

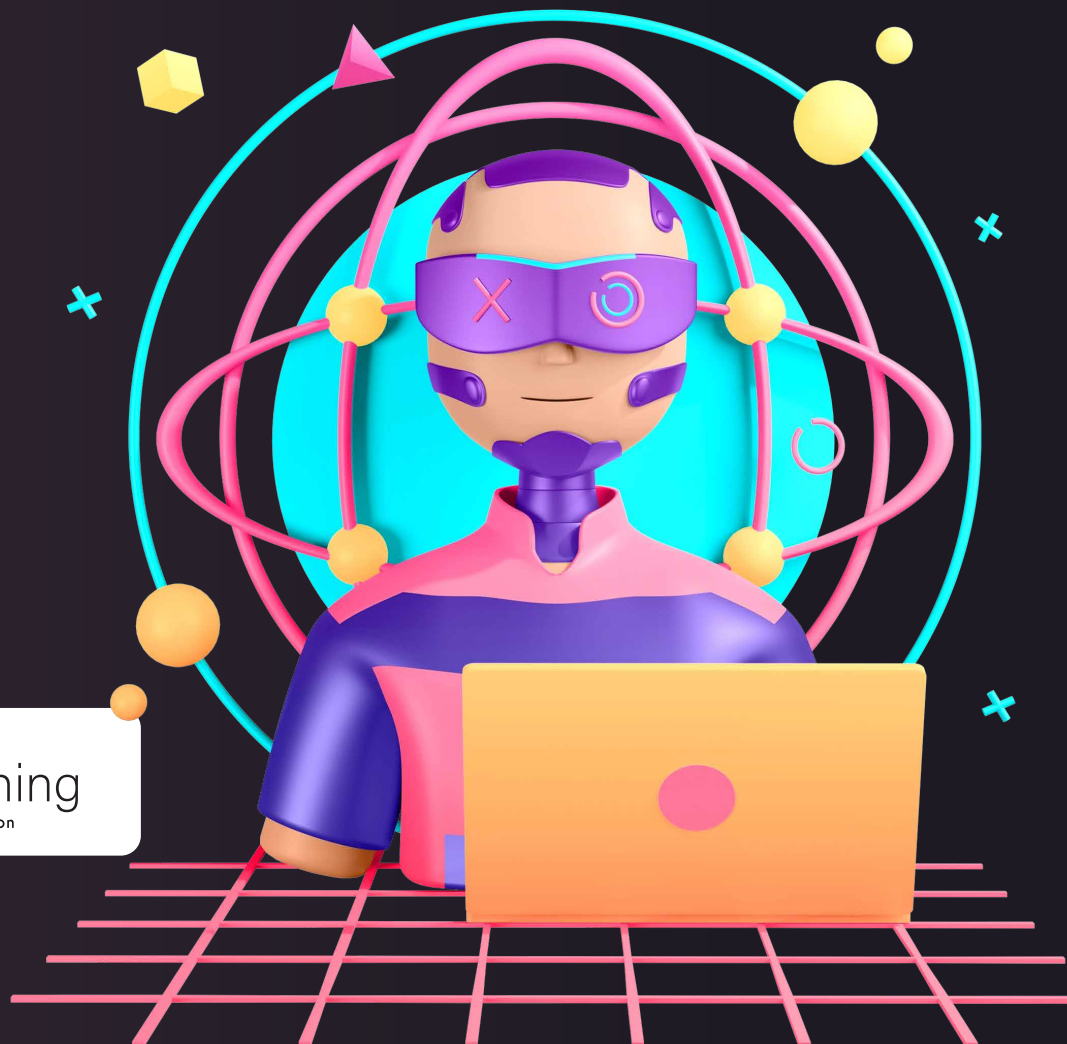
# Machine Learning Algorithms for Intelligent Data Analytics

---

Dr. S. Balamurugan

CEng. Radhey Shyam Meena

Dr. Ramasamy V



**Echnoarete**<sup>®</sup> Publishing  
Integrating Researchers to Incubate Innovation

**Dr.S.Balamurugan,  
CEng. Radhey Shyam Meena  
Dr.Ramasamy V  
Editors**

# **MACHINE LEARNING ALGORITHMS FOR INTELLIGENT DATA ANALYTICS**

# Editors

**Dr S. Balamurugan, SMIEEE,**

ACM Distinguished Speaker,

Director- Albert Einstein Engineering & Research Labs (AEER Labs), India

Vice Chairman- Renewable Energy Society of India

**CEng.Radhey Shyam Meena**

Ministry of New & Renewable Energy (MNRE)

Government of India, New Delhi, India

**Dr.Ramasamy V, PhD,**

Associate Professor, Department of CSE,

Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology (Deemed to be University),

Chennai, Tamil Nadu, India.

**Machine Learning Algorithms for Intelligent Data Analytics**

**ISBN : 978-93-92104-07-7**

DoI : <https://dx.doi.org/10.36647/AAIMLH/2022.01.Book1>

Published on December, 2022

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Technoarete Publishers.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. This Technoarete imprint is published by Technoarete Publishers registered under the company Technoarete Research and Development Association. The registered company address is: Rais Towers, 2054/B, 2nd Floor, West block, 2nd Ave, Anna Nagar, Chennai 600040.

*We are honoured to dedicate the Machine Learning Algorithms for Intelligent Data Analytics book to all the authors, contributors, reviewers, and editors.*



# Preface

A huge amount of potential is vested in Machine Learning and Data Analytics. This book aims to cover a wide range of applications of Machine Learning Algorithms for Intelligent Data Analytic techniques. Machine learning analytics is an entirely different process. Machine learning automates the entire data analysis workflow to provide deeper, faster, and more comprehensive insights. Machine learning automates the creation of analytical models. It is a subfield of artificial intelligence founded on the notion that machines are capable of learning from data, spotting patterns, and making judgments with little assistance from humans. Machine learning is the study of algorithms that get more efficient over time. There is a role for artificial intelligence. Without the need for human intervention, a machine learning system automatically learns from data and applies the learning.

The process of cleaning, examining, modelling, and changing data for the purpose of discovering insightful information, guiding conclusions, and boosting the decision-making process is known as data analytics. Drawing intelligent inferences from the given data is the goal of data analytics. Businesses use data analytics to help them make better decisions about a range of topics, including marketing, production, etc. Data analytics can be used to extract usable information from raw data.

We would like to offer our sincere thanks to all the authors for their timely support and for considering this book for publishing their quality work. We also thank all reviewers for their kind cooperation extended during the various stages of processing the manuscript. Finally, we would like to thank Technoarete publications for producing this volume.

S.Balamurugan  
Radhey Shyam Meena  
Ramasamy V  
December 2022

# Acknowledgements

The editors of the Machine Learning Algorithms for Intelligent Data Analytics book would like to thank all the authors, co-authors, and contributors for their timely support and for considering this book for publishing their quality work.

We would like to express our gratitude to all the reviewers for their kind cooperation extended during the various stages of processing the manuscript. Finally, we would like to thank Technoarete publications for producing this volume.

## About the Editors

**Dr.S.Balamurugan Ph.D., SMIEEE**, ACM Distinguished Speaker is the Director of Albert Einstein Engineering and Research Labs. India. He received his B.Tech., Degree from PSG College of Technology, Coimbatore, India, M.Tech., and Ph.D. Degrees from Anna University, India. He has published more than 60 books, 300 articles in international/national journals/conferences, and 200 patents. He is also the Vice-Chairman of the Renewable Energy Society of India (RESI). He serves as a research consultant to many companies, startups, SMEs, and MSMEs. He is the series editor of several book series and serves in various editorial capacities of several international journals. He has received numerous awards for research at national and international Levels.

A few of them include:

- Rashtriya Vidhya Gourav Gold Medal Award and The Best Educationalist Award from Hon.Justice O.P Saxena, Supreme Court, New Delhi, India.
- Three Lifetime Achievement Awards
- Dr.A.P.J.Abdul Kalam Sadhbhavana Award from Hon. Balmiki Prasad Singh, Former Governor of Sikkim. India
- Jewel of India Award from Mr. Gurpreet Singh, General Secretary, India
- Star of Asia Award from Mr. Korn Debbaransi, Former Deputy Prime Minister, of Thailand
- Pride of Asia Research Excellence Award from Hon. Anant. V.Sheth, Deputy Speaker- Goa Legislative Assembly, India
- CSI Young IT Professional Award
- National CSI Youth Award

**Dr. Ramasamy V** received his B.E Degree from Anna University Chennai, in 2006 an M.E Software Engineering from Anna University, Tiruchirappalli in 2009, and a Doctor of Philosophy in Computer Science, from Anna University Chennai, in 2021. He is currently working as an Associate Professor in the Department of CSE at Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology (Deemed to be University), Chennai, Tamilnadu, India. His area of interest includes Mobile Cloud Computing, IoT, Data Science, Artificial Intelligence, and Machine Learning. He is the author of several scholarly research papers in national and international journals and conferences. He is the Editor in Chief for Technoarete Transactions on Advances in Computer Applications in Technology (TTACAT) journal. He is the Book Editor of a few books for Technoarete publisher. He is the Guest Editor for Innovations in Future IoT Communications in MMTC Communications - Frontiers (IEEE COMSOC). He is the Evaluation Committee Member for IFERP Innovative Project Seed Funding Scheme. He is the organizing Chair of the International Conference on Advanced Computing and Intelligent Engineering ICACIE and organizing committee member of Advanced Communications and Machine Intelligence - MICA.

## About the Contributors

Dr.Sayed Abdulhayan , P.A.College of Engineering, Mangalore, Karnataka, India.

Sivakumar.V, Assistant Professor, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India.

Saravana kumar.R, Associate Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bangalore, Karnataka, India.

R.Swathi Assistant Professor, Department of Computer Science, Sree Abiraami College for Women, Thiruvalluvar University, Tamil Nadu, India.

Dr. Anitha T N, Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.

Dr. Jayasudha K, Associate Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.

Dr. Nur Fadzilah Mohamad Radzi, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.

Assoc. Prof. Dr. Azura Che Soh, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.

Assoc. Prof. Dr. Asnor Juraiza Ishak, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.

Assoc. Prof. Ir. Dr. Mohd Khair Hassan, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia

K. Sujatha , Department of Electrical and Electronics Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.

N.P.G. Bhavani , Department of ECE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. India.

V. Srividhya , Department of Electrical and Electronics Engineering, Meenakshi College of Engineering, Chennai, India

T. Kalpalatha, Department of ECE, S.V. Engineering College for Women, Karakambadi, Tirupati, India.

B. Latha, Department of Physics, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.

U. Jayalatsumi, Department of ECE, Dr. MGR Educational & Research Institute, Chennai, Tamil Nadu, India

T.Kavitha, Department of Civil Engineering, Dr. MGR Educational & Research Institute, Chennai, Tamil Nadu, India

A. Ganesan, Department of EEE, RRASE College of Engineering, Chennai, Tamil Nadu, India.

A. Kalaivani, Department of CSE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India.

Su-Qun Cao, Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, China.

Varsha Naika , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India

Dr. Rajeswari Kannanb , PCCoE, Pimpri Chinchwad College of Engineering, Pune, India.

Snehalraj Chugha , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India

Ahbaz Memona , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India

Himanshu Chaudharia, MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India

Dr.D.Kalaivani, Associate Professor and Head, Department of Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, India

Aman, Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India

Rajender Singh Chhillar, Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India  
Preeti Thareja, Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India  
Dewi Syahidah, Research Centre for Veterinary Science, National Research and Innovation Agency of Indonesia (BRIN), Indonesia  
Bernadetta Rina Hastilestari, Research Centre for Genetic Engineering, BRIN, Indonesia  
Atul Anil Kumar Kumbhar, Research Scholar, DSATM, Research Centre, Karnataka, India  
Dr G Manjula, Associate Professor, Dept. of ISE, DSATM, Karnataka, India  
Dr Roopa R Kulkarni, Associate Professor, Dept. of ECE, SATM, Kerala, India  
Dr. Prashant P. Patavardhan, Professor, Dept. of ECE, DRVITM, Karnataka  
Sasi Kumar M, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Sasi Kumar V, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Samyukthaa LK, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Gokul Karthik S, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Abirami A, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Lakshmanaprakash S, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
Vinothraja R, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India  
S. Menaga, Assistant Professor, Department of Electronics and Communication Engineering, Jai Shriram Engineering College, Tiruppur, India  
Dr.G.Kalaiarasi, Associate Professor, Department of Electronics and Communication Engineering, VSB Engineering College, Karur, India  
Dr. R.Vanithamani, Professor, Department of Biomedical Instrumentation Engineering, School of Engineering, Avinashilingam Institute for Home science and Higher Education for Women, India  
M.Nivetha, Assistant Professor, Department of Electronics and Communication Engineering, Jai Shriram Engineering College, Tiruppur, India  
Dr.D.Satheesh Kumar, Associate Professor, Dept. of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India  
Sai Raam V, Department of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India

# Table of Contents

<b>Preface</b> .....	<b>iii</b>
<b>Acknowledgement</b> .....	<b>iv</b>
<b>Chapter 1</b>	
<b>Data Analytics for Cloud – IOT Systems</b>	
<b>dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch001</b> .....	<b>1</b>
<i>Dr.Sayed Abdulhayan , P.A.College of Engineering, Mangalore, Karnataka, India.</i>	
<b>Chapter 2</b>	
<b>Fertilizers Usage in Agriculture and Crop Prediction Using ML Techniques</b>	
<b>dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch002</b> .....	<b>15</b>
<i>Sivakumar.V, Assistant Professor, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India.</i>	
<i>Saravana kumar.R, Associate Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology &amp; Management, Bangalore, Karnataka, India.</i>	
<i>R.Swathi Assistant Professor, Department of Computer Science, Sree Abiraami College for Women, Thiruvalluvar University, Tamil Nadu, India.</i>	
<b>Chapter 3</b>	
<b>Applications of IOT using Deep Learning</b>	
<b>dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch003</b> .....	<b>32</b>
<i>Dr. Anitha T N, Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.</i>	
<i>Dr. Jayasudha K, Associate Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.</i>	
<b>Chapter 4</b>	
<b>Machine Learning Algorithms for Herbs Recognition Based on Physical Properties</b>	
<b>dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch004</b> .....	<b>41</b>
<i>Dr. Nur Fadzilah Mohamad Radzi, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.</i>	
<i>Assoc. Prof. Dr. Azura Che Soh, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.</i>	
<i>Assoc. Prof. Dr. Asnor Juraiza Ishak, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia.</i>	
<i>Assoc. Prof. Ir. Dr. Mohd Khair Hassan, Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia</i>	
<b>Chapter 5</b>	
<b>Forecasting COVID-19 from Lung X-Ray Images</b>	
<b>dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch005</b> .....	<b>59</b>
<i>K. Sujatha , Department of Electrical and Electronics Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.</i>	
<i>N.P.G. Bhavani , Department of ECE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. India.</i>	
<i>V. Srividhya , Department of Electrical and Electronics Engineering, Meenakshi College of Engineering, Chennai, India</i>	
<i>T. Kalpalatha, Department of ECE, S.V. Engineering College for Women, Karakambadi, Tirupati, India.</i>	
<i>B. Latha, Department of Physics, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.</i>	
<i>U. Jayalatsumi, Department of ECE, Dr. MGR Educational &amp; Research Institute, Chennai, Tamil Nadu, India</i>	
<i>T.Kavitha, Department of Civil Engineering, Dr. MGR Educational &amp; Research Institute, Chennai, Tamil Nadu, India</i>	
<i>A. Ganesan, Department of EEE, RRASE College of Engineering, Chennai, Tamil Nadu, India.</i>	
<i>A. Kalaivani, Department of CSE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India.</i>	
<i>Su-Qun Cao, Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, China.</i>	



## Chapter 6

### Twitter Sentiment Analysis of Covid-19 Vaccination Using Deep Learning

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch006](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch006) .....75

*Varsha Naika , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*

*Dr. Rajeswari Kannanb , PCCoE, Pimpri Chinchwad College of Engineering, Pune, India.*

*Snehalraj Chugha , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*

*Ahbaz Memona , MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*

*Himanshu Chaudharia, MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*

## Chapter 7

### An Intrusion Detection System Based on Data Analytics and Convolutional Neural Network in NSS-KDD dataset

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch007](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch007).....93

*Dr.D.Kalaivani, Associate Professor and Head, Department of Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, India*

## Chapter 8

### Disease Prediction using Deep Learning Algorithms in Healthcare Sector

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch008](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch008) .....108

*Aman, Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

*Rajender Singh Chhillar, Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

## Chapter 9

### Applications of Deep Learning Models in Bioinformatics

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch009](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch009) .....116

*Preeti Thareja, Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

*Rajender Singh Chhillar, Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

## Chapter 10

### Machine Learning Approach for Early Detection of Plant and Fish Diseases

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch010](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch010) .....127

*Dewi Syahidah, Research Centre for Veterinary Science, National Research and Innovation Agency of Indonesia (BRIN), Indonesia*

*Bernadetta Rina Hastilestari, Research Centre for Genetic Engineering, BRIN, Indonesia*

## Chapter 11

### State-of-the-Art Analysis and Research Direction towards Secure Mobile Edge Computing in Transport System

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch011](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch011) .....137

*Atul Anil Kumar Kumbhar, Research Scholar, DSATM, Research Centre, Karnataka, India*

*Dr G Manjula, Associate Professor, Dept. of ISE, DSATM, Karnataka, India*

*Dr Roopa R Kulkarni, Associate Professor, Dept. of ECE, SATM, Kerala, India*

*Dr. Prashant P. Patavardhan, Professor, Dept. of ECE, DRVITM, Karnataka*

## Chapter 12

### Knowledge Discovery and Intelligent Data Mining

[dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch012](https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch012) .....145

*Sasi Kumar M, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Sasi Kumar V, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Samyukthaa LK, Department of Computer Science and Engineering, Bannari*

*Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Gokul Karthik S, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Abirami A, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Lakshmanaprakash S, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

## Chapter 13

### Data Visualization Techniques

**dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch013 .....165**

*Sasi Kumar M, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Sasi Kumar V, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Samyukthaa LK, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Vinothraja R, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Abirami A, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

*Lakshmanaprakash S, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India*

## Chapter 14

### Data Analytics for Disease Prediction

**dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch014.....185**

*S. Menaga, Assistant Professor, Department of Electronics and Communication Engineering, Jai Shriram Engineering College, Tiruppur, India*

*Dr.G.Kalaiarasi, Associate Professor, Department of Electronics and Communication Engineering, VSB Engineering College, Karur, India*

*Dr. R.Vanithamani, Professor, Department of Biomedical Instrumentation Engineering, School of Engineering, Avinashilingam Institute for Home science and Higher Education for Women, India*

*M.Nivetha, Assistant Professor, Department of Electronics and Communication Engineering, Jai Shriram Engineering College, Tiruppur, India*

## Chapter 15

### Integrating Smart Wearables and Exploratory Data Analysis for Disease Prediction

**dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch015.....203**

*Dr.D.Satheesh Kumar, Associate Professor, Dept. of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India*

*Sai Raam V, Department of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India*



# Chapter - 1

## Data Analytics for Cloud – IOT Systems

Dr.Sayed Abdulhayan <sup>1</sup>

<sup>1</sup> P.A.College of Engineering, Mangalore, Karnataka, India.

Email id: sabdulhayan@gmail.com

*Abstract— The Internet of Things (IoT) links any "thing" that produces data to the Internet, including computers, wearable technology, video games, automobiles, home appliances, satellites, and aircraft. In summary, the IoT connects a wide variety of sensors and devices that continuously produce data. According to IoT big data statistics, as the technology becomes more widely adopted, devices will produce exponentially more data globally in the next years. Data generated by automated systems and devices, such as smart thermostats and automatic lights, is referred to as automation data.*

*Keywords-Data Analytics, System Logs, CRM (Customer relationship management) data, SCM (Supply Chain Management ) data, ERP(Enterprise resource planning) data.*

### I. INTRODUCTION

The Internet of Things (IoT) links any "thing" that produces data to the Internet, including computers, wearable technology, video games, automobiles, home appliances, satellites, and aircraft. In summary, the IoT connects a wide variety of sensors and devices that continuously produce data. Anyone may regulate a home's lighting, temperature, or intrusion with a smartphone. It predicts that there will be 250,000 linked automobiles by the year 2020. We refer to this as the "Internet of Things" (IoT). There are connected devices all around the world. without human interaction, the generation of data and the development of a sentient, independent, and usable system. More connected gadgets than people are present on the planet (8.3 billion). Up to 2025, 75 billion devices are expected to connect. IoT data generation provides the volume and speed that big data applications need. IoT and big data go hand in hand.

### II. SMARTPHONE GENERATED DATA

Anyone may regulate a home's lighting, temperature, or intrusion with a smartphone. It predicts that 250,000 automobiles will be online by the year 2020.

#### A. Smartphone

Smart phones produce BIG data to address BIG issues. We generate and store more than 335 Exabyte of data per year with Smartphone alone, if you multiply the 60 gigabytes of data generated annually by each Smartphone user by the six billion devices (excluding notebooks, notepads, and other devices). The Smartphone collects context information about the user thanks to the technologies outlined above (e.g., Geolocation, Climate data, Network Traffic data etc).

#### B. Geolocation data:

Any type of information that makes it possible to pinpoint the position of a person or an object on Earth with some degree of accuracy is referred to as geolocation data. Typically, a signal from an electronic device, such as a mobile phone, linked vehicle, or smart watch, is used to create this data. Geolocation describes the process of identifying and following the movements of linked electronic devices by using location technologies like GPS or IP addresses. Geolocation is frequently used to track people's movements and whereabouts for surveillance because these devices are frequently worn by individuals.

Here are examples of formats that work:

1. 41.40338, 2.17403 in decimal degrees (DD).
2. The coordinates for this location are 41°24'12.2"N 2°10'26.5"E.
3. 41 24.2028 degrees, 2 10.4418 decimal minutes (DMM).

### C. *Data Analytics for Geolocation data*

Here the final output will be in terms of latitude, longitude, and altitude. We use here Linear regression Model or Logistic Regression Model for estimating the three entities i.e., latitude, longitude, and altitude.

### D. *Temperature data:*

For estimating the energy consumption of the unit operations for space heating, space cooling, and ventilation, temperature sensor data are especially helpful. The efficiency of heat exchangers can also be determined using data from temperature measurements. Imagine, for instance, that our doctor had a history of our typical body temperature going back several years. Our average (or normal) body temperature is determined by analysing historical data to be 37°C. We refer to this average temperature of 37°C as information.

### E. *Data Analytics for Temperature Sensors data*

Temperature data is being calculated in degree Celsius and hence analysis and reporting can be made in terms of extremes of temperature levels and forecast. Here we can use ID3, Random Forest etc machine learning Algorithms for estimation and prediction.

### F. *Pressure Sensors data*

A device or gadget that monitors pressure in gases or liquids is called a pressure sensor. A pressure sensor consists of an output signal generator as well as a pressure-sensitive part that can detect applied pressure. Absolute pressure, gauge pressure, differential pressure, and sealed pressure are only a few of the several types of pressure.

### G. *Data Analytics for Pressure Sensors data*

As mentioned above we get different pressure and evaluate the state of Object based on these pressure values. Here we may use Find-S and candidate Elimination Algorithm of machine learning to evaluate the values.

### H. *Touch Sensors data*

The SAW touch sensor, as its name implies, detects disturbances in ultrasonic waves passed across a glass layer's surface. Such sensing is made possible by piezoelectric crystals that are fixed to the glass layer of the LCD panel. Data in this case could be expressed as capacitance or resistance. Capacitive and resistive touch sensors/screens are the two most popular varieties.

### I. *Data Analytics for Touch Sensors data*

Here we are estimating the position of touch screen display and selecting the required apps and tools based of capacitance and resistance. Hence, we need to map these values with existing display, so we K-mean clustering or estimating-simplification based of machine Learning Algorithms.

### J. *Operating System Logs*

Syslog is a log of operating system events also referred to as the system log. Starting messages, system changes, unanticipated shutdowns, problems, and warnings are all included in this, along with other critical tasks. Windows, Linux, and macOS all generate syslog.

### K. *Data Analytics for Operating System Logs*

It will be possible to carry out data analysis and find solutions for issues with application and system level programming as well as hardware level problems with the help of diagnoses of start-up messages, system changes, unanticipated shutdowns, errors and warnings and their repetitions, time duration, and other factors.

### L. *Web server Logs*

The source of the issue can be quickly identified and fixed by using web server logs. A server log file is a straightforward text document that records all the activity of a particular server over a set amount of time (e.g., one day).

### M. *Data Analytics for Web server Logs*

Data Analysis will help to diagnose the problem and find a solution at web hosting, maintenance etc.

### N. *Telecom calls Logs*

Call logging is the process of gathering, analysing, and reporting technical and statistical information about incoming and

outgoing calls. Call logs give a broad overview of the real-time usage of a communications network. By giving the busiest parts of the network priority, that data can be leveraged to save expenses. Call logs can guarantee the most efficient use of maintenance resources to maintain a network's top performance.

#### *O. Data Analytics for Telecom calls Logs*

Data Analysis will help to diagnose the various problems in telecommunication, switching and networking. This will not only help in diagnosis of problems but also provides solution applicable to it.

#### *P. Network Management Logs*

A document that contains a history of activity within the application. It includes a history of all user and process access requests made to objects, along with attempts made at authentication. Here we collect the records of sender, recipient host number, IP number, MAC Address, time of interaction, number of interactions, number of attempts to interact, Network load and traffic etc.

#### *Q. Data Analytics for Network Management Logs*

From the data of Network logs, we will handle network Security issues and address the congestion issues. But before finding these solutions we need to have a clear-cut data Analytic procedure for Network management Logs. We can find out the grave threats of Cyber Security by data Analytics of Network management Logs.

### **III. SMART STREET TRAFFIC DATA**

IoT technology can be used by sensors placed in key locations to collect data on traffic, rerouting vehicles away from certain areas. This data can be analysed by IoT Big Data solutions, which can then be used to find alternate routes and enhance traffic signals. Radars, magnetic or piezo-sensor invasive approaches, human surveyors in the field or from a video, as well as basic machine vision image analysis algorithms, are all examples of traditional methods for gathering traffic data.

#### *A. Magnetic sensor data*

Magnetic sensors can detect moving ferrous metal. Magnetic sensors can detect moving ferrous metal. A wire wound around a permanent magnet makes up the most basic magnetic sensor. The magnetic flux across the coil is altered as a ferrous object approaches the sensor, producing a voltage at the coil terminals.

#### *B. Piezo-sensor data*

A piezoelectric sensor efficiently monitors compression utilising the piezoelectric effect since "piezo" is Greek for "press" or "squeeze."

#### *C. Radar data*

The range, altitude, direction of motion, and speed of objects are all determined by radar, an object detecting device that works with radio waves. Numerous techniques, such as the use of induction loops, pneumatic road tubes, and piezoelectric sensors, can be used to collect traffic data. These methods can gather traffic flow, but since they can't record vehicle speed and position, they can't be used with algorithms that analyse traffic flow.

#### *D. Pneumatic road tubes data*

Rubber hoses are extended across the road and joined to a data logger at one end. The tube's opposite end is sealed. Traffic volume and composition data can be collected.

#### *E. Induction loop data*

A moving magnet or an alternating current are used to create an induction in a device known as an inductive loop, which is used for electromagnetic communication or detection. For observing traffic, inductive loop sensors are a tried-and-true technique.

#### *F. Speed sensor:*

Wheel speeds are measured by vehicle speed sensors, which then send that information to the car's ECU (Electronic Control Unit). The ECU regulates the ignition timing transmission shift points, and air/fuel ratios, of the vehicle. The electronic control units (ECUs) in your car are then sent this measurement as an analogue signal or a low voltage square wave signal.

#### *G. Data Analytics for Street traffic management data*

From the sensors mentioned above we will be able to find out stress on roads, bridges, and tracks. Traffic Volume, speed of vehicles, direction of vehicles, peak hour traffic, slag time in traffic etc can be found out using the data from sensors mentioned. We need to process the data after getting the output from each of the sensors and finally applying it to the estimation, prediction,

deduction, visualization Etc Algorithm.

#### IV. COMPUTER GENERATED DATA

Information that is automatically produced by a computer Programme, application, or other mechanism without the active involvement of a human is referred to as machine-generated data. Although the phrase has been around for more than fifty years, there is now significant disagreement regarding its definition. The countless system logs that the operating system and other infrastructure software regularly produce, as well as the click stream and request logs produced by Web servers, are examples of machine data. Machine-generated data also includes telecom call detail records and network management logs. Machine data analytics is the process of collecting, analysing, and displaying data generated by software from many sources, such as personal computers, smart phones, and other gadgets.

##### A. *Operating System Logs*

Syslog is a log of operating system events also referred to as the system log. problems, warnings, unanticipated shutdowns, system changes, and Starting messages, are all included in this, along with other critical tasks. Linux, macOS, and Windows all produce syslog.

##### B. *Web server Logs*

Using web server logs, you may quickly identify the issue and promptly fix it. A server log file is a straightforward text document that records all the activity of a particular server over a set amount of time (e.g., one day).

##### C. *Telecom calls Logs*

Call logging is the process of gathering, analysing, and disclosing technical and statistical information about phone calls. Call logs give a broad overview of the real-time usage of a communications network. By giving the busiest parts of the network priority, that data can be leveraged to save expenses. Call logs can guarantee the most efficient use of maintenance resources to maintain a network's top performance.

##### D. *Network Management Logs*

A document that contains a history of activity within the application. It records user and process access requests to objects, authentication attempts, and other activity. Here we collect the records of sender, recipient host number, IP number, MAC Address, time of interaction, number of interactions, number of attempts to interact, Network load and traffic etc.

##### E. *Network Data Traffic*

To identify and address security issues, network traffic analysis (NTA) involves intercepting, recording, and analyzing network traffic communication patterns.

##### F. *Data Analytics for Operating System Logs*

Diagnoses of Start-up messages, system changes, unexpected shutdowns, errors and warnings and their repetitions, time duration will help to carry out data Analysis and find out solutions for problems in Application and System level programming or at hardware level problems.

##### G. *Data Analytics for Web server Logs*

Data Analysis will help to diagnose the problem and find a solution at web hosting, maintenance etc.

##### H. *Data Analytics for Telecom calls Logs*

Data Analysis will help to diagnose the various problems in telecommunication, switching and networking. This will not only help in diagnosis of problems but also provides solution applicable to it.

##### I. *Data Analytics for Network Management Logs*

From the data of Network logs, we will handle network Security issues and address the congestion issues. But before finding these solutions we need to have a clear-cut data Analytic procedure for Network management Logs. We can find out the grave threats of Cyber Security by data Analytics of Network management Logs.

#### V. SMART TV DATA

Netflix, Amazon Prime, and Now TV apps frequently make the claim that they only utilise data for essential functions like recommendations or credit checks. But this might also contain information like device identifiers, geolocation, browser type, email address, and payment details.

### *Data Analytics for Smart TV data*

By data analysis of the Smart TV data, we can find out the payment history, proportion of choice of people etc. Device Model, version history associated with geo location etc can be analysed. This is only possible if we have Cloud service available for computing and execution.

## **VI. SMART CLOCK DATA**

It can facilitate relaxation and improve sleep quality because it was created to limit Smartphone screen time at night. Additionally, it can manage your schedule, control your smart home, and play your preferred music around the house.

### *A. Alarm data*

This data is collected along with various other environments conditional data is collected and made to act intelligently.

### *B. Schedule Information data*

This data is collected with other conditional items and made to work intelligently.

Home automation data: This data is collected and disseminated to different items in house for automatic working based on time schedule.

### *C. Data Analytics for Smart Clock data:*

Here whatever data we are getting we need to classify based on binary classification of Machine learning Algorithm and watch the threshold being crossed or not. To do so we need Data Analysis.

## **VII. SMART WATCH DATA**

An application for smart watches gathers sensor monitor data as well as patient- and user-reported results. It transforms the gathered information into interpretable variables and sends them to the distant server.

### *A. Blood pressure data*

Like a fraction, it is measured using two numbers, one at the top (systolic) and one at the bottom (diastolic). For illustration, 120/80 mm Hg. Millimetres of mercury are used to measure blood pressure. The mm/Hg stands for that. Bio-Medical devices being connected to smart watch will help in getting status of Blood pressure of patient anytime.

### *B. Heartbeat data*

Wearing a heart rate monitor (HRM) allows you to continuously check your heart rate. Heart rate monitors can be worn on the wrist or on the chest. Each heartbeat is detected, and the data is sent to a receiver such a watch, fitness gear, or phone app. The information is presented as a beats per minute count. You can estimate this by deducting your age from 220. For illustration, a 30-year-old would do the following to determine what their maximum heart rate is:  $220 - 30 = 190$  bpm. The electrical heart signals are detected by cardiac-rate monitors. They are delivered to a watch or a data storage facility. With the help of the computer-based data analysis offered by many models, you may analyse your workout and better understand the benefits of your activity.

### *C. Geolocation data*

Geolocation data is information that can be used to pinpoint the precise location of an electronic device.

Here is example of effective format.

Degrees in decimal form (DD): 41.40338, 2.17403.

### *D. Data Analytics for Geolocation data*

Here the final output will be in terms of latitude, longitude, and altitude. We use here Linear regression Model or Logistic Regression Model for estimating the three entities i.e., latitude, longitude, and altitude.

### *E. Temperature data*

For estimating the energy consumption of the unit operations for space heating, space cooling, and ventilation, temperature sensor data are especially helpful. The efficiency of heat exchangers can also be determined using data from temperature measurements. Imagine, for instance, that our doctor had a history of our typical body temperature going back several years. Our average (or normal) body temperature is determined by analysing historical data to be 37°C. We refer to this average temperature of 37°C as information.

### *F. Data Analytics for Temperature Sensors data*

Temperature data is being calculated in degree Celsius and hence analysis and reporting can be made in terms of extremes

of temperature levels and forecast. Here we can use ID3, Random Forest etc machine learning Algorithms for estimation and prediction.

#### *G. Touch Sensors data*

The SAW touch sensor, as its name implies, detects disturbances in ultrasonic waves passed across a glass layer's surface. Such sensing is made possible by piezoelectric crystals that are fixed to the glass layer of the LCD panel. Here, the information might be expressed as capacitance or resistance. Capacitive and resistive touch sensors/screens are the two most popular varieties.

#### *H. Data Analytics for Touch Sensors data*

Here we are estimating the position of touch screen display and selecting the required apps and tools based of capacitance and resistance. Hence, we need to map these values with existing display, so we K-mean clustering or estimating-simplification based of machine Learning Algorithms.

#### *I. Data Analytics for Smart Watch data*

With the data Analysis carried with sensor generated data from smart watch, we can keep track of people and their health conditions in real Time Scenarios. This is only possible when we collect data from sensors and carry out data Analysis.

### **VIII. IOT DATA FOR GAMING INDUSTRY**

Companies are transitioning to big data cloud storage because it decreases implementation costs because of an increase in demand for data storage. To continue playing games with a friend whenever and wherever you wish, you may connect your gaming device—be it a console, desktop, laptop, smartphone, or tablet—to their gaming device using the Internet of Things (IoT). The gadgets' electronics and sensors are utilised to establish online connections with other gadgets. It would also be accurate to argue that the Internet of Things (IoT) has sparked this revolution in how people play games nowadays. The gaming industry has advanced thanks to the use of virtual reality, augmented reality, richer graphics, and animations, as well as high levels of connectivity.

The physical and digital worlds are connected by the Internet of Things (IoT). With the help of technologies like augmented reality (AR) and virtual reality, data from IoT-enabled devices is gathered, transformed into information, and made visible in real time (VR). This technology has increased access to internet games while also freeing up storage on computers and gaming consoles. A gamer can access their favorite games using cloud-based technologies without having to spend a lot of money on PCs and gaming consoles.

According to Newzoo's Global Games Market Report 2021, the games industry will reach \$218.7 billion by 2024 and continue to grow at a steady rate of 8.7 percent annually. The number of players also keeps growing. Around 3 billion people played video games worldwide in 2021, up 5.3% from the previous year, according to Newzoo. The equipment used in gaming is designed specifically for a certain type of game; for example, some games use wearables and sensor-based equipment to make the character move in the game exactly like the player moves. AR/VR gadgets gather a lot of biometric information that can be used to identify people and deduce other details about them. While improving immersive experiences, this data might also increase privacy problems.

#### *A. AR Sensor data*

Depending on the type, AR can gather information about the user's environment using depth sensors, accelerometers, cameras, gyroscopes, and light sensors.

#### *B. Depth sensors data*

Depth sensors are a type of three-dimensional (3D) range finder that collects distance information from multiple points over a sizable Field-of-View (FoV). One or more sensors (such a Lidar) with relatively tiny Fields-of-View are often used in conventional distance sensing techniques to measure distance. Depth sensing is also a key to navigation, localisation, and mapping and collision avoidance.

#### *C. Gyroscope data*

The gyroscope calculates the rotational speed in rad/s around the x, y, and z axes of a device. A gyroscope detects changes in a device's orientation, and when combined with an accelerometer, it is a powerful tool for determining how an object is oriented in three-dimensional space. Gyroscopes calculate angular velocity, which is usually expressed in radians per second.

#### *D. Light sensor data*

Light sensors pick up on the presence of light and translate that energy into an electrical signal. The radiant energy within the source's infrared to ultraviolet light frequency spectrum can subsequently be measured after being transformed into



electrical energy.

#### *E. Accelerometer data*

Wearable accelerometers measure the accelerations of the body part to which the monitor is fastened. The monitor typically filters and pre-processes the signal to produce activity counts, or accelerations brought on by body movement. The accelerometer measures angular acceleration instead of linear. Acceleration is the rate at which anything moves faster or slower in one dimension.

1. Increased velocity is indicated by positive values.
2. Negative numbers represent a decline in velocity.
3. Zero values signify continuous speed (which might not be zero).

#### *F. Camera data*

EXIF information is often stored in every photo file produced by contemporary cameras. EXIF data is metadata; essentially, it keeps track of the time and date the image was created as well as the camera settings that were used to take it. These EXIF data from one of my images are shown here in Adobe Bridge. Information is taken and kept as data rather than actual images in the world of digital photography. Light is transformed into a series of 1s and 0s by the input device (camera), which then stores the data on a disc or memory card. Then, output devices (like printers and monitors) transform this digital data back into images. They typically store photographs in either TIFF, which is an uncompressed format, or JPEG, which is a compressed format; however, some employ RAW format. Most cameras store images in JPEG format, and they occasionally provide quality settings (such as medium or high).

#### *G. Nyquist sampling data*

The frequency  $f_{Nyq} = d_{scan} / 2$  is called the Nyquist frequency. By definition  $f_{Nyq}$  is always 0.5 cycles/pixel. The Nyquist frequency can be visualized as the frequency that has two samples per cycle. Lower frequencies (more than two samples per cycle) can be reproduced exactly, but higher frequencies cannot.

#### *H. Noise camera*

There are several underlying sources of internal camera image noise. Electricity, heat, and sensor illumination levels are the three main culprits. When the sensor is over-volted (ISO pushed) under low-light conditions, each pixel has very little light wave variation to record before being magnified.

#### *I. Colour camera*

A camera with a unique design for creating colour-separation negatives (such as a one-shot or beam-splitter camera). First, according to science, the human eye is capable of distinguishing or seeing around 10 million different colours. However, even in the most sophisticated digital cameras, the camera sensor can only discern roughly 3 different colours (red, green, and blue).

#### *J. Bit-depth*

The number of bits used to represent each image pixel. It can be misleading because the phrase is frequently used to refer to bits per pixel and other times to refer to the total amount of bits used multiplied by the total number of channels.

#### *K. Intensity scaling*

A scalar value is multiplied by all intensities. Evidently, this will alter the histogram, but it has no equivalent effect to histogram equalisation (it will just scale the x-axis of the histogram, barring some binning effects).

#### *L. Tone*

Tone describes the brightness levels in the image, ranging from completely black to completely white. Highlights are bright tones, whereas shadows are dark tones. Most images of nature feature a variety of tones, from dark or almost black to white or nearly white.

#### *M. Resolution*

How many pixels are displayed per inch of an image is sometimes referred to as PPI, or pixels per inch. More pixels per inch (PPI) and richer, more detailed images are also benefits of higher resolutions.

#### *N. Magnification*

The optimal procedure for picture magnification is one that essentially improves image resolution in order to highlight implicit information that was included in the original image but wasn't immediately obvious. It can be viewed as a shift of scale.

#### *O. Data Analytics for AR in Games*

A computer-generated image is used in augmented reality, which dramatically changes how the real world is displayed. As technology and computers advance, augmented reality will cause a significant shift in how people perceive the real world. This is only possible if we have Cloud service available for computing and execution.

#### *P. VR Sensor data:*

VR systems may be used to gather sensitive data, such as sensitive data like facial muscle movements that can be used to identify users' emotions or state of health. A basic VR system uses an inertial measurement unit (IMU), which can include an accelerometer, gyroscope, and magnetometer.

#### *Q. Accelerometer data*

Wearable accelerometers measure the accelerations of the body part to which the monitor is fastened. The monitor typically filters and pre-processes the signal to produce activity counts, or accelerations brought on by body movement. The accelerometer measures angular acceleration instead of linear. Acceleration is the rate at which anything moves faster or slower in one dimension.

- Increased velocity is indicated by positive values.
- Negative numbers represent a decline in velocity.
- Zero values signify continuous speed (which might not be zero).

#### *R. Gyroscope data*

The gyroscope calculates the rotational speed in rad/s around the x, y, and z axes of a device. A gyroscope detects changes in a device's orientation, and when combined with an accelerometer, it is a powerful tool for determining how an object is oriented in three-dimensional space. Gyroscopes calculate angular velocity, which is usually expressed in radians per second [12].

#### *S. Magnetometer data*

The magnetometer sensor detects the magnetic field in T along each of the three physical axes (micro-Tesla). Two new interfaces are defined by this specification: and a magnetometer that reports calibrated magnetic field measurements. Magnetometer that measures the magnetic field but is not calibrated. Geophysical surveys routinely use magnetometers to measure the Earth's magnetic field, identify various magnetic anomalies, and determine the dipole moment of magnetic materials. In the attitude and heading reference system of an aircraft, they are typically used as a heading reference.

