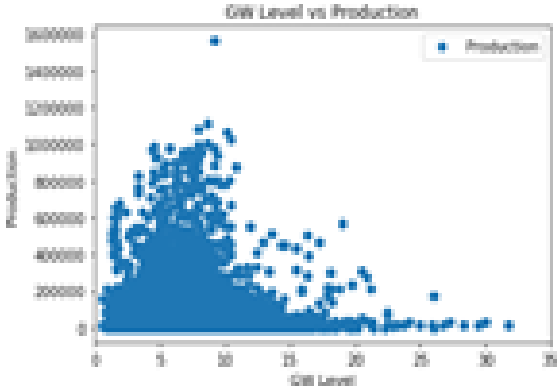


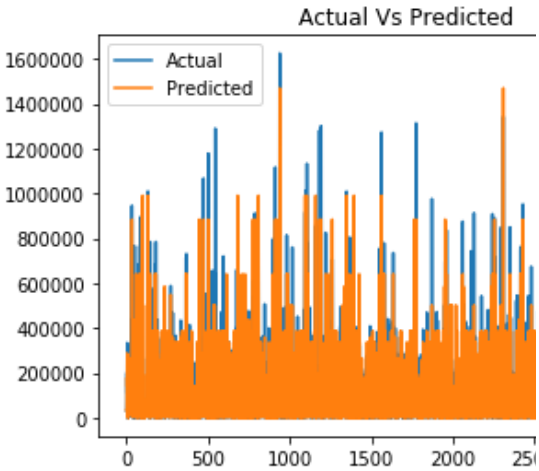
Fig. 7 Effect of ground-water level on yield production

The above analysis shows the suitable conditions for rice crop production. As we can see in the graph when the temperature and area are increasing then the production of the rice crop is also increase. Whereas rainfall and ground-water level show the appropriate ranges of water for rice yield production. Once the model is trained efficiently it is tested on a test dataset that differs from the training data in sample values.

5. RESULT

In this paper, we have used jupyter notebook as an open-source, multi-platform integrated development (IDE) environment for python scientific programming. Model is evaluated through R^2 accuracy matrix which used to estimate the efficiency of the decision tree model. This model gives 0.85 (i.e. 85%)

accuracy. The following figure shows the effectiveness of the model that how much it is correctly predicting the actual data.



.

Fig. 8 Performance of model for predicting crop yield

6. CONCLUSION

Various machine learning algorithms have been applied to agricultural data to evaluate the best performing method. This paper successfully demonstrated the use of a supervised learning model, which is a decision tree on a dataset containing various parameters related to obtaining expected rice yields. Agriculture is playing a major role in the Indian economy. Nowadays, machine learning algorithms are used to increase performance according to their data in almost every field. As we know that food is the basis of lifestyle and everyone needs it too. Machine learning is gradually spreading its wings in the agricultural sector as well. Through this we can improve food quality and yield production. One of the best benefits of a decision tree is it is transparent in nature. The dataset is divided into two parts, which are training datasets and testing datasets with a 70:30 ratio, respectively. The decision tree model gives 85% accuracy with a low error value. Thus, this model achieves the best possible results for predicting crop yield.

REFERENCES

[1] John V. Stafford , Silsoe Solutions, Ampthill, "Implementing Precision Agriculture in the 21st Century", at Journal of Agriculture Engineering, Vol. 76(3), Pages: 267-275, 2000.

[2] Samir Kumar Sarangi and Dr. VivekJaglan, "Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques", at International Journal of Emerging Technologies in Computational and Applied Sciences, ISSN:2279-0047, 2013.

[3] S. Kumudini, F. H. Andrade, K. J. Boote, G. A. Brown, K. A. Dzotsi, G. O. Edmeades, T. Gocken, "Predicting Maize Phenology: Intercomparison of Functions for Developmental Response to Temperature", at Agronomy

Journal, Vol. 106(6), Pages: 2087-2097, 2014.

[4] M. Iwańska, A. Oleksy, M. Dacko, B. Skowera, T. Oleksiak, E. Wójcik-Gront, "Use of classification and regression trees (CART) for analyzing determinants of winter wheat yield variation among fields in Poland", Biometrical Letters, Vol. 55(2), Pages: 197-214, 2018.

[5] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques", International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, Pages: 1-4, 2016.

[6] Surya A. Venkaiah, Kondajji Swati Sunitha, "A new approach for predicting crop yield prediction using data mining techniques", in International Journal of Engineering, IT and Scientific research, Vol. 3(1), 2019.

[7] Moor, H., Hylander, K., & Norberg, J., "Predicting climate change effects on wetland ecosystem services using species distribution modeling and plant functional traits", at AMBIO, Vol. 44(S1), Pages: 113-126, 2015.

[8] P. Kumar, A. Sarangi, D. K. Singh, S. S. Panhar, "Simulation of salt dynamics in the root zone and yield of wheat crop under irrigated saline regimes using SWAP model", at Agricultural Water Management, Elsevier, Vol. 148, Pages: 72-83, 2015.

[9] Deepak K. Ray, James S. Gerber, Graham K. MacDonald & Paul C. West, "Climate variation explains a third of global crop yield variability", at Nature Communications, Vol. 6(1), Pages: 1-9, 2015.

[10] C. Rudin, K.L. Wagstaff, "Machine learning for science and society", at Machine Learning, Vol. 95(1), Pages: 1-9, 2014.

[11] C.J. Hsu, C.Y. Huang, "Comparison of weighted grey relational analysis for software effort estimation", Software Quality Journal, Vol. 19(1), Pages: 165-200, 2011.

[12] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection", at International Conference on Machine Learning (ICML), Vol. 1, Pages: 74–81, 2001.

[13] Anyanwu MN, Shiva SG., "Comparative analysis of serial decision tree classification algorithms", at International Journal of Computer Science and Security, Vol: 3(3), Pages: 230-40, 2009

IFERP International Conference
IFERP Explore
https://icdsmla.net/ | info@icdsmla.net

UPCOMING CONFERENCES

Technoarete® Group
Integrating Researchers to Incubate Innovation

SUPPORTED BY

ALL CONFERENCE ALERT
Academics conferences at a glance

ECA
Get World wide Engineering Conferences Update..

Conference Alerts
Global Academic Conferences

INTERNATIONAL CONFERENCE ALERTS