#### *T. Data Analytics for VR in Games*

Utilizing computer technology, virtual reality (VR) creates interactive virtual worlds that may be experienced while wearing a headset. Users are essentially immersed "within" a virtual world, which could result in a better sense of immersion than that offered by a "conventional" flat screen. provides players with enticing virtual goods. Provide a real-time join option for players. modern additions that enhance the gaming experience. a 24-hour augmented reality service. This is only possible if we have access to cloud computing and execution services.

## **IX. IOT DATA FOR CARS**

Telematics log data, which includes performance utilisation, infotainment system data, speed information, battery consumption management, and odometer readings, are just a few examples of the data collected about your car from your car.

#### *A. Telematics log data*

By using GPS and on-board diagnostics (OBD), telematics is a technique for keeping track of vehicles, trucks, equipment, and other assets. The movements of the asset are then displayed on a computerised map. Fuel consumption, idle time, location, speed, abrupt acceleration or braking, and vehicle issues are all common components of telematics data.

#### *B. Infotainment system data*

The telematics and infotainment systems in cars retain a tonne of information, including call records, contact lists, SMS messages, emails, images, videos, social network feeds, and the navigation history of all the places the car has visited.

#### *C. Speed information*

Speed race information data recorded in database and further used for data analysis.

#### *D. Battery usage management*

Battery level usage information data recorded in database and further used for data analysis.



### E. Odometer readings

The simple mathematical formula to find out the mileage with an odometer is to divide the distance travelled by the amount of fuel used.

### F. Data Analytics for IOT data for Cars

Car's state, condition can be evaluated, monitored, and tracked by doing data analysis of the above-mentioned parameters. This is only possible if we have Cloud service available for computing and execution.

## X. DATA FOR SMART REFRIGERATORS

These "smart fridges" could make shopping lists and connect to mobile apps, enabling users to adjust the temperature remotely, receive alerts if the door was left open, and access internet recipes depending on the contents of their refrigerator. The features on a smart refrigerator can vary depending on the make and model, but here is a list of some common features and their benefits.

- Wi-Fi,
- a touchscreen interface,
- an interior camera,
- a shopping list,
- a built-in browser,
- a recipe database,
- entertainment, and
- a whiteboard

are some of the features available.

### Data Analytics for Smart Refrigerators

Condition tracking and monitoring can be done by getting the real-time values of sensor data and computed it by using AI-ML programming by carrying data Analysis of recorded data in different database tables.

## XI. ATM (AUTOMATIC TELLER MACHINE) DATA

IoT was first used in ATMs in the 1970s (or cash machines). They are regarded as one of the earliest instances of the Internet of Things being employed in our culture. Even while the technology underlying ATMs is now quite simple, they were nevertheless linking gadgets to the internet at the time. Your account information is captured when the card is swiped or pushed on the reader, so the host processor receives the card's data (server). Thus, the host processor makes use of this information to obtain information from cardholders. In addition to cash withdrawal and checking account balances, ET Wealth highlights 9 helpful services that may be accessed through an ATM.

- Open or withdraw a fixed deposit. (Bank Details Data) ...
- Recharge your mobile. (Service Provider Data) ...
- Pay income tax. (Income tax Data) ...
- Deposit cash. (Bank Details data) ...
- Pay insurance premium. (Policy Insurance Data) ...
- Apply for personal loan. (Bank details and Data) ...
- Transfer cash. (Bank Details and data) ...
- Pay your bills. (Electric bill, Water bill, house bill Information data)

### Data Analytics for ATM data

Conditional Evaluation, tracking and monitoring of people and their respective accounts with Banks, Electricity Boards, Water Boards, Insurance bodies can be carried out effectively by storing data in clouds and carrying out data Analysis of data of people and their associated service records by generating reports. This is only possible if we have Cloud service available for computing and execution.

## XII. IOT DATA FOR SATELLITES

In addition to other weather, climate, and environmental monitoring applications, polar-orbiting satellites collect data on precipitation, sea surface temperatures, atmospheric temperature and humidity, sea ice extent, forest fires, volcanic eruptions, global vegetation analysis, and search and rescue. Climate Data (Temperature data, humidity data, wind speed data, precipitation data etc). A climatic data element is a measured parameter that aids in describing the climate of a particular place

or region.

- Temperature,
- Wind
- Direction,
- Humidity

Thermometers, rain gauges, and other tools are used by people from many walks of life to keep track of their local weather. Additionally, around the world, automated networks of scientific devices continuously track the weather and climate.

#### A. Humidity Sensor Data

Instead of the absolute level of humidity in the air, relative humidity is often given for humidity in weather analysis. Relative humidity is the difference between the amount of water vapour in the air and the amount that the air can contain at the current temperature.

Data from the wind speed sensor show how wind speed is determined. Anemometer factor  $\times$  instantaneous shaft speed equals instantaneous wind speed. Average Wind Speed is calculated as follows: Anemometer Factor  $\times$  (Time / Turns). The research indicates that the most likely wind speeds for the hub heights under consideration are 5.881 and 6.775 m/s, respectively, and the wind speeds for maximum energy are 6.630 m/s and 7.439 m/s. The location has yearly mean energy densities of 110.006 kWh/m<sup>2</sup> and 160.430 kWh/m<sup>2</sup> at hub heights of 25 m and 65 m, respectively.

#### B. Precipitation sensor data

Rain gauges quantify the amount of precipitation at a specific location. Measurements from a single rain gauge are frequently used to depict the amount of precipitation in broader areas, i.e., between gauge locations. Rainfall is typically measured using rain gauges that are manually read or that automatically provide real-time data to a telemetry network using wired or wireless transmission. It is a tested technology that performs well for monitoring when there are numerous rain gauges in the area to be monitored.

#### C. Sea surface temperatures data

Data sets on sea surface temperatures (SST) are a crucial tool for tracking and comprehending climatic variability and change. SSTs make up the majority (about 71 percent) of the input into combined global land-ocean surface temperature data packages.

#### D. Atmospheric temperature data

Thermometers are used to measure the air's temperature. Common thermometers consist of a glass tube with liquid inside that is fastened to a scale. Degrees Celsius (°C), Degrees Fahrenheit (°F), or both might be used to indicate (graduate) the scale. A liquid is delivered into the tube from a reservoir, or "bulb," at the thermometer's base.

#### E. Sea ice extent data

The amount of ice that is now covering the Arctic Ocean is known as the sea ice extent. Sea ice is crucial for several processes, including controlling ocean and air temperatures, moving ocean water, and preserving animal habitats. This data is collected with the help of sensors.

#### F. Forest fires data

There are numerous fire detection algorithms available with various fire detection approaches. The region affected by the fire is forecasted in the current work processes using satellite photos. This data is collected with the help of sensors.

#### G. Volcanic eruptions data

While monitoring Earth's volcanoes from the ground will continue, satellite data will make it possible for researchers to employ sensors and follow the eruption of hot, toxic gases, ash, lava, and rock that can cause catastrophic loss of life and property, particularly in densely populated areas.

#### H. Global vegetation analysis data

A technique for examining the species makeup and organisation of a plant community is vegetation analysis. To record, map, or analyse vegetation, a transect is a cross section of the area in question. There may be two types: I A line transect: This method involves running a piece of tape parallel to the ground. Any species that meets the line separating the two lines is noted.

#### I. Search and rescue Operation data

Emergency services conduct search and rescue (SAR) operations to find people who are believed to be in danger, lost, ill,

or injured, either on land or at sea, in remote or challenging-to-reach areas like mountains, deserts, or forests. SAR operations frequently involve the assistance of well-trained volunteers.

#### *J. Data Analytics for IOT data for Satellites*

With different sensors and various special cameras, we get to know the present situation of climate etc. Then we will keep on recording these data values and carryout data analysis with these values. With huge data of previous records extending from years together in cloud we will be able to analyze the situation and evaluate the condition. This is only possible if we have Cloud service available for computing and execution.

### **XIII. IOT DATA FOR AIRPLANES**

IoT-enabled products like barcodes or chips can assist travellers in tracking their bags as they head to the airport to board their trip. From their smartphones, travellers may use these tools and capabilities to track the whereabouts and condition of their luggage in real time.

#### *A. Barcode data*

Used to tag luggage and other entities by IOT devices and stored in cloud.

#### *B. Chip data*

Used to tag important entities by IOT devices and stored in cloud.

#### *C. RFID Tag data*

Used to tag and monitor employee movement and is stored in cloud.

#### *D. Data Analytics for IoT data of Airplanes*

This data is very useful whenever we move in flights to keep track of employees and our luggage. Hence its necessity to be stored in cloud to be monitored at different stations and locations.

### **XIV. IOT DATA FOR FACTORIES**

Manufacturing can undergo a digital transformation with the help of the Industrial Internet of Things (IIoT). Industrial IoT collects vital production data via a network of sensors, and cloud software then transforms this data into illuminating knowledge about how efficiently manufacturing operations are conducted.

#### *A. Production Analytics*

Production Analytics allows your operators to be experts by giving them more time to fix underlying issues, optimize performance and innovate throughout the oilfield. The solution vastly reduces the time operators and engineers need to monitor wells by surfacing evolving problems directly to them. Productivity analytics is the application of data analytics to optimise the efficiency of a manufacturing facility. This involves preventing elements like unscheduled machine downtime (when machines are idle), which have a major negative influence on a manufacturing enterprise's profitability.

#### *B. Descriptive Analytics*

A statistical technique called descriptive analytics is used to search and summarise historical data to find patterns or meaning.

#### *C. Data Analytics for Factory/Industry Descriptive data*

E.g., Factory/Industry Descriptive data, Production data, modeling data, product rating data etc.

Using descriptive statistics, characteristics of a sample or data set, such as the mean, standard deviation, or frequency of a variable, are summarised or described.

#### *Tools for Descriptive Statistics*

- Line of Best Fit Scatter Plot Chart Maker (Offsite)
- Calculator for mean, median and mode.
- Calculator for variance.
- Calculate standard deviation here.
- Calculator for the coefficient of variation.
- Calculator for percentiles
- Calculator for the interquartile range.
- Calculator for Pooled Variance.

With the above techniques we will predict and estimate the situation and condition. This is only possible if we have Cloud service available for computing and execution.

#### *D. Diagnostic Analytics*

Diagnostic analytics' goal is to ascertain the underlying reason for an event or trend. A trend is frequently discovered utilising a previous descriptive analysis stage. The business might use diagnostic analytics to ascertain the cause of the trend.

E.g., Factory/Industry Diagnostic data, Production data, modeling data, product rating data etc.

The diagnostic performance data of diagnostic tests are frequently used for evaluation, comparison, and marketing purposes. These data are based on a comparison between test results and an independent evaluation of the actual disease state.

#### *E. Data Analytics for Factory/Industry Diagnostic data*

Diagnostic evaluation can be carried out manually, automatically, or using statistical tools (such as Microsoft Excel). Before getting into diagnostic analytics, it is important to comprehend the following ideas: a diagnostic regression analysis, the distinction between correlation and causation, and hypothesis testing. This is only possible if we have Cloud service available for computing and execution.

#### *F. Predictive Analytics*

Businesses utilise predictive analytics to identify threats and opportunities by using trends in this data.

E.g., Factory/Industry Predictive data, Stock exchange rating data, Weather data, Financial Crisis data etc.

#### *G. Data Analytics for Factory/Industry Predictive data*

This predictive model is then used to forecast future events or to suggest a plan of action for the best outcomes using recent data.

#### *H. Prescriptive Analytics*

Data analytics known as prescriptive analytics offer recommendations for what action should be taken next. Descriptive, diagnostic, and predictive analytics are all tied to one another.

E.g., Factory/Industry Prescriptive data: Stock exchange rating data, Weather data, Financial Crisis data etc.

#### *I. Data Analytics for Factory/Industry Prescriptive data*

Machine learning is used in prescriptive analytics to assist businesses in choosing a course of action based on predictions made by a computer programme. Predictive analytics, which makes use of data to forecast short-term outcomes, complements prescriptive analytics. This is only possible if we have Cloud service available for computing and execution.

Manufacturers can:

- Boost productivity and uptime by using IoT, Cloud, and advanced analytics technologies.
- Boost the efficiency of the procedure.
- Quicken innovation.
- Cut down on asset downtime.
- Boost operational effectiveness.
- Establish complete operational visibility.
- Boost product caliber.
- Lower operating expenses.

## **XV. CRM DATA (CUSTOMER RELATIONSHIP MANAGEMENT)**

CRM systems can use the Internet of Things (IoT, a network of connected objects, devices, and equipment that allows data to be transferred over a network without requiring human contact) to improve end-to-end business operations. The information is obtained through a variety of client interaction programmes, including surveys and help desks. Some of the types of information that are included in a CRM database include name, title, email address, social profiles, contact history, lead score, order history, recent news, and personality attributes.

#### *A. Visiting card data (for Business Malls)*

A business card normally contains the giver's name, the name of their organisation or business affiliation (often with a logo), and their contact details, including their street address, phone number(s), fax number, e-mail address(es), and website. Prior to the invention of electronic communication, telex information was sometimes included on business cards.

#### *B. Mobile Number data*

Data that can be used to identify a specific individual is known as personally identifiable information (PII) or personal data.

A phone number, national ID number, email address, or any other piece of information that can be used to get in touch with, locate, or identify a person is considered personally identifiable information (PII).

#### *C. Debit/Credit Card data*

This information includes the card number, the identity of the card issuer (organization/store address), the transaction amount, the transaction number, the transaction date and time, the transaction type (deposits, withdrawals, purchases, or refunds), the type of account being debited or credited, and the identity of the terminal (company name).

#### *D. Discount Coupons data*

A discount code or voucher code is only a string of letters and numbers that, in most cases, only partially explain the deal being offered. For a given period, brands will receive a fixed discount, and the voucher may be utilised to get this cash back.

#### *E. Data Analytics for CRM data*

Analytics for customer relationship management (CRM) includes all software that analyses customer data and delivers it to support and streamline better business decisions. CRM analytics may use data mining and can be categorised as an instance of online analytical processing (OLAP). This is only possible if we have Cloud service available for computing and execution.

### **XVI. ERP DATA (ENTERPRISE RESOURCE PLANNING)**

Businesses may increase data availability, which can result in operational excellence, by connecting IoT and ERP. The IoT sensors' gathered data will be delivered straight into the ERP programme. Real-time updates on any process modifications are provided. Instantaneous access to critical business information is made possible for enterprises by integrating ERP with IoT data. Businesses may undertake real-time analysis using the continuous data stream from IoT sensors and devices, giving them practical insights to help them make better decisions. Modules that concentrate on various business metrics are used to present data gathering for ERP to the user. Significant commercial operations including product planning, distribution, accounting, marketing, human resources, finance, and inventory are likely served by them.

#### *Data Analytics for ERP data*

Rapid identification, capture, validation, and transmission of pertinent data points across a complete supply and sales chain are all capabilities of ERP data analysis. This makes ERP users more foresighted than rivals and enables rapid increases in overall revenue efficiency over time. This is only possible if we have Cloud service available for computing and execution.

### **XVII. SCM DATA (SUPPLY CHAIN MANAGEMENT)**

Even for tiny businesses, it is a practise that is expanding. Such waste can be reduced using IoT based on connected sensing technology. By enabling supply chain visibility, it enables the stakeholders to exchange crucial real-time information, minimising interruption. Inventory is a significant factor in the supply chain that is challenging to control. AI-driven supply chain optimization software magnifies important decisions by recommending the optimum course of action based on cognitive predictions. The effectiveness of the entire supply chain might increase as a result. Additionally, it helps manufacturers think through alternative scenarios' possible time, cost, and financial impacts.

#### *Data Analytics for SCM data*

Analytics is the ability to make judgments based on a summary of reliable, relevant data, frequently employing graphs for display. Regardless of the industrial use cases, the research suggests that data analytics should be a core component of the supply chain application development process. To get the most out of their data resources, manufacturing and supply chain organizations should also concentrate on developing BI applications. This is only possible if we have Cloud service available for computing and execution.

The following services offered by cloud-based gadgets will transform logistics:

- Real-time condition monitoring,
- Connected fleet management,
- Asset tracking solutions,
- Improved supply chain visibility and transparency,
- Increased agility to demand or supply fluctuations,
- Data mastery using AI and Big Data,
- New levels of risk management.

## XVIII. IOT CLOUD

Three crucial factors influence firms' decision to move to the cloud: business agility, scalability, and cheaper cost of ownership.

Why we need Cloud?

The Internet of Things uses cloud computing to share resources to store IoT data and make it available when needed. It's important to keep in mind that sending massive data packets produced by the IoT via the Internet is made simple using cloud computing. To support the agility needed by IoT devices, cloud enabled IoT hosting providers do not need to rely on any type of hardware or equipment.

Best IoT Cloud Platforms are as below:

- IoT Platform by Amazon Web Services. Amazon is the market leader in consumer cloud.
- Oracle IoT,
- Sales force IoT,
- Bosch,
- IBM Watson IoT Cloud Platform,
- Google IoT Cloud Platform,
- Microsoft Azure IoT Hub, and
- Cisco IoT Cloud Connect.

Cloud computing gives businesses the flexibility, scalability, and connection they need to store, manage, and process data across platforms that are cloud-enabled.

## REFERENCES

- [1] Submitted to Solihull College, West Midlands, Student Paper
- [2] repository.iitr.ac.in, Internet Source
- [3] Submitted to NCC Education, Student Paper
- [4] en.wikipedia.org, Internet Source
- [5] Submitted to Sim University, Student Paper
- [6] Submitted to Fiji National University, Student Paper
- [7] Submitted to University of Greenwich, Student Paper
- [8] Submitted to Florida Virtual School, Student Paper
- [9] Submitted to The Scientific & Technological, Research Council of Turkey (TUBITAK), Student Paper
- [10] Submitted to University of Lincoln, Student Paper
- [11] enam.uac.bj, Internet Source
- [12] Submitted to University of Kentucky, Student Paper
- [13] www.sensortips.com, Internet Source
- [14] Submitted to Asia Pacific University College of Technology and Innovation (UCTI), Student Paper
- [15] Submitted to Victorian Institute of Technology, Student Paper
- [16] Submitted to Ashoka University, Student Paper
- [17] Submitted to UOW Malaysia KDU University College Sdn. Bhd, Student Paper
- [18] Submitted to University of Ulster, Student Paper
- [19] Submitted to University of Bolton, Student Paper
- [20] Submitted to Coventry University, Student Paper
- [21] Submitted to Liverpool John Moores University, Student Paper
- [22] Submitted to Middlesex University, Student Paper
- [23] scholarworks.uni.edu, Internet Source
- [24] Submitted to Federation University, Student Paper
- [25] www.entrepreneur.com, Internet Source
- [26] Submitted to Durban University of Technology, Student Paper
- [27] Submitted to University of Wollongong, Student Paper
- [28] www.sensorsuae.com, Internet Source
- [29] Atonu Ghosh, Koushik Majumder, Debashis De. "Chapter 2 A Systematic Review of Digital,Cloud and IoT Forensics"
- [30] Springer Science and Business Media LLC, 2021.



# Chapter - 2

## Fertilizers Usage in Agriculture and Crop Prediction Using ML Techniques

Sivakumar.V <sup>1</sup>, Saravanakumar.R <sup>2</sup>, R.Swathi <sup>3</sup>

<sup>1</sup>Assistant Professor, School of Computer Science and Engineering (SCOPE),  
Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India.

<sup>2</sup>Associate Professor, Department of Computer Science and Engineering,  
Dayananda Sagar Academy of Technology & Management, Bangalore, Karnataka, India.

<sup>3</sup>Assistant Professor, Department of Computer Science, Sree Abiraami College for Women,  
Thiruvalluvar University, Tamil Nadu, India.

Email: [sivakumarvgym@gmail.com](mailto:sivakumarvgym@gmail.com), [saravanakumar.rsk28@gmail.com](mailto:saravanakumar.rsk28@gmail.com), [rswathimca@gmail.com](mailto:rswathimca@gmail.com)

*Abstract— India being a farming country, its cost-cutting measure for the most part workers depend on cultivation of crop and its growth connected with agricultural manufacturing products. In India, crop growing is for the most part subjective by rainwater which is extremely volatile. Crop growing also depends on diverse soil parameters, namely Phosphorus, Nitrogen, Potassium, Soil moisture, Crop rotation, and Surface temperature and also on weather aspects which include temperature, rainfall, etc. In India now is quickly progressing in the direction of technological growth. Technology will show to be favourable to agriculture which will boost crop production follow-on in better yields to the cultivator. The proposed article provides a clarification for Smart Agriculture by monitoring the farming field which can support the cultivator in increasing effectiveness to a massive coverage. Weather forecast data such as temperature and rainfall and soil parameters obtained from Indian Meteorological Department (IMD) this repository gives just round the corner into which crops are appropriate to be advanced at a particular area.*

*Keywords— Smart Agriculture, Crop prediction, Machine learning, advanced farming techniques, Fertilizers, yield prediction.*

### I. INTRODUCTION

Agriculture is the first born among all profession as it is the definitive source of living for all humans. India being an agrarian country, 50% of the country's workforce is involved in this occupation and contributes nearly 17%–18% of the Gross domestic product (GDP) [1]. This sector significantly impacts the country's economy due to its contribution to exporting and the wide range of stakeholders involved. Moreover, food safety and security are paramount for a highly populated country like India. The United Nations has set up Zero hunger as one of its Sustainable Development goals to achieve a better and sustainable future Today security measures are used to protect the core factors of cyber security which are integrity, confidentiality, and accessibility of data. Many new expertise's are being use in the health sector in order to prevent attacks and reduce damage.

The recent climate changes vary often. So, it will be quite a task to cultivate crops by analysing the weather forecast that's needs to be implemented using various technologies to discover the agricultural information and educate the farmers to cultivate various crops and fertilizers. The fertilizers cannot be used adequately. As of now in the real world, the need of using new methods for improvisation is good. The usage of new methods is utilizing the data to develop a well-defined representation that can be included at the crop prediction. The mechanism works in the same way which humans react. Mainly humans get ideas from their experiences. The more knowledge we acquire, easily we can predict. By analogy, the chances of success are lower in an unknown circumstance than in a known situation. This model gets to know at what crop is grown.

India is a highly populated country and randomly change in the climatic conditions need to secure the world food resources. Framers face serious problems in drought conditions. Type of soil plays a major role in the crop yield. Suggesting the use of fertilizers may help the farmers to make the best decision for their cropping situation. The number of studies Information and Communication Technology (ICT) can be applied for prediction of crop yield. By the use of Data Mining, we can also predict the crop yield. By fully analyse the previous data we can suggest the farmer for a better crop for the better yield.

For the better yield we need to consider soil type and soil fertility things and also one of the major factors rainfall and groundwater availability if it is dry land it is better to go for cash crops and if it is wetland it is better to go for wheat and sugarcane. There are 15 agro-climatic regions in India these regions are divided on the bases of a type of the land. Each agro-climatic region can grow some specific crops. Based on that we need to suggest the farmer that which crop is best among those crops which belong to those climatic regions. Achieving the maximum crop at minimum yield is the ultimate Aim of the project. Early detection of problems and management of those problems can help the farmers for better crop yield. Crop yield prediction is the important research which helps to secure food. For the better understanding of the crop yield, we need to study of the huge data with the help of machine learning algorithm so it will give the accurate yield for that crop and suggest the farmer for a better crop. Improving the quantity of the crop is the key goal of precision agriculture means obtaining a better understanding of the crop using the information technology methods. The main goal of precision agriculture is profitability and sustainability.

From ancient times agriculture has become the backbone of our country. Nowadays climatic conditions vary very often. So, it is hard to grow crops by understanding weather conditions. We need to use some technology to find or understand the crop details and guide the farmers to grow crops accordingly and moreover fertilizer also one of the major factors to grow crops accordingly. If fertilizer is used more or less in the field the soil may lose its fertility and crop may not give the expected yield. So, fertilizer also becomes the major factor in it mostly understanding the temperature conditions are much necessary for India.

Crop yield prediction is an important agricultural problem. Every farmer always tries to know how much yield will be produced and whether it meets their expectations. In the past, yield prediction was calculated by analyzing a farmer's previous experience on a particular crop. The Agricultural yield is primarily dependent on weather conditions pests and planning of harvest operation. Accurate information about the history of crop yield is an important thing for making decisions related to agricultural risk management.

## II. PROBLEM STATEMENT

The efficient way of yield prediction will be a major decision for the people at national and regional levels for making quick decision. A proper way and good knowledge of crop yield prediction model will give an idea to farmers how they can decide to grow and when they can grow.

Currently in India now also farmers follow the old methods which they got knowledge from their previous generations. Main drawback is previously climate was very healthy and all the things would take place accordingly. Nowadays all the things are changing because of global warming and many and different reasons. Mainly in India the problem is lack of rainfall is not expected. In this method they have taken the help of the ensemble technique which is currently in use and a detailed study is being done. The given framework will make sure that it gives the proper information which is taken from old data which applies machine learning techniques. A method called Multiple Linear Regression is used; proper reasonable yields are being directed to current natural conditions. This mainly revolves with characterization of soil utilization where different calculations can be done.

Indian agriculture is a huge market which gives jobs in different forms of farming field. Mainly India grows many kinds of crops every year, but it mainly produces large quantity of rice and wheat. Indian farmers specifically also farm urad, jowar, small millets, sugarcane and non food crops like hemp, jute, cotton. In recent days weather conditions are unpredictable and farmers face difficulties to depend upon the traditional old ways of farming.

This project mainly focuses on helping of farmers for taking effective decision while predicting the crops. Mainly to increase the accuracy along with the given data, some required data for temperature and humidity is also collected from government website and stored accordingly. Also historic rainfall data is recorded and saved.

- Collect the weather data, crop yield data, soil type data and the rainfall data and merge these datasets in a structured form and clean the data. Data Cleaning is done to remove inaccurate, incomplete and unreasonable data that increases the quality of the data and hence the overall productivity.
- Perform Exploratory Data Analysis (EDA) that helps in analyzing the complete dataset and summarizing the main characteristics. It is used to discover patterns, spot anomalies and to get graphical representations of various attributes. Most importantly, it tells us the importance of each attribute, the dependence of each attribute on the class attribute and other crucial information.
- Divide the analyzed crop data into training and testing sets and train the model using the training data to predict the crop yield for given inputs.
- Compare various Algorithms by passing the analyzed dataset through them and calculating the error rate and accuracy for each. Choose the algorithm with the highest accuracy and lowest error rate.
- Implement a system in the form of a mobile application and integrate the algorithm at the back end.
- Test the implemented system to check for accuracy and failures.



### III. RELATED WORKS

Author says that [2] “Analysis and Forecasting of Electrical Energy a Literature Review” this research proposes about the analysis of the various sources of electricity generation and to predict the electricity generation to meet the future demand on energy using different data mining techniques.

Dirichlet Reputation Systems [3] this paper discusses various probability distribution analysis and compute reputé score based on the ratings from several different parties of the various sources of electricity generation and to predict the electricity generation to meet the future demand for energy-using different data mining techniques.

Author says that [4] “Demand Side Management of a University Load in Smart Grid Environment” this paper recommends a Demand Side Management (DSM) technique. The authors are suggested a bottom-up load modeling procedure to learning the arrangement of electrical energy shortage in Motilal Nehru National Institute of Technology (MNNIT) Institute, Allahabad, India. Additionally supporting, it recommends a load scheduling performance to form the shortage curvature by reorganizing the lecture schedules.

Implementation of Wireless Fidelity (WiFi) based single phase smart meter for Internet of Things (IoT) [6] proposes a system that utilizes the ESP8266 WiFi Module in order to connect to the internet and transmit data as opposed to the previous paper which would require a stable mobile network. According to the paper, this meter can correctly and reliably read the energy meter parameters such as demand value, load profile, and total energy consumption.

Regression Analysis for Prediction of Residential Energy Consumption [7] this research work, simple linear regression analysis and multiple linear regression analysis methods achieved in additionally a quadratic regression analysis be there implemented taking place daily or hourly records from a firm. The phase interval for the perceived statistics was shown to be a significant feature that well-defined the importance of the model.

Validation of a Distributed Energy Management Approach for Smart Grid Based on a Generic Colored Petri Nets Model [8] this paper proposed the Petri Nets Model with an intelligent neighborhood-based energy management approach that precedes assistance of neighborhood power leftover. The introduced approach allows consumers to insist on a dependence reason on their neighbors in order to control the preeminent alternate to fulfil their necessities.

Multi objective Optimization Technique for Demand Side Management with Load Balancing Approach in Smart Grid in this paper [9] they used a multi-objective evolutionary procedure, which outcomes in the charge saving for power convention and decreases the waiting time for machine performance.

Development of Arduino Based IoT Metering System for On-Demand Energy Monitoring [10] discuss the implementation of an Internet of Things based Energy meter using an Arduino Uno, Hall Effect Sensor and a SIM 800 L Global System for Mobile communication (GSM) Component data transmission. To connect the internet and data relay using General Packet Radio Services (GPRS), GSM technologies in real-time. This structure was shown to be popular in energy consumption, measuring energy and moreover handling charges acquired by the consumer [11]. These metrics are interconnecting charges to the cloud server and power depletion.

Power Grid System Management through Smart Grid in India [12] in this research work concerted happening the inverter for system crossing point can effectively be assistance in the direction of accomplishing devolution of dynamic power acquired from the sustainable asset, stack quick to respond energy demand support, energy noises compensation next to PCC and energy unbalance and dispassionate present wage if around ought to be an existence of 3-stage 4-wire outline.

Implementation of Machine Learning Algorithm for Predicting User Behavior and Smart Energy Management [13-15] this paper explores the preeminent varieties of methodologies that have been inspected to solve the load disaggregation badly behaved, specifically, the feedback data for efficient and improved energy management.

Forecasting Residential Energy Consumption: Single Household Perspective [16] this paper investigates fifteen anonymous individual household’s electricity consumption forecasting using SVR (support vector regression) modeling approach is pragmatic to both daily and hourly data granularity.

#### 1. Title: CROP YIELD PREDICTION AND FERTILIZER RECOMMENDATION [18,21,22,23]

**Summary:** Agriculture is a major source of the economy of the country. In this paper, we attempt to provide a precise and accurate decision in predicting crop yield and deliver the end-user with proper recommendations about the required fertilizer ratio based on parameters such as atmospheric and soil parameters of the land which enhance the increase of crop yield and increase farmer revenue.

**Main contributions and strengths:** The proposed system is useful for the agriculture department to predict crop yield and to suggest the suitable fertilizers if yield is low. It is useful to farmers to know the crop yield and required fertilizers to improvise yield. In this proposed system there is no need to analyze manually. Two efficient algorithms have been employed. Naive Bayes Algorithm for Crop yield prediction. This algorithm provides us high accuracy and KNN Algorithm is used for Fertilizer Recommendation.

**Main Weaknesses:** The KNN makes no assumptions about the data, though data scaling is a must. The NB is scalable with large datasets, but does not work well if the training data is not representative of the population.

**Ideas for how to improve it:** the input data is pre-processed to find the missing values, eliminate redundant data, standardize the dataset, and convert target attributes into factor attributes.

**2. Title:** CROP YIELD PREDICTION AND EFFICIENT USE OF FERTILIZERS [37,40,44]

**Summary:** Yield prediction is a very important issue in agriculture. Any farmer is interested in knowing how much yield he is about to expect. Analysing the various related attributes like location, pH value from which alkalinity of the soil is determined. Along with it, percentage of nutrients likes Nitrogen (N), Phosphorous (P), and Potassium (K)

**Main contributions and strengths:** with the use of various suitable machine learning algorithms helps to create a model. The structure comes with a model to be precise and accurate in predicting crop yield and deliver the end user with proper recommendations about required fertilizer ratio based on atmospheric and soil parameters of the land which enhance to increase the crop yield and increase farmer revenue.

**Main Weaknesses:** The KNN makes no assumptions about the data, though data scaling is a must.

**Ideas for how to improve it:** The optimized attributes need to have classification techniques applied to them, prior to which the dataset is split into training and testing phases.

**3. Title:** “CROP YIELD PREDICTION AND EFFICIENT USE OF FERTILIZERS” [24, 28]

**Summary:** India being an agriculture country, its economy predominantly depends on agriculture yield growth and agroindustry products. Data Mining is an emerging research field in crop yield analysis. Yield prediction is a very important issue in agriculture. Any farmer is interested in knowing how much yield he is about to expect.

**Main contributions and strengths:** They take various data from the previous years to estimate future data. They used SMO classifiers in WEKA to classify the results. The main factors that take into consideration are minimum temperature, Maximum temperature, average temperature, and previous year’s crop information and yield information. Using the SMO tool, they classified the previous data into two classes that are high yield and low yield.

**Main Weaknesses:** SMO classifier gives less accuracy, Ideas for how to improve it: crop yield prediction with Naïve Bayes and Bayesian network gives high accuracy when compared to SMO classifier and forecasting the crop yield prediction in different climate and cropping scenarios.

**4. Title:** PREDICTION OF CROP YIELD AND FERTILIZER RECOMMENDATION USING MACHINE LEARNING ALGORITHMS [37-40]

**Summary:** This paper proposes and implements a system to predict crop yield from previous data. This is achieved by applying machine learning algorithms like Support Vector Machine and Random Forest on agriculture data and recommends fertilizer suitable for every particular crop. The paper focuses on creation of a prediction model which may be used for future prediction of crop yield. It presents a brief analysis of crop yield prediction using machine learning techniques.

**Main contributions and strengths:** SVM calculation has a regularization parameter, which stays away from over-fitting. SVM calculation utilizes the portion trap, so you can construct master learning about the issue. The random forest algorithm is not biased, since there are multiple trees and each tree is trained on a subset of data. Random Forest algorithm is stable if a new data point is introduced in the dataset the overall algorithm is not affected.

**Main Weaknesses:** The computational burden has to be reasonable thus very limited to on range.

**Ideas for how to improve it:** the mappings are utilized by the SVM plan to guarantee the tiny items will be figured as far as the variable in the first degree, for that a bit capacity  $k(x, y)$  chosen to get the ideal computational time.

**5. Title:** PREDICTING YIELD OF THE CROP USING MACHINE LEARNING ALGORITHMS [41-47]

**Summary:** This paper uses R programming with Machine Learning techniques. R is the leading tool for statistics, data analysis, and machine learning. It is more than a statistical package; it’s a programming language, so you can create your own objects, functions, and packages. It’s platform independent, so it can be used on any operating system and it’s free. R programs explicitly document the steps of our analysis and make it easy to reproduce and/or update analysis, which means it can quickly try many ideas and/or correct issues. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield were selected for the present research. The dataset contains the following parameters: rainfall, season, and temperature and crop production.

**Main contributions and strengths:** The Random Forest algorithm achieves a largest number of crop yield models with the lowest models. It is suitable for massive crop yield prediction in agricultural planning. The dataset used for modelling here includes the climatic factors as well i.e., rainfall and temperature.

**Main Weaknesses:** dataset includes very few attributes that would not give accurate predictions.

**Ideas for how to improve it:** Split the loaded data sets into two sets such as training data and test data in the split ratio of 67% and 33%. Then calculate Mean and Standard Deviation for needed tuples and then summarize the data sets. Compare the summarized data list and the original data sets & calculate the probability. Based on the result the largest probability produced is taken for prediction. The accuracy can be predicted by comparing the resultant class value with the test data set. The accuracy can range from 0% to 100%.

## 6. Title: HEURISTIC PREDICTION OF CROP YIELD USING MACHINE LEARNING TECHNIQUE

**Summary:** The paper says, vast research has been done and several attempts are made for application of Machine learning in agricultural fields. Major challenge in agriculture is to increase the production in the farm and deliver it to the end customers with best possible price and good quality. It is found that at least 50 percent of the farm produce never reaches the end consumer due to wastage and high-end prices. Machine learning based solutions developed to solve the difficulties faced by the farmers are being discussed in this work. The real time environmental parameters of Telangana District like soil moisture, temperature, rainfall, humidity are collected and crop yield is being predicted using KNN Algorithm.

**Main contributions and strengths:** This paper includes a comparative study of the KNN, SVM and Linear Regression giving KNN as the most appropriate one with maximum accuracy. The climatic as well as soil properties are analyzed to predict the yield. This paper does not include recommendation of fertilizers or crops based on the soil, climate and location.

**Main Weaknesses:** An element is distinguished by a larger proportion of its neighbours vote, with the entity being divided among its closest neighbours to the most regular class.

**Ideas for how to improve it:** In future, providing other factors that greatly influence the crop yield is our concern, also more data of all these parameters of different seasons in the state will be added to make this model more accurate and efficient.

## IV. IMPLEMENTATION METHODOLOGY

The system uses machine learning to make predictions of the crop and Python as the programming language since Python has been accepted widely as a language for experimenting in the machine learning area. Machine learning uses historical data and information to gain experiences and generate a trained model by training it with the data. This model then makes output predictions [2]. The better the collection of dataset, the better will be the accuracy of the classifier. It has been observed that machine learning methods such as regression and classification perform better than various statistical models [37-48].

Crop production is completely dependent upon geographical factors such as soil chemical composition, rainfall, terrain, soil type, temperature etc. These factors play a major role in increasing crop yield. Also, market conditions affect the crop(s) to be grown to gain maximum benefit. We need to consider all the factors altogether to predict the yield. Hence, using Machine Learning techniques in the agricultural field, we build a system that uses machine learning to make predictions of the production of crops by studying the factors such as rainfall, temperature, area, season, etc.

Machine Learning is undeniably one of the most influential and powerful technologies in today's world. Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data [10]. This mass of data is useless; we analyze it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making. To learn the rules governing a phenomenon, machines have to go through a learning process, trying different rules and learning from how well they perform. Hence, why it's known as Machine Learning.

This trained model gives an insight about collection of live data and it builds the model which creates a interface of the user which gives the proper inputs. Firstly processing of data is done. When the previous step is finished, using the algorithm model prediction can be done. The test data is sent for prediction.

The given trained model will be put to test with random input values which will be based on accuracy values and mistakes generated while trail. Unless the problem is solved this process keeps on repeating.

Machine Learning depends heavily on data. It's the most crucial aspect that makes algorithm training possible. It uses historical data and information to gain experiences. The better the collection of the dataset, the better will be the accuracy. The first step is Data Collection. For this project, we require two datasets. One dataset is used for modelling the yield prediction algorithm and other one for predicting weather i.e. Average Rainfall and Average Temperature. These two parameters are predicted so as to be used as inputs for predicting the crop yield. The sources of our datasets are: <https://en.tutiempo.net/> for weather data and <https://www.kaggle.com/srinivas1/agriculture-crops-production-in-india> for crop yield data[49-50].

The yield prediction module dataset requires the following columns: State, District, Crop, Season, Average Temperature, Average Rainfall, Soil Type, Area and Production as these are the major factors that crops depend on. Production's the dependent variable or the class variable. There are eight independent variables and 1 dependent variable. We achieved this by merging the datasets. The datasets were merged taking the location as the common attribute in both. We are considering only two states here, Maharashtra & Karnataka as the suicide rates in farmers in these two States were found to be very high.

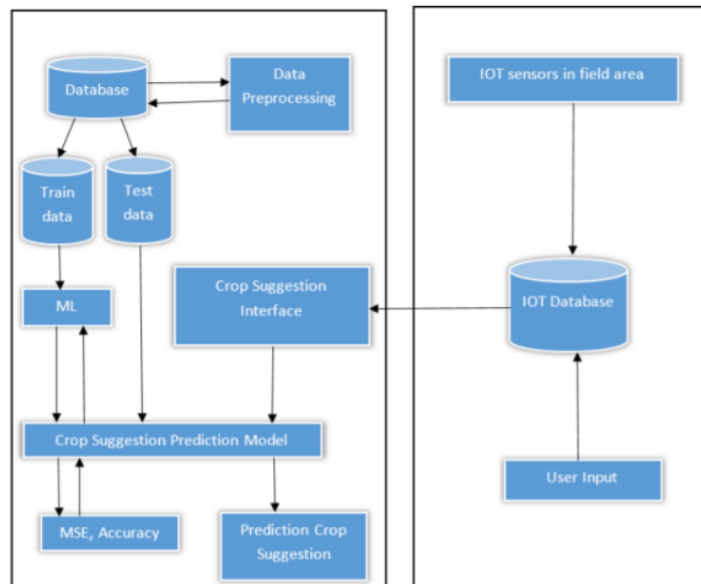
The implementation was divided into two categories .i.e. crop yield prediction and rainfall prediction (for fertilizers module).

This model will give output as the predicted production of crops based on the user's input. If the user wants to know the production of a particular crop, the system takes the crop as the input as well. Else, it will give a list of crops along with their production as output.

These are the following steps of the algorithm implemented:

- **Step 1** : Choose the functionality i.e., crop prediction or yield prediction.
- **Step 2** : If the user chooses crop prediction:-
  - Take soil type and area as values.
  - These inputs are given as input to the random forest implementation in the backend and the corresponding predictions are returned.
  - The algorithm returns a list of crops along with their production predicted.
- **Step 3** : If the user chooses yield prediction:-
  - Take crop, soil type and area as inputs.
  - These values are given as input to the random forest implementation in the backend and the corresponding crop yield prediction is returned.
  - The algorithm returns the predicted production of the given crop.

Fig.1 shows that flowchart for the Crop Prediction in Agriculture proposed system. After pre-processing of data's are stored in the database. Later database have split the train data and test data separately. Train data and Test data is used to compute the accuracy of your model. Machine Learning will create a form to expect the result of assured measures. To measure if the model is good enough, so here used a method is called Train and Test.



**Fig.1.** Flowchart for Proposed System

The first method is to collection of Data from the Field Area. The main parameters need to be included Such As Moisture, Temperature, Humidity the data which is being collected and then it is Stored and gives input To GUI. The Content of the water soil contains will be counted by the usage of Moisture Sensor where the soil is being utilized. This specific parameter is essential.

The implementation of the system can be divided into two, i.e., frontend and backend implementation. The frontend is implemented using the ionic development tools. Ionic Framework is an open source UI toolkit for building performance, high-quality mobile and desktop apps using web technologies — HTML, CSS, and JavaScript — with integrations for popular frameworks like Angular and React. Ionic Framework focuses on the frontend UX and UI interaction of an app — UI controls, interactions, gestures, animations. It integrates with other libraries or frameworks, such as Angular, React, or Vue and thus can be used on any platform.

Ionic is the only mobile app stack that enables web developers to build apps for all major app stores and the mobile web from a single codebase. And with Adaptive Styling, Ionic apps look and feel at home on every device. Thus, the performance of the system is enhanced. Ionic is built to perform and behave great on the latest mobile devices with best practices like efficient hardware accelerated transitions, and touch-optimized gestures. Ionic is designed to work and display beautifully on all current mobile devices and platforms. With ready-made components, typography, and a gorgeous (yet extensible) base theme that adapts to each platform, you'll be building in style. Ionic emulates native app UI guidelines and uses native SDKs, bringing the UI standards and device features of native apps together with the full power and flexibility of the open web. Ionic uses Capacitor (or Cordova) to deploy natively, or runs in the browser as a Progressive Web App. Thus, our system is web optimized.

The system can be built and deployed across multiple platforms, such as native iOS, android, desktop, and the web as a Progressive Web App - all with one code base. We use the Ionic CLI to perform cloud builds and deployments, and administer our account.

To build the frontend resources use npm. npm is the world's largest Software Registry. The registry contains over 800,000 code packages. Open-source developers use npm to share software. Many organizations also use npm to manage private development. To build the system and install ionic we execute the following command:

```
npm install -g @ionic/cli --save
```

The data needed for the system or resources is hosted on firebase. Firebase Storage provides secure file uploads and downloads for Firebase apps, regardless of network quality, to be used for storing images, audio, video, or other user-generated content. All the files will be supported by firebase.

We can add the platforms where we want our app to run using the following command:

```
$ cordova platform add ios
```

```
$ cordova platform add android
```

To build the app we run the following command:

```
$ cordova build
```

On successful build we can run the application on the configured host and use

```
#include <WiFi.h>
#include <FirebaseESP32.h>
#include "DHT.h"
#define DHTPIN 13
#define WaterSensorPIN 15
#define DHTTYPE DHT11
DHT dht(DHTPIN, DHTTYPE);
#define FIREBASE_HOST "mytestproject-e7991-default-rtdb.firebaseio.com"
//Do not include https:// in FIREBASE_HOST
#define FIREBASE_AUTH "vufAoDoMbG044iSLuycBKIUcZVQCUZZ0qRTtxRoJ"
#define WIFI_SSID "rama"
#define WIFI_PASSWORD "mockingjay"
FirebaseData firebaseData; //Define FirebaseESP32 data object

void setup() {
  // put your setup code here, to run once:
  Serial.begin(115200);
  pinMode(WaterSensorPIN,INPUT);
  dht.begin();
  Connect_Wifi();
}

void loop() {
  // put your main code here, to run repeatedly:
  Serial.println("Sensor details ");
  float th = dht.readHumidity();
  float tt = dht.readTemperature();
  int ss = digitalRead(WaterSensorPIN);
  Serial.print("Temperature : ");
  Serial.println(tt);
  Serial.print("Humidity : ");
  Serial.println(th);
  Serial.print("Water Sensor : ");
  Serial.println(ss);
  Firebase.setString(firebaseData, "/temperature",String(tt));
  Firebase.setString(firebaseData, "/humidity",String(th));
  Firebase.setString(firebaseData, "/Water",String(ss));
  delay(2100);
}
```



```

}
void Connect_Wifi()
{
  WiFi.begin(WIFI_SSID, WIFI_PASSWORD);
  Serial.print("Connecting to Wi-Fi");
  while (WiFi.status() != WL_CONNECTED)
  {
    Serial.print(".");
    delay(360);
  }
  Serial.println();
  Serial.print("Connected with IP: ");
  Serial.println(WiFi.localIP());
  Serial.println();

  Firebase.begin(FIREBASE_HOST, FIREBASE_AUTH);
  Firebase.reconnectWiFi(true);

  //Set database read timeout to 1 minute (max 15 minutes)
  Firebase.setReadTimeout(firebaseData, 1000 * 60);
  //tiny, small, medium, large and unlimited.
  //Size and its write timeout e.g. tiny (1s), small (10s), medium (30s) and large (60s).
  Firebase.setwriteSizeLimit(firebaseData, "tiny");
}

```

## V. PROPOSED METHODOLOGY

In the current climatic conditions are vary often. So the method is difficult to extend flora with the necessary aid of the use of climate condition. The main result of crop yield mainly is dependent on certain parameters such as variety of crop, seed types and environmental parameters which is affecting such as sunlight, soil, water, rainfall and humidity. With the information of the soil and atmosphere at specific region the best crop will have more crop yield; the exact crop yield can be predicted accordingly. This specific Temperature device is developed, to get the details of The Temperature of the Soil. Mainly levels of humidity is Used To consider Air Temperature and Moisture of the given soil. For obtaining the required data From the Sensors, Microcontroller is mainly used which is mainly used For Collecting data from the detector. The stored data From The device Is automatically saved Using wireless connection.

We propose system for Smart Management of Crop Cultivation using Machine Learning– a smart system that can assist farmers in crop management by considering sensed parameters (temperature, humidity) and other parameters (soil type, location of farm, rainfall) that predicts the most suitable crop to grow in that environment. Fig.2 shows that proposed model for collecting sensor data values and store it to the database.

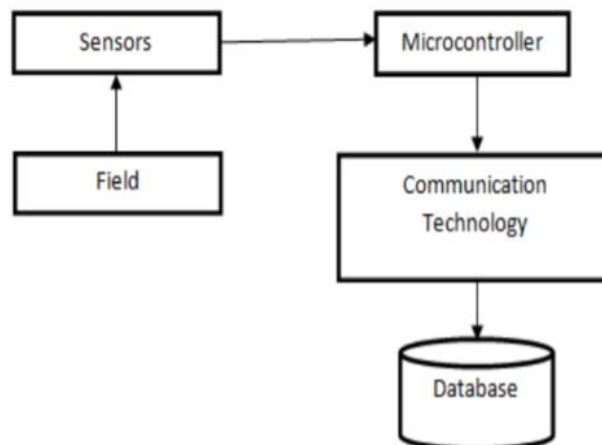


Fig.2. Model for Proposed System

The main factor for cultivation of plants is nutrients required. Usually soil will have diverse category of nutrient which are

helpful. Soil will be of assistance in preserve the water. This gives appropriate nutrient to the respective soil. This helps in proper understanding of usage of humidity content for the exacting soil which is used. Using heat sensor, the hotness of the soil is always being calculated it gives an indication of the optimal temperature for plant growth which is needed.

## VI. SENSORS USED

Arduino controller is also equipment where programming takes place. This makes a note of a huge innovative idea and network at a low price, which extends its usage with on point innovation. This model is a motherboard for making a relationship with different articles and reasonable computer programming IDE. Fig.3 shows that sample DHT11 Sensor component. Fig.4 shows that humidity Sensor component.

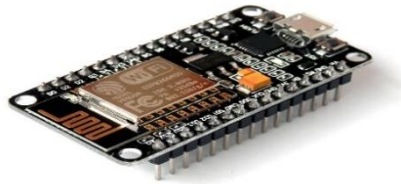


Fig.3. DHT11 SENSOR

The usage of this Sensor takes care of an application termed as FIRE BASE CONTROL using Arduino sensor tool, which acts as a Ethernet Shield to all of its libraries which are enabled. The User can login as it is safe over cloud platform to take care of the levels of humidity at the present location.. The controller controls the mashups caused while setting up the system. This enables the program and gives request to each line which makes the server turn on. Clients can easily login to this web server. The main usage of this Sensor gives an way to detect moistness with computerized temperature and it shows how damp this module is. This makes to justify that the module has been taken care off or flag yielding of the condition and the other sensors. The device which is shown below contains a capacitive sensor, wet parts and a high precision condition which gives accurate results, estimation equipment, which is linked with a known device. This sensor has got great quality, and it has quick reaction and can be done at some provided expense.

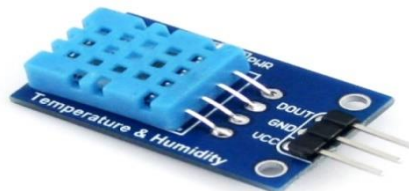
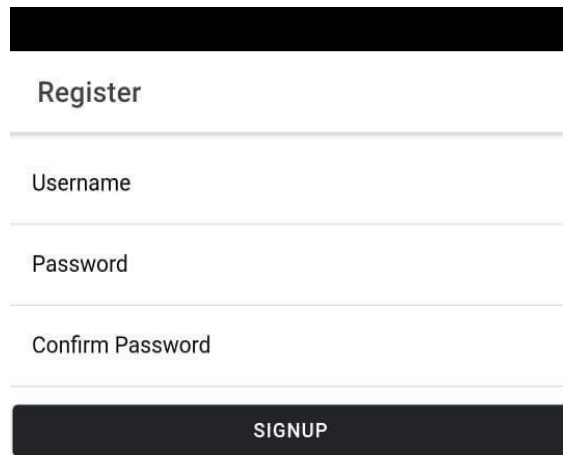


Fig.4. Humidity Sensor

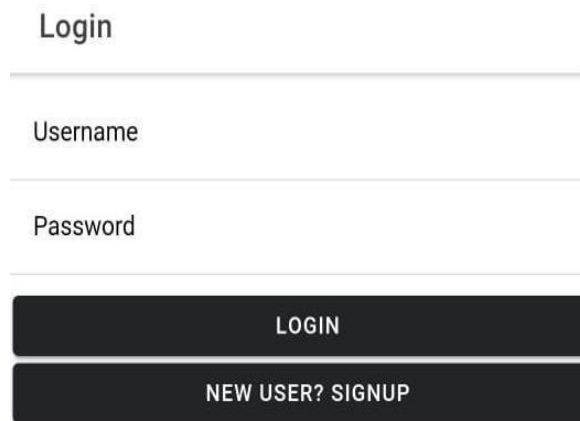
## VII. PERFORMANCE ANALYSIS & RESULT

At the implementation of this application the opening screen is user can view their login page. The user can register their details or login using their credentials into this application as seen in below mentioned the Fig. 5 and Fig. 6.



The Register screen features a black header bar at the top. Below it, the word "Register" is centered. There are three input fields: "Username", "Password", and "Confirm Password", each with a horizontal line below it. At the bottom, there is a black button with the text "SIGNUP" in white.

Fig. 5. Register Screen

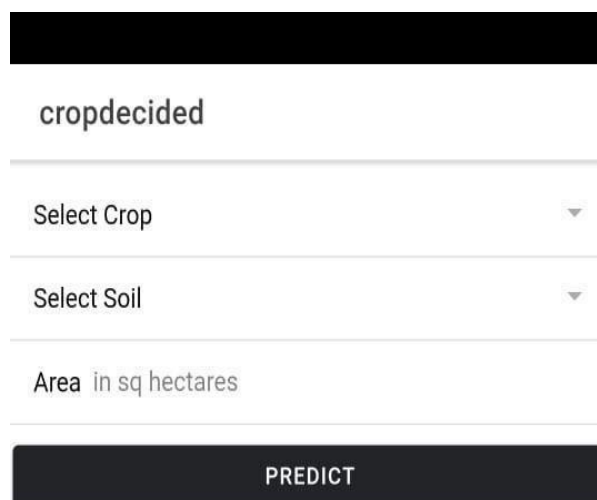


The Login screen features a black header bar at the top. Below it, the word "Login" is centered. There are two input fields: "Username" and "Password", each with a horizontal line below it. At the bottom, there are two black buttons: the top one says "LOGIN" and the bottom one says "NEW USER? SIGNUP", both in white text.

Fig. 6. Login Screen

This model classify with three major features:

- i) **Yield Prediction:** This application obtain the required values to guess the yield of the given crop. The required values to be given like our crop type, soil type and area as shown in the Fig. 7. The system returns a screen with the predicted yield as seen in the Fig. 8.



The Yield Prediction screen features a black header bar at the top. Below it, the text "cropdecided" is centered. There are three input fields: "Select Crop", "Select Soil", and "Area in sq hectares", each with a horizontal line below it. At the bottom, there is a black button with the text "PREDICT" in white.

Fig. 7. Yield Prediction Screen



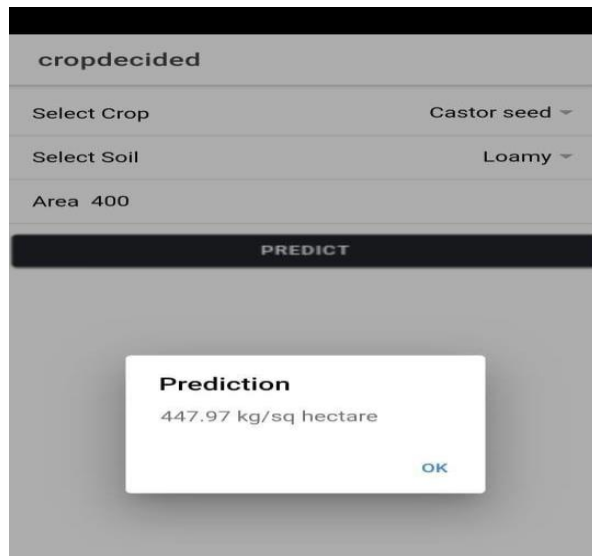


Fig. 8. Yield Predicted Screen

ii) **Crop Prediction:** This part of the system capture the required inputs i.e., soil type and area type in the Fig. 9. This application returns a screen with the list of crops with their

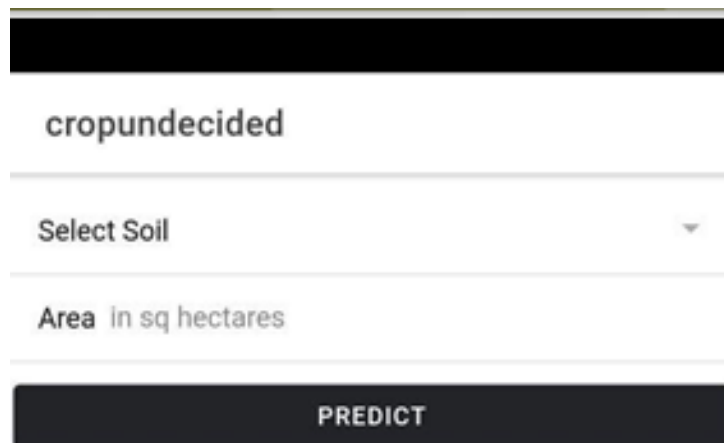
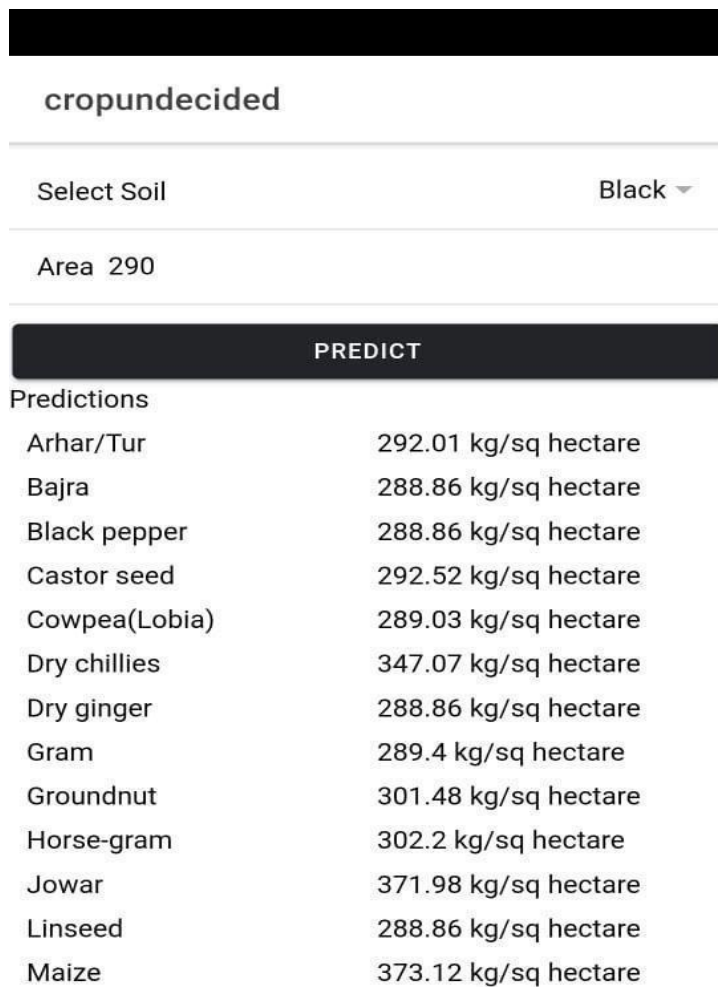


Fig.9. Crop Prediction Screen



**Fig.10.** Crops and their productions predicted

Given format of the code is being stored in the software application and the MCU board which is been changed to work properly. The main outlook of Firebase a specific profile will be created then the Data base will be created instantly. With the information we have a data base is created. After that threshold values will be inserted according to the corresponding factors of temperature and humidity.

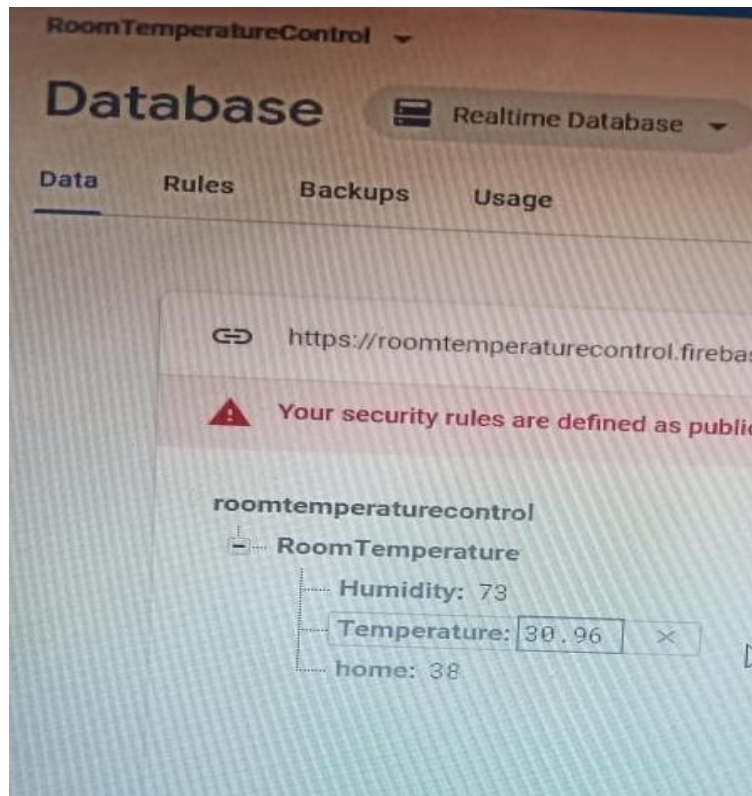


Fig.11. Display in Database

In the previous talk we had the utility which directs us to temperature factor and humidity factor is shown accordingly to what the work has been done. The information which can be seen in the project it helps us to know at what specific time dew point goes above the limit which is present.



Fig.12. Data reached to the smart-phone

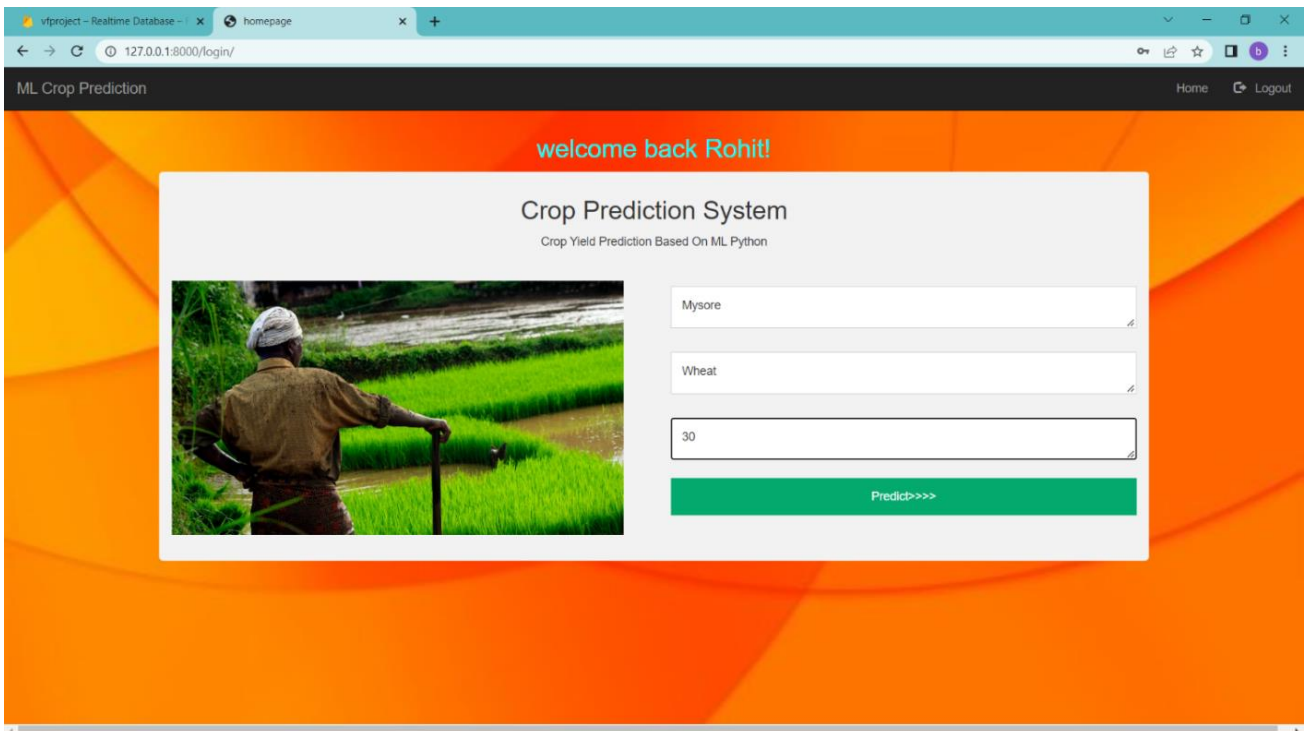


Fig.13. Enter the crop data details

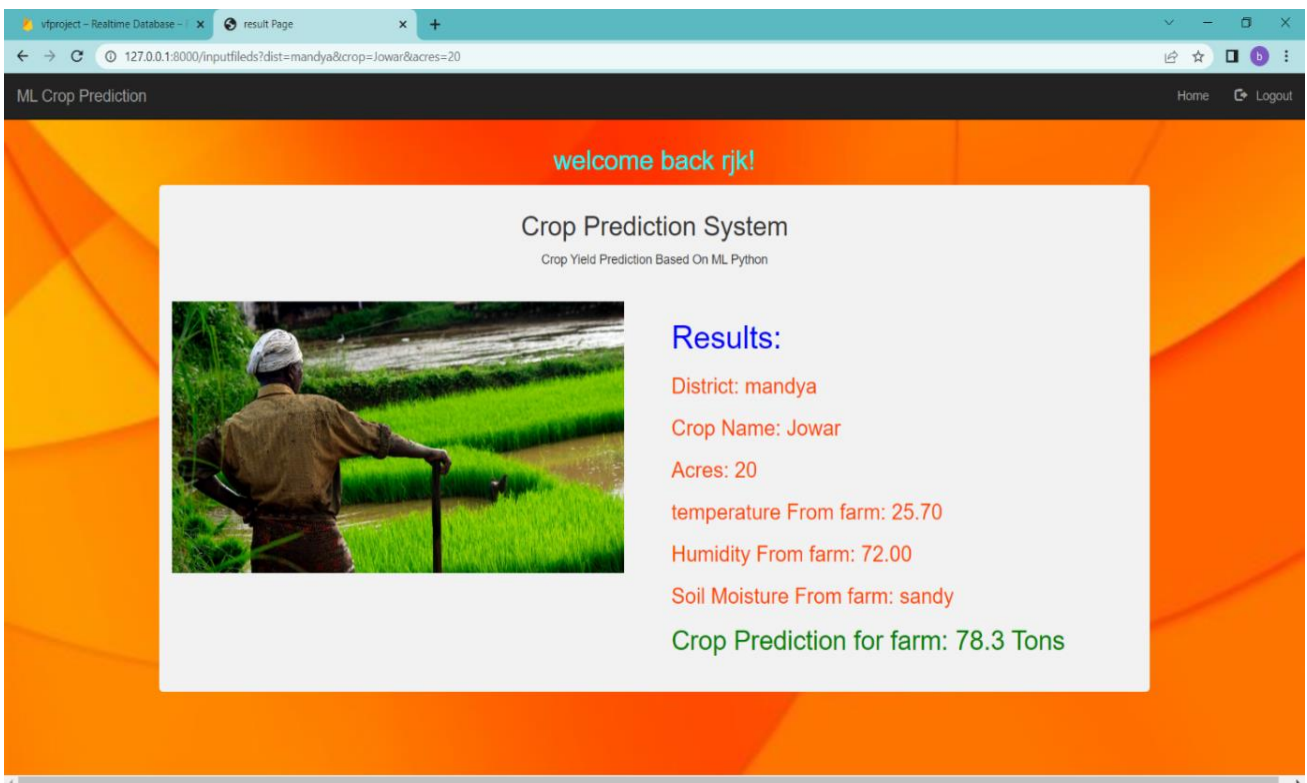


Fig.14. Crop detail result

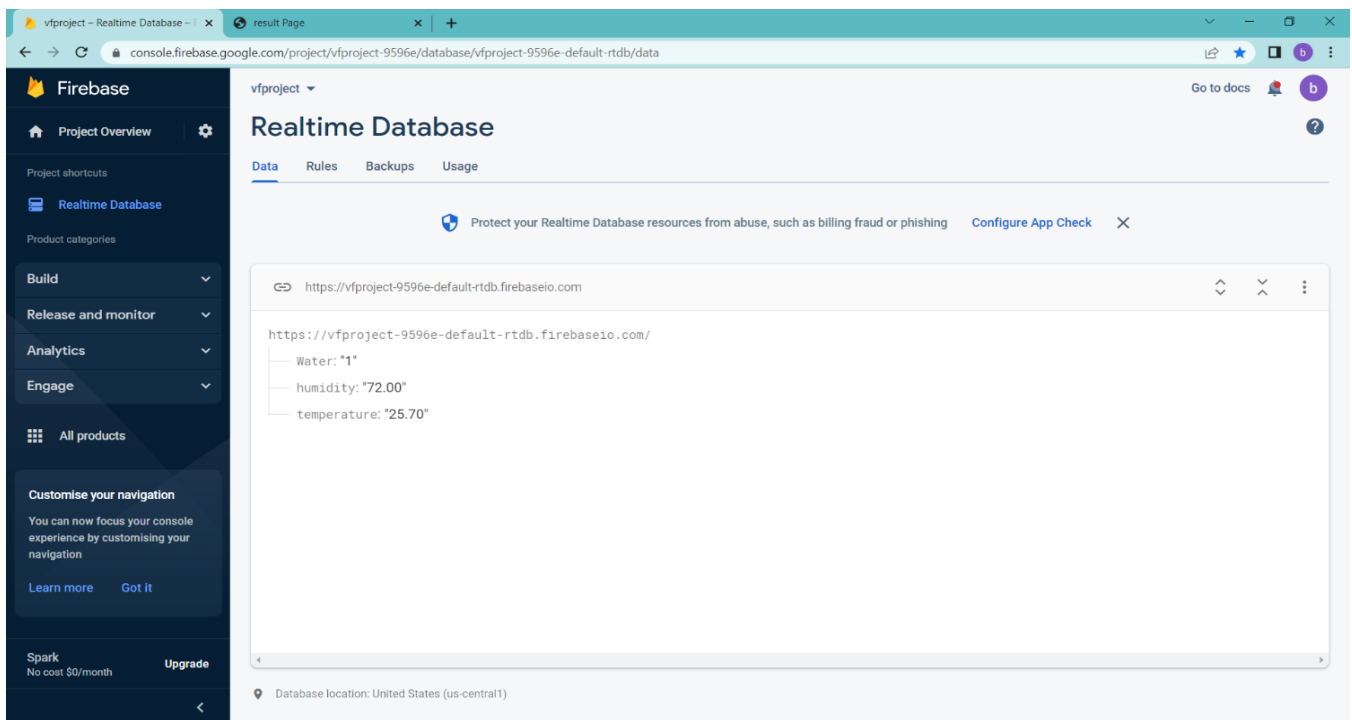


Fig.15. Cloud platform result

## VIII. CONCLUSION

The proposed system identifies the given framework which gives an effective approach of the entire framework for observing the parameters. The next step can take place. Mainly this observation of field just not only enables client to have less burden of work, it even allows user to examine properly the proper changes to take place in order for the use of environmental purpose and for developing a correct decision. The farmer is given details of the of a crop considering the important factors such as land , rainfall, temperature and district using this model techniques. It also predicts the future market price of crops which are present now by visiting previous crop price and predicted yield data is duly noted.

This is mainly proposed to arrangement with the decreasing rate of farmer suicides and to help them to grow financially stronger. The yield recommender system helps the crop cultivator to predict the yield of a given crop and also helps them to make a decision which crop to grow. Moreover, it also tells the user the right time to use the fertiliser.

Proper datasets were used, studied thoroughly, tested and trained using different machine learning tools. This system tracks the user's location and fetches needed information from the backend based on the location. Thus, the user needs to provide limited information like the soil type and area.

This system contributes to the field of agriculture. Mainly important and novel contributions of the method is recommending to the user at the correct time to use the fertiliser, this is done by predicting the weather of the next 14 days. Also, the system provides a list of crops with their productions based on the climatic conditions.

## IX. FUTURE WORK

The future work is focused on providing the sequence of crops to be grown depending on the soil and weather conditions and to update the datasets time to time to produce accurate predictions. The Future Work targets a fully automated system that will do the same. Another functionality that we are trying to implement is to provide the correct fertiliser for the given crop and location. To implement this through study of fertilisers and their relationship with soil and climate is required. We are also aiming to predict the crisis situation in advance like the recent hike of onion prices.

The prediction of crop yield based on location and proper implementation of algorithms have proved that the higher crop yield can be achieved. From above work that conclude for soil classification Random Forest is good with accuracy 86.35% compare to Support Vector Machine. For crop yield prediction Support Vector Machine is good with accuracy 99.47% compare to Random Forest algorithm. The work can be extended further to add following functionality. Mobile application can be build to help farmers by uploading image of farms. Crop diseases detection using image processing in which user get pesticides based on disease images.

## ACKNOWLEDGEMENT

We thank our college for providing us with resources to bring this research and survey, and faculty of Computer science and engineering department for providing insight and expertise. We take this chance to recognize the encouragement & support from our family and friends

## REFERENCES

- [1] Akash Raj N, Balaji Srinivasan, Deepit Abhishek D, Sarath Jeyavanth J, Vinith Kannan A, "IoT based Agro Automation System using Machine Learning Algorithms", International Journal of Innovative Research in Science, Engineering and Technology November 2016, pp. 19938-19342
- [2] Anita, Priscilla, Mary, M., & Josephine, M.S. (2018), Analysis and Forecasting Of Electrical Energy a Literature Review. International Journal of Pure and Applied Mathematics, 119(15), 289-293.
- [3] Audun, Josang. & Jochen, Haller. (2007, April). Dirichlet Reputation Systems, Paper presented at the Second International Conference on Availability, Reliability and Security (ARES'07), Vienna, Austria
- [4] Basumatary, Jwngsar., Pratap, Singh, Brijendra., Gore, M. M. (2018, January). Demand Side Management of a University Load in Smart Grid Environment, Paper presented at the Workshops ICDCN '18, Varanasi, India.
- [5] Hao, Hu., Rongxing, Lu., Zonghua, Zhang. (2015, December). Vtrust: A robust trust framework for relay selection in hybrid vehicular communications, IEEE Global Communications Conference, GLOBECOM 2015, San Diego, CA, USA.
- [6] Hlaing, Win., Thepphaeng, Somchai., Nontaboot, Varunyou., Tangsun, Natthan., Sangsuwan, Tanayoot., Chaiyod, Pira. (2017, March). Implementation of WiFi-based single phase smart meter for Internet of Things (IoT), International Electrical Engineering Congress (iEECON), Pattaya, Thailand
- [7] Fumo, Nelson., & Biswas, Rafe, M.A. (2015). Regression analysis for prediction of residential energy consumption. Elsevier Renewable and Sustainable Energy Reviews, 7(47), 332-343.
- [8] Mhadhbi, Zeineb., Zairi, Sajeh., Gueguen, Cedric., Zouari, Belhassen. (2018). Validation of a Distributed Energy Management Approach for Smart Grid Based on a Generic Colored Petri Nets Model, Journal of Clean Energy Technologies, 6(1), 20-25.
- [9] Muralitharan, K., Sakthivel, R., Shi, Y. (2015). Multiobjective Optimization Technique for Demand Side Management with Load Balancing Approach in Smart Grid, Elsevier Neurocomputing, 177, 110-119.
- [10] Okafor, K.C., Ononiwu, G.C., & Precious, U. (2017). Development of Arduino Based IoT Metering System for On-Demand Energy Monitoring. International Journal of Mechatronics, Electrical and Computer Technology, 7(23), 3208-3224.
- [11] Rajeshwari, Sundar., Santhosh, Hebbar., Varaprasad, Golla. (2015). Implementing Intelligent Traffic Control System for Congestion Control, Ambulance Clearance, and Stolen Vehicle Detection, IEEE Sensors Journal, 15(2), 1109 – 1113.
- [12] Rajput, Rashika., & Gupta, Amit. (2018). Power Grid System Management through Smart Grid in India. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 5(1), 17-26.
- [13] Ramanan, Rajasekaran, G., Manikandaraj, S., Kamaleshwar, R. (2017, February). Implementation of Machine Learning Algorithm for Predicting User Behavior and Smart Energy Management, International Conference on Data Management, Analytics and Innovation, Pune, India
- [14] Rashmi, Hegde., Rohith, Sali, R., Indira, M. S. (2013). RFID and GPS based automatic lane clearance system for ambulance, International Journal of Advanced Electrical and Electronics Engineering, (IJAE), 2(3), 102–107.
- [15] Vignesh, G., Vishal, Narayanan., Prakash, S., Sivakumar, V. (2016, May). Automated Traffic Light Control System and Stolen Vehicle Detection, 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India. URL: <http://ieeexplore.ieee.org/document/7808101/>
- [16] Zhang, Monica, Xiaoou., Grolinger, Katarina., Capretz, Miriam, A.M. (2018, December). Forecasting Residential Energy Consumption: Single Household Perspective, 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA.
- [17] Niketa Gandhi et al. (2016), "Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques", IEEE International Conference on Advances in Computer Applications (ICACA).
- [18] K.E. Eswari. L. Vinitha. (2018) "Crop Yield Prediction in Tamil Nadu Using Bayesian Network", International Journal of Intellectual Advancements and Research in Engineering Computations, Vol-6, Issue-2, ISSN: 23482079.
- [19] V. Sivakumar, R Swathi, Yuvaraj., "An IoT-Based Energy Meter for Energy Level Monitoring, Predicting" "Challenges and Opportunities for the Convergence of IoT, Big Data, and Cloud Computing", IGI Publisher, Chapter No. 4, Pages 48-65, 2021. DOI: 10.4018/978-1-7998-3111-2.ch004 or <https://www.igi-global.com/chapter/an-iot-based-energy-meter-for-energy-level-monitoring-predicting-and-optimization/269556>
- [20] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, Sonali Idate. (2018) "Use of Data Mining in Crop Yield Prediction" IEEE Xplore ISBN: 978-1-5386-0807-4; Part Number: CFP18J06.
- [21] Anna Chlingaryana, Salah Sukkarieha, Brett Whelan (2018) — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture 151 61–69, Elsevier.
- [22] Dakshayini Patil et al (2017), "Rice Crop Yield Prediction using Data Mining Techniques: An Overview", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5.
- [23] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh" 978-1-4799-86767, IEEE SNPD.
- [24] Snehal S. Dahikar, Dr. Sandeep V. Rode (2014), "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach", International journal of innovative and research in electrical, instrumentation and control engineering, volume 2, Issue 2.
- [25] Sivakumar Venu; A. M. J. Md. Zubair Rahman, "Effective Routine Analysis in MANET's Over FAODV" 2017 IEEE International Conference on "Power, Control, Signals and Instrumentation Engineering (ICPSCI)", ISBN: 978-1-5386-0813-5 on 21st & 22nd Sep 2017 Published in IEEE Conference publications, page no. 2016-2020 <https://ieeexplore.ieee.org/document/8392068/>
- [26] Sivakumar Venu, Zubair Rahman, "Energy and cluster based efficient routing for broadcasting in mobile ad hoc networks", Springer Cluster Computing, 2018, Vol. 22, pp. 661-671. <https://doi.org/10.1007/s10586-018-2255-3>
- [27] Dr. V. Sivakumar, Bakkachenna Ranadeep, Swathi, "IoT enabled Agriculture in Smart Drip Irrigation System" in Grenze International Journal of Engineering and Technology (GIJET), ISSN: 2395-5295, 2022 January, Volume no: 8, Issue No: 1, Page No: 581-586 URL: <http://thegrenze.com/index.php?display=page&view=journalabstract&absid=1082&id=8> OR <http://thegrenze.com/pages/servej.php?fn=70.pdf&name=IOT%20Enabled%20Agriculture%20in%20Smart%20Drip%20IrrigationSystem&id=1082&association=GRENZE&journal=GIJET&year=2022&volume=8&issue=1>
- [28] Ramesh A. Medar (2014) "A survey on data mining techniques for crop yield prediction", International Journal of advance in computer science and management studies, ISSN: 2231-7782, volume 2, Issue 9.



- [29] S.kanaga Subba Raju et al.(2017), Demand based crop recommender system for farmers, International Conference on Technological Innovations in ICT For Agriculture and Rural Development.
- [30] Yadav, T. & Reddy, Dr & Prasad, Ram & Gopal, Pradeep. (2020). CROP YIELD AND FERTILIZERS PREDICTION USING DECISION TREE ALGORITHM. International Journal of Engineering Applied Sciences and Technology. 5. 187-193.
- [31] V. Sivakumar, Anburajan. M. N, Aravind. R, ArunPrasath. R, Muniyasamy. K, "Packet Loss Detection in MANETs Using Modified Fine Grained Approach" in International Journal of Management, Technology And Engineering, Volume 9, Issue 4, April 2019, ISSN NO : 2249-7455 DOI:16.10089.IJMTE.2019.V9I4.19.27093 OR <https://app.box.com/s/y4gOpt4yf07canj8xk07q0k88g2efjsd>
- [32] V.Sivakumar, J.Kanimozhi, B.Keerthana, R.Muthu lakshmi "Capacity Enhancement using Delay-Sensitive Protocol in MANETs", Springer Lecture Notes in Networks and Systems, "Inventive Communication and Computational Technologies" Proceedings of ICICCT 2019, entitled Volume 89, Pages 901-910, Publisher: Springer, Singapore, ISSN 2367-3370 <https://www.springer.com/series/15179>
- [33] Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R L, "Prediction of Crop Yield using Machine Learning," International Research Journal of Engineering and Technology (IRJET) Feb 2018, pp. 2237-2239
- [34] Radhika, Narendiran, "Kind of Crops and Small Plants Prediction using IoT with Machine Learning," International Journal of Computer & Mathematical Sciences April 2018, pp. 93-97
- [35] Shridhar Mhaiskar, Chinmay Patil, Piyush Wadhai, Aniket Patil, Vaishali Deshmukh, "A Survey on Predicting Suitable Crops for Cultivation Using IoT," International Journal of Innovative Research in Computer and Communication Engineering January 2017, pp. 318- 323
- [36] T Raghav Kumar, Bhagavatula Aiswarya, Aashish Suresh, Drishti Jain, Natesh Balaji, Varshini Sankaran, "Smart Management of Crop Cultivation using IOT and Machine Learning," International Research Journal of Engineering and Technology (IRJET) Nov 2018, pp. 845- 850
- [37] S. Bhanumathi, M. Vineeth and N. Rohit, "Crop Yield Prediction and Efficient use of Fertilizers," 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.
- [38] N. Gandhi and L. J. Armstrong, "Rice crop yield forecasting of tropical wet and dry climatic zone of India using data mining techniques," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 357-363. doi: 10.1109/ICACA.2016.7887981
- [39] S. Mishra, P. Paygude, S. Chaudhary and S. Idate, "Use of data mining in crop yield prediction," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2018, pp. 796-802. doi: 10.1109/ICISC.2018.8398908
- [40] S. Sahu, M. Chawla and N. Khare, "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 53-57. doi: 10.1109/CCAA.2017.8229770
- [41] Jig Han Jeong et al., "Random Forests for Global and Regional Crop Yield Predictions", PLOS-ONE, June 2016.
- [42] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data", Springer journal, 2017.
- [43] S. G L, N. V and S. U, "A Review on Prediction of Crop Yield using Machine Learning Techniques," 2022 IEEE Region 10 Symposium (TENSYPMP), Mumbai, India, 2022, pp. 1-5. doi: 10.1109/TENSYPMP54529.2022.9864482
- [44] J. R, H. D and P. B, "A Machine Learning-based Approach for Crop Yield Prediction and Fertilizer Recommendation," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1330-1334. doi: 10.1109/ICOEI53556.2022.9777230
- [45] D. Sharma and A. Sai Sabitha, "Identification of Influential Factors for Productivity and Sustainability of Crops Using Data Mining Techniques," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 322-328. doi: 10.1109/SPIN.2019.8711630
- [46] D S. Jambekar, S. Nema and Z. Saquib, "Prediction of Crop Production in India Using Data Mining Techniques," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5. doi: 10.1109/ICCUBEA.2018.8697446
- [47] U. Inyaem, "Construction Model Using Machine Learning Techniques for the Prediction of Rice Produce for Farmers," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 2018, pp. 870-874. doi: 10.1109/ICIVC.2018.8492883
- [48] M. Manjunatha and A. Parkavi, "Estimation of Arecanut Yield in Various Climatic Zones of Karnataka using Data Mining Technique: A Survey," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-4. doi: 10.1109/ICCTCT.2018.8551083
- [49] dataset are: <https://en.tutiempo.net/> for weather data
- [50] dataset are: <https://www.kaggle.com/srinivas1/agriculture-crops-production-in-india> for crop yield data.



# Chapter - 3

## Applications of IOT using Deep Learning

Dr. Anitha T N<sup>1</sup>, Dr. Jayasudha K<sup>2</sup>

<sup>1</sup>Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.

<sup>2</sup>Associate Professor, ISE Department, Atria Institute of Technology, Bengaluru, India.

Email: <sup>1</sup>[anitha.tn@atria.edu](mailto:anitha.tn@atria.edu), <sup>2</sup>[jayasudha.k@atria.edu](mailto:jayasudha.k@atria.edu)

*Abstract-Deep learning, a branch of machine learning and a branch of artificial intelligence, focuses on simulating the human brain in settings involving data collecting and processing. Because the neural networks used by deep learning to learn have many deep layers, the term "deep learning" has been coined. It employs a programmable neural network, which enables machines to decide correctly without the assistance of humans.*

*The network of physical objects (things) that are integrated with sensors, software, and other technologies for the purpose of communicating and exchanging data with other devices and systems over the internet is described by the term "Internet of Things" (IoT). IoT devices are pieces of hardware, such as sensors, actuators, gadgets, appliances, or machines, that may communicate data over the internet or over other networks and are designed for particular uses.*

*Artificial intelligence includes deep learning, whereas IoT refers to internet-connected technologies that link and exchange data with other systems and devices. IoT is rapidly expanding in the fields of science and engineering. On IOT devices, deep learning applications usually have strict real-time requirements. In order to support a new realm of interactions between people and their physical surroundings, this article offers a survey of deploying deep neural networks to Internet of Things (IoT) devices.*

*Keywords: Applications, Deep learning, Internet of Things, Smart*

### I. INTRODUCTION

For a range of wireless platforms, including smart phones, sensor networks, and unmanned aerial vehicles among others, applications for Internet of Things (IoT) technologies are currently being developed. It is essential to develop an application for IOT data analysis and employ conventional techniques. For many of the learning techniques employed in such systems, shallow architectures are often used, which have very limited modelling and representational capabilities. A more powerful analytical tool and deep learning are absolutely need to properly realize the potential of the precious raw data generated in various IoT applications.

Deep learning (DL) is a method of machine learning that trains computers to perform. Deep learning, an important method utilized in many applications, enables cars without drivers to recognize a stop signal or identify a lamp pole pedestrian. DL is a key technique in data science, that also covers prediction and statistics. This process is done faster and simpler by deep learning, that is highly useful for data scientists for collection, evaluation, and interpreting large quantity of data.

#### 1.1 Deep Learning in an IOT Applications

Deep learning has more potent powers in generalizing the complex relationship of large amounts of raw data in a variety of IOT applications. Deep learning is able to characterize for more complex situations and extract more sophisticated hidden features. For predicting weather events like earthquakes and tsunamis, rain DL is frequently used. It helps in maintaining the required safety measures. Deep learning allows machines to understand speech and produce the intended results. It permits the machines to identify the photos of individuals and things that are contributed to them. In addition, DL models help advertisers in real-time auction and advertising display targets.

### 1.2 Significance of Deep Learning

In contrast to Machine Learning (ML), which exclusively uses semi-structured and structured data sets, deep learning makes use of both organized and unstructured data.

- Deep learning algorithms are more powerful at handling challenging problems when compared to machine learning algorithms.
- Machine learning algorithms begin to slow down as size of data rises, suggesting the use of deep learning techniques to maintain model performance.
- Deep learning accepts huge quantities of information as load and analyses this load to extract properties out of an object, in contrast to ML algorithms that utilize annotated test data to extract designs, is the significance of DL.

## II. DEEP LEARNING APPLICATIONS

Deep Learning is a part of machine learning that aids in providing insightful solutions to challenging problems. Deep Learning is built on an understanding of the composition and function of the human brain. Deep learning uses artificial neural networks to analyze data and make predictions. It can be used in almost every sector of business. Many real-world applications of deep learning are described, some of which are depicted in figure 1.

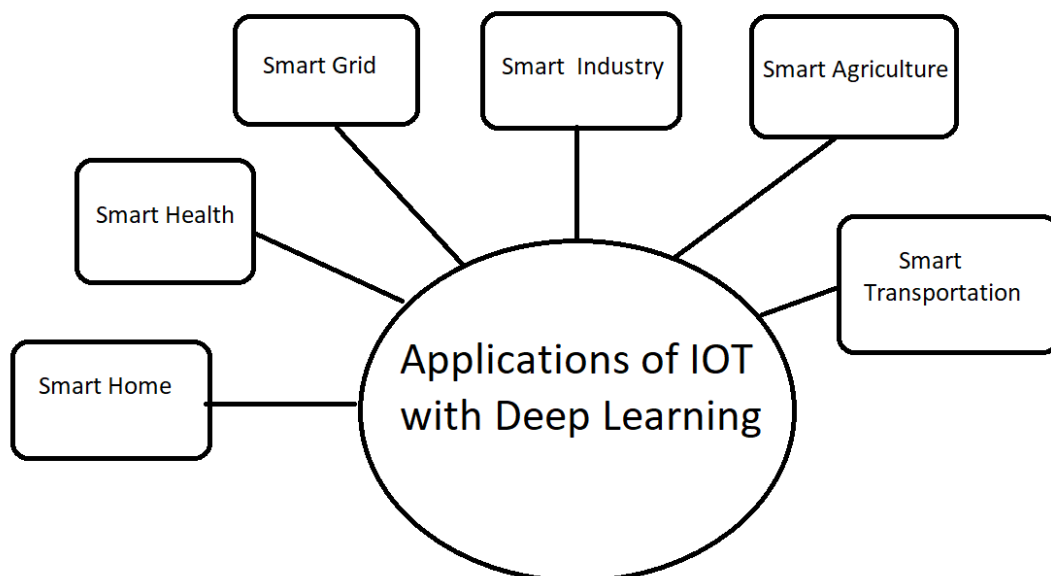


Figure 1: Applications of IOT with Deep Learning

### 2.1 Deep Learning Methods

It was formerly said that the model training phase of deep learning challenges was time-consuming and required significant computer resources. Due to the development of specialized hardware and well-organized training algorithms, deep models can be created instead of utilizing conventional techniques to evaluate complex problems and handle data.

### 2.2 UNSUPERVISED LEARNING

The two types of Deep Learning (DL) are models trained with labelled data and models learned with unlabeled data (unsupervised learning) (supervised learning).

### 2.3 RESTRICTED BOLTZMANN MACHINES

In order to handle vast amounts of unlabeled data, unsupervised learning must be used in addition to traditional learning methods. For back propagation, steady initialization, and global correction during training, Restricted Boltzmann Machine's (RBM) or stacked Auto Encoders (AE) can be utilized.

### 2.4 AUTOENCODER

An "auto encoder" is a neural network that can copy input to output (AE). An AE has three layers instead of the two seen in RBMs: the input layer, the hidden layer, and the output layer. The hidden layer's output is a reconstruction of the input that is explained by a code that corresponds to the input. The encoder function (f), which extracts the dependencies from the input, and the decoder function (g), which builds a reconstruction, are the two major components of the network.

## 2.5 SUPERVISED LEARNING

In contrast to unsupervised learning, supervised learning builds a system model using a labelled training set. The model comprehends the relationship between input, output, and system parameters. The main method used in supervised learning is the back propagation algorithm.

## 2.6 CONVOLUTIONAL NEURAL NETWORKS (CNNs)

A particular type of neural network known as a CNN (Convolution Neural Network) is utilized for processing data using a grid-like layout. The CNN building is depicted in figure 2. Convolution layers, pooling layers, and fully linked layers make up its three main layers. A 3D image file with dimensions of width, depth, and height serves as the input. A patch of output connects each layer to the one below it.

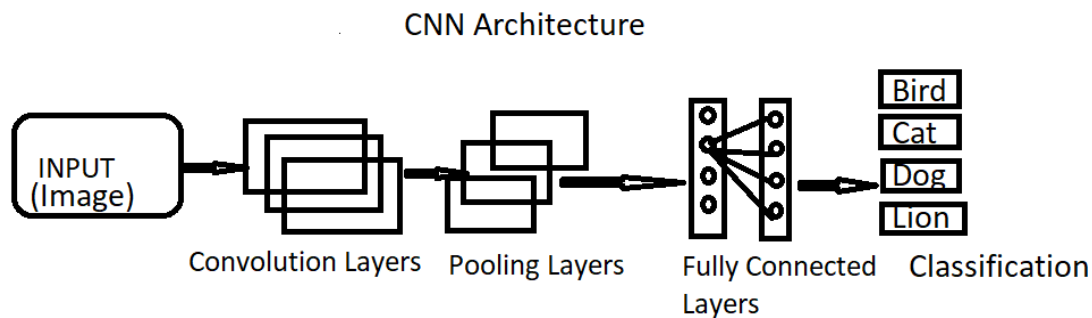


Figure 2: CNN Architecture

The only source of energy for CNNs is Receptive Field, a concept created through study of the cat visual brain. In order to help a machine learning system, perform better, CNN condenses three key ideas: sparse interactions, parameter sharing, and equip-variant representations. The heart of the CNN design consists of convolutional and pooling layers, which are optionally followed by a fully connected layer for classification or prediction. In comparison to traditional neural networks, CNNs efficiently reduce the amount of network parameters and the effect of the gradient diffusion problem, enabling the effective training of deep models with more than 10 layers.

## III. SMART HOME

### 3.1 Introduction

A "smart home" is a place where people can live that integrates bright control and management of different lifestyle systems through a network using cutting-edge network communication technology, computer technology and other combination of personal needs to make home life more comfortable and practical. The release of the Microsoft Kinect depth sensor in November 2010 enabled hardware support for the creation of smart homes based on attitude recognition technology [1]. A number of studies have been done on the technology and theory of the smart home, and literature [2–5] alludes to the associated design pattern and theory of somatosensory interaction. Sensory integrates a variety of smart home appliances utilizing wireless Bluetooth technology, allowing consumers to utilize a mobile phone client to instantly operate the appliances. In literature [6–7], wireless Bluetooth technology is used to connect various smart home appliances, allowing consumers to immediately control the appliances in actual time via a mobile phone client. The combination of smart grid and smart home was researched in literature [8], but it mostly focused on controlling home appliances, with no mention of interactive technology study. The somatosensory device utilized in this system as a model of Microsoft's Kinect somatosensory peripheral. The Kinect system consists of a color camera, an infrared device, a number of microphones, a motor, a logic circuit and other components. Kinect contains three cameras: a Red Green Blue (RGB) color camera in the center, infrared transmitters for the left-right lenses, and Complementary Metal Oxide Semiconductor (CMOS) infrared cameras with three-dimensional depth sensors. These cameras can capture both color and depth images simultaneously. Because Kinect extracts the point cloud image [9–10] from a dark room, as we've seen, the system can be employed late night or at night.

### 3.2 Methodology

Processing software and Kinect are integrated with a cloud-based attitude recognition control system for smart homes. The wireless communication module sends the decided data to the Arduino CPU. Different home appliances are managed via the XBee access technology.

The cameras on the Kinect can simultaneously capture color and depth images thanks to the RGB color camera in the center, infrared transmitters for the left and right lenses, and CMOS infrared cameras with 3D depth sensors. Because Kinect extracts the point cloud image from a dark room, the system can be used late night or at night.

The hardware component of the Arduino Uno is the Arduino Uno circuit board, and the other component is the Arduino Uno Integrated Development Environment (IDE), a programming environment that is saved on a computer. The 8 bit Advanced Technology for Memory and Logic (ATMEGA) 328 microcontroller that powers Arduino Uno, an open source hardware development platform, has 14 digital input and output ports as well as six analogue input pins. It is capable of supporting Universal Serial Bus (USB) data transfer.

An image and video recognition technology is called a convolutional neural network. In order to detect two-dimensional images, the CNN model generates 64 feature datasets. The trained dataset used to train and validate the convolution neural network model.

## IV. SMART HEALTH CARE

### 4.1 Introduction

Sensitive medical data is tracked and monitored by IoT-based healthcare systems by gathering data in real time. A deep learning application for IoT in the healthcare industry gathers unstructured data and applies machine learning methods to it. Classifying objects and movies is a capability of deep learning. Deep learning's hidden architecture assists in sifting through input data to identify relevant traits. The Internet of Things (IoT) is currently one of the innovative technologies employed in the healthcare sector. IoT tools and software gave the healthcare industry more intelligence.

### 4.2 Methodology

IoT's three tier design is utilized to gather real-time data. They gather a variety of data and transmit it to the closest edge server. The edge server has the ability to process data quickly and maybe remove noise from it. Deep learning enables the healthcare sector to examine data at extraordinary speed without sacrificing accuracy. It is a sophisticated blending of two systems, not a machine learning or artificial intelligence technique. Deep learning techniques are representation learning strategies built from basic non-linear models. The multi-layer deep learning algorithms perform classification tasks to find abnormalities in medical images, grouping patients with similar traits into cohorts based on risk from enormous amounts of unstructured data. Less data preprocessing is needed for deep learning.

The network itself handles a number of filtering and normalization tasks using machine learning. Deep learning shortens the supervision period and speeds up the dataset processing. The analysis of the Magnetic Resource Imaging (MRI) results is ideally suited for deep learning utilizing convolution neural networks. Applications for IoT-based health care employing DL are listed in Table 1.

TABLE 1: Summary of IoT based healthcare applications

Addressed Problem	DL Used	Sensor	Data set
Authentication and malware	MLP,	LSTM	N/A
Malicious traffic detection and data clustering	MLPLSTM	N/A	SPDSL-II, Waikato VIII, WIDE-18
Preventing security	DBN	N/A	N/A
Intrusion detection on IoT devices	DNN	N?A	Simulated Data

## V. SMART GRID

### 5.1 Introduction

A cutting-edge technology for the future and an intriguing topic of study is energy management for smart grids. In order to distribute electrical energy among various consumer groups, including smart businesses and smart residences, smart grids are the safe and dependable places to do so. This electrical energy is produced in power plants, distributed through smart grids, and used in either profit-driven structures and industrial sectors [11] or household consumption [12]. The quantity of energy generated at power plant and distributed among grids is utterly amazed by how it is used at the consumer level. The majority of consumers do not understand how to request energy from electric networks, which guarantees financial loss an There are numerous unsolved research problems as a result of a thorough analysis of the literature on workload prediction.

Acquiring accuracy in the forecasting perfection is the most crucial and difficult challenge while showcasing a new energy prediction technique. Implementation of the suggested algorithm across edge nodes, which results in relevant communication between connected devices in an Internet of Things (IoT) network for energy consumption, presents another significant problem that is little covered in the existing literature. Producers seek to get a higher level of energy formation while lowering the cost, therefore there are numerous unsolved research problems as a result of a thorough analysis of the literature on workload prediction.

Acquiring accuracy in the forecasting perfection is the most crucial and difficult challenge while showcasing a new energy prediction technique. Execution of the suggested algorithm across edge nodes, which results in relevant communication between connected devices in an Internet of Things (IoT) network for energy consumption, presents another significant

problem that is little covered in the existing literature.

In recent times, resource-constrained IoT domains have revealed top-level insights in a variety of fields, including video analytics [13], healthcare [14], and many others [15]. Along with these difficulties, a significant worry is the reduced time complexity of energy display methods, particularly when dealing with the issue of short-term load forecasting (STLF).

Future energy prediction literature will employ less of the cloud [16] and fog computing [17-20] prototypes, that are credible programs for big data analysis and instantaneous decision-making, acting as anomalous energy demand forecasting. Hence, a unique edge smart-based energy prediction platform for managing energy in smart grids is presented with the contributions listed below in order to tackle such problems effectively and appropriately in IoT networks.

- 1) Variations in energy demand are managed using a trustworthy edge intelligence-based new and flexible framework that guides energy consumers and producers with a comfortable platform for efficient transmission based on the algorithm's successful forecasts.
- 2) A system to make use of resource-constrained devices at various customer locations (smart homes or smart businesses), which are linked via an IoT network and under the supervision of a cloud server, in order to meet their present needs and predict their future ones. The correct amount of energy is sent through smart grids in response to home and commercial requests received from the cloud server, ensuring efficient energy management. Each request is filtered out by the cloud server in order to account on the enormous energy demands from consumers. It contains a storage device for energy predicting data that it features a storage device for energy predicting data that is used for in-depth study in the future.
- 3) We demonstrate our platform to be a paradigm for upcoming edge-intelligence-based energy predicting techniques based on our comprehensive experiments.

### 5.2 DL Methods for Load Prediction in Future

Scientists employed deep learning for energy forecasting [21] to get more accurate prediction results as it evolved in computer vision, security, IoT, healthcare, etc. Additionally, forecasting residential structures is the primary emphasis of DL approaches in the literature relating to energy prediction. The author of Article [22] describes a hybrid method for predicting the energy consumption of residential structures. They combined Long Short-Term Memory (LSTM) with DL and genetic algorithms to present a reduced objective function and hidden neurons. This method is tested using data from residential and commercial buildings to anticipate Village Social Transformation Foundation (VSTF), and the results outperform currently used traditional forecasting algorithms. In light of this, a recent article [23] constructed ensemble structures for Short Term Load Forecast (STLF) using wavelet neural networks and CNN, as well as LSTM and CNN. There is a vast amount of DL literature on energy prediction, with a primary focus on sequential data processing techniques.

The edge nodes are currently not altered by sequential learning methods with any appreciable exactness. The Pennsylvania New Jersey Maryland (PJM) dataset [24] is a well-known commercial dataset that we use to implement an energy forecast platform that is operational on resource-constrained devices to address this issue. It is gathered by PJM Interconnection LLC, a regional transmission company in the United States (PJM). PJM is a component of the Eastern Interconnection system, in charge of supplying energy to 14 various regions, like Illinois, Delaware, Indiana, and so on.

## VI. SMART INDUSTRY

### 6.1 Introduction

The term "smart industry" describes how industrial processes have advanced using IoT technologies to guarantee and maximize production activity. It covers a wide range of topics, including improving production, worker and product safety, logistics, maintenance, and quality control using smart manufacturing methods that are IoT enabled, sometimes referred to as Industrial IoT or IIoT. In fact, this makes use of sensor networks to gather production data, which is then transformed into insightful information utilizing cloud software and deep learning. Intelligent learning and processing abilities make this possible. Predictive analysis can be used by DL models to locate various production loss and industrial operations bottlenecks problems. Software that is DL-based is used to enhance product develop DL-enabled computer vision algorithms can help to ensure quality assurance. DL models are useful for studying the schedule of product deliveries and current driver information in the context of logistics. The following are just a few of the key DL applications in the smart industry:

- (i) Product creation
- (ii) Predictive analysis
- (iii) Maintenance
- (iv) Logistics
- (v) Robotics
- (vi) Supply Chain Management (SCM)

This concerns about how product design is applied in product development. This is accomplished by using DL-based software to enhance material and cost-related aspects of product design. Without employing pricey models, this software utilizes the most recent designs. To get ideal performance, this method is becoming more and more common and design (ex: racing cars). Products with flaws can be identified using application of predictive analysis is useful in forecasting outcomes based on historical data. Utilizing statistical and data modeling approaches, the historical data is examined. Large volumes of data are processed accurately, and different patterns are also recognized. To maintain the condition of industrial equipment, smart industries must do maintenance. DI in maintenance enables automatic fault detection and diagnosis, hence boosting component usage rates.

Logistics in smart industries aid in route optimization and hence lower shipping costs, increasing revenues. While reducing time, DL algorithms do indeed build datasets, gather data on the routes, and make quick conclusions. During the transportation of goods, logistics also considers atypical events that might occur.

Intelligent industries use robotics to enable autonomous cars navigate between production floors as they transport cargo. They are able to navigate more effectively and learn new things thanks to DLs with feedback systems. The use of intelligent robots to improve the concept of a smart factory is a big opportunity for small and medium-sized businesses.

In clever industries, Supply Chain Management (SCM) unquestionably boosts business. SCM's current status is that it lacks transparency, falls short of customer expectations, and has complicated operational issues. Five steps make up SCM: procurement and productivity. Deep learning algorithms are utilizing in a survey of IIoT applications by author Shahid Latifet.al [25] discusses upcoming difficulties as well. Deep learning techniques and use examples related to smart manufacturing, smart agriculture, and other topics are covered by author Rahul Amin [26]. A substantial part of a smarter production process is played by sensors and smart devices, according to Chen et al. in article [27]. IIoT solutions are being adopted by new industrialists in order to increase profitability and production. The authors Sadeghi, Perera, and Zheng state in articles [28], [29], and [30] that there is a strong alliance between stake holders and IIoT applications, and as a result, businesses from all over the world engage in the IIoT industry, which is anticipated to expand by 2030 with production, storage, logistics, and customer service. In terms of outlining a plan of action to address the problems, the DL algorithm in SCM shows promise. The use of algorithms enables SCM to join scalable and knowledgeable digital supply networks.

## 6.2 Methodology

By processing data samples, an intrusion detection system uses a deep learning technique called Auto Encoder (AE). It uses network traffic data and sensory data as datasets. Worker activity recognition, manufacturing monitoring, and inspection are all done using convolution neural networks (CNN). For conditional prediction, Deep Belief Network (DBN) is employed. For equipment analysis, long short-term memory (LSTM) is used. Utilizing sensory data are CNN, DBN, and LSTM. In the aviation business, a few DI algorithms are used. Foretelling potential risks in aviation systems, Recurrent Neural Networks (RNN) are employed. Choice by Pairs Business price estimation is done using Markov Choice (PCMC-Net). Flight delays are caused when trajectories deviate, anomaly detection is done using Denoising Auto Encoders (DAE).

# VII. SMART AGRICULTURE

## 7.1 Introduction

Farmers are looking forward to newest and latest technological advancements that increase crop production, with fewer expenses with effective usage of available resources. Smart agriculture is the process of converting traditional agriculture into smart agriculture using automation and IoT technologies. For smart farming different sensors are utilized to check humidity, temperature, light, soil moisture etc. for better monitoring of crop yield. Drone technology is also helpful in farming as it makes easy to scan and analyze the crops and with high quality images. It also gives an idea to farmers whether to harvest the crops or not. Smart agriculture concerns about the data collected at diversified sources like sensors, actuators, collecting geographical information via satellite position and monitoring agriculture in real time. It helps farmers to make right decisions at right time. Some of the applications of deep learning in smart agriculture are:

- (i) Classification of crops
- (ii) Identification of plant disease
- (iii) Prediction of crop yield
- (iv) Counting of fruits
- (v) Identification of weeds

Classification of crops in smart agriculture concerns with harvesting the crop at right time to avail market demands. DL makes harvesting easier by monitoring various like whether, humidity, soil type, scheduling of fertilizers etc. Added advantage is that farmers can observe their yield from any part of the world.

Identification of plant disease in smart agriculture deals with identification of fungus, bacteria and microbes that affects the



growth of the plant. If this is not recognized at early stages, it leads to a great loss to farmers. With the aid of DL algorithms, plant disease detection becomes easy.

Prediction of crop yield helps farmers to decide what to cultivate at what time. It paves the way for researchers to explore more innovative methods so as to predict suitable crop yield pertaining to varying seasons.

## 7.2 Methodology

The algorithms used in smart agriculture are: CNN, RNN, DBM and LSTM. LSTM is used for weather prediction by collecting dataset from syngeta. CNN is used for fruits counting, weed mapping, crop yield prediction, disease prediction of plants, soil segmentation and root segmentation. Datasets used are multirotor UAV and tomographic images of soil. RNN uses sensory data to record temperature details. DBN algorithm predicts soil and moisture contents.

Counting of fruits concerns with keeping count of ripened, not ripened and half ripened fruits. Such that farmers get a rough idea about their profit or loss about their yield. DL algorithm has made it easier with less effort compared to manual counting with labor cost.

Identification of weeds in smart agriculture is to help farmers by capturing thousands of weed images. Later they are separated from crops. This technique is already in practice in developed countries. Robots are playing a major role in identification of weeds.

Author Maha Altalak et.al [31] discusses most important techniques such as CNN and RNN, contributions in various fields leading to smart agriculture. Comparisons have been made with respect to datasets, methods and models. Author Disha Garg et.al [32] summarizes about various factors related to smart agriculture like water management, soil management, crop cultivation, crop disease, weeds removal, crop distribution etc. In paper [33], Kirtan Jha explains how ANN is used in smart agriculture with better usage and management of water. Authors Zhu et.al [34], Pouyanfar et.al [35] and Ajit et.al discusses the details of CNN algorithm and its architecture in smart agriculture. CNN is mainly used for image classification, fragmentation, voice recognition, object detection and such related functions.

## VIII. SMART TRANSPORTATION

### 8.1 Introduction

Smart transportation goals are oriented towards integrated application of modern technologies to provide different transport modes, traffic management, and incident management and speed control improvement ensuring safety to customers. Lot amount of research work can be focused on smart parking applications and smart lighting. If sensors are embedded into the vehicles, it is possible to optimize the routes, improve parking facilities, and prevent accidents with autonomous driving. There are plenty of opportunities for research in this domain. Sensors such as Radio Frequency Identification (RFID) cards, cameras and electronic gadgets can monitor the weather condition and help the people to manage daily life activities. Also, sensors can monitor air pollution in the cities. Few applications of deep learning in smart transportation are:

- (i) Traffic management
- (ii) Accident detection
- (iii) Prediction of car parking

Traffic management is one of the important applications in DL. It aids driver to select most feasible route thus efficiently manages transportation. It includes traffic flow, traffic speed and travelling time. One is used to predict the other feature.

Accident detection and prevention is required utmost to save human life in any type of cities. If the driver is alerted throughout the driving period, accidents can be avoided. Visual recognition tasks like obstacles can be detected using camera-based image systems. These aids autonomous driving of vehicles with good roadway infrastructure. Thus, it can detect forefront obstacles and leverages occurrence of accidents. Even highway roads are mounted with sensors that can send images of traffic congestion. It all depends on different parameters like weather, days of the week, geographical position etc.

In urban cities traffic is a major problem. Hence there is huge demand for parking spaces. Car parking is tedious and sustainable issue. An intelligent parking prediction in one of the interesting areas of research, where more of parking lots need to be allotted that minimizes the search time for parking space. One way is to collect the images in real time to solve the parking problem. This can be achieved by wireless sensors or IR sensors. Different regression techniques were also used to solve the problem.

Author Dogra et.al [36] discusses about effective usage of ML algorithms along with IoT devices in development of new applications for making smart transportation. Author Ozbayoglu M et.al [37] explains that if road side sensors are monitored in real time, then accidents can be detected by using KNN algorithms and FF-NN regression tree algorithms. In article [38], author Zontalis et.al, uses the term umbrella that covers all aspects related to smart transportation like road anomalies, route optimization, street lights, parking, accident detection etc. In paper [39] the author Jain et.al, focuses on IoT mobile to mobile communication indeed helping vehicle to vehicle communication encouraging to exchange useful information. Author Sarika



et.al, [40] focuses on smart parking using IoT devices, ultrasonic sensors and smart signboard to check parking places. This information is sent to cloud server using Wi-Fi model. Sign boards will have LED displays with Raspberry pi that can collect and show information about availability of parking lots and up to limited distances.

## 8.2 Methodology

Various DL algorithms used in smart transportation are LSTM, AE, RNN, DBN and GAN. LSTM is used for mode detection, human mobility and short-term traffic prediction. The datasets used for this are floating car data, GPS data and transport network data. AE algorithms are used to predict road accident prediction using traffic videos. RNN is used for occupancy prediction and traffic speed prediction using trajectory dataset. DBN is used for traffic flow prediction making use of historical road traffic flow. GAN is used for path planning using localization dataset.

## IX. CONCLUSION

The topic is elaborate on several technical method in order to boost Deep Learning through IoT applications. Also intends to find out various aspect of deep learning based IoT technologies and algorithms to the deep learning foundations. Furthermore, comprehensive research can be placed on activities about how deep learning can address various aspects like tele-health, ambient assisted living systems, machine health monitoring systems, human activity recognition, collecting vital signs of patients and data fusion. However, there are few potentials for optimizing QoS parameters, privacy and deployment that can be addressed in future works. Overall, the topics discussed are anticipated to be helpful to research scholars, engineers, healthcare professionals and policy makers who work in the field of IoT in deep learning.

## REFERENCES

- [1] Qu, C., J. Sun, and J. Z. Wang. "Automatic detection of the fall of old people based on Kinect sensor." *Journal of sensor technology* 29, no. 3 (2016): 378-383.
- [2] Peng, Yanfei, Jianjun Peng, Jiping Li, and Ling Yu. "Smart home system based on deep learning algorithm." In *Journal of Physics: Conference Series*, vol. 1187, no. 3, p. 032086. IOP Publishing, 2019.
- [3] Benyue, Su, Wang Guangjun, and Zhang Jian. "Smart home system based on internet of things and Kinect sensor." *Journal of Central South University (Science and Technology)* 44, no. Suppl 1 (2013): 182-184.
- [4] Yu Zesheng. "Research and Design of Smart Home System Based on Kinect attitudeRecognition". Liaoning University of Science and Technology, 2017.
- [5] Guo Zhe, Chen Peitou, Hu Mengkai, et al. 2016. "Kinect-based Smart Home System", *Modern Electronic Technology*, 2016, 39 (18): 149-152.
- [6] Hou Yuyi, Yang Dongtao, Liu Yan, et al. 2016. "Smart Home Life and Security System Based on Wireless Bluetooth Technology". *Journal of Jiaying University*, 2016, 34 (5): 36-40.
- [7] Li Tao. 2014. "Design and implementation of Android based smart home APP". Suzhou: Soochow University.
- [8] Naglic M, Souvent A. 2013. "Concept of Smart Home and Smart Grids integ ration". *Energy ,International Youth Conference on.IEEE*, 2013:1-5.
- [9] Smisek J, Jancosek M, Pajdla T. 2013. 3D with Kinect[J]. "Advances in Computer Vision & Pattern Recognition", 2013, 21(5):1154-1160.
- [10] Kepski M, Kwolsek B. 2013. "Human Fall Detection Using Kinect Sensor[J]". 2013, 226:743-752.
- [11] Klemenjak C, Egarter D, Elmenreich W. 2016. "YoMo: the Arduino-based smart meteringboard[J]". *Computer Science - Research and Development*, 2016, 31(1-2):97-103.
- [12] Barbon G, Margolis M, Palumbo F, et al. 2016. "Taking Arduino to the Internet of Things: TheASIP programming model[J]". *Computer Communications*, 2016, s 89-90:128-140.
- [13] Krauss R. 2016. "Combining Raspberry Pi and Arduino to form a low-cost, real-time autonomousvehicle platform[C]", *American Control Conference. IEEE*, 2016:6628-6633.
- [14] Dai XiGuo. 2017. "Research on human attitude recognition based on Convolutional neuralNetwork"
- [15] Chengdu University of Technology. Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia, "An ensemble forecasting method for the aggregated load with subprofiles," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3906-3908, Feb. 2018.
- [16] Y.-F. Zhang and H.-D. Chiang, "Enhanced ELITE-load: A novel CMPSOATT methodology constructing short-term load forecasting model for industrial applications," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2325-2334, Apr. 2020.
- [17] P. Zhuang and H. Liang, "Hierarchical and decentralized stochastic energy management for smart distribution systems with high BESS penetration," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6516-6527, Nov. 2019.
- [18] G. Erne, D. Dovžan, and I. Škrjanc, "Short-term load forecasting by separating daily profiles and using a single fuzzy model across the entire domain," *IEEE Trans. Ind. Electron.*, vol. 65, no. 9, pp. 7406-7415, Sep. 2018.
- [19] Y. Huang et al., "LoadCNN: A efficient green deep learning model for day-ahead individual resident load forecasting," 2019. [Online]. Available: arXiv:1908.00298.
- [20] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multi-view video summarization using CNN and bi-directional LSTM," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 77-86, Jan. 2020.
- [21] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72-81, Sep. 2019.
- [22] T. Hussain, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Intelligent embedded vision for summarization of multi-view videos in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2592-2602, Apr. 2020.
- [23] T. Hussain, K. Muhammad, S. Khan, A. Ullah, M. Y. Lee, and S. W. Baik, "Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers," *J. Artif. Intell. Syst.*, vol. 1, no. 15, p. 2019, 2019.
- [24] K. Muhammad, S. Khan, V. Palade, I. Mehmood, and V. H. C. De Albuquerque, "Edge intelligence-assisted smoke detection in foggy surveillance environments," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1067-1075, Feb. 2020.
- [25] Latif, Shahid, Maha Driss, Wadii Boulila, Zil E. Huma, Sajjad Shaukat Jamal, Zeba Idrees, and Jawad Ahmad. "Deep learning for the industrial internet of things (iiot): A comprehensive survey of techniques, implementation frameworks, potential applications, and future directions." *Sensors* 21, no. 22 (2021): 7518.

- [26] Khalil, Ruhul Amin, Nasir Saeed, Mudassir Masood, Yasaman Moradi Fard, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri. "Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications." *IEEE Internet of Things Journal* 8, no. 14 (2021): 11016-11040.
- [27] Chen, B.;Wan, J. Emerging trends of ml-based intelligent services for industrial internet of things (iiot). In *Proceedings of the 2019 Computing, Communications and IoT Applications (ComComAp)*, Shenzhen, China, 26–28 October 2019; pp. 135–139.
- [28] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and Privacy Challenges in Industrial Internet of Things," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015, pp. 1–6.
- [29] C. Perera, C. H. Liu, and S. Jayawardena, "The Emerging Internet of Things Marketplace from an Industrial Perspective: A Survey," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 585–598, 2015.
- [30] P. Zheng, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarak, S. Yu, X. Xu et al., "Smart Manufacturing Systems for Industry 4.0: Conceptual Framework, Scenarios, and Future Perspectives," *Front. Mech. Eng.*, vol. 13, no. 2, pp. 137–150, 2018.
- [31] Altalak, Maha, Amal Alajmi, and Alwaseemah Rizg. "Smart Agriculture Applications Using Deep Learning Technologies: A Survey." *Applied Sciences* 12, no. 12 (2022): 5919.
- [32] Disha Garg, Samiya Khan, and Mansaf Alam, "Integrative Use of IoT and Deep Learning for Agricultural Applications", Springer, pp. 521–531, 2020.
- [33] Kirtan Jha, Aalap Doshi and Poojan Patel, "Intelligent Irrigation System Using Artificial Intelligence And Machine Learning: A Comprehensive Review", Vol. 6, Issue 10, pp. 1493-1502, 2018.
- [34] Zhu, N.; Liu, X.; Liu, Z.; Hu, K.; Wang, Y.; Tan, J.; Guo, Y. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *Int. J. Agric. Biol. Eng.* 2018, 11, 32–44. [CrossRef]
- [35] Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.-L.; Chen, S.C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. (CSUR)* 2018, 51, 1–36. [CrossRef]
- [36] Dogra, Ajay Kumar, and Jagdeep Kaur. "Moving towards smart transportation with machine learning and Internet of Things (IoT): A review." *Journal of Smart Environments and Green Computing* 2, no. 1 (2022): 3-18.
- [37] Ozbayoglu M, Kucukayan G, Dogdu E. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. *2016 IEEE International Conference on Big Data (Big Data)*; 2016 Dec 5-8; Washington, DC, USA. 2016.p. 1807-13.
- [38] Zantalis, Fotios, Grigorios Koulouras, Sotiris Karabetos, and Dionisis Kandris. "A review of machine learning and IoT in smart transportation." *Future Internet* 11, no. 4 (2019): 94.
- [39] Jain, B.; Brar, G.; Malhotra, J.; Rani, S.; Ahmed, S.H. A cross layer protocol for traffic management in Social Internet of Vehicles. *Future Gen. Comput. Syst.* 2018, 82, 707–714.
- [40] Saarika, P.; Sandhya, K.; Sudha, T. Smart transportation system using IoT. In *Proceedings of the 2017 IEEE International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, Bangalore, KA, India, 17–19 August 2017; pp. 1104–1107.

## Machine Learning Algorithms for Herbs Recognition Based on Physical Properties

<sup>1</sup>Dr. Nur Fadzilah Mohamad Radzi, <sup>2</sup>Assoc. Prof. Dr. Azura Che Soh, <sup>3</sup>Assoc. Prof. Dr. Asnor Juraiza Ishak, <sup>4</sup>Assoc. Prof. Ir. Dr. Mohd Khair Hassan

<sup>1,2,3,4</sup> Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia

Email: <sup>1</sup>[nfmr86@gmail.com](mailto:nfmr86@gmail.com), <sup>2</sup>[azuracs@upm.edu.my](mailto:azuracs@upm.edu.my), <sup>3</sup>[asnorji@upm.edu.my](mailto:asnorji@upm.edu.my), <sup>4</sup>[khair@upm.edu.my](mailto:khair@upm.edu.my)

*Abstract— Currently, herbs recognition system has become a promising method to identify herbs species. Misuse of herbal medicine can cause serious health problems due to toxicological effects of phytochemical. As a result, a system that able to distinguish the types of herbs is needed. Most herbs recognition systems available in the market are dependent on experts. In this research, the concern is to identify the herbs compounds within the same group species where the physical appearance and aroma are similar. The work mainly focuses on herbs recognition system that intended for researchers and medical practitioners use without the need for experts. Electronic Nose (E-Nose) devices have been used extensively to differentiate and characterize the herb species based on their unique odour. Electrical signal generated from the gas sensor array is one of the physical properties studied. The robustness test of the proposed herbs recognition systems is performed via four classification models based on machine learning algorithm: Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Multinomial Logistic Regression (MLR), and Gaussian Radial Basis Function (RBF) Kernel. The performance of the classification accuracy using KNN shows a better result within the family group from 92.15% to 100% compared to the others method.*

*Keywords— Electronic Nose, Gaussian Radial Basis Function Kernel, Herbs Recognition System, Herbs Physical Properties, K-Nearest Neighbours, Multinomial Logistic Regression, Odours Pattern Recognition, Support Vector Machine.*

### I. INTRODUCTION

Herbs are plants with savoury or herbal properties and are used for a number of purposes including therapeutic, flavouring, fragrances, colouring, culinary, and spiritual. Herbs are often known as special plants with nearly equal physical and aroma properties among the same family group species. Although the aroma produces by the herbs' leaves are almost similar, it is still distinctive and unique. Thus, various attempts have been made in recent decades to establish automated strategies that could be used to identify herb species from the same kind. Amongst the significant approaches are by utilising chemical and physical properties of the herbs' leaves. Herbs' leaves have been extensively used for classification purposes. For instance, shape of herb leaf, outline of leaf measurement and vein structure in the form of image can be utilised to classify herb [1]-[7]. On the other hand, [8],[9] had considered odour of leaves as sample for discrimination of herbs types due to its authentication [4],[10],[11]. In short, leaves becoming researchers' favourite to identify herbs.

Physical characteristics such as odours, tastes, and physical characteristics (shapes, colour, texture, and size) may all be used to differentiate herb types. The most challenging part while differentiating herbs is when they have similar aromas and characteristics, particularly those of the same family community genus. However, the growth in recognition system currently trying to address this challenge. The most notable technique used in recognition system is E-Nose [8],[9], [12]-[15]. The ability to identify the sample based on the sensors' signal response is demonstrated by an electronic device. This device mimics the discrimination of the human olfactory system utilising several combinations of sensor array for identification of gas or chemical presents along with an effective pattern recognition system [16]. For more than a decade, scientists and researchers have been involved in this invention as it has the potential to solve the limitations in chemical experiments. E-Nose is an invention that was created to imitate the human nose. As the sensor array reacts to the emitted gases, E-Nose generates electrical signals. This physical approach is studied in this research. Although the human nose can classify smell, the judgments can be bias. Furthermore, the human nose has detection limits for various gases. Thus, human nose cannot be used as a universal instrument for smell-related classification and discrimination. A gas sensor array, E-Nose, is used to distinguish particular volatiles, which pattern recognition algorithms may use to discriminate and classify them from the fingerprint response it produces.

In a pattern recognition system, a classification model is a supervised learning algorithm that produces a set of classifications, or labels, for a given set of data. Generally, the system was modelled using labelled training data and the testing data was used to calculate the model performance. The proposed model was tested with unseen data or new dataset to substantiate the robustness of the system. In machine learning algorithm, four types of classification model are usually used for pattern recognition system which are KNN, SVM, MLR, and Gaussian RBF Kernel.

Back in the 1990s, the SVM model was developed and it became the most popular model at that time. SVM is a supervised learning model which requires label data. Technically, the SVM algorithm is to get the maximized margin of separation for optimal hyperplane between two or more classes [17],[18]. Data points that are closer to the hyperplane are defined as support vectors. The SVM cost function measures the error between predicted and expected values. SVM has been one of the favourites for plant classification. Researchers [19] uses texture and colour of leaves as features for classification. The authors chose SVM and Multi-Layer Perceptron (MLP) classifier to identify the leaves where MLP outperformed SVM with 94.5% accuracy. On the other hand, [20],[21] employ SVM by using information extracted from leaves' images i.e., leaves' texture and shape to classify the plant. SVM showed a satisfactory result with 93.3% and more than 90% of accuracy respectively. Comparison of accuracy has been made between ANN and SVM where SVM produced a better result [22]. In this work, SVM has showed a remarkable performance in classifying plant with more than 90% accuracy and also suitable for combination of feature selection.

KNN is a lazy learning algorithm that is often used to find similar items or patterns in data. Without learning anything, the data remains unchanged [23]. Data from testing will be classified into classes based on its similarity to the most similar features from training data. Euclidean distance was applied in KNN to measure the distance between the new data with all data points in all classes. New data will classify into the class with the highest number of nearest neighbours. Even though this model is the simplest classification model, the computation cost is high due to the requirement of frequent update of value  $k$ . Ghosh et al. [24] had used KNN as well as SVM classifiers to identify types of plants where the author found out that both classifiers had a comparable performance and accuracy. Other researchers [25] had conducted an experiment by using shape of leaves as feature and implemented KNN for classification purposes and the result obtained was 91.5% recognition accuracy. However, KNN faced some difficulties when the leaves have deficiencies and changes in value of nearby distance. Separately, [26] had conducted comparison between ANN and KNN classifiers for identifying leaves using its colour and shape. ANN outperformed KNN with 93.3% recognition accuracy whereas KNN only 85.9%. KNN is the simplest classifier but it is susceptible to noise.

Other methods used for classification techniques are MLR and Gaussian RBF Kernel model. MLR basically is a method that extends logistic regression to solve a multiclass problem. MLR is a statistical method for predicting the probabilities of a contingent nominal variable [27] which independent variables have a major effect on the dependent variable. Gaussian RBF Kernel model is the improvement model derived from SVM model. SVM model is a linear hyperplane. It became impossible for SVM to find a hyperplane for non-linear classification boundary. However, separation becomes easy when data is projected into higher dimensions. Original data can be transformed into a suitable space using Kernel trick in SVM.

In this research, the sample herbs used only focus on aromatic herbs families. Therefore, E-Nose herbs recognition algorithm is proposed in this study to differentiate herbs leaves based on emitted odours. E-Nose is a mimicking the human sensory using several types of gas sensors for odour detection. Voltage signal considered as physical properties is collected from E-Nose experiment using 5 types of gas sensors. The unique odours extracted from the voltage signal using moving average techniques and signal filtering were used to develop the database of physical herbs properties. The discriminant techniques were applied to increase the boundary gap between the group clusters by reducing the dimension. Finally, the classification of the herbs species was conducted using four difference classification technique based on machine learning algorithm. A comparison among classification performance of herbs recognition system based on physical properties are analysed and discussed.

## II. DATA COLLECTION

Nineteen types of herb species from five different family groups of herbs, as listed in Table 1, are the selected herb samples used in this investigation. The choice of herbs samples was made after consultation with a botanist at Universiti Putra Malaysia (UPM). Each sample was gathered from the Agricultural Conservatory Park located at Institute of Bioscience (IBS), UPM.

Table 1: List of 19 types herb samples

Group Species	Herb Name		
	Scientific Name	Local Name	Abbreviation
Family <i>Lauraceae</i>	<i>Cinnamomum Iners</i>	Medang Teja	F1S1
	<i>Cinnamomum Verum</i>	Kayu Manis	F1S2
	<i>Cinnamomum Porrectum</i>	Medang Wangi	F1S3
	<i>Litsea Elliptica</i>	Medang Kesing	F1S4

Family <i>Myrtaceae</i>	<i>Syzygium Aromaticum</i>	Cengkih	F2S1
	<i>Syzygium Polyanthum</i>	Daun Salam	F2S2
	<i>Melaleuca Alternifolia</i>	Gelam Wangi	F2S3
	<i>Rhodomyrtus Tomentosa</i>	Kemunting	F2S4
Family <i>Zingiberaceae</i>	<i>Scaphochlamys Kunstleri</i>	Tepus Hutan	F3S1
	<i>Etlintera Triorgyalis</i>	Kantan Hutan	F3S2
	<i>Elettariopsis Curtisii</i>	Lempoyang	F3S3
	<i>Zingiber Zerumbet</i>	Halia Kesing	F3S4
Family <i>Annonaceae</i>	<i>Cananga Odorata</i>	Bunga Kenanga	F4S1
	<i>Goniothalamus Tapis</i>	Kenarak	F4S2
	<i>Goniothalamus Umbrosus</i>	Kenerek	F4S3
Family <i>Rubiaceae</i>	<i>Prismatomeris Glabra</i>	Tongkat Ali	F5S1
	<i>Ixora Grandifolia</i>	Jarum Hutan	F5S2
	<i>Morinda Elliptica</i>	Mengkudu Kecil	F5S3
	<i>Morinda Citrifolia</i>	Mengkudu Besar	F5S4

Fresh leaves from the plants were used in this study. To ensure maximum freshness, the leave samples were taken at 8:30am to 9:30am in the morning. The advantage of having aromatic herb types as samples was that the Volatile Organic Compounds (VOCs) would emit strong odor. Therefore, the gas sensor would efficiently capture the respond. An expert botanist from IBS would monitor the process of samples collection from the IBS botanical garden to verify the herbs species.

### III. ELECTRONIC NOSE (E-NOSE) RECOGNITION SYSTEM

This section constructs the method of herbs recognition system using voltage signal that produced by the gas sensors. Each signal was pre-processed to extract the information from complex signals in the presence of noise. Moving average was applied in order to reduce the noise. Only the signal in respond time period was taken to build the herbs classification model. Then, in feature extraction process, two methods were applied for dimension reduction, namely Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA). Signal data were transformed onto a different set of orthogonal axes to reduce the cluster overlapping by finding the maximum spread. Four different classification models were proposed and the accuracy was studied. In conclusion, Figure 1 shows the proposed herbs recognition system using E-Nose data for this study.

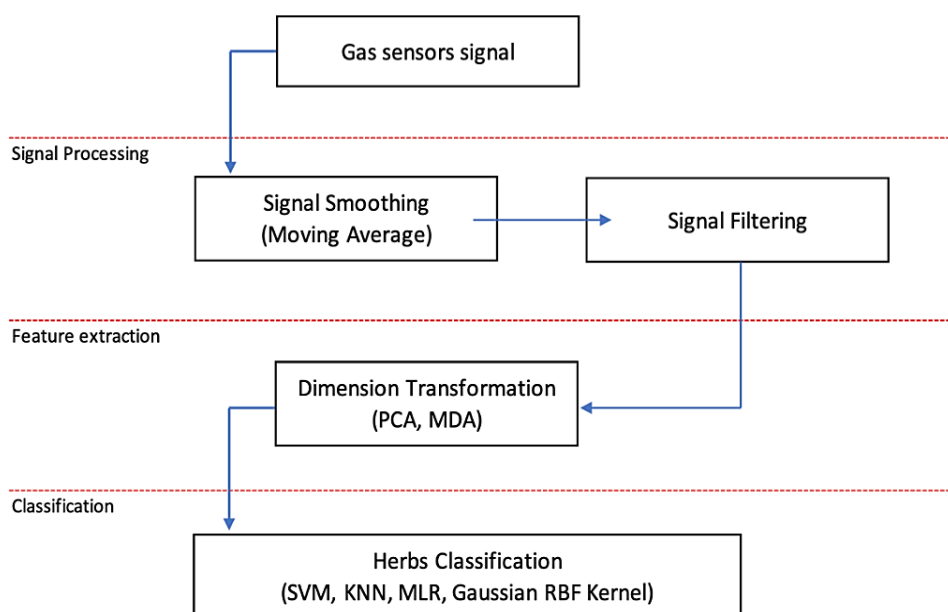


Figure 1: Herbs recognition system using E-Nose data

#### 3.1 Experiment Setup

E-Nose has been developed at Control and Automation Laboratory, Faculty of Engineering, UPM. The whole setup of the developed E-Nose system for data collection is as shown in Figure 2. The system consists of a blender with a container for



herbs sample, a slot for the gas sensor array, a National Instrument data acquisition system, NI USB6009, and the graphical user interface. Five different types of Metal Oxide Semiconductor (MOS) gas sensor as listed in Table 2 were used.

The procedure for data collection, first step was to preheat the sensor in the E-Nose to stabilise the signal. The voltage for the gas sensor common baseline was kept between 0 V and 1 V. Next, 15 grams of fresh leaves was blended for 15 seconds in the sample container.

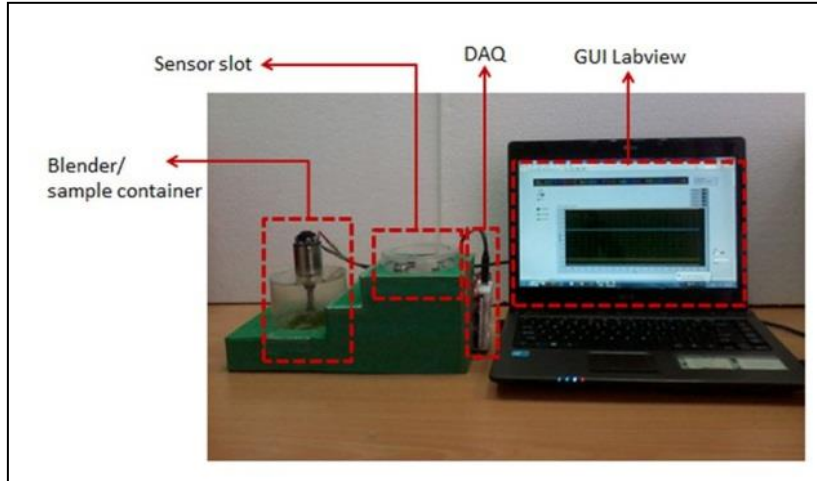


Figure 2: Complete configuration of the E-Nose system.

Table 2: MOS gas sensor for E-Nose

Sensor Type	Code Sensor	Type of gas detection
TGS 2610	S1	Butane, propane, liquefied petroleum gas
TGS 2611	S2	Methane, natural gas
TGS 2620	S3	Alcohol, toluene, xylene, volatile organic compound
TGS 823	S4	Organic solvent vapors
TGS 832	S5	Halocarbon, Chlorofluorocarbon

The signal measurement process begins with clicking the ‘START’ button on GUI. The complete electrical response signal need to captured from the E-Nose experiment has 3 slot time which are baseline, response time and recovery time as shown in Figure 3. During the first 60sec, the baseline voltage was captured. The container of the sample was then put in the sensing slot and the response time was set at 120sec. The container was taken out from the sensor slot. The recovery time would be in next 120sec. The complete signal response will be captured in 300 sec. The simplified process flow of the response signal capturing is shown in Figure 4.

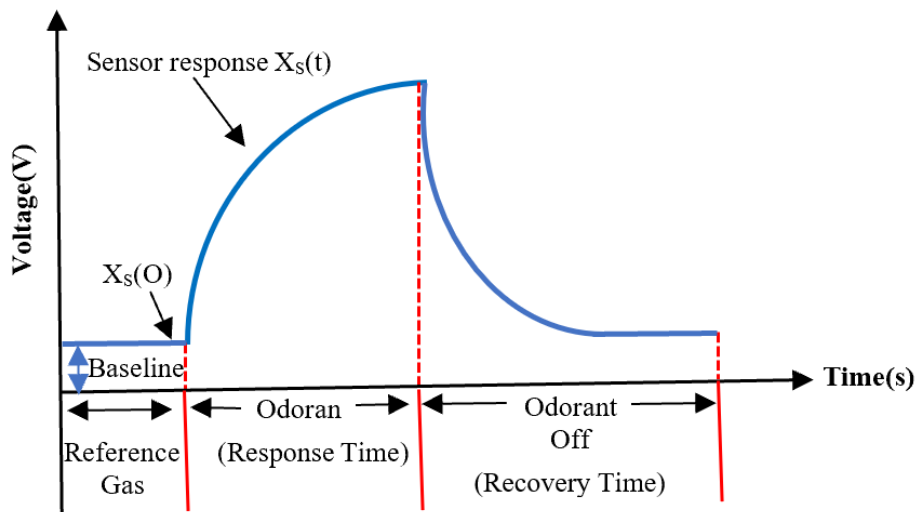
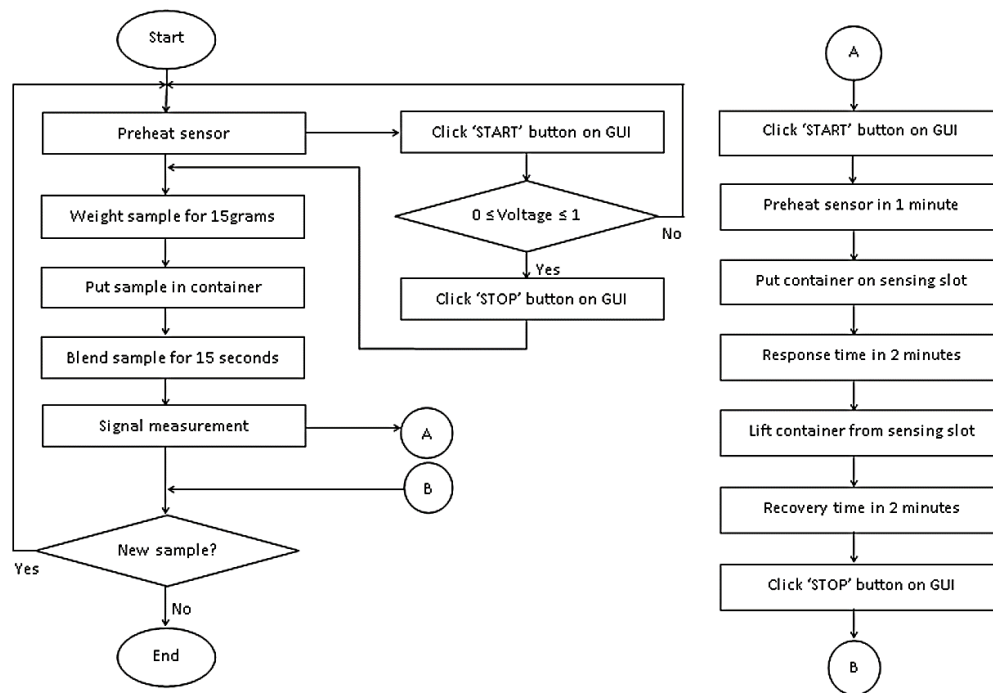


Figure 3: Complete electrical signal response from gas sensor

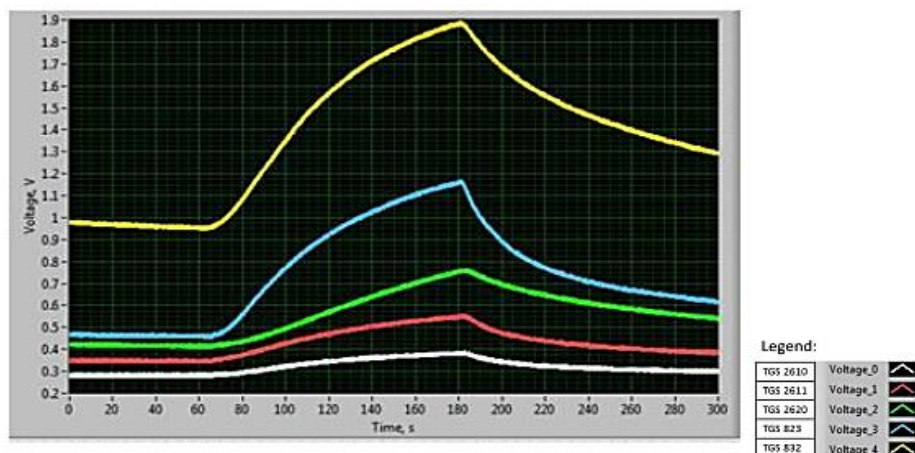


**Figure 4:** Flow chart of data collection using E-Nose experiment

### 3.2 Signal Pre-processing Based on Physical Properties

The pre-processing stage of the raw data was done to improve the data quality and to ease the mining process. Signal processing technique was used for data pre-processing stage preparation. Throughout data gathering process, sensor drift normally happened in signal processing. The benefit of pre-processing was it compensated and enhanced the data. It also compressed the transient response of the sensor array and reduced sample to sample variations [28].

Complete signals captured by the E-nose system for a single sample is shown as in Figure 3. Under most practical situations, gas sensors array signal would start to increase from the baseline when it detected the existing of respective gases. The pattern of signal response was unique for each gas depending on the concentration of each type of gas. During the response time, gas emitted from the herbs was detected and collected. Figure 5 shows five different gas sensors signal responses for a single herb sample.



**Figure 5:** Electrical signal response from 5 gas sensor of E-Nose

The signal observed contain noise as shown in Figure 6. Then, the dimension of each collected data signal was smoothed and reduced by means of moving average technique. The moving average technique formula used is shown in Eq. (1).

$$\hat{x} = \frac{1}{n} \sum_{i=1}^n x_i = \left(\frac{1}{n}\right) x_1 + \left(\frac{1}{n}\right) x_2 + \dots + \left(\frac{1}{n}\right) x_n \quad (1)$$



Where,  $n$  is number of sample data.

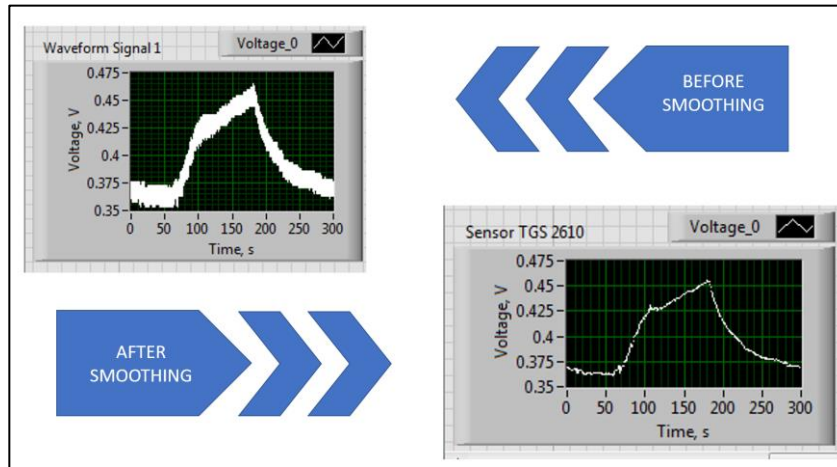


Figure 6: Smoothing process for electrical signal response

In practice, there was a process delay in sensor to respond during placing the sample container from blender to sensor slot. Herbs true signals were required for precision improvement. The signal of herbs was recorded in two minutes from time 60sec to 180sec. Therefore, the unwanted signal before time 60sec and after 180sec needed to be removed. Only the signal in response time as shown in Figure 3 was required in building the classification model as illustrated in Figure 7. The output for the signal processing is explained in Eq. (2).

$$y = x_t|_{t=60}^{t=180} - \left( \frac{1}{n} \sum_{t=1}^{60} x_t \right) \quad (2)$$

Where;  $x_t$  is signal of herbs

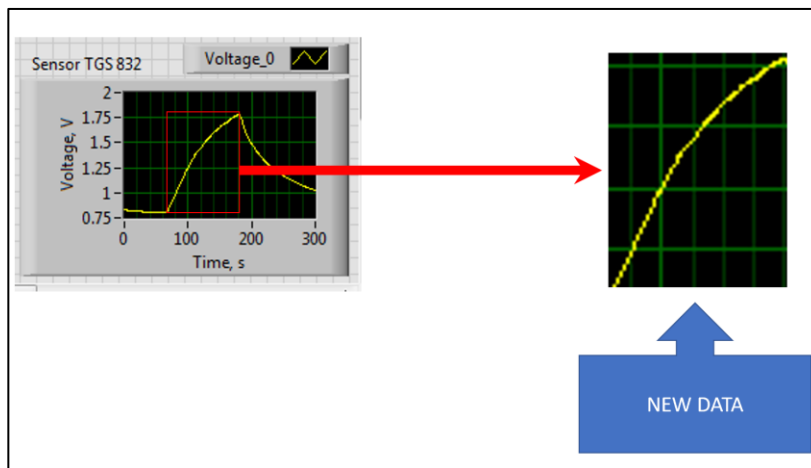


Figure 7: Signal filtering

#### IV. DISCRIMINANT ANALYSIS BASED ON PHYSICAL PROPERTIES

Most of recognition systems use PCA to reduce the high dimensional data [29]-[31]. PCA is an unsupervised learning. Ignoring the classes separation, PCA method solely focuses on dimension with maximum variance for all data points from all classes. In this research, the naming of the sample of herbs species were known beforehand with the guidance from IBS scientist. Therefore, dimension reduction method for supervised learning was more suitable than unsupervised learning.

Another dimensional reduction method, MDA was proposed. MDA is a discriminative technique used to solve the multiclass problem. This supervised learning technique is stimulated from Fisher's linear discriminant analysis which intended to solve the discrimination between two-class problems. The MDA projects data in the same way as the PCA, where the projected data is based on the best separation in the least-square sense. The ratio of between-class scatter to the within-class scatter projection must still be maximized while MDA searches for optimal transformation.

The mapping of discrimination between classes is well separated and the amount of information losses is also lesser. The

Fisher criterion function is shown below in Eq. (3).

$$J(W) = S_w^{-1} S_B = \frac{|W^T \cdot S_B \cdot W|}{|W^T \cdot S_w \cdot W|} \quad (3)$$

where  $S_B = (\mu_1 - \mu_0) \cdot (\mu_1 - \mu_0)^T$  and  $S_w = (S_1 + S_2 + \dots + S_n)$  are respectively the ratios of between-class scatter matrix and within-class scatter matrix. The scattering matrix for class n is  $S_1 + S_2 + \dots + S_n$ , while  $W$  is the projection matrix, and  $J(W)$  is an eigenvalue that measures the scattering matrix within the class and differentiates the class averages. By assuming the final matrix of eigenvectors as  $\omega^T$ , then the new projected data,  $x$  for E-Nose data  $x_{enose}$  will be given as:

$$x = \omega^T x_{enose} \quad (4)$$

## V. CLASSIFICATION BASED ON PHYSICAL PROPERTIES

There are varieties of classification methods for supervised machine learning. In this research, herbs were classified into its species group using four classification models based on machine learning algorithms which are SVM, KNN, MLR and Gaussian RBF Kernel.

### 5.1 Classification using Support Vector Machine (SVM)

The powerful technique used for classification is SVM. The approach is based on statistical analysis and is suitable for supervised classification applications. SVM is a more accurate classification method than other classification methods. The cost and kernel parameters depend on how the parameters are set [32]. So as to obtain the optimal parameter, K-fold cross validation was applied in SVM. It then searched for the maximum geometric margin and minimized classification error as shown in Figure 8 [33].

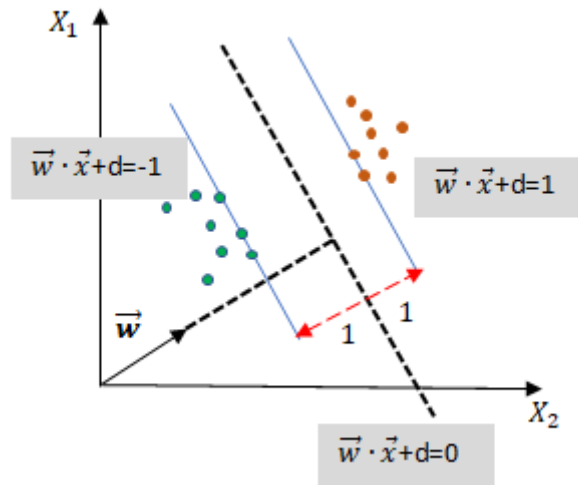


Figure 8: Example of maximum geometric margin for two classes

SVM will find the maximum distance between two hyperplanes that separate the data. Larger margins of these hyperplanes indicate better generalization error of the classifier. Parallel hyperplane is defined in Eq. (5) where  $w$  is width or margin,  $b$  is a constant, and  $f(x_{pca1}) = 0$  is a decision boundary that completely separates the 2 classes,  $f(x_{pca1}) > 0, \forall x_{pca1}$  of class red, and  $f(x_{pca1}) < 0, \forall x_{pca1}$  of class green.

$$f(x_{pca1}) = wx_{pca1} + b \quad (5)$$

Support Vectors (SV) are the data points along the hyperplanes. While vector theta must be perpendicular to decision boundary. Multiple iteration of weight,  $w$  updates are required in order to get optimal hyperplane which could best separate the data. The final separation would give the minimum cost function. To train the SVM, the structure of hypothesis in Eq. (6) and the cost function,  $\sigma$  in Eq. (7) were utilized.

$$h_{\theta}(x_{pca1}) = \frac{1}{1 + e^{-\theta^T x_{pca1}}} \quad (6)$$

$$\sigma = \min_{\theta} C \sum_{i=1}^n \left[ y_i \text{cost}_1(\theta^T x_{pca1_i}) + (1 - y_i) \text{cost}_0(\theta^T x_{pca1_i}) \right] + \frac{1}{2} \sum_{j=1}^d \theta_j^2 \quad (7)$$

given the maximum margin;

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^d \theta_j^2 ; \begin{cases} \theta^T x_{pca1_i} \geq 1 & \text{if } y_i = 1 \\ \theta^T x_{pca1_i} \leq -1 & \text{if } y_i = -1 \end{cases} \quad (8)$$

In classification of non-linear separable data, the success of SVM depended on the tuning of the cost parameter,  $C$ , and the kernel parameters,  $\gamma$  and  $d$ . Grid-search method is a good way to find the best parameters and RBF kernel in cross validation. While tolerating the outlier and solving the constraints of the optimization problem, a soft margin SVM was applied.

### 5.2 Classification using K-Nearest Neighbours (KNN)

KNN is a supervised learning algorithm used in this research to classify herbs. This is an instance-based learning which is non-parametric model. The model can be trained by memorizing the training dataset. The KNN algorithm is a special case of instance-based learning where the cost of learning process is zero [34].

KNN ranks the sample by taking a majority vote from its neighbours. K-Based on the algorithm concept, the sample is assigned to the class that its  $k$  closest neighbours are the foremost common class. This algorithm needs training data and pre-defined  $k$  value in order to determine the class. The value of  $k$  is typically a small positive value integer. By means of a similarity measure of a distance metrics, the algorithm will look around the training sample area for the  $k$ -most similar samples [35]. The distance metrics is one factor that could affect the classification performance. In this study, the distance between existing training dataset and the new data point was defined using Euclidean distance. Euclidean distance analyzes the root of square differences between coordinates of two objects. For each feature  $x_i$  computes the Euclidean distance to all other features in sample. The formula in Eq. (9) was used to calculate the Euclidean distance  $d(x, y)$  between features  $x_{pca_i}$  and  $y_{pca_i}$ .

$$d(x, y) = \sqrt{\sum_{i=1}^d (x_{pca_i} - y_{pca_i})^2} \quad (9)$$

The KNN algorithm uses Euclidean distance to determine the class of the new data as shown in Figure 9. Parameter ' $k$ ' in KNN is referring to the number of nearest neighbors to be used in the majority voting process. The new data to be determined was marked as "X" and the parameter  $k = 4$  was represented by the grey arrow showing the Euclidean distance computation of four nearest neighbors. The Euclidean distance computed the algorithm and classified marked "X" to the nearest circle of class blue. In this thesis, a majority of 5 group species had 4 species under the same family, therefore the value of  $k = 4$  was applied in this algorithm.

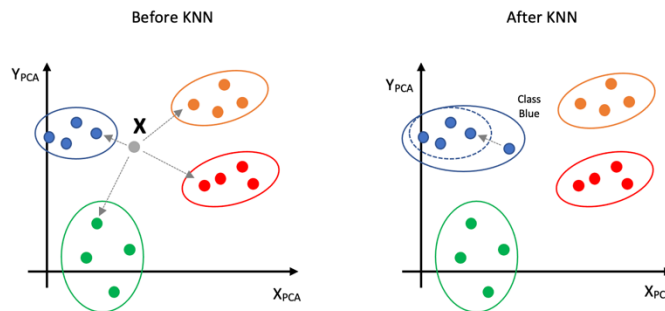


Figure 9: The classification of KNN algorithm with using Euclidean distance

### 5.3 Classification using Multinomial Logistic Regression (MLR)

MLR is a statistical model that uses logistic function to model dependant variable for multiclass classification. It uses maximum possibility estimation to examine the probability of categorical membership as defined below:-

$$\pi(x) = P(Y = j|x) \quad (10)$$

The sum of probabilities of five categories of herbs species that belong to the dependent variable,  $x$

$$P(Y = 0|x) + P(Y = 1|x) + P(Y = 2|x) + P(Y = 3|x) + P(Y = 4|x) = 1 \quad (11)$$

The multinomial probability,  $P(Y)$  in MLR model can be expressed in Eq. (12). Meanwhile, the MLR model is expressed as in Eq. (13).

$$\pi(x_k) = \frac{e^{(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)}}{1 + e^{(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)}} \quad (12)$$

$$\log(\pi_j(x_k)) = \frac{e^{(\alpha_{0j} + \beta_{1j} x_{1k} + \dots + \beta_{pj} x_{pk})}}{1 + \sum_{j=1}^{k-1} e^{(\alpha_{0j} + \beta_{1j} x_{1k} + \dots + \beta_{pj} x_{pk})}} \quad (13)$$

Where  $k$  is the dependent variable,  $j$  is the dependent variable category, the parameter  $\beta$  refers to the effect of  $x_k$  on the log, and  $x$  represent the projected features in Section 4.0.

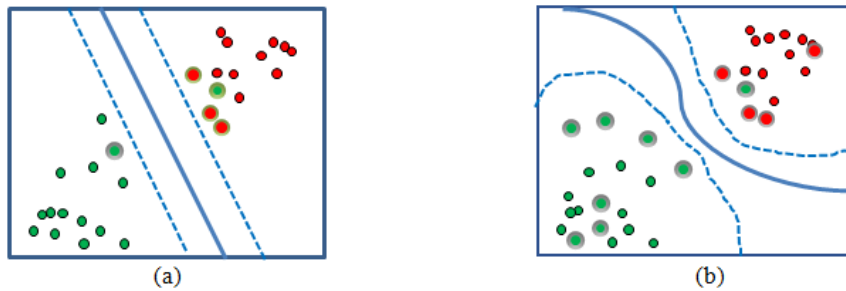
#### 5.4 Classification using Gaussian Radial Basis Function (RBF) Kernel

SVM algorithm is associated in determining the hyperplane which data is linearly separable based on maximum margin. However, not all data are indeed linearly separable. Using SVM to separate the data set which is linearly inseparable tends to have misclassification problem. Therefore, kernel trick was applied in the SVM. Dot product of support vectors were converted to dot product of mapping function by Kernel trick. The idea was to generate nonlinear combinations of the original features to project them onto a higher-dimensional space via a mapping function, where the data turns out to be linearly separable.

One of the most used types of kernel function is RBF Kernel. It has confined and limited response along the entire x-axis. Gaussian RBF is a kernel in the form of radial basis function that has similarity to the Gaussian distribution. This kernel was applied in SVM model to improve and speed up the SVM classification. Gaussian RBF Kernel function was formulated as:

$$K_{RBF}(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0 \quad (14)$$

Where,  $x_i$  denotes projected features in Section 4.0,  $\gamma = \frac{1}{2\delta^2}$  is a parameter that sets the spread of the kernel, and  $\delta$  is a constant that defines the kernel width. Finding the decision boundary was an important hyper-parameter for the SVM. Kernel Trick utilized existing features, made some alterations, and formed new features. SVM discovered the nonlinear decision boundary from those new features. Radial Basis Function kernel acted as a converter that generated new features by calculating the distance between  $x_i$  to a specific center denotes as  $x_j$ . Figure 10 illustrates the difference of hyper-plane between linear SVM and SVM with Gaussian RBF Kernel.



**Figure 10:** Linear and non-linear SVM decision boundary (a) Decision boundary for SVM and (b) Decision boundary for SVM with Gaussian RBF Kernel

## VI. RESULTS AND DISCUSSIONS

This study used E-Nose device to collect gas signal response from the odour released by the herbs. Five different types of Metal Oxide Semiconductor (MOS) gas sensor were used which are TGS 2610(S1), TGS 2611(S2), TGS 2620 (S3), TGS 823(S4) and TGS 832(S5). Figure 11 and Figure 12 show the two samples of signal responses from *Lauraceae* and *Rubiaceae* families. From the figures, the gas sensors captured in 300sec. It can be observed that the responses of gas sensor in terms of voltage value are increasing between time 60sec until 180sec. When the chamber containing herb sample was removed from the sensors array slot, the responses started to decrease after time 180sec, as no gases were detected. All the complete signal has been compiled as E-Nose herbs database.

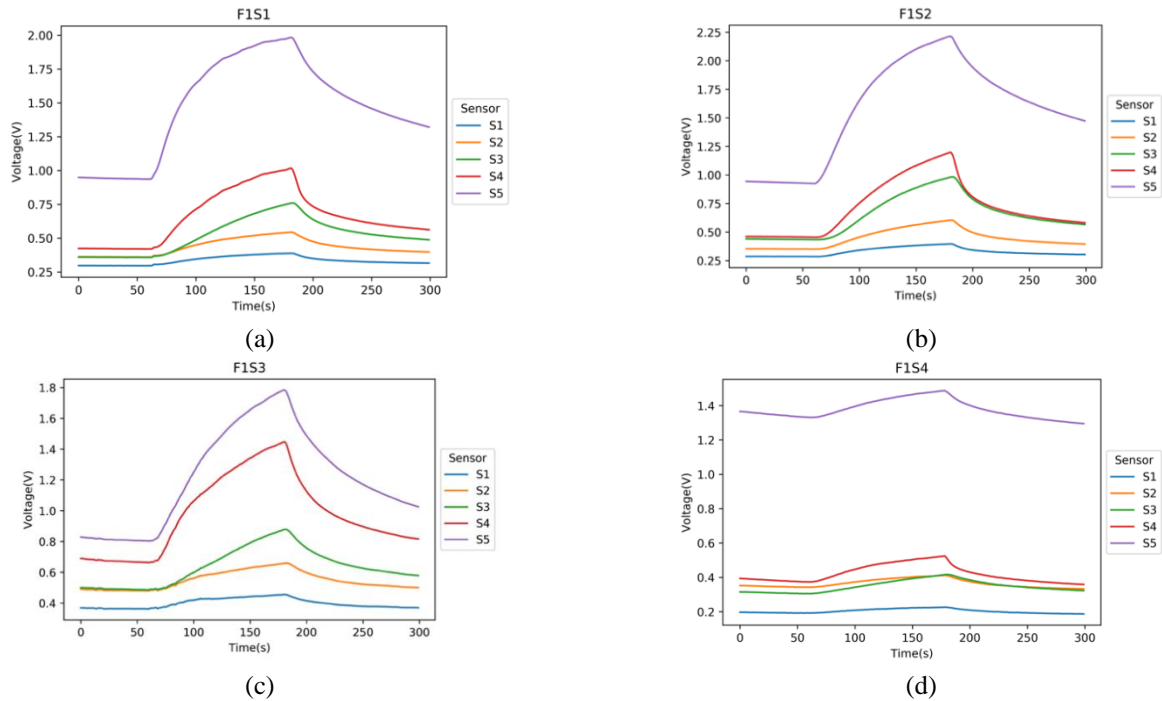


Figure 11: Complete signal response from five MOS gas sensor array for Family Lauraceae

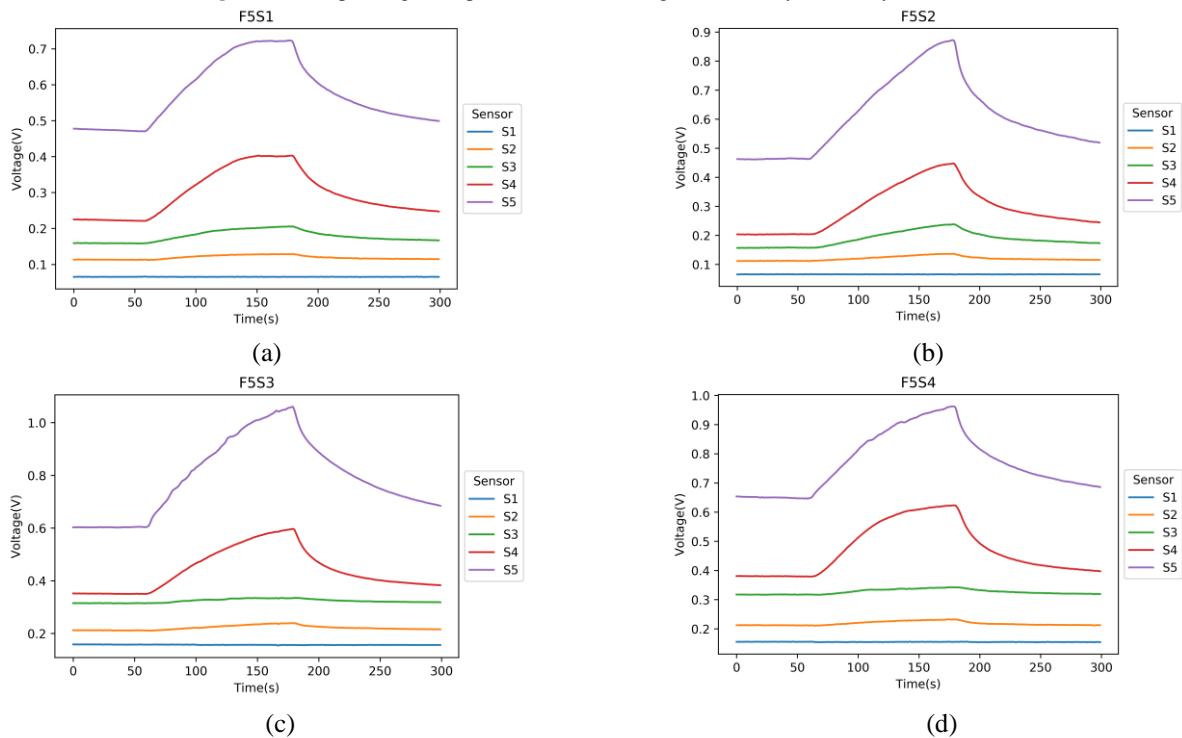


Figure 12: Complete signal response from five MOS gas sensor array for Family Rubiaceae

### 6.1 Discriminant Analysis of Physical Properties

The discriminant analysis techniques were used to discriminate the multiclass problems of PCA and MDA in this study. Table 3 lists the percentage of PCA and MDA principal components, (PCA<sub>1</sub> and PCA<sub>2</sub>) and (MDA<sub>1</sub> and MDA<sub>2</sub>), as well as the percentage of total information lost after projecting the data within same family group to a lower dimensional subspace.

**Table 3:** The percentages of PCA and MCA

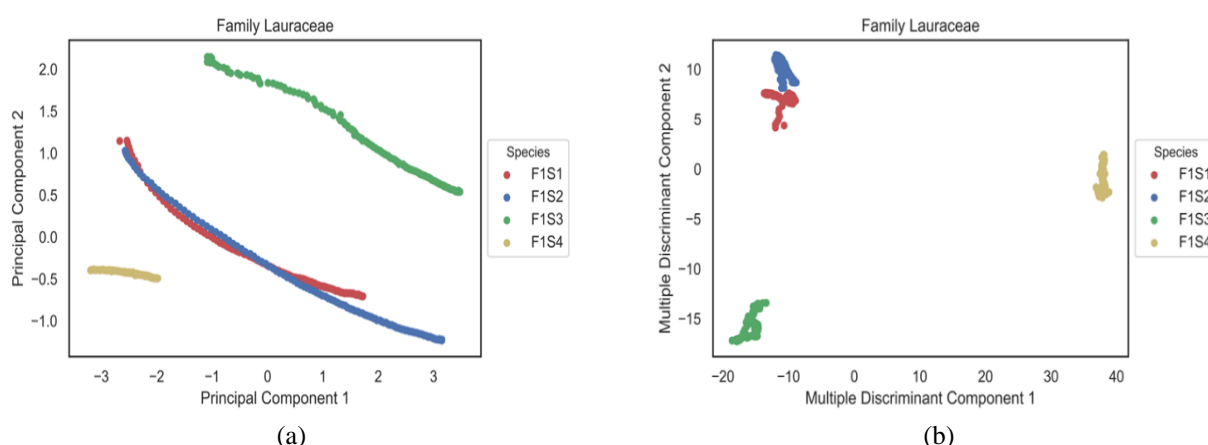
Group Species	PCA (%)			MDA (%)		
	PCA <sub>1</sub>	PCA <sub>2</sub>	Information Lost	MDA <sub>1</sub>	MDA <sub>2</sub>	Information Lost
<i>Lauraceae</i>	82.08	14.99	2.93	82.17	17.39	0.44
<i>Myrtaceae</i>	73.29	22.83	3.88	91.32	8.31	0.37
<i>Zingiberaceae</i>	78.83	19.74	1.43	97.04	2.86	0.1
<i>Annonaceae</i>	85.39	14.18	0.43	99.99	0.00	0.01
<i>Rubiaceae</i>	88.50	10.48	1.02	99.60	0.38	0.02

From Table 3, it can be observed that MDA method is carrying more amount of information than PCA method, where the total of information lost are between 0.01% to 0.44% and 0.43% to 3.88% respectively. Each component in the table signifies the amount of data information remained after the transformation. In theory, the first component (PCA<sub>1</sub> and MDA<sub>1</sub>) shall contain more information compared to the other component (PCA<sub>2</sub> and MDA<sub>2</sub>). Total information carried forward for projected data can be calculated by summing up the first and second components. From the results, by taking the first two components, more than 90% of previous data information is carried after the projection. Therefore, only the first two components data were taken into account. Overall, for the first component, data of MDA showed higher percentage than PCA for each species of family group: *Lauraceae*, *Myrtaceae*, *Zingiberaceae*, *Annonaceae* and *Rubiaceae*. Comparison of total information percentage between PCA and MDA is shown in Table 4. Projected data of MDA method holds the information around 0.42% to 3.51% higher than PCA method.

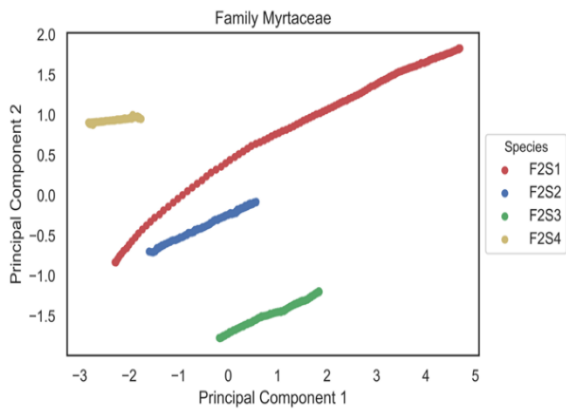
**Table 4:** Comparison percentage of two components between PCA and MDA

Group Species	PCA (%)	MDA (%)	Difference (MDA-PCA) (%)
<i>Lauraceae</i>	97.07	99.56	2.49
<i>Myrtaceae</i>	96.12	99.63	3.51
<i>Zingiberaceae</i>	98.57	99.90	1.33
<i>Annonaceae</i>	99.57	99.99	0.42
<i>Rubiaceae</i>	98.98	99.98	1.00

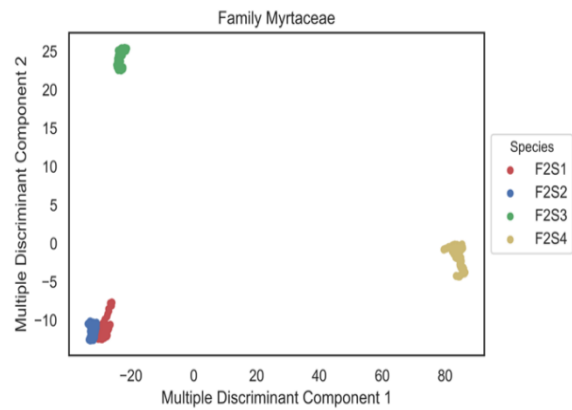
Figure 13-17 show the lower dimensional subspace projected data results of herb species from 5 group families using PCA and MDA. Some of the species are seen to have overlapping problem. This can be proved by calculating the minimum distance between data points from two herbs species as shown in Table 5. Range of minimum distance for PCA projected data is between 0 to 4.15 while for MDA projected data is between 1.34 to 270.33. From the results, MDA projected data showed no overlapping between these species compared to PCA projected data. This is because MDA method brought the data points under the same herbs species spreading close to each other while at the same time telling them to move farther from as many different species as possible. Therefore, MDA method was proposed in E-Nose system.



**Figure 13:** New projection for Family *Lauraceae*: (a) with PCA method (b) with MDA method

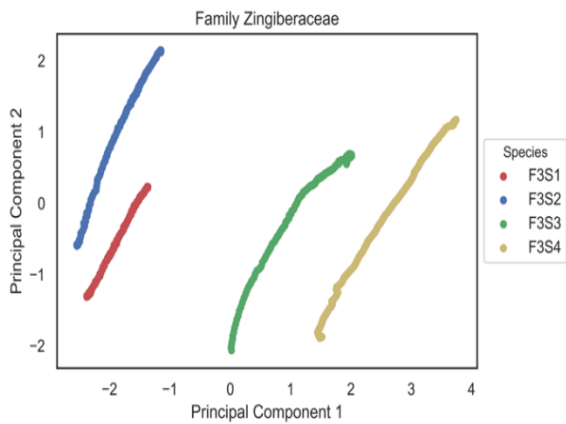


(a)

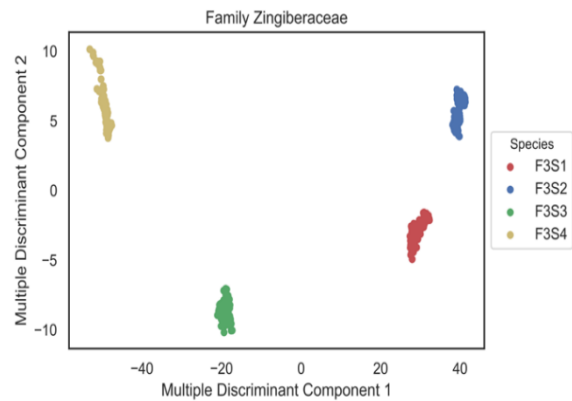


(b)

**Figure 14:** New projection for Family *Myrtaceae*: (a) with PCA method (b) with MDA method

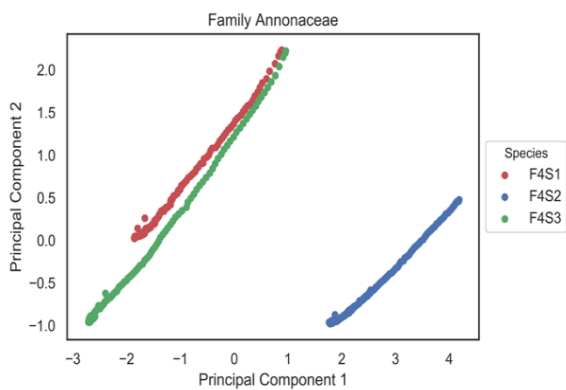


(a)

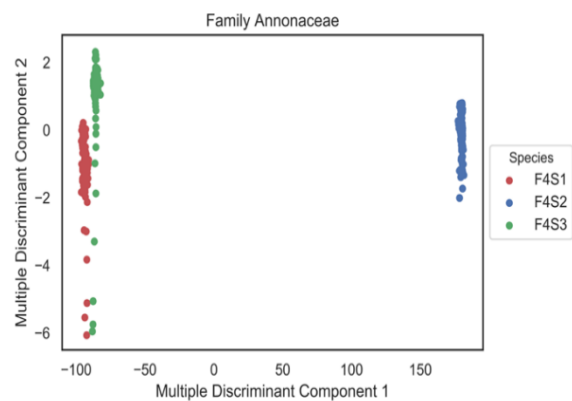


(b)

**Figure 15:** New projection for Family *Zingiberaceae*: (a) with PCA method (b) with MDA method



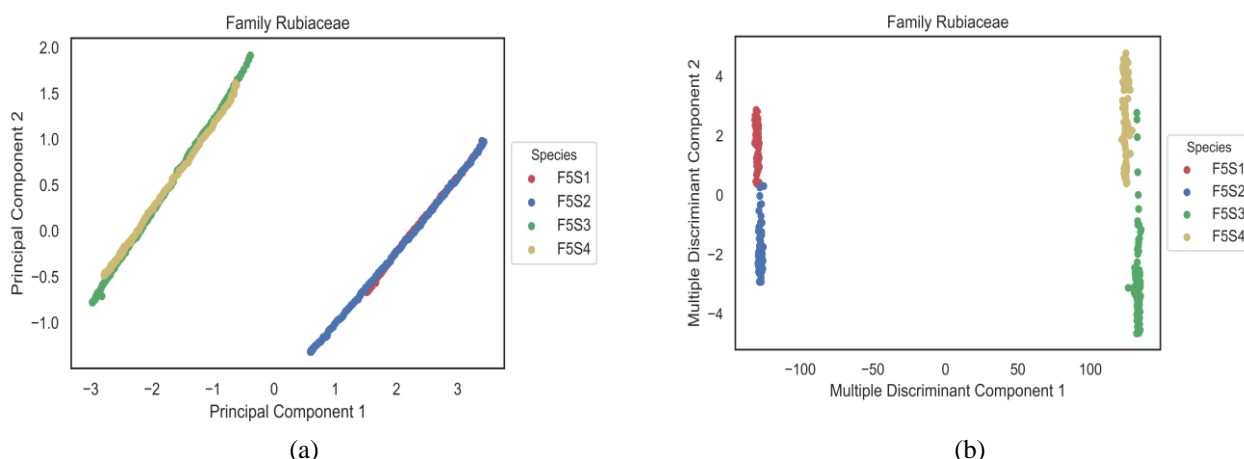
(a)



(b)

**Figure 16:** New projection for Family *Annonaceae*: (a) with PCA method (b) with MDA method





**Figure 17:** New projection for Family *Rubiaceae*: (a) with PCA method (b) with MDA method

**Table 5:** Comparison of minimum distance between data points for two herbs

Family	From	To	Minimum Distance	
			PCA Data	MDA Data
<i>Lauraceae</i>	F1S1	F1S2	0.01	2.26
	F1S1	F1S3	1.76	17.97
	F1S1	F1S4	1.39	47.15
	F1S2	F1S3	1.79	22.09
	F1S2	F1S4	1.41	46.88
	F1S3	F1S4	3.23	53.35
<i>Myrtaceae</i>	F2S1	F2S2	0.39	1.34
	F2S1	F2S3	2.01	30.34
	F2S1	F2S4	1.68	108.95
	F2S2	F2S3	1.69	33.85
	F2S2	F2S4	1.98	112.73
	F2S3	F2S4	3.70	105.02
<i>Zingiberaceae</i>	F3S1	F3S2	0.73	9.22
	F3S1	F3S3	2.44	46.02
	F3S1	F3S4	3.8	76.4
	F3S2	F3S3	2.83	57.73
	F3S2	F3S4	4.15	86.19
	F3S3	F3S4	1.21	31.57
<i>Annonaceae</i>	F4S1	F4S2	3.71	270.33
	F4S1	F4S3	0.07	4.12
	F4S2	F4S3	3.64	261.73
<i>Rubiaceae</i>	F5S1	F5S2	0	1.48
	F5S1	F5S3	3.83	256.01
	F5S1	F5S4	3.90	251.43
	F5S2	F5S3	3.53	252.16
	F5S2	F5S4	3.46	249.03
	F5S3	F5S4	0.02	3.42

## 6.2 Classification Analysis Based on Physical Properties

Four classification methods, namely the SVM, KNN, MLR, and Gaussian RBF Kernel were applied to investigate the classification performance between two discriminant analyses which are new projected PCA data and new projected MDA data. Two types of datasets for each family group species were used for two different purposes as shown in Figure 18. Dataset 1 was split into 70% for training and 30% for testing to build the herbs recognition system. Dataset 2 was the unseen data used to validate the system performance. Repeated K-fold Cross Validation (CV) method inspired by [36] as presented in Figure 19 was applied.

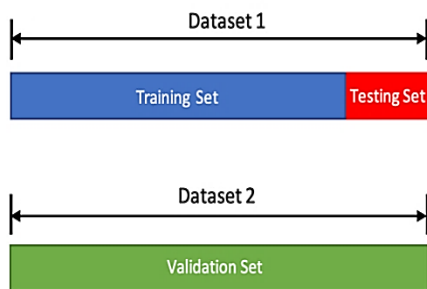


Figure 18: Training, Testing, and Validation datasets

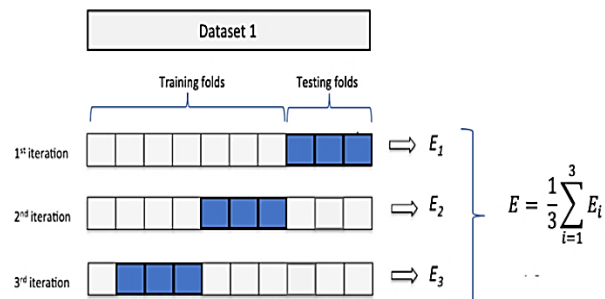


Figure 19: Repeated K-fold cross validation process

Dataset 1 was subjected to a 10-fold CV repeated 3 times. Dataset 1 was also divided into 10 equal parts which 7 of the 10 parts were treated as training set and the remaining 3 parts were testing set. This process was repeated for 3 times. The performance metrics were measured for each iteration. The final classification accuracy was computed by averaging the total of performance metrics. The robustness of the recognition system was validated with the unseen data (dataset 2). The accuracy performance for training, testing, and validation using all classification methods for five group species are shown in Table 6, Table 7, Table 8, and Table 9, respectively.

From the SVM results in Table 6, it shows that the classification accuracy with implementation of MDA is better than PCA for three families: *Lauraceae* (95.76%), *Zingiberaceae* (100.00%) and *Rubiaceae* (96.81%). Meanwhile, for family *Myrtaceae* (85.83%) and *Annonaceae* (92.59%) show the opposite results. The reason why some accuracy for the model with PCA is better than that with MDA is because SVM is a linear hyperplane which separates data into  $n$ -classes. Misclassification will increase if there is serious overlapping between two or more classes. For example, SVM decision boundary for Family *Lauraceae* is shown in Figure 20. SVM with PCA was not able to classify species FIS1 and FIS2 properly because a serious overlapping happened. Some data points of FIS1 fell into FIS2 boundary, and vice versa. Meanwhile, decision boundary for SVM with MDA showed some reduction in misclassification because of the new projected data had brought the data closer to each other. Hence, reducing the overlapping problem. However, it can be observed that MDA performed very well in validating the unseen data for all families.

Table 6: Classification accuracy for SVM

Group Species	Discriminant Analysis	Training Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)
<i>Lauraceae</i>	PCA	75.20	74.93	76.32
	MDA	96.03	95.76	93.26
<i>Myrtaceae</i>	PCA	86.09	85.83	82.01
	MDA	81.03	79.93	90.14
<i>Zingiberaceae</i>	PCA	93.08	92.85	89.38
	MDA	100.00	100.00	97.64
<i>Annonaceae</i>	PCA	92.73	92.59	93.33
	MDA	92.01	91.57	100.00
<i>Rubiaceae</i>	PCA	51.45	45.28	71.04
	MDA	97.29	96.81	87.01

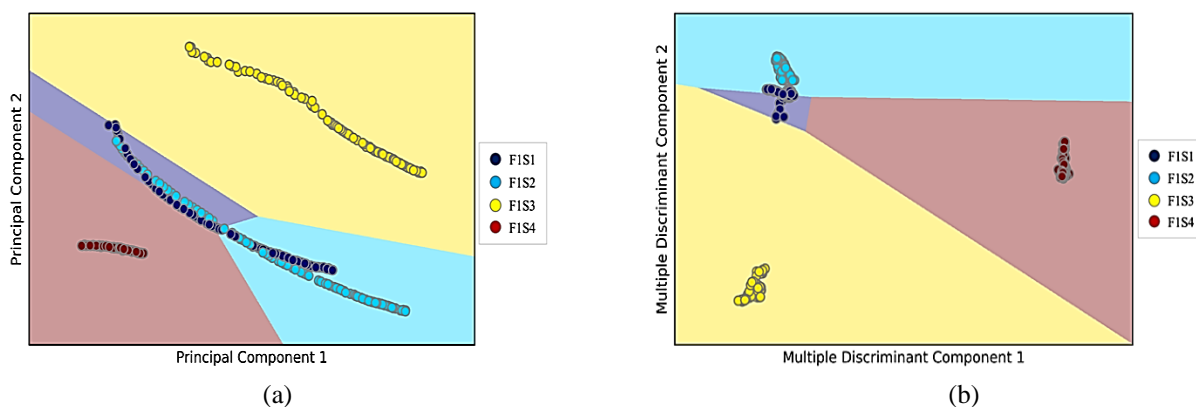


Figure 20: Decision boundary which separates species between Family *Lauraceae*: (a) with PCA method (b) with MDA method

For KNN method depends on the selection of  $k$ . Therefore,  $k$ , ranging from 3 to 10 were tested to define the best value of  $k$ . Table 7 summarizes the accuracy from one family versus of  $k$  value for system with PCA and MDA. From Table 7, it can be said that  $k = 4$  is the best value for KNN method.

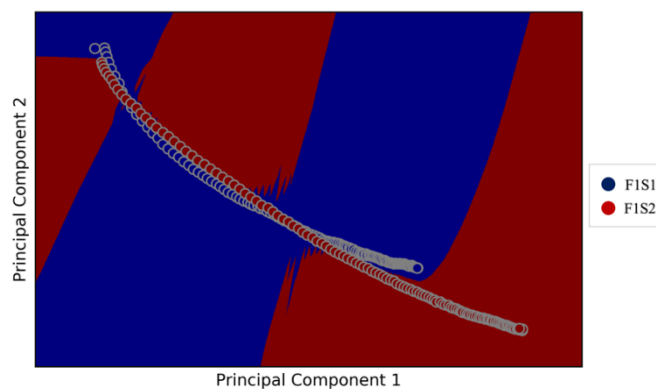
**Table 7:** Performance of  $k$  value for Family *Annonaceae*

$k$	Accuracy (%)	
	PCA	MDA
3	98.70	100
4	99.07	100
5	98.61	100
6	95.19	100
7	95.37	100
8	93.15	100
9	94.44	100
10	91.20	100

From the KNN results in Table 8, all classifications accuracy with MDA for all families show a better result compared to the system with PCA. Meanwhile, in PCA, Family *Myrtaceae* and *Zingiberaceae* shows 100% classification accuracy. The identification was accurate because all species in Family *Myrtaceae* and *Zingiberaceae* were well separated as shown in Figure 14(a) and Figure 15(a), respectively. The case is different for Family *Lauraceae*, *Annonaceae* and *Rubiaceae* in Figure 14(a), Figure 16(a) and Figure 17(a), respectively. Close proximity is the main characterization of KNN. Data points are clustered together with its nearest neighbors. The more the closeness of data points from two or more classes, the more it leads to misidentification. For example, some of data points for species FIS1 and FIS2 in Family *Lauraceae* were seen very close and overlapped. Some data points of FIS2 were misclassified as FIS1 because it fell into FIS1 boundary as shown in Figure 21. In validation, model with MDA showed better results in identifying the herbs species from unseen data compared to the classification model with PCA.

**Table 8:** Classification accuracy for KNN

Group Species	Discriminant Analysis	Training Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)
<i>Lauraceae</i>	PCA	93.74	87.78	95.83
	MDA	100.00	100.00	100.00
<i>Myrtaceae</i>	PCA	100.00	100.00	92.15
	MDA	100.00	100.00	99.65
<i>Zingiberaceae</i>	PCA	100.00	100.00	100.00
	MDA	100.00	100.00	100.00
<i>Annonaceae</i>	PCA	98.51	93.33	99.07
	MDA	99.99	99.91	100.00
<i>Rubiaceae</i>	PCA	80.05	67.22	90.69
	MDA	99.93	99.58	99.38



**Figure 21:** Decision boundary which separates FIS1 and FIS2, Family *Lauraceae*

Table 9 shows that MLR model with MDA performed better in testing and validation when compared to the model with PCA for all species. It can be observed that MDA helps improving the classification accuracy. Classification performance using Gaussian RBF Kernel is shown in Table 10. From the results, classification accuracy was better improved with the application of MDA compared to PCA for all species. The system using MDA projected data was able to identify the herbs species from unseen data, with the majority validation accuracy higher than the system using PCA for all group species.

As a conclusion, overall classification performance shows that herbs recognition systems using all four models (SVM, KNN, MLR, Gaussian RBF Kernel) with MDA are improved compared to the new data projected by PCA. New data projected by MDA can reduce the overlapping problem between classes which can lead to the misrecognition of herbs species.

**Table 9:** Classification accuracy for MLR

Group Species	Discriminant Analysis	Training Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)
<i>Lauraceae</i>	PCA	77.89	77.85	96.46
	MDA	99.58	99.58	100.00
<i>Myrtaceae</i>	PCA	96.47	96.46	89.79
	MDA	100.00	100.00	99.58
<i>Zingiberaceae</i>	PCA	100.00	100.00	100.00
	MDA	100.00	100.00	100.00
<i>Annonaceae</i>	PCA	94.23	93.98	94.07
	MDA	100.00	100.00	100.00
<i>Rubiaceae</i>	PCA	52.03	45.76	83.61
	MDA	99.54	99.38	99.17

**Table 10:** Classification accuracy for Gaussian RBF Kernel

Group Species	Discriminant Analysis	Training Accuracy (%)	Testing Accuracy (%)	Validation Accuracy (%)
<i>Lauraceae</i>	PCA	85.08	84.38	83.33
	MDA	99.93	99.72	99.93
<i>Myrtaceae</i>	PCA	98.17	97.85	90.97
	MDA	100.00	100.00	99.38
<i>Zingiberaceae</i>	PCA	100.00	100.00	100.00
	MDA	100.00	100.00	98.89
<i>Annonaceae</i>	PCA	85.20	84.91	85.37
	MDA	99.03	97.59	99.54
<i>Rubiaceae</i>	PCA	62.75	55.42	69.44
	MDA	99.73	99.24	98.19

## VII. EVALUATION OF ACCURACY PERFORMANCE WITH PREVIOUS WORK

This section discusses the performance comparison of this proposed E-Nose herbs recognition system with the developed E-Nose system done by [8]. The previous E-Nose system has been built from three group species samples of Family *Lauraceae*, *Zingiberaceae*, and *Myrtaceae*. Thus, accuracy under among these three families (12 herbs species) were compared as shown in Table 11.

**Table 11:** Comparison accuracy between previous work with the proposed system

E-Nose System	Classifier	Accuracy (%)
[8]	ANN (with PCA)	91.7
	ANFIS (with PCA)	94.8
Proposed system	KNN (with MDA)	100.0
	SVM (with MDA)	60.6
	MLR (with MDA)	96.4
	Gaussian RBF Kernel (with MDA)	90.0

PCA method has been applied in the previous work to project the data. Meanwhile, in the proposed system, PCA method was replaced with MDA method. Four classification techniques in proposed system are performed for 12 herbs species (Family

Lauraceae, Zingiberaceae, and Myrtaceae). Increasing the number of species that need to identify will reducing the system performance [37]. From Table 11, E-Nose system with KNN and MLR performed well compared to the previous work. Among all six classification techniques, KNN was concluded as the most robust proposed E-Nose system by given the highest percentage of accuracy for 100% compared to the other classification techniques.

## VIII. CONCLUSIONS

In this research, 19 herbs species from Family *Lauraceae*, Family *Myrtaceae*, Family *Zingiberaceae*, Family *Annonaceae* and Family *Rubiaceae* were used to analyse the performance of classification using SVM, KNN, MLR and Gaussian RBF Kernel techniques. The gas released from the herbs leaves using E-Nose system was the subject tested by these classification algorithms. The signal processing technique used for feature extraction had simplified the classification and produced optimal results. The PCA and MDA were the two discriminant techniques employed to distinguish different herbs species of the similar family group. As a result, both discriminant techniques extracted most appropriate information with less information losses and projected vector (dataset) onto a smaller dimensional space. Even though both techniques managed to discriminate and classify several herb group species, MDA was found to show better herbs species classification even for those in the same family group compared to PCA.

As a conclusion, overall classification performance shows that herbs recognition systems using all four models (SVM, KNN, MLR, Gaussian RBF Kernel) with MDA are improved compared to the new data projected by PCA. New data projected by MDA can reduce the overlapping problem between classes which can lead to the misrecognition of herbs species. Classification model for herbs recognition system based on physical properties was successfully designed using SVM, KNN, MLR and Gaussian RBF Kernel. The accuracy of E-Nose herbs recognition system had been successfully improved with MDA. Among all the classification methods, KNN had shown the highest performance. Furthermore, performance comparison between this proposed E-Nose system with previous work of E-Nose system had been evaluated. In comparison with the previous work, the proposed E-Nose system with KNN and MLR had performed better by 5.2% - 8.3% and 1.6% - 4.7% respectively. Herbs recognition system using KNN classification model compared to the other techniques, showed the most reliable performance with higher accuracy in herbs species classification. CV method was employed to investigate the classification performance of the model. The study's findings are helpful for researchers, as they can learn or identifying the plant species without the help of botanists or forest rangers.

## REFERENCES

- [1] Bodhwani, V., Acharjya, D. P., & Bodhwani, U. 2019. Deep residual networks for plant identification. *Procedia Computer Science*, 152: 186–194.
- [2] Naresh, Y. G., & Nagendraswamy, H. S. 2016. Classification of medicinal plants: An approach using modified LBP with symbolic representation. *Neurocomputing*, 173: 1789–1797.
- [3] Mustafa, M. S., Husin, Z., Tan, W.K., Mavi, M. F., & Farook, R. S. M. 2020. Development of automated hybrid intelligent system for herbs plant classification and early herbs plant disease detection. *Neural Computing and Applications*, 32:11419-11441.
- [4] Muneer, A., & Fati, S.M. 2020. Efficient and automated herbs classification approach based on shape and texture features using deep learning. *IEEE Access*, 8 :196747-196764.
- [5] Shabanzade, M., Zahedi, M., & Aghami, S. A. 2011. Combination of local descriptors and global features for leaf recognition, signal and image processing. *Signal & Image Processing: An International Journal (SIPIJ)*, 2(3): 23-31.
- [6] Vo, A.H., Dang, H.T., Nguyen, B.T., & Pham, V.-H. 2019. Vietnamese herbal plant recognition using deep convolutional features. *International Journal Machine Learning Computing*, 9(3): 363–367.
- [7] Zhang, W., & Wen, J. 2021. Research on leaf image identification based on improved AlexNet neural network. *Journal of Physics*, 2031:1-13.
- [8] Mohamad Yusof, U. K. 2015. Development of electronic nose for herbs recognition based on artificial intelligent techniques. Unpublished Master Thesis, Faculty of Engineering, Universiti Putra Malaysia, Serdang, Selangor, Malaysia.
- [9] Xu, M., Wang, J., & Zhu, L. 2021. Tea quality evaluation by applying E-nose combined with chemometrics methods. *Journal of Food Science and Technology*, 58(4): 1549–1561.
- [10] Haryono, Anam, K., & Saleh, A. 2020. Autentikasi daun herbal menggunakan convolutional neural network dan raspberry pi. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 9(3): 278 – 286.
- [11] Prasad, S., Kumar, P. S., & Ghosh, D. 2017. An efficient low vision plant leaf shape identification system for smart phones. *Multimedia Tools & Applications*, 76(5): 6915–6939.
- [12] Chiu, S. W., & Tang, K. T. 2013. Towards a chemiresistive sensor-integrated electronic nose: a review. *Sensors*, 13(10): 14214-14247.
- [13] Cui, S., Inocente, E. A. A., Acosta, N., Keener, H. M., Zhu, H., & Ling, P.P. 2019. Development of fast e-nose system for early-stage diagnosis of aphid-stressed tomato plants. *Sensors*, 19(3480): 1-14.
- [14] Jia, W., Liang, G., Jiang, Z., & Jihua, W. 2019. Advances in electronic nose development for application to agricultural products. *Food Analytical Methods*, 12: 2226–2240.
- [15] Tan, T., & Xu, J. 2020. Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: A review. *Artificial Intelligence in Agriculture*, 4: 104-115.
- [16] Harvey, B. S., & Flores-Sarnat, L. 2019. Development of the human olfactory system. *Handbook of Clinical Neurology*, 164: 29-45, 2019.
- [17] Huang, S., Cai, N., Pedro, P.P, Narrandes, S., Wang, Y., & Xu, W. 2018. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics & Proteomics*, 15(1): 41-51.
- [18] Yan, X., & Jia, M. 2018. A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. *Neurocomputing*, 313: 47-64.

- [19] Manojkumar, P., Surya, C. M., & Varun, P. G. 2017, Identification of ayurvedic medicinal plant by image processing of leaf samples, Proceeding of International Conference on research in computational intelligence and communication network, 3-5 November 2017, Kolkata, India: 351-355. USA: IEEE.
- [20] Kan, H. X., Jin, L., & Zhou, F. L. 2017. Classification of medical plant leaf image based on multi-feature extraction. Pattern recognition and analysis, 27(3): 581-587.
- [21] Prabhakar, P., Shyamdew, K., Philip, V. S., Kishore, P., & Roopashree, S. 2016. Robust recognition and classification of herbal leaves. International Journal of Research in Engineering and Technology, 6(4):146-149.
- [22] Basavaraj, S. A., Suvarna, S. N., & Govardhan, A. 2010. A combined color, texture and edge features-based approach for identification and classification of Indian medical plants. International Journal of Computer Applications, 6(12): 45-51.
- [23] Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. 2016. Efficient kNN classification algorithm for big data. Neurocomputing, 195(C):143-148.
- [24] Ghosh, S., Singh, A., K., Jhanjhi, N. Z., Masud, M., & Aljadhali, S. 2022. SVM and KNN based CNN architectures for plant classification. Computers, Materials & Continua, 71(3): 4257-4274.
- [25] Bhardwaj, A., Kaur, M., & Kumar, A. 2013. Recognition of plants by leaf image using moment invariant and texture analysis. International Journal of Innovation and Application Studies, 3(1): 237-248.
- [26] Satti, V., Satya, A., & Sharma, S. 2013. An automatic leaf recognition system for plant identification using machine vision technology. International Journal of Engineering, Science and Technology, 5(4): 874-879.
- [27] Connelly, L. 2020. Logistic regression. Medsurg Nursing: Pitman, 29(5): 353-354.
- [28] Borah, J. W. G. S., Hines, E. L., Leeson, M. S., Iliescu, D. D., & Bhuyan, M. 2008. Neural network based electronic nose for classification of tea aroma. Univ. Warwick Institutional Repos, 2(1): 7-14.
- [29] Abdolvahab, E.R., & Kumar, Y.H.S. 2010. Leaf recognition for plant classification using GLCM and PCA methods. International Journal of Computer Science & Technology, 3(1): 31-36.
- [30] Kaur, P., Robin, Mehta, R.G., Balbir, S., & Arora, S. 2019. Development of aqueous-based multi-herbal combination using principal component analysis and its functional significance in HepG2 cells. BMC Complement Alternative Medicine, 19(18):1-17.
- [31] Rana, P., Liaw, S. Y., Lee, M. S., & Sheu, S. C. 2021. Discrimination of four Cinnamomum species with physico-functional properties and chemometric techniques: application of PCA and MDA models. Foods, 10(11): 2871, 2021.
- [32] Srivastava, D.K., & L. Bhambhu, L. 2010. Data classification using support vector machine. Journal of Theoretical and Applied Information Technology, 12(1): 1-7.
- [33] Ben, J. M. Jason, M. D., Naomi, S. B., Mitzi, L. D., & Dudley, R. A. 2014. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. Journal of the American Medical Informatics Association, 21(5): 871-875.
- [34] Zhang, Z. 2016. Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine, 4(11):1-7.
- [35] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. 2017. Learning k for kNN classification. ACM Transactions on Intelligent Systems and Technology, 8: 1-19.
- [36] Sontakke, S., Lohokare, J., Dani, R., & Shivagaje, P. 2018. Classification of cardiocography signals using machine learning, Proceedings of the 2018 Intelligent Systems Conference, 6-7 September 2018, London, UK: 1-6. USA: IEEE.
- [37] Daniel, S. P., Ferri, C., & Ramirez, M. J. 2017. Improving performance of multiclass classification by inducing class hierarchies. Procedia Computer Science, 108: 1692-1701.



# Chapter - 5

## Forecasting COVID-19 from Lung X-Ray Images

K. Sujatha<sup>1</sup>, N.P.G. Bhavani<sup>2</sup>, V. Srividhya<sup>3</sup>, T. Kalpalatha<sup>4</sup>, B. Latha<sup>5</sup>, U. Jayalathsumi<sup>6</sup>, T.Kavitha<sup>7</sup>,  
A. Ganesan<sup>8</sup>, A. Kalaiivani<sup>9</sup>, Su-Qun Cao<sup>10</sup>

<sup>1</sup> Department of Electrical and Electronics Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.

<sup>2</sup> Department of ECE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. India.

<sup>3</sup> Department of Electrical and Electronics Engineering, Meenakshi College of Engineering, Chennai, India.

<sup>4</sup> Department of ECE, S.V. Engineering College for Women, Karakambadi, Tirupati, India.

<sup>5</sup> Department of Physics, Dr. M.G.R. Educational and Research Institute, Chennai, Tamilnadu, India.

<sup>6</sup> Department of ECE, Dr. MGR Educational & Research Institute, Chennai, Tamil Nadu, India

<sup>7</sup> Department of Civil Engineerin, Dr. MGR Educational & Research Institute, Chennai, Tamil Nadu, India

<sup>8</sup> Department of EEE, RRASE College of Engineering, Chennai, Tamil Nadu, India.

<sup>9</sup> Department of CSE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India.

<sup>10</sup> Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, China.

Email: <sup>1</sup> [sujathak73586@gmail.com](mailto:sujathak73586@gmail.com), <sup>2</sup> [sbreddy@gmail.com](mailto:sbreddy@gmail.com), <sup>4</sup> [drkalpalatha.thokala@gmail.com](mailto:drkalpalatha.thokala@gmail.com),

<sup>8</sup> [dragmephd@gmail.com](mailto:dragmephd@gmail.com), <sup>9</sup> [kalaiivaniarbarasan@rediffmail.com](mailto:kalaiivaniarbarasan@rediffmail.com)

*Abstract— Presently, the diagnosis of Corona Virus – 2019 (COVID-19) is a challenging task worldwide as the disease is spreading at a very faster rate. Several people are detected with COVID-19 and the data analyst say that the rate of spread of the disease is increasing exponentially. This investigation has facilitated the need for diagnosing the disease within a short duration of time from the X-ray images of the lungs. Artificial intelligence like deep learning algorithms is deployed to diagnose COVID-19 by maintaining social distancing. Real time data sets are gathered from the government hospitals for healthy as well as those who are affected by COVID-19. On development of a smart phone Application the patients themselves will record the respiratory sounds. The features are extracted using Discrete Wavelet Transform (DWT), where a threshold is applied to extract useful coefficients used to train the Deep learning Neural Networks (DLNN) using Fast Recurrent Convolutional Neural Networks (F-RCNN). The respiratory audio signals are captured to detect patients affected by Corona Virus by a way of non-contact, non-intrusive approach. This mobile phone App is effective in diagnosing the COVID-19 from the X-ray images of the Lungs. Even low income people can also use this technology. The effectiveness of the proposed system which uses DWT and thresholding has a F-measure of 96–98%. The forecasted results were in the range of 89%-95% for the above said algorithms. It is significant from the above results that the severe impact of COVID-19 can be diagnosed using a non-invasive mobile phone App using X-ray images.*

*Keywords— Artificial Intelligence, respiratory sounds, Convolutional NeuralNetwork, COVID-19, mobile phone App, Discrete Wavelet Transform, Thresholding.*

### ABBREVIATIONS

Fast Recurrent Convolutional Neural Networks (F-RCNN)

Discrete Wavelet Transform (DWT)

World Health Organization (WHO)

Computed Tomography (CT)

Magnetic Resonance Imaging (MRI)

Deep learning Neural Networks (DLNN)

## I. INTRODUCTION

The novel coronavirus was reported in 2019 in Wuhan City, China. The World Health Organization (WHO) Director - General, Dr. Tedros Adhanom Ghebreyesus. The WHO Director-General, Dr. Tedros Adhanom Ghebreyesus, declared that the disease was affected by the new virus named COVID- 19, in February 2020. However, the first case was reported on 11<sup>th</sup> January 2020 and healthcare workers were also infected due to the rapid spread from the patients. As per report the 11 million population of Wuhan City was restricted to travel and found to be under lockdown to avoid the drastic spread. To control the wide spreads of the COVID -19, most of the countries were under lockdown.

---

© 2022 Technoarete Publishing

K. Sujatha – “Forecasting COVID-19 from Lung X-Ray Images”

Pg no: 59 – 74.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch005>

The breaking news was that the cases continued to increase and spread to other countries Brazil, South Korea, Indian, Italy and Iran. Thus leads to pandemic worldwide serious health risk and subsequently WHO also raised the threat to the CoV epidemic to the increased rate of spread on 28th February 2020. The death rate was too high in China as compared to other countries by 13 times higher. However, the infection gets transmitted from an asymptomatic individual, which happens before the onset of symptoms. The spread crossed many countries and as of June 20, 2020 about 121,000 deaths in the U.S, 50,000 deaths in Brazil which made a great concern by the WHO to declare the COVID-19 a pandemic. The death rate was underrated due to the confines of investigation and screening for the presence of virus. Although the lower lethal rate of COVID-19 has been reported compared to SARS and MERS epidemics, but the transmissions of the SARS- COV-2 virus is widely spread than the other viruses mentioned above. World Wide it has been identified that about one in five individuals are at increased risk of severe COVID-19 disease which might be due to comorbidity [1]. Lockdown was announced for most of the countries and international airports were closed to avoid the spread. By mid-March, the number of cases was increased all over India and first death was reported on the same month 12th 2020. Meantime, the disease spread throughout the national except Sikkim. In India from 23th March 2020 lockdown was declared national level, to evade the transmission of deadly virus. During the one month duration, there have been 135,163 deaths globally and 507 deaths in India (Available at: <https://covid19.who.int/>. Accessed April 17, 2020, <https://www.mohfw.gov.in/> Accessed April 23, 2020).

Till date, seven human CoVs (HCoVs) have been identified. Only 2% of the population are healthy carriers of a CoV and 5% - 10% responsible for acute respiratory infections. Common human CoVs are HCoV-OC43, and HCoV-HKU1 (beta CoVs of the A lineage); HCoV-229E, and HCoV-NL63 (alpha CoVs). These viruses can cause common cold upper respiratory infections in immune competent individuals. Other categories of human CoVs are SARS-CoV-2, and MERS- CoV (beta CoVs of the B and C lineage). These viruses cause variable clinical severity featuring respiratory and high mortality rates up to 10% and 35% respectively.

Corona viruses are + SS RNA, enveloped, ranges from 60nm to 140nm in diameter and spike which gives the appearance like a crown under the electron microscope, hence the name forename as corona viruses. It was named by international committee based on the taxonomy of viruses as severe Acute respiratory syndrome corona virus 2. Shocking this virus causing pandemic disease is termed as COVID-19 by the WHO on the 11th February 2020. Recently studies showed that snakes, bats and pangolins would be the hosts for SARS-COV-2 [2]. It has been seen in two occasions from the past two decades that transmission of animal beta corona viruses to humans. The very first incidence of the virus was observed on 2002-2003 From the tracking of the transmission of virus, it was observed that the origin from bats which crossed over to human through an intermediate host of cats [3]. Still in Asia the L strain is very predominant but the other strains are slowly increasing as the world trade and transport comes back in action after lockdown. As the G, GH, GR and S strains are increasing the L and V strains are gradually disappearing. Apart from the strains recognized above there are new strains discovered on a day today basis by researchers. The variation has been observed to be less than 1% in most of the sequences [4, 5].

## II. LITERATURE REVIEW

Corona viruses belong to the order Nidovirales in the subfamily coronavirinae and classified into 4 genera Alpha Coronavirus, Beta Coronavirus, Delta corona virus and Gamma Corona virus [6]. The table represents the corona viruses types Table.1. There are six serotypes of coronaviruses are spread among the birds, mammals and bats being most for the variety of genotypes [7]. These viruses possess the largest viral genomes, length of 27 to 32Kb, the spike glycoprotein spikes are known to interact/ bind to the host membrane.

**Table : 1** Types of Corona viruses

	Types of viruses	Infects	Receptor
Alpha Corona virus	HCoVs, HCoV-229E and HCoV-NL63	Mammals	Amino peptidase N (APN) (Yeager CL et al., 1922)
Beta Coronavirus	SARS-CoV MERS- CoV SARS-CoV-2 and some HCoVs Non-SARS human species – HCoV-OC43 and HCoV-HKU1	Mammals	Angiotensin – converting enzyme (ACE-2) (Hofmann H et al., 2005), Hemagglutinin- esterase activity and utilize sialic acid residues as receptor (Vlasak R et al., 1988)
Delta Corona virus	Bottlenose dolphin corona virus HKU22	Primarily infects birds and humans	
Gamma Corona virus	Avian infectious bronchitis	Primarily infects birds and humans	-

The seasonal variations are also plays a major role in the infections. Depending on the seasonality especially based on the

climate: Temperate and subtropical regions. In the temperature climates, infections occurs in the winter, also a smaller peaks seen during the spring even the infections can occur at any time of the year [8]. The eight years study report from Michigan, U.S revealed that infections were identified between December and May, peak in January and February and a fall in spread only by 2.5% of infections during June and September [9].

A detailed study from Scotland using molecular testing for the respiratory viruses was performed with 74,000 samples which includes adults and children from 2005 to 2017. The output of the study gave the idea about the incidence of age criteria and seasonality of the infections. The three species in their age incidence patterns, such as OC43 most commonly found in infants, young children and elderly, 229E was found in adults of the all ages and NL63 was found in infants (under a year of age) [10]

Another survey was taken in Sør-Trøndelag County hospital, Norway. Samples were collected from under 16 years of age children. It was been found that both the strains HCoV-OC43 and HCoV-NL63 are the reason for the infections and also been predicted that were epidemic every other winter and 229E strain was unusual [11]. In the case of subtropical regions, Seven year research was done in Guangzhou, China. From there study, it was clearly said that outbreaks can happen any time of year however, predominantly spread seen during the spring. In other survey, it has been reported that HCoV-OC43, HCoV-NL63, HCoV-229E, and HCoV-HKU1 predominate unpredictably in certain years and in certain parts of the world [12]. The major transmission mode is from human- to human transmission. The mode of the transmission is by the droplets during cough and sneezing or by fomite and virus also get released in the stool [13].

### 2.1 Mechanisms of viral infection

The spike protein of SARS-CoV mediates binding with membrane fusion through the ACE-2 receptor. The spike proteins of the virus are responsible for receptor binding and cell membrane fusion by S1 and S2 domain respectively. ACE-2 receptors are expressed in many tissues, but high levels of receptor are expressed in the alveolar epithelial type II cells. After the binding and fusion, down regulates the ACE-2 intracellular signaling this causes inflammation, vasoconstriction and fibrosis in the lung. Many studies also showed that, patients with infections have a tendency of elevated concentration of pro-inflammatory cytokines, procalcitonin, C-reactive protein [14]. However the COVID-19 severity is based on the clinical features, the infection can be asymptomatic, mild, moderate, severe and critical. In the case of asymptomatic, there would be no symptoms and sign, but the running nose, sneezing, acute upper respiratory tract infection, fever, and fatigue in few cases patient might not have fever, but have nausea, abdominal pain, diarrhea and vomiting for the mild COVID 19 infections. For the moderate cases, dry cough, fever and wheezing and CT shows lung lesions. The symptoms for the severe patient have fever, cough and associated with gastrointestinal symptoms such as diarrhea, oxygen saturation which becomes < 92%. In the case of critical stage, acute respiratory distress syndrome (ARDS), which is the respiratory failure, seen with septic shock and finally leads to organ failure, can be life threatening [15]. We all have to know about the immunity against infection, however this novel virus is very new to the entire global population hence the development of immunity is a great question. The whole population is in a risk until our immune cell develops herd immunity or either through vaccination to limit the spread of the virus. From the know report, herd immunity is approximately 70%, thereby immunity would be preferably best by taking the vaccination [16].

The samples were collected from nasopharyngeal, oropharyngeal swabs of the suspected patients. The samples were detected by the presence of antigen by the immunological method by rapid assay. Another reliable method is real-time polymerase chain reaction (RT-PCR) to identify by the molecular level characterization. The investigation in the laboratories showed the presence of white blood cell count, platelet count usually normal or low. However, CSR and ESR are elevated and procalcitonin level found to be normal. In the presence of bacterial co-infection, high level of procalcitonin was observed along with elevated level of ALT/AST, creatinine, CPK, LDH, prothrombin time with severe disease (Tanu Singhal et al., 2020). Once the patient showed positive for the COVID -19, treatment was given according to the severity. They will be discharged from the isolation ward after the two consecutive RT-PCR test. However, the health workers were instructed to wear N95 mask and PPE suits and as they were in direct contact with patients, thereby need to be monitored for development of symptoms of COVID-19.

However, all types of respiratory viral infections including COVID-19 will show similar symptoms it would not be possible to differentiate the novel virus by the routine lab test. One of the important criteria is travel history, but in the epidemic spreads the travel history would become irrelevant..

### 2.2. Novelty/Innovation

The proposed online based COVID-19 detection smart phones App is built, to remotely measure and monitor the patients affected by COVID-19 by using a non-invasive geo-spatial method. This painless method enables user friendly measurements from the X-ray images of lungs by embedding intelligent signal processing algorithms which will process the X-ray images of the lungs captured by the camera in the smart phone. This smart phone App extracts the features and shape of the X-ray images of the lungs and will classify the COVID-19 at onset, medieval and chronic stage with specific and accurate measurements along with the geo-spatial information instantly. In addition, geo-tagged signals and the estimated location of abnormalities

were recorded during diagnosis. All these recorded data and associated information were used for the accuracy assessment. On the other dimension, this novel technology will place an end to the challenge involved in the disposal of biomedical waste, thereby offering a non-invasive measurement system together with the geo-spatial information about the patients so that they can tracked and monitored in a better way during this pandemic, COVID-19 situation.

### 2.3. Study Objectives

The objectives of this project are listed below:

- i. To implement a non-invasive, remote system for detection of COVID-19 by screening the X-ray images of the lungs along with the geo-spatial information about the patients.
- ii. To relieve the patients from getting exposed to harmful radiation for longer duration during Computed Tomography (CT), X-ray imaging and Magnetic Resonance Imaging (MRI) scanning techniques repeatedly.
- iii. To minimize the cost involved in laboratory analysis which will require the usage of chemicals and reagents.
- iv. To eliminate the biochemical hazards involved in disposal of related biomedical waste.
- v. To nullify the time delay involved in laboratory scale analysis to identify COVID-19 and the geo-spatial information about the patients, facilitates to track and monitor them consistently.
- vi. To enable a remote, online and continuous monitoring system to screen the patients affected by COVID-19 at the earlier stage with their Geo-spatial information which can be prevented from spreading further by analyzing at least 10,000 samples within no matter of time.

## III. METHODOLOGY

The study is planned to be carried out in the rural areas of Chennai. Area to be focused are Tiruvallur and Chengalpattu districts which has a population of 70 lakhs.

Inclusion criteria – Both male and female adults who are COVID positive.

Adults in the age group of 18 and above will be included.

Written informed consent will be obtained.

Exclusion criteria – Non COVID infected adults.

Sample size- 20,000 samples to be collected from rural areas

Study Design -Cross sectional study.

The primary objective is to create a low cost app that will help detect Covid virus and its variants. The data collected is correlated with results obtained using RT-PCR. Identification of COVID-19 virus by RT-PCR method is very accurate in the detection of virus as well as identification of specific mutant. The sensitivity of RT-PCR for very small fragments of virus material is very high and also it can continue to detect the fragments of SARS mutant.

Prolonged infection in immune compromised individuals can occur for a month. Healthy people can become re-infected and show no symptoms. They are asymptomatic carriers and can be easily go undetected by RT-PCR. But the real time data provided by the app is very powerful to detect such individuals and helps to prevent further transmission of the virus.

The COVID testing facility will be set up in A.C.S medical college and Hospital. A high accuracy Chiagen RNA extraction kit will be set up in the testing lab.

Health workers will be instructed to wear N95 mask and PPE suits and as they come in direct contact with patients, thereby need to be monitored for development of symptoms of COVID-19.

Dr MGR university has 3 medical colleges and the faculty of Dr. M.G.R. Educational and Research Institute with the good blend of Academia and Research expertise will be able to complete the project successfully with due expectation of ICMR.

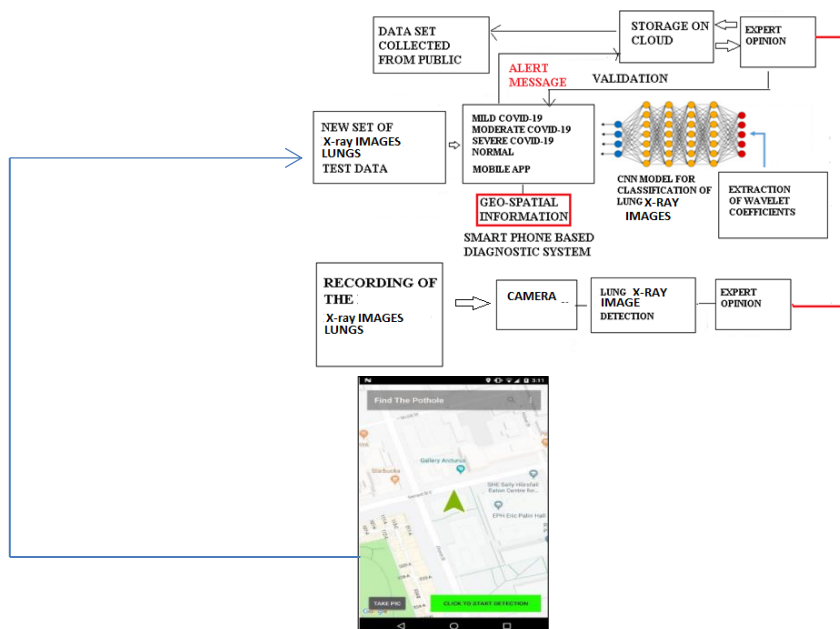
### 3.1. Study Design

Mobile phone App for screening and detection of COVID-19 from the lung sound during respiration along with the Geo-spatial information. The scheme for intelligent X-ray images of the lung diagnosis system is proposed here. The proposed scheme comprises of many stages with compatibility to share, store and access the related database by the physicians, patients and researchers from any part of the world unbounded by time limits along with their geo-spatial information as illustrated in Figure 1. Such kind of flexible systems will facilitate the development of robust algorithms for diagnosis of COVID-19 without human intervention there preventing the spread of the disease.

Patients suffering from various types of respiratory disease develop similar symptoms as that of COVID-19 which includes cough, increase in body temperature, running nose, congestion of lungs and difficulty in respiration. All these symptoms are related to any type of viral infection and so it becomes difficulty to differentiate Corona virus from other types of viruses responsible for other respiratory diseases. Hence investigation has shown that extraction of wavelet coefficients from the X-ray images of the lungs will enable a non-invasive, instantaneous, early and accurate diagnosis of the COVID-19 without

encountering time delay. The need of the hour has enabled the researchers to think in a new direction to develop a smart phone App based technology with signal and intelligent classification techniques. This state of the art technology will be a freely available technology on the Google, where even a layman possessing a smart phone can download the App and use it for himself alone. As a result, each and every individual possessing a smart phone can download this App and use it for detecting the COVID-19 with less human intervention during this pandemic. The Android API provides the best available location information based on Android 1.5 (API Level 3) location providers such as Wi-Fi and GPS (Global Positioning System) available on each smart phone. Furthermore, the API can deliver the accuracy of the reported location information determined by the available location providers. Also, this concept is easily portable and does not incur any transportation charges. The main highlight is that there is no direct contact between the patient and physician, which will control the spread of Corona virus when people come into contact. To throw more light on the technology of the proposed smart phone App is that, it is absolutely free of cost and user-friendly immaterial of the knowledge possessed by the users thereby enabling early diagnosis along with the geo-spatial information about the patients, so that it becomes easy for the physicians and data analysts to track and monitor them.

Presently, the smart phone-based App is used to capture the X-ray images of the lungs to infer the effect of COVID-19 along with the geo-spatial information about the patients. The researchers have not developed the technology to infer the COVID-19 by using a non-invasive method. Hence from our side, we have focused to develop a robust smart phone App which will use intelligent image processing algorithms to detect COVID positive from the X-ray images of the lungs along with the geo-spatial information about the patients. Some other researchers have developed an embedded system-based sensor module to detect COVID-19 from lung sounds. Even with such kind of portable devices or miniature instruments, it requires a dedicated device to detect COVID-19 alone. This proposed technology, will be a smart phone App which will be installed in the mobile phone so that even a layman, when capturing the X-ray images of the lungs using the camera in mobile phone and will be able to determine the presence of COVID -19 along with the geo-spatial information, thereby eliminating the usage of separate device for monitoring COVID-19. The proposed smart phone App once developed will be capable to infer COVID-19 from the X-ray images of the lungs by a non-invasive technique which relieves the patients from pain. This technique will also avoid the time delay involved in sampling the samples in the laboratory without use of any chemicals or reagents. Moreover, the proposed technology is a layman approach, which will avoid the direct contact of the patients in the Laboratory in spite of COVID-19 enabling social distancing along with geo-spatial information about the patients, so that it becomes easy to track and monitor them.



**Figure 1.** Schematic for Smartphone App development for identification of patients affected by COVID-19 from X-ray images of lungs using PCA and Deep learning technique along with geo-spatial information

A clinical assessment was conducted at A.C.S Medical College, Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu, India to capture the respiratory sounds of the lungs and relate its wavelet coefficients by clinical lab test (swab test) so as to determine COVID positive patients. Patients with various histories of various lung sounds during breathing were analyzed. Samples pertaining to three categories (onset stage, medieval stage and chronic stage) of COVID levels were recorded from Laboratory scale swab test. Also, few samples pertaining to normal X-ray images of the lungs were gathered. Preliminary analysis included collection of samples on a small scale with 51 samples spread out in three categories as



discussed earlier and roughly around for COVID positive (mild, moderate and severe) and another 25 samples of COVID negative cases for whose X-ray images are gathered. Quality control measures were incorporated to eliminate the noise that had affected the quality of the breathing sound is eliminated using thresholding technique available in wavelet tool box in MATLAB. Pulmonologist and Lab technician estimate the various conditions of COVID positive (mild, moderate and severe) from the extracted wavelet coefficients as inputs to the Fast Recurrent Convolutional Neural Networks (F-RCNN). All the samples were extracted using swab test. After this the patient's X-ray images of the lungs must be captured. At the same time the location finder is also enabled on the smart phone by Internet activation which will enable the data analysts and physicians to identify the geo-spatial information about the patients affected by COVID-19. The flowchart for training , testing and validation the samples on large scale on real time is illustrated in Figure 5 and 6 respectively.

All the X-ray images of the lungs were captured with an Apple iPhone 11 Pro Max with 64GB memory storage with default settings in the voice recorder. Before the X-ray images of the lungs are being captured, they are preprocessed for noise removal.

The X-ray images of the lungs are captured during clinical investigation inside a room, where the conditions are favorable to differentiate the X-ray images of the lungs made by the other internal organs in the human body. The technology behind the development of detection and monitoring the COVID patients with their respective geo-spatial information is shown in the Figure 2. Similarly Figure 3 depicts how the signal processing and deep learning algorithms will be embedded in the Geographic Information system so as to track and monitor the status of COVID-19 affected people along with their location.

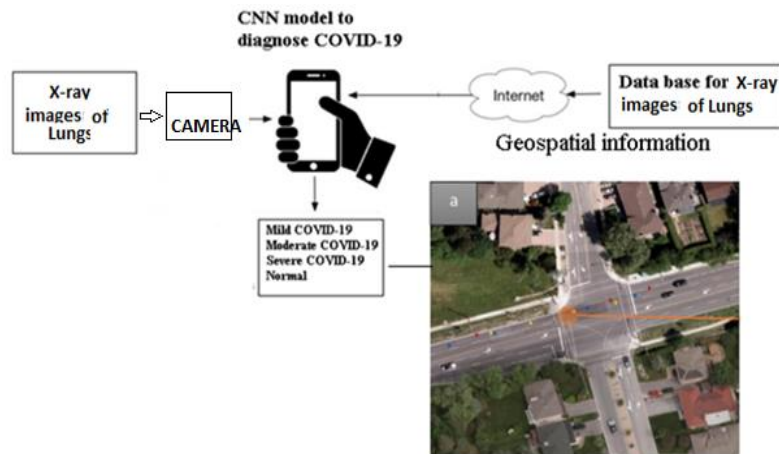


Figure 2. Technology for COVID-19 detection with geo-spatial information using smart phone App

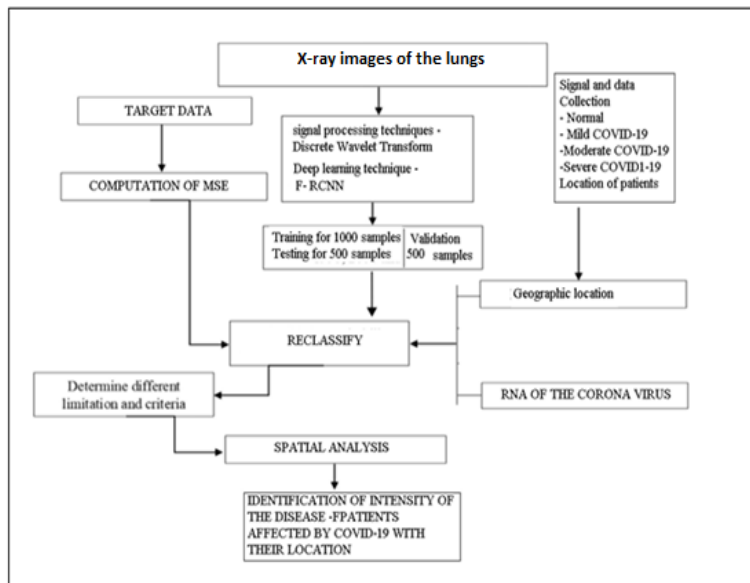


Figure 3. Mobile App for COVID-19 detection along with the Geo-spatial information from X-ray Images of Lungs



### 3.2. Sample Size

The total sample size is around 20,000 samples. 10,000 samples of the lung X-ray images along with the RT\_PCR test and swab test will cover mild, moderate and severe COVID-19 cases along with the persons who are normal. For testing the proposed algorithm, 5,000 samples will be used and the remaining 5,000/- samples will be used for validation. The flow diagram is represented in Figure 4.

The principle of operation of the customized smart phone App for detection of COVID-19 along with geo-spatial information of the patients is depicted in Figure 5. The proposed algorithm will have high sensitivity and precision with excellent screening capacities. Once the smart phone App is developed, then the calibration will be done by asking the patient to install the App onto their smart phone. After that, the patient needs to open the App, which will in turn enable the camera in smart phone to be switched ON. At the same time, an App which is installed in prior, will measure the light intensity to open, measure the value of the light intensity at that instant and make necessary auto adjustments to match the illumination level in the surroundings. The images of the lung X-rays are captured and then processed by the app in the smart phone which uses intelligent image processing algorithms without the need for sampling the blood extracted from the patient's body using conventional invasive methods. The flowchart for validation is depicted in Figure

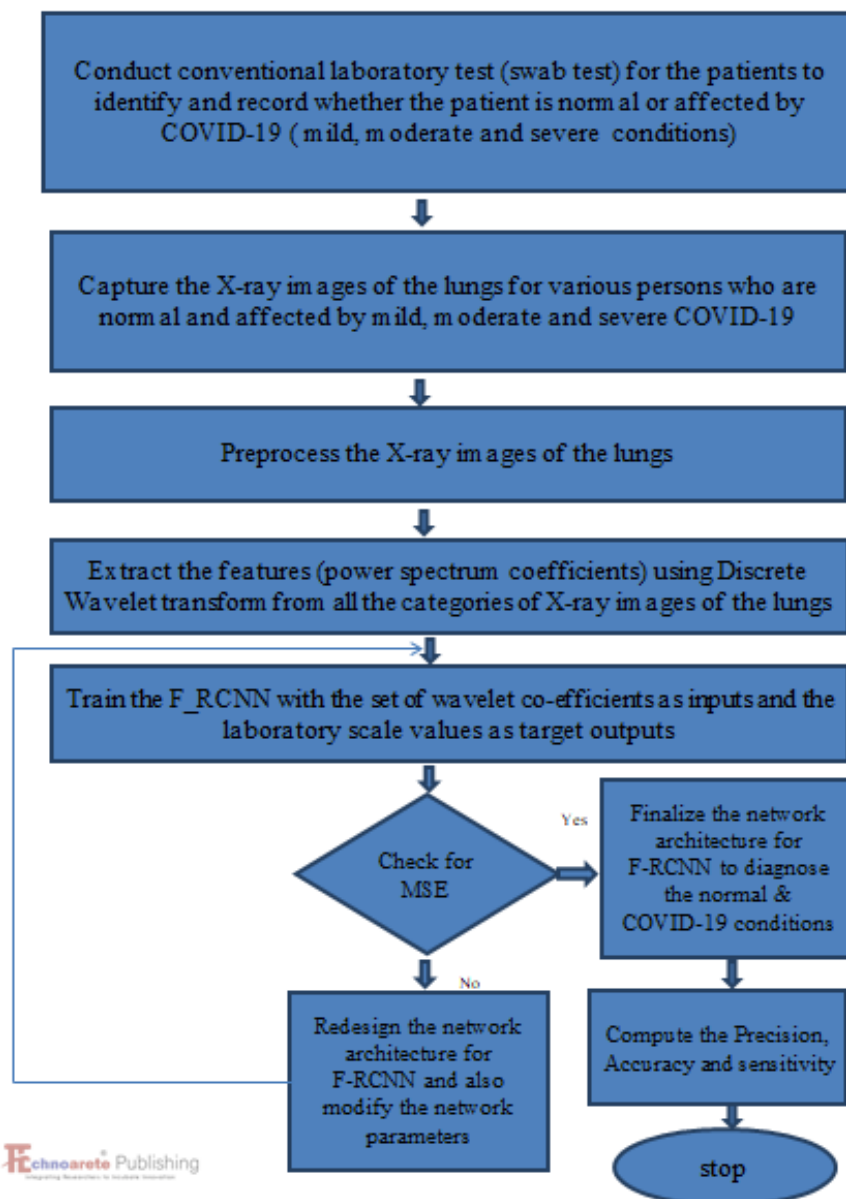
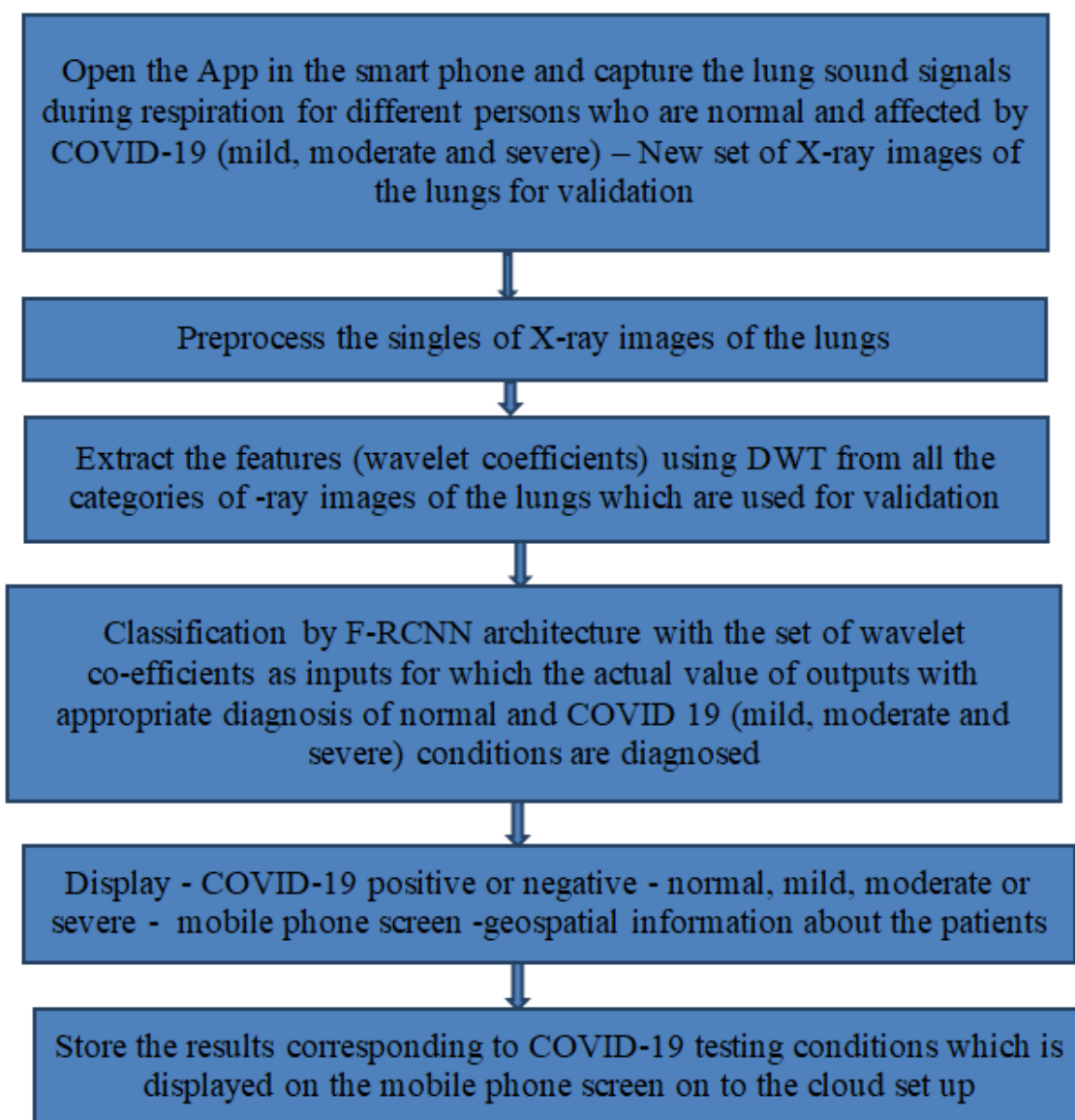


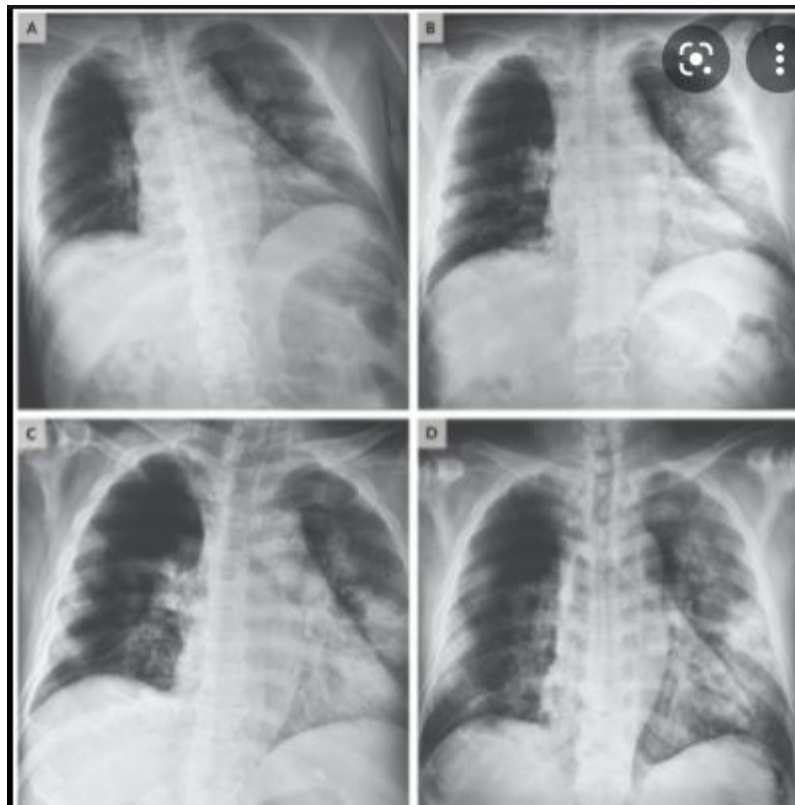
Figure 4. Flow chart for testing the F-RCNN algorithm



**Figure 6.** Flowchart for Validation after the development of Smartphone App

### 3.3. Data collection & statistical analysis plan:

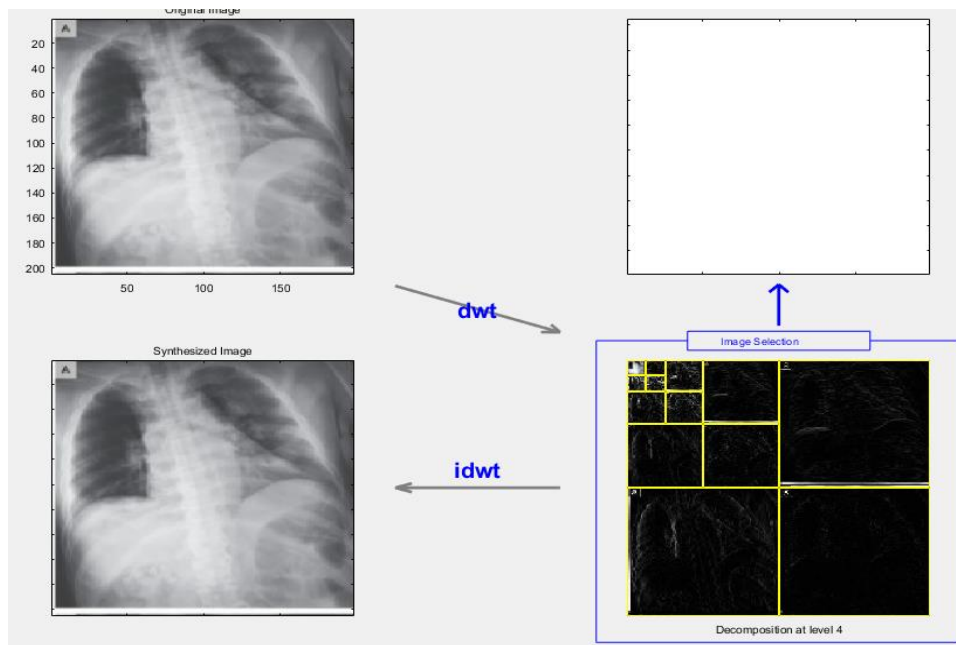
A correct indication of the COVID positive can be detected from the X-ray images of the lungs. All the organs in the human body create a unique sound during their normal functioning periods. When there arises an abnormal situation, these sounds vary which will have a unique variation in the amplitude and spectrum levels. This concept, indeed finds an important application in detection of COVID-19 from the X-ray images of the lungs in Figure 6. This smart phone based App can produce equivalent results as that of the swab test through laboratory analysis. This kind of smart phone App will facilitate the patients to monitor and detect the presence of Corona Virus by them ensuring a remote, non-contact, non-invasive method. This smart phone App helps the patients to diagnose the COVID and infer whether the patient is in need of critical emergency care. This smart phone-based App testing can be used by Health workers and can be used for testing in low-income countries like India and Nigeria, where there is high smart phone penetration and prevalence of COVID-19.



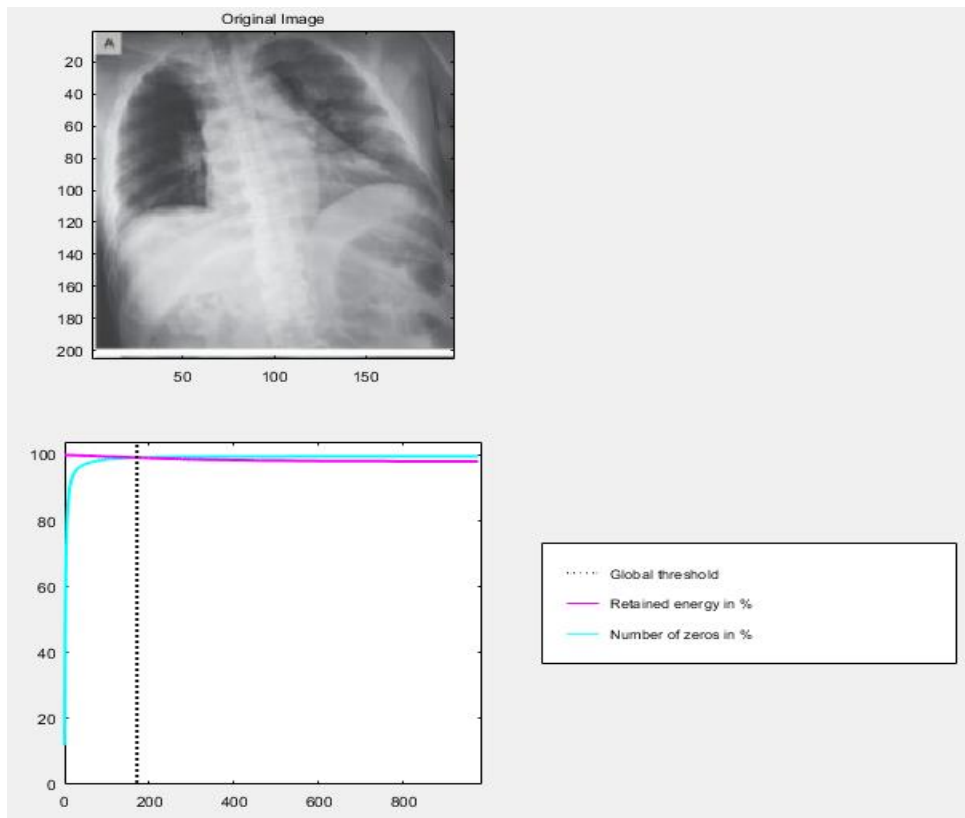
**Figure 6.** X-ray images of Lungs – Sample images

### 3.4. Wavelet Transform

This signal decomposition may not serve all applications, as in the case of certain images relating to biomedical applications. Wavelets are obtained when the signal is decomposed using the standard set of functions. Scale (or dilation) defines how “stretched” or “squished” (Figure 7) a wavelet is which is related to frequency. The spatial and temporal coordinates of the wavelets are identified and represented graphically.

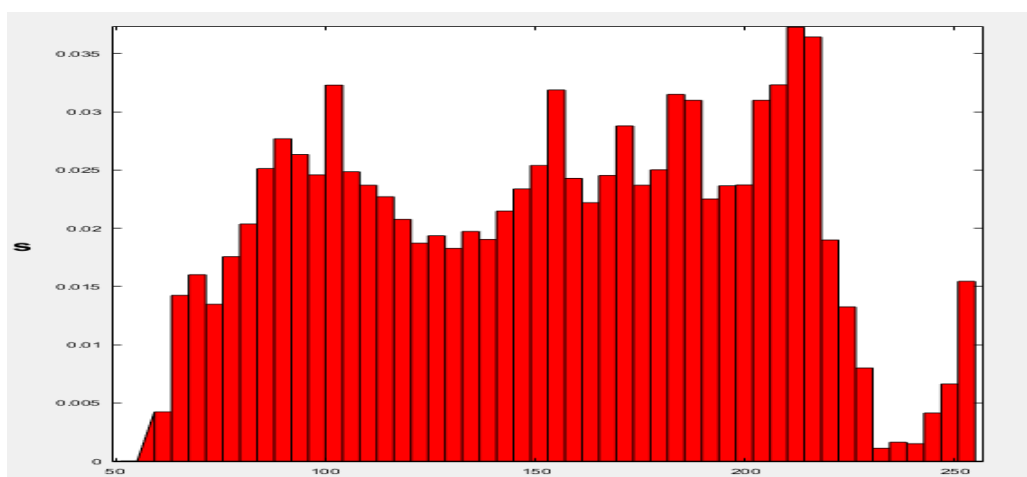


**Figure 7(a).** Simulation results for DWT and IDWT

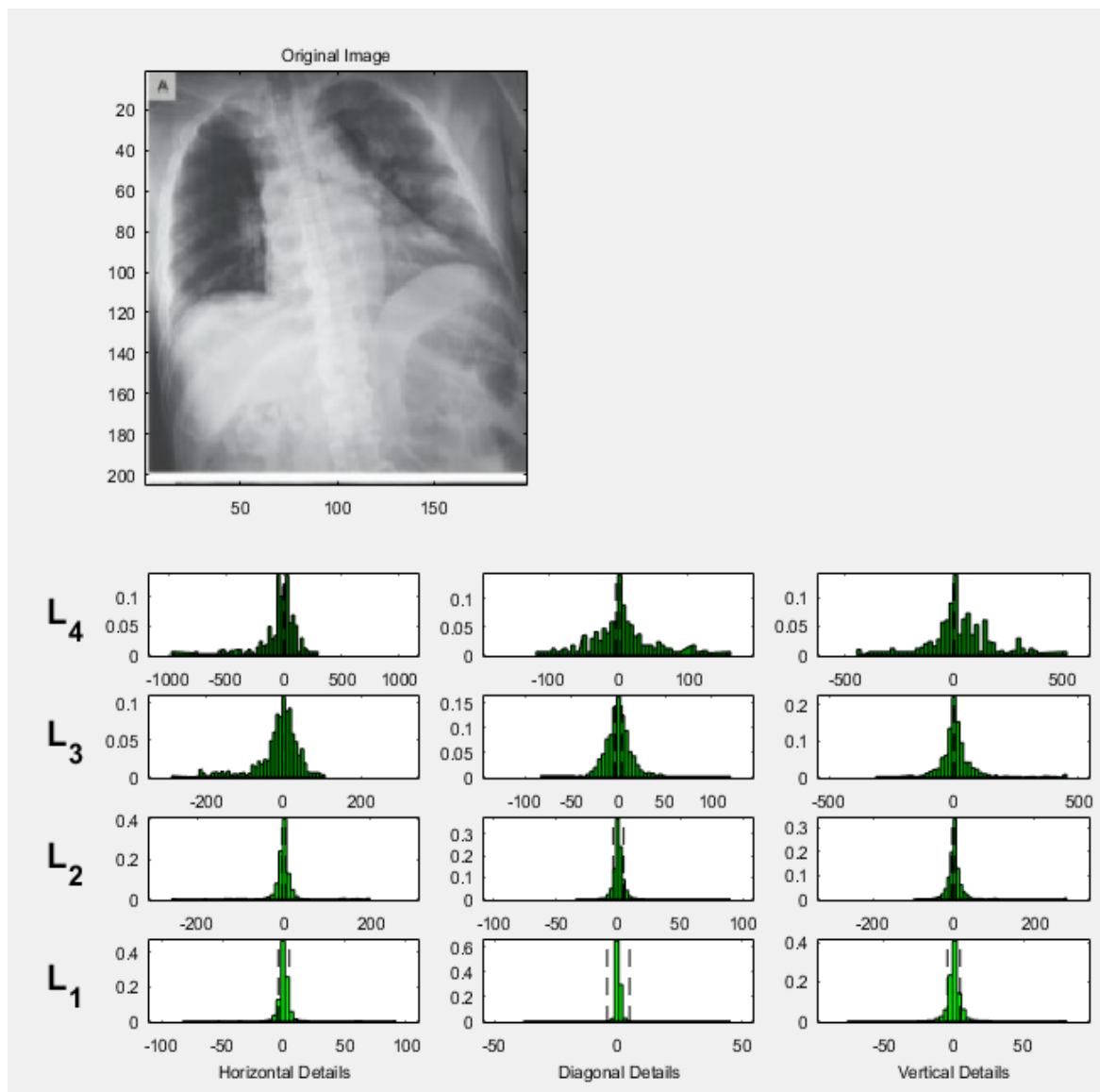


**Figure 7(b).** Graphical representation for various thresholding using DWT for X-ray images of the lungs

The factor “a” is defined as the scaling factor. The waveform gets squished for decrease in the scaling factor and tracks the higher frequency components. Waveform for the wavelet is stretched, if the scaling factor is increased and tracks the low frequency information. The factor “b” denotes the position of the wavelet. The wavelet is shifted to its left if the value of “b” is decreased and vice-versa. Like the continuous and discrete signals, the Wavelet Transforms are also classified as continuous and discrete transforms. The scaling factor and the position are countless in the case of Continuous Wavelet Transform (CWT). Conversely, for Discrete Wavelet Transform (DWT), the scaling factor and the positions are available only for discrete instants of time. The various types of wavelets are graphically depicted in Figure 8.



**Figure 8(a).** Histogram Analysis



**Figure 8(b).** Graphical representation for various features from the X-ray images of the lungs using DWT Fast Recurrent Convolution Neural Network (F\_RCNN)

The Fast\_RCNN algorithm is done using three different steps. Feature maps are generated then region proposals are produced. Region of interest pooling is done along with RCNN to predict the class of the bounding box.

The VGG network architecture is used to get the feature maps. The VGG network architecture contains different convolution layers with max pooling, fully connected layers and is using a softmax classifier. The required feature maps are drawn using the VGG architecture. Using the feature maps obtained by the VGG architecture the region proposed network proposes the random regions.. The Region Proposal Network gives the anchors with two things in mind, first one is the objectness score, if any object is present in the particular anchor box and the second one is the regression of the bounding box for adjusting the anchors with the correct position for the object. After the Region Proposal Network the obtained proposals undergo non-max suppression for removing the overlapped proposals.

Then the region of interest pooling is done using the obtained proposals, the image is cropped and fed into a region-based convolutional network, by extracting features the class label will also be predicted. The pseudocode is shown in Table 1.

**Table 1:**Pseudocode for the Fast\_RCNN

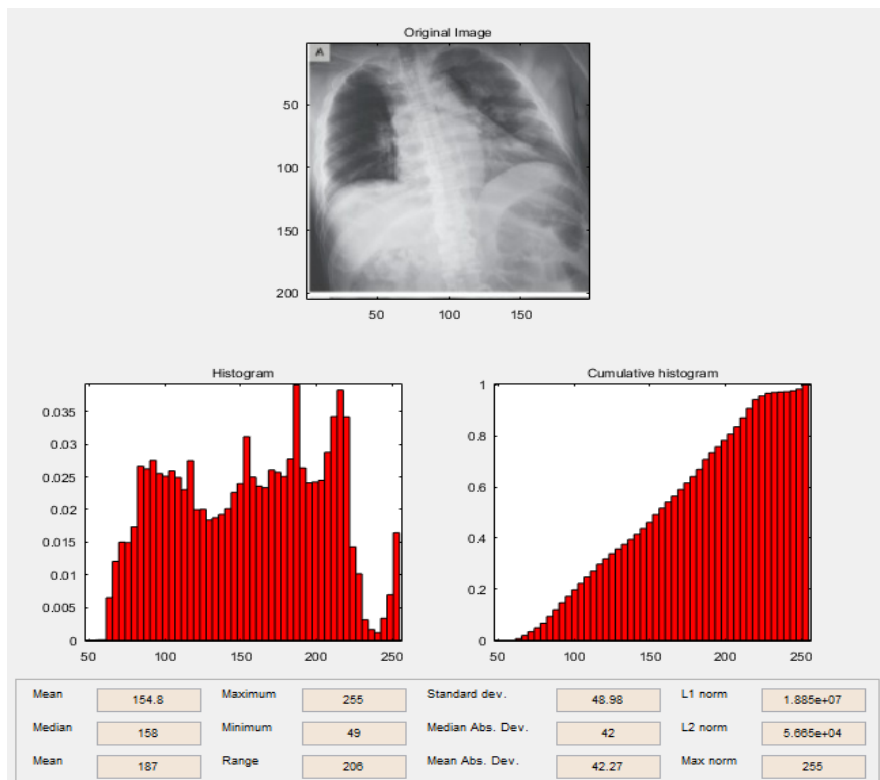
//Input
I :Labelled. image file along with the XML file.
1. Wavelet Coefficients are fed into the VGG architecture.
2. Feature maps are generated
3. Features are fed into the network and operations are done
4. Region proposal is generated using bounding boxes
5. Non-Max suppression is performed
6. Region of interest pooling is done on the remaining proposals
7. Bounding box is formed for the given input values
//Output
Test images of X-rays for lung with the bounding boxes along with the label names and accuracy.

**3.5. Pre-processing**

The noise removal is also done using the wavelet tool box. De-noising helps to improve the quality of the signal, so that appropriate and exact values of the features alone can be extracted. A high Pass Filter (HPF) allows the signal portion corresponding to high frequency values alone to pass through it. The transfer function for the HPF in discrete domain is given by  $1 - 0.99z^{-1}$ .

**3.6. Feature extraction using Wavelet Transform (WT)**

An extensive variety of features are extorted from the lung sound signal during respiration using one dimensional discrete wavelet transform. The preferred feature set serves as the input and includes the details and approximation components to train and test the F-RCNN as shown in Figure 9.



**Figure 9.** Wavelet Analysis for X-ray images of the lungs with COVID positive

**3.7. Identification using F-RCNN**

A supervisory segmentation scheme, to trace and capture the X-ray images of the lungs during respiration even in a noisy environment can be facilitated by implementing F-RCNN to detect the abnormality in the training phase by using 70% of the collected signals itself, so that an appropriate model for X-ray images of the lungs analysis can be launched. During testing of F-RCNN, only 20% of the collected sample X-ray images of the lungs are used for detection of abnormality. The remaining 10% can be used for validation after the development of the mobile phone app.



Fast Recurrent Convolution neural networks (F-RCNN) includes a variety of applications in the domain of machine vision, Big data analytics and a lot of other classification applications where enormous quantity of data is be processed and classified. Similar to the conventional Artificial Neural Network (ANN), the architecture has a number of interconnected layers of processing elements that are related by random numbers called weights. Since convolution operation is performed in the layers of CNN, it performs the function of the filter which facilitates the noise removal. CNNs learn the filters during the training process, which can be thought of a way to generate important features out of the data. The CNN requires apriori knowledge about the data base for obtaining accurate classification.

### 3.8. Development of mobile App

Once the simulation using MATLAB is over, the next stage is the development of smart phone App.

#### Mobile Application Development Architecture

The following block diagram explains a high-level architecture of the mobile application (to be developed). This architecture will have –

- Mobile Client – which will act as a user-interface, and can be installed on their mobile phones
- Server – which will be deployed on AWS as an API and this will be responsible for the data processing and notifications.

The following are the various components present on the mobile application -

Mobile Application - is the container which enables the user to authenticate, and capture his/her X- ray images of lungs. This application will consume the various API's developed, to process the captured signals and detail out the deficiencies along with the geo-spatial information based on the algorithm. Server APIs are the connectors which bridge the mobile application and the data processing engine hosted on the application server. These API utilize its own authentication mechanisms for a secured data transport.

Data processing engine is developed with a pre-approved algorithm to process the captured X-ray images, match it with the pre-built models, and provide a response back with the details (Chinmay C., et al, 2014). This module also has its data ingestion module for the administrator to upload more samples, or tweak sample results pertaining to the detection of COVID-19 along with the Geo-spatial information about the patients. Error handling module provides the base framework to capture erroneous data, or any server related downtime. This module will be responsible for providing the user readable messages. Application Servers are the containers to host the server APIs.

## IV. RESULTS AND ITS RELATED DISCUSSION

The main motive behind developing accurate technology is to diagnose the COVID-19 maintaining social distancing thereby preventing the spread of the viral infection. Health conditions were recognized by the physician who scrutinized the patients for which various laboratory testing have been carried out to identify the exact reasons.

### 4.1. Extraction of Wavelet Co-efficient

The proposed technology was validated with nearly 10% of the total X-ray images of lungs collected were treated as a predominant factor for remote diagnosis of COVID-19. The outcome was predictable from the segment of the X-ray images of lungs captured using the mobile phones. Then these X-ray images of lungs were pre-processed for noise removal and split up manually in consultation with the experts. The main benefit of this method is that the proposed F-RCNN is robust enough that it is capable of restoring even the lost portion of the X-ray images of the lungs along with the geo-spatial information as in Figure 10(a) and (b) respectively.

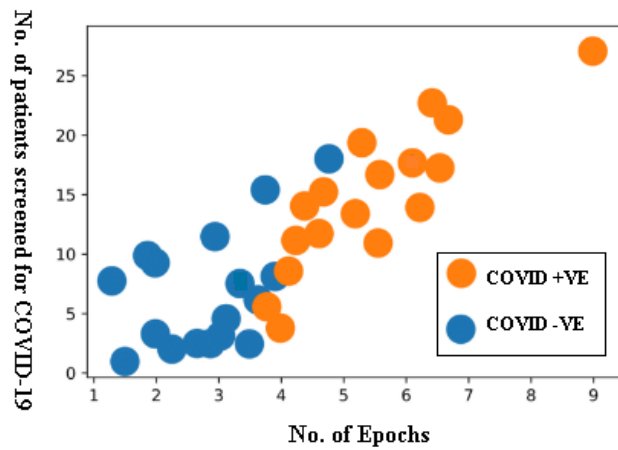


Figure 10(a). Detection of COVID-19 by F-RCNN

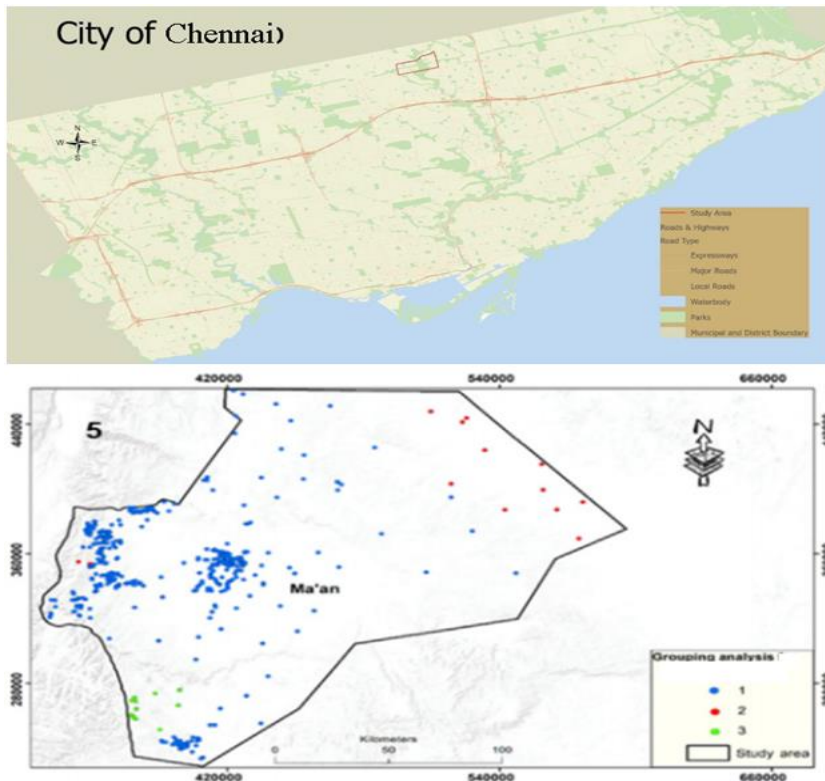
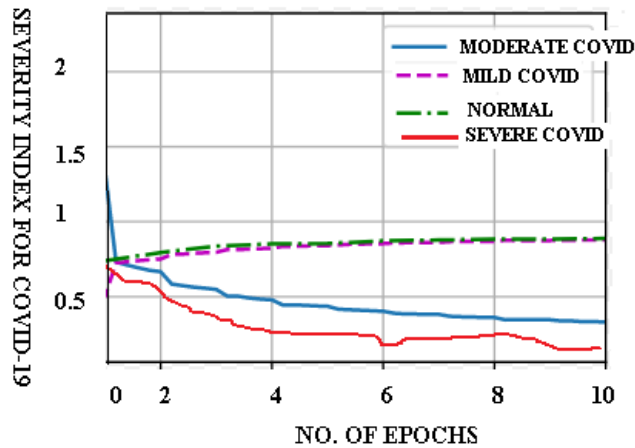


Figure 10(b). Identified Patient's location along with geo-spatial information on Smart phone



**Figure 11.** Determination of Various conditions of COVID from X-ray images of the lungs using F-RCNN

The training of F-RCNN is done using Gradient Descent Rule (GDR) with optimal values of learning rate and momentum. The change in weights after each iteration is calculated and each time, the new weights are found. The training includes the filter interpretation which is derived at the first layer of the F-RCNN. The variation in the signal amplitude is tracked and captured during this phase. Once the network is trained and the F-RCNN architecture is finalized, the same network topology can be used for testing and validation.

The measures like true positive (TP), false positive (FP), true negative (TN), and false negative (FN) were found to suitably address the need. TP denotes the correct number of lung sound signals detected; FP denotes all X-ray images of lungs that are erroneously detected. TN denotes the number of lung sound signal rejected correctly; FN denotes all the X-ray images of lungs that are not all detected. The performance evaluation is done by calculating the metrics (Table 3) like True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR) and accuracy. It is inferred from Table 3 that, by the proposed method (WT-FRCNN) for analysis of X-ray images of lungs, the TPR value is 98.23% and 85.56% during training and Validation. The overall accuracy is calculated to be nearly 96.79% from Figure 11.

Once the smart phone App is developed, entirely new set of images corresponding to the X-ray images of lungs along with the geo-spatial information will be used for detection of COVID-19 (non-contact) using the proposed smart phone App which is installed in the smart phone device (Apple make) will be used as the master piece for calibration. Calibration of the developed smart phone App can be done by cross verifying with the results which are obtained from the hospitals recorded for diagnostics. A comparative analysis will be done as indicated for various models and make of the mobile phones. If the deviation is within tolerance on consultation with the recognized, medical council authorized physician, the smart phone App will be hosted in the server for the free usage of the public who need assistance in diagnosing COVID-19 during this pandemic situation frequently. During the testing phase, randomly few brands of the smart phone will be used to detect COVID-19 by capturing the X-ray images of lungs and then identifying the appropriate reasons using the image processing techniques and will be compared with the diagnosis done by physician from swab test. Also the location of the patient can be found out using the proposed smart phone App.

The proposed technology is to be developed is an App for smart phone can cater as many numbers of users as possible and will be able to download the App from the internet sources at free of cost. They only have to install the App in their smart phone, open the App to capture the baby’s cry signals and simply use it. The major advantage of the proposed technology for detection of COVID-19 from the X-ray images of lungs along with geo-spatial information about the patient recorded using a voice recorder. Smart phone App once developed, does not require a separate device but instead the App which is available in the internet sources at free of cost can be downloaded on a smart phone with voice recorder and Android technology.

**Table 3.** Performance Metrics

%	Training the F-RCNN			Testing the F-RCNN			Validation of F-RCNN			Accuracy
	TPR	FPR	FNR	TPR	FPR	FNR	TPR	FPR	FNR	
WT-FRCNN	98.23	8.12	3.97	84.56	19.65	9.44	97.36	3.13	1.74	96.79
K-means	88.12	24.7	12.69	91.25	29.72	10.65	82.85	29.6	17.15	85.64
FFT-GMM	93.03	18.52	6.97	84.56	29.65	15.44	95.63	10.36	3.37	91.01
EMD-HMM	88.32	23.6	11.68	90.35	27.74	9.65	92.85	19.6	7.15	88.94
FFT-GMM	95.6	10.32	4.4	89.52	20.46	10.48	97.26	3.23	2.74	94.29
EMD+HMM	93.7	15.08	6.3	92.47	18.21	7.53	95.07	7.84	4.93	92.16

In the present scenario internet is being used to connect various medical related devices to increase the diagnosis efficiency, reduced cost and achieve better results in the domain of healthcare. This proposed technology is a wireless technology, with high computing levels so as to integrate with internet and develop an Internet of Medical Things technology. This, MIIoT technology is capable to gather, transmit and track the medical data related to COVID-19 so as to evolve a state of the art technology.

## V. CONCLUSION

Segmentation of X-ray images of lungs along with the geo-spatial information and segmenting the homogenous. This non-contact type of Remote COVID Diagnostic System is used to detect the region of interest from X-ray images of lungs along with the geo-spatial information about the patients. The performance can be considerably increased even if the images are corrupted with noise. The key objective behind this work is to identify the useful component and the noisy component from the lung sounds during respiration. In future, this method of X-ray image analysis can be used for developing the database that facilitates the following applications like differentiating the diversified features present in the X-ray images of lungs and to categorize the other kinds of respiratory diseases adhering social distancing, thereby preventing the spread of Corona Virus.

The proposed technology which is to be developed is an App for smart phone using Android. So, as many numbers of users as possible will be able to download the App from the internet sources at free of cost. They only have to install the App in their smart phone, open the App which will request the users to enable their location finder to capture the X-ray images of lungs with the geo-spatial information about the patients. Any common man who possesses an Android Mobile can install this App free of cost and use it anytime from any part of the world. Once the smart phone App is developed, entirely new set X-ray images along with their geo-spatial information will be used for detection of COVID-19 using the proposed smart phone App which is installed in the smart phone device from Apple which will be used as the master piece for calibration. Calibration of the developed smart phone App can be done by cross verifying with the results which are obtained from the diagnostic centers recorded in the first column of Table 3 against the “Laboratory scale results” and a comparative analysis will be done as indicated in Table 3 for various models and make of the mobile phones. If the deviation is within tolerance on consultation with the recognized, medical council authorized physician, the smart phone App will be hosted in the server for the free usage of the people who need health checkup frequently along with their geo-spatial information. During the testing phase, randomly few brands of the smart phone will be used to detect COVID-19 by recording the lung sounds during respiration and the results obtained from the proposed App along with the geo-spatial information about the patients, simulation results and swab test results will be compared to check the efficiency of the proposed smart phone App.

Once the smart phone App is developed, entirely new set of X-ray images will be used for detection of COVID-19 using the proposed smart phone App which is installed in the smart phone device from Apple which will be used as the master piece for calibration. Calibration of the developed smart phone App can be done by cross verifying with the results which are obtained from the diagnostic centers like Shree Test Tube Baby and Stannis Rea diagnostics recorded against the “Laboratory scale results” and a comparative analysis will be done as indicated in various models and make of the mobile phones will be used for this purpose. If the deviation is within tolerance on consultation with the recognized, medical council authorized physician, the smart phone App will be hosted in the server for the free usage of the people who need health checkup frequently. During the testing phase, randomly few brands of the smart phone will be used to detect the COVID-19 along with the location of the patients by capturing the X-ray images will be compared with the laboratory scale values.

## REFERENCES

- [1] A literature review of 2019 novel coronavirus (SARS-CoV2) infection in neonates and children Matteo Di Nardo, Grace van Leeuwen, Alessandra Loreti, Maria Antonietta Barbieri, Yit Guner, Franco Locatelli and Vito Marco Ranieri. *Pediatric research*. 2020
- [2] Wang, Y., Wang, Y., Chen, Y. & Qin, Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *J. Med. Virol.* 92, 568–576 (2020).
- [3] Clinical and immunological features of severe and moderate coronavirus disease 2019. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, Wang T, Zhang X, Chen H, Yu H, Zhang X, Zhang M, Wu S, Song J, Chen T, Han M, Li S, Luo X, Zhao J, Ning Q *J Clin Invest.* 2020 May 1; 130(5):2620-2629.
- [4] Human Coronavirus in Hospitalized Children with Respiratory Tract Infections: A 9-Year Population-Based Study from Norway. Heimdal I, Moe N, Krokstad S, Christensen A, Skanke LH, Nordbø SA, Døllner H, *J Infect Dis.* 2019;219(8):1198.
- [5] Medical reviews. Coronaviruses. Monto AS, *Yale J Biol Med.* 1974;47(4):234
- [6] Coronavirus Occurrence and Transmission Over 8 Years in the HIVE Cohort of Households in Michigan. Monto AS, DeJonge PM, Callear AP, Bazzi LA, Capriola SB, Malosh RE, Martin ET, Petrie JG, *J Infect Dis.* 2020;222(1):9
- [7] Seroepidemiologic studies of coronavirus infection in adults and children. McIntosh K, Kapikian AZ, Turner HC, Hartley JW, Parrott RH, Chanock RM, *Am J Epidemiol.* 1970;91(6):585
- [8] Epidemiology of Seasonal Coronaviruses: Establishing the Context for the Emergence of Coronavirus Disease 2019. Nickbakhsh S, Ho A, Marques DFP, McMenamin J, Gunson RN, Murcia PR, *J Infect Dis.* 2020;222(1):17.
- [9] Clark A, Jit M, Warren-Gash C, Guthrie B, Wang HHX, Mercer SW, Sanderson C, McKee M, Troeger C, Ong KL, Checchi F, Perel P, Joseph S, Gibbs HP, Banerjee A, Eggo RM., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob Health.* 2020 Aug;8(8):e1003-e1017
- [10] Human aminopeptidase N is a receptor for human coronavirus 229E. Yeager CL, Ashmun RA, Williams RK, Cardellicchio CB, Shapiro LH, Look AT, Holmes KV, *Nature.* 1992;357(6377):420
- [11] Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. Hofmann H, Pyrc K, van der Hoek L, Geier M, Berkhout B, Pöhlmann S, *Proc Natl Acad Sci U S A.* 2005;102(22):7988. Epub 2005 May 16.
- [12] Human and bovine coronaviruses recognize sialic acid-containing receptors similar to those of influenza C viruses. Vlasak R, Luytjes W, Spaan W, Palese P, *Proc Natl Acad Sci U S A.* 1988;85(12):4526.
- [13] Farhana Parvin, Sk Ajim Ali, S. Najmul Islam Hashmi, Ateeque Ahmad, Spatial prediction and mapping of the COVID-19 hotspot in India using geo-statistical technique, *Korean Spatial Information Society* 2021.
- [14] Hari Shankar Gangwar, P.K. Champati Ray, Geographic information system-based analysis of COVID-19 cases in India during pre-lockdown, lockdown, and unlock phases, *International Journal of Infectious Diseases*, 2021.
- [15] Ivan Franch-Pardo, Brian M. Napoletano, Fernando Rosete-Verges, “Spatial analysis and GIS in the study of COVID-19. A review”, *Science of the Total Environment*, 2020.
- [16] Felix Nikolaus Wirth, Marco Johns, Thierry Meurers, Fabian Prasser, “Citizen-Centered Mobile Health Apps Collecting Individual-Level Spatial Data for Infectious Disease Management: Scoping Review”, *JMIR MHEALTH AND UHEALTH*, 2020.

# Chapter - 6

## Twitter Sentiment Analysis of Covid-19 Vaccination Using Deep Learning

Varsha Naika <sup>1</sup>, Dr. Rajeswari Kannan<sup>2</sup>, Snehalraj Chugh<sup>3</sup>, Ahbaz Memona <sup>4</sup>, Himanshu Chaudharia <sup>5</sup>

<sup>1,3,4,5</sup> MIT-WPU, Dr. Vishwanath Karad's MIT World Peace University, Pune, India.

<sup>2</sup> PCCoE, Pimpri Chinchwad College of Engineering, Pune, India.

Email: <sup>1</sup> [varsha.powar@mitwpu.edu.in](mailto:varsha.powar@mitwpu.edu.in), <sup>2</sup> [kannan.rajeswari@pccoepune.org](mailto:kannan.rajeswari@pccoepune.org), <sup>3</sup> [snehalchugh2016@gmail.com](mailto:snehalchugh2016@gmail.com),

<sup>4</sup> [ahbazmemon0@gmail.com](mailto:ahbazmemon0@gmail.com), <sup>5</sup> [himanshuchaudhari2346@gmail.com](mailto:himanshuchaudhari2346@gmail.com)

*Abstract— Covid-19 had consequential social, economic, and extreme mental outcomes on the community, where media platforms like Twitter increasingly became essential networking mediums generating information with a large volume of reports, views, opinions, and information shared by individuals and authorized outlets. In 2021, when the second wave of COVID emerged in India, we recognised the fastest outbreak with more than 20 lakh cases in April's first half. Until then, India distributed over one billion vaccine units with two producers: Bharat Biotech, producing Covaxin, and Covishield, OxfordAstraZeneca's vaccine, by SII (Serum Institute of India).*

*We collected datasets for analysis, and applied our novel algorithms for preprocessing, i.e., removal of URLs, @, #, contracted words, punctuations, numbers, POS, etc. Converted tweets into tokenized words, used stemming & lemmatizations, then applied neural spellchecker. Using our in-house algorithm, we cleaned around 500 tweets in just 0.5 seconds, getting rid of duplicate and redundant tweets. A word cloud with classes: Positive, Negative, and Neutral was constructed which then used neural network to predict them, resulting in 97% training and 99% testing accuracy. Results aid in improved policy design, keeping citizens' perspectives in mind, and an aware government about issues like vaccination shortages, food, poverty, etc.*

*Keywords— Covid-19; Pandemic; Sentimental Analysis; Word Cloud; K-Means Clustering; Natural Language Processing; Vaccinations; Twitter; Covaxin; Covishield.*

### I. INTRODUCTION

In almost every country, the new coronavirus is spreading rapidly (COVID-19). Thousands of citizens have been infected, and huge numbers have died from the illness worldwide. Twitter activity has also increased by roughly 25% within the same period.

It used to be widely accepted that the virus was not infectious until the beginning of January 2020. However, researchers later identified it to be the new coronavirus as the source of the sickness and found that it could transmit from individual to individual. Subsequently, the city of Wuhan [1], a metropolis region with 11 million inhabitants, was ordered to remain closed down, and the province of Hubei swiftly followed suit. The illness eventually led numerous Chinese regions to be placed in quarantine in February. To limit the radiation, China halted its commerce from February to March of 2009 [2], [3]. Wuhan, a metropolitan region with 11 million citizens, was soon put on lockdown with severe orders for everyone to stay inside, and the state of Hubei quickly fell under the lockdown. This subsequently triggered the implementation of lockdowns in several Chinese regions in February. China needed to stop the economy from running for the majority of February and the first two months of March to limit the spread [4]

China's recent efforts to control the epidemic since late January 2020, while preventing the spread of the disease to other countries, has only succeeded in spreading the sickness throughout the world [2]. As every evidence proves, the virus is naturally occurring and has originated from bats or may have arisen through an intermediary mammal species. However, transferability between individual to individual is uncertain. In particular, from animals and humans, the transferability is unknown as well yet. This worldwide catastrophe sparked an epidemic proclamation by WHO on the 11th of March [5], and several national emergencies followed. Using physical distance (including school closures, nightclubs, eateries, cinemas, and encouraging companies to only have their executives work from home). It's indeed strongly discouraged or prohibited to even have major public assemblies such as concerts, graduation ceremonies, and sports activities. It is said that the economic effect of mitigation has decimated countless companies, but according to reports, over 40 million individuals in India have applied for initial unemployment benefits [6].

---

© 2022 Technoarete Publishing

Varsha Naika – “Twitter Sentiment Analysis of Covid-19 Vaccination Using Deep Learning” Pg no: 75 – 92.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch006>



Individuals utilize the media to learn more about their personal health decisions. Because of the amount of data available [7], this may be even more relevant in the case of the COVID-19 outbreak. Despite the fact that fresh information is always flowing in, the primary questions of viral transmission, post-recovery antibodies, and medication therapy remain unanswered [6], [8], [9]. In light of the increasing amount of information, many people resort to social media to get clarity. Several studies have found that vaccination material is extensively distributed throughout social networking sites, with particular attention paid to how it is depicted on the Internet. Social media conversations around vaccines have grown following current occurrences in the news. Using platforms that allow people to discuss issues around vaccination, content appears to show up throughout individuals who have similar attitudes about vaccinations [8], but hardly ever among those with opposing beliefs.

To promote vaccination on social media [2], [9], the public health service may be prevented from increasing its spread by ideological isolation. Much research has gathered Twitter data following the COVID-19 epidemic to help comprehend public responses & debates concerning COVID-19. The volume of anti-vaccine material disseminated throughout social media is impressive. The current research, although early, shows that exposure to this type of information may impact vaccination attitudes and, as a result, vaccination delay [10]. Confusion and misinformation are both spreading quickly as individuals try to comprehend the best way to defend themselves, their loved ones, and also to post as many provocative comments as possible, which hinders one's reading comprehension.

Phase one of the vaccine campaign focused on making sure that 30 million healthcare providers and 2.7 billion priority population members were aware of the immunization opportunities [10]. It was anticipated to be finished by July [11]. Although the Indian government has recently launched two vaccines for the nation's huge campaign, the Ministry of Health has revealed that even though the vaccines are in high demand and logistical issues will make it difficult for individual people to pick and then choose the antibiotics, they will have to go through with the government's decision. Although everyone just above the age of eighteen would be entitled to receive COVID-19 vaccinations in India, which is something the federal government stated on Monday, May 1st [9], [10], [12] not everyone will receive vaccinations. New expenditures and programs relating to immunizations for the citizens in India also were announced. Since getting relevant, reliable, and high news from these sites means that there will be precise and up-to-date data upon the COVID-19 epidemic and the vaccine review for people in that age group bracket of 18 to 45, this will help keep the pandemic at bay. Some Indians can't be vaccinated as they wait, and thus they're also using Twitter to voice their views.

The goal of this research is to analyze the feelings and attitudes through Twitter of the Indian people towards both Covishield and Covaxin vaccines for individuals aged 18 and above, which was authorized since [4], [11]. The public social media data published by people worldwide will be utilized to discover the major ideas, beliefs, emotions, and subject matter that individuals have around the COVID-19 pandemic vaccines. Information such as this can assist public politicians, healthcare providers, and citizens in identifying important concerns and offering better solutions.

## II. DATA GATHERING

This research intends to explore the public discussion and sentiments linked to the COVID-19 spread by analysing tweets gathered with the use of Tweepy Python package for leveraging Twitter's Streaming API [12], [13] in order to extend the knowledge on public responses. Twitter specifies the language of each tweet when using its streaming API. Not surprisingly, given worldwide Twitter usage, the majority of tweets (57.1%) are in the English language thus we focus on collecting only English language tweets.

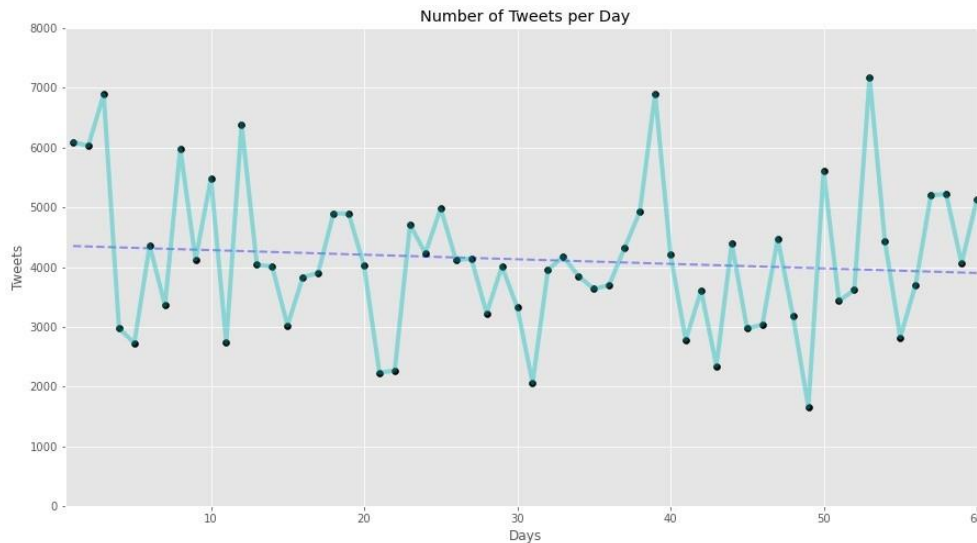
We started by analyzing the number of average tweets we could retrieve, thus we started collecting tweets from February 25, 2021, to April 26, 2021 across India except Jammu & Kashmir, as we had received the least tweets from that State, given in the Figure 1. While the approach proposed in this paper can be extended to adding parameters and searching for a variety of tweets, in the present study all the tweets without any filters have been retrieved, approximately retrieving in a total of about 2,47,627 tweets. The number of tweets that have been daily retrieved has been plotted in the above graph. This helped in understanding the average amount of tweets that are taken daily, which was giving a threshold value of 4128 tweets on average, Figure 1.

As in our research, we are aiming to perform an analysis of the sentiments of people around the globe related to COVID [14]. The vaccinations were granted to them after May 1, 2021. We mapped out to prepare a dataset of 50,000 tweets, thus understanding the average tweets we get daily (4128) it took us approximately 13 days to download and retrieve. As an allowance of vaccination was provided and there was a hike of citizens reacting with a variety of emotions, it was perfect to perform the retrieval of tweets from May 3, 2021, to May 16, 2021, Figure 1. In conclusion, we acquired a dataset of around 49,345 rows of tweets filled with hashtags and about a variety of topics.

Furthermore, we then created an algorithm that performs a keyword & hashtag extraction over a certain tweet, this helped us in retrieving a particular number of tweets just related to our topic. We noticed that there were in total approximately 2,657 unique hashtags and keywords on general topics, used in all the tweets in our dataset. After understanding our topic elaborately through various websites, we created a list of 724 keywords out of 2,657 hashtags related to COVID-19 and VACCINATION. We then scanned the corpus with our skewed keyword list, obtaining a set of 15,174 related to our problem statement tweets by



people across the globe. These filtered tweets were further sorted in time-ascending order and converted to a bag-of-word representation. All analyses were performed using Tweepy & Python (3.7.0.0)



**Figure 1.** A threshold quantity of tweets streamed over a period of 60 days, starting on February 25, 2021, and ending on April 26, 2021, using that information gives us an approximate daily stream of tweets.

### III. DATA DESCRIPTION

From March 3, 2021, to March 16, 2021, these tweets were collected from around the globe. The data stream we retrieved had 15174 rows and 10 columns, with the majority of them being categorical and objects, Figure 2. For our suggested statement, all of them must be categorical. This dataset has 10 attributes: 'Tweets,' 'User,' 'User statuses count,' 'user followers,' 'User location,' 'User verified,' 'fav count,' 'rt count,' and 'tweet date.' Tweets are text strings that include emoticons, hashtags, usernames, and other characteristics. User is the username, User Status Count is the number of tweets issued by the user, User Followers is the number of followers each user has, and User Location is the user's location from where the tweet was tweeted. However, since the majority of Twitter users do not allow geotagging for tweets, the location was retrieved from the profile of the tweet's username using the Tweepy Python module. They are from different places and not from a particular location, having in total about 3240 unique locations. User Verified refers to a user's verification, which is a blue tick that each user receives after their Twitter account has been verified. The number of times this tweet has been favorited is shown by the Fav Count. The retweet count indicates how many times the tweet was retweeted, while the tweet date displays the date and time the tweet was posted. As a result, the dataset includes 75% of the retweet count as 1. The total number of verified and non-verified individuals is 1486 and 13688, respectively.

### IV. DATA PREPROCESSING

Preprocessing of the raw experimental data is critical since it improves the quality of the raw data. Bringing our tweet into a reliable and computationally efficient shape is simply referred to as preprocessing. To make the tweet more acceptable for machine learning and neural network algorithms, we should modify the tweet and preprocess it.

#### 4.1 Preprocessing of Tweets

When we discovered the data, we had to cleanse it and ensure that it was free of repetitions or mistakes so this dataset had no error. We went ahead and completed this by utilizing our processing algorithms that were not just for cleaning data but for removing repetitive tweets as well. Many techniques are used in order to prepare the tweets for modeling sentiment analysis [13]. The system's success rate and runtime will both drop if data cleansing and removal of duplicates are not performed. The rise in the success rate occurs when you get rid of words and terms that are useless.

##### 4.1.1 Cleaning of Tweets

**Transformation of Cases and WhiteSpaces:** For consistency, all capital letters are transformed into lowercase. Machine learning techniques are case sensitive and so, the words "Vaccination", "VACCINATION", "vaccination", and #vaccination is all rendered as "vaccination". To eliminate unnecessary spaces, we use strip () to completely remove all white spaces in the tweets.

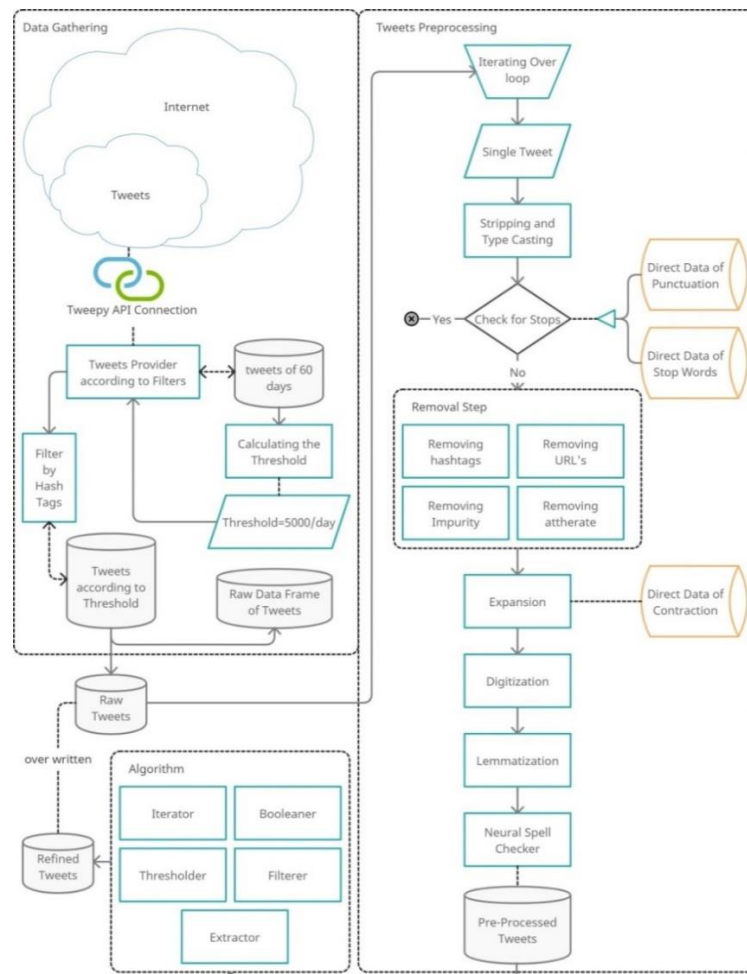
**Removal of URLs and hyperlinks:** Many of our tweets contained URLs or hyperlinks which were not helpful either. Due

to the nature of links in tweets, they cannot be used for sentiment analysis, thus they must be deleted. A large number of scripting languages, online resources, and text-based patterned expressions tools make use of regular expressions. They are widely employed in statistical analysis, data analysis, the transformation of data, and in a wide variety of other data-cleaning and data-transformative processes. Data cleaning is a critical procedure that should be completed for the machine to enhance learning [15].

Regular expressions were used to remove the URL (.org, HTTPS, .com) and links from the tweets. When doing this operation, we encountered a snag: the algorithm was unable to offer us 100 percent accuracy in the removal of hyperlinks and URLs from the dataset string whenever executed. For example, when the phrase "HTTP(s)" was removed, the letter "s" in the phrase remained. We streamlined the process by developing an efficient function that eliminated all the unnecessary URLs, and so created an ideal outcome. It did this with the least amount of effort while delivering excellent accuracy.

**Removal of Usernames & Hashtags:** The hashtag "#" makes all the Tweets to which it is related readily available. It acts as an indicator that a piece of information about a certain topic or categorized under a certain topic could be provided [15]. A user's profile is marked by the @ symbol at the beginning of a handle. For example, if tags such as @harbour or #vaccine, remain in the tweets when there is no data cleaning performed then this might lead to model inaccuracy and lower the success rate. Instead of repeating the elimination step using Regular Expressions, we developed an optimized function for this as well to take advantage of irregularities and maximize outcomes.

**Removal of Contraction of words & Punctuations:** Truncated forms of sentences or words are called contractions. They are often encountered in English, whether that be in writing or verbal. For many words, we omit the vowels while forming the contractions. Eliminating contractions helps in standardizing texts and is beneficial when dealing with Twitter data. It also has the added benefit of improving sentiment analysis by making sure the words don't contradict one other. Thus, looking at words like "I'm," "Could've," "Wouldn't," and so on that have punctuation, which can't be completely erased or deleted since the words' meanings could get affected.



**Figure 2.** The diagram above depicts all of the processes involved in collecting data, preprocessing tweets, passing them through the algorithm, and resulting in refined tweets. All these tweets are overwritten after the processes in refined tweet

The library provides “Contractions” as a variety of terms and made it possible for the user to input more words, making it suitable for our use case. Therefore, we utilized this library, but in addition, we built our dictionary of words. The greatest pleasure here is that the library, as a result of a multitude of additions, was able to carry out the straightforward process of replacing the contractions without making the search or replacement process longer. Machine processing is hindered by the punctuation characters, such as ‘!’, ‘~’, ‘’’, and ‘?’’, and thus these characters are eliminated as well.

**Converting numbers to string values:** Additionally, we also observed the presence of fractions, whole numbers, and decimals, all of which supported our overall problem statement. To better match our results, we transformed the figures to words using the num2words package. To provide an example, 16, One Six, sixteen, etc. gets unanimously converted to sixteen.

**Tokenization:** The process of tokenization divides tweets into a list of words. This approach separates words, sentences, and other linguistic units from the text into pieces called the tokens. This is a procedure where we systematically work our way through each tokenized phrase and filter words whose length is less or more than a set threshold. We decided to utilize the word\_tokenize function from NLTK [12]. Twitter-specific tokenization provides a better way of verifying that URLs and hashtags are completely separated in the tweet.

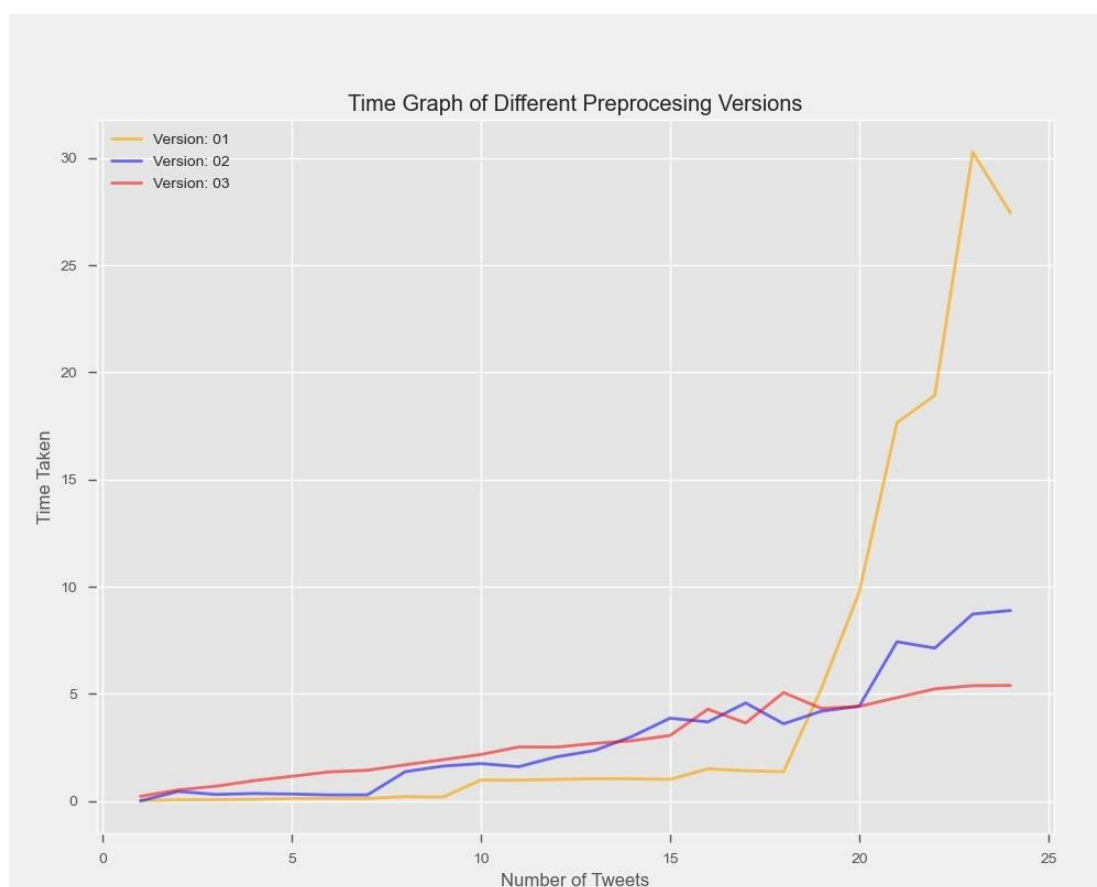
**Removal of Stop Words:** Certain terms are abundant in tweets, such as, ‘and’, ‘a’, and so on. We may focus on the relevant aspects of our content by eliminating unnecessary words. By removing stop-words, we significantly minimize the clutter in tweets, which is unnecessary for issue analytics. We utilized the library called stop-words present in NLTK to clean these tweets and it aided in removing words such as the, an, with, at, etc. [16].

**POS Parts of Speech:** This approach can identify nouns, adverbs, adjectives, subjects, and objects in a phrase, and it further categorizes the part of speech of an individual word to aid in analyzing sentence construction. It is additionally used to determine the word’s raw form of connotation disambiguation. A large number of features are available by completing this phase. By doing this step, the words representing the structure of the phrase and the content of the message in the field to which the sentence belongs can be obtained. To build an approach and also to get word meaning for just tweet tags, a technique created using the NLTK class called POS tagger was utilized [17]. To accurately identify the sentiment score of a phrase, you must first arrange the sentence according to part of speech. The accuracy of attributing emotion polarity to sentences will be impaired as consequence, meaning favorable, unfavorable, or neutral.

**Stemming and Lemmatization:** In the normalization process, words are stemmed and lemmatized to create a normalized version of the term in the text. Word stemming is a methodology that recovers the word’s basis in the sentence. A novel manner to normalize a word is to remove its suffix from the term. To accurately determine sentiment, just stemming words with a length larger than two (“a”, “is”, “an”, and “the”) are used, since words like “an”, “the”, “is”, and “are” are not included into the sentiment vocabulary while measuring the word’s polarities.

The word lemmatization technique transforms affixation and/or changes a vowel from either the base or dictionary form of a word. A lemma is the result of the process of word making. The lemmatized word is the entrance to the WordNet since the lemmas are the words that already have a core definition of the phrase that is sought. Computing lemmatization using an approach thus creates a lemma, which is then transferred to WordNet dictionaries to obtain a new sentiment for the term. Using the WordNetLemmatizer classes (accessible via the WordNet stem package) within the Python NLTK stem bundle, one may conduct lemmatization in terms of the corresponding character-by-character [12]. The use of WordNetLemmatizer produces the lemma (root meaning) of such input sentences while ensuring accurate representation. For example, “Vaccines” is a variant of the root word “Vaccine”, that is defined throughout the WordNet vocabulary.

Preprocessing is essential as it influences the accuracy of models of learning. Figure 3 illustrates the same thematic of our approach. We can see here that the time required to clean from version 1 to version 3 is enormous. Our versions have been improved during the course to ensure that it required as little time as feasible. For the model to learn effectively, data cleaning is an essential step. The model will thus get a better result because of the lack of redundant words and phrases in the data set. Now, we start by reviewing each function we used to clean our dataset, detailing how our accuracy and time to complete each step were polished to their highest possible quality for each iteration, getting to version 3.



**Figure 3.** The following graph is a comparison of three different versions throughout time, the number of tweets where the most optimum results were received in our final version 3.

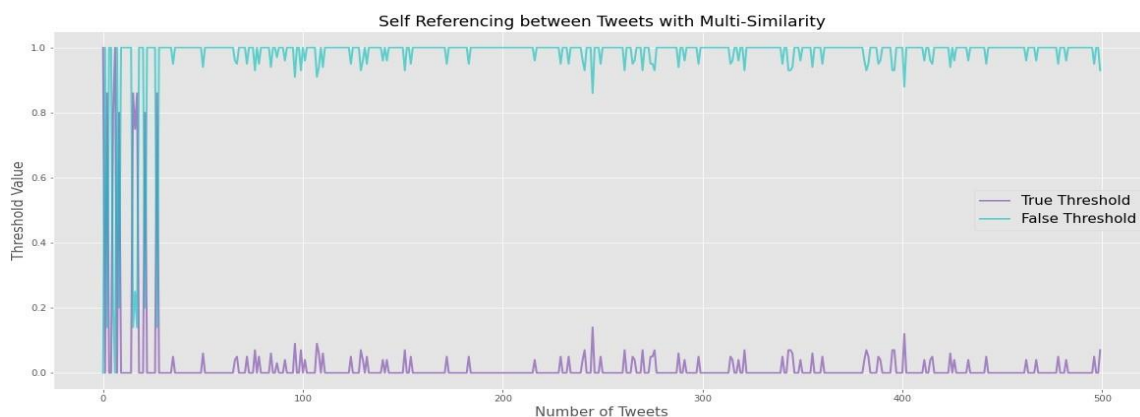
#### 4.1.2 Algorithm

Every Twitter user has access to the features of the user’s tweet, like retweet and reply count. A tweet may be described as the ability to affect someone emotionally, with an opinion, or through influencing their conduct. For Twitter, a favorite is an indicator of reader approval for a tweet. Tweeting and retweeting indicate that the user agrees with the sentiment expressed in the tweet and is eager to share it with his followers. Commenting (re-tweeting) and disseminating further the opinion of a tweet demonstrates that the reader wants to discuss and share their views.

However, people may settle on merely copying someone’s tweet and altering it slightly to make it distinct. According to this approach, there were many, many tweets identified similarly as shown in Figure 4. Specifically, the correlation between the support for a tweet and the prediction affects the prediction, resulting in an incorrect prediction. Because there were so many redundant tweets in the dataset, we built an algorithm to prune these posts, so we could make more accurate prediction.

When we started, we developed an algorithm that examines a tweet and runs a set of operations to generate a dataset of unique tweets, which also comprises the least-repeated terms and each tweet is unique inside the dataset. Using the algorithm, we found several areas to modify since we didn’t get the precision we anticipated in the removal. Everything is detailed down to the smallest possible detail in five different versions, and all of these versions included numerous modifications which enabled us to achieve our objective of an optimal level of accuracy when we finally arrived at our final design, version 5 Figure 7. Below we have explained how we began with our first version and with incremental improvements culminated in our objective.

So, we started by implementing the iterator function in our algorithm, which locates all of the tweets in the dataset and afterward compares each of them to all of the other tweets in the dataset. It takes the tweets and then tokenizes both of them, i.e., the one which is comparing itself with all the tweets and the other one which is being compared. When we tokenize both of these tweets, we pass them via another function called boolenear. The function boolenear takes each tokenized word and checks each word in the tweet to see whether it matches any word in the other tweet. Following all the iterations, the algorithm produces a list of booleans that imply whether the first word, second word, third word, and so on of the tweet are all equivalent to the first word, second word, third word, and so on of the other tweets.



**Figure 4.** This graph shows the location of words in two tweets being similar at places before cleaning, here tweets are compared and if two tweets included the same words in the same locations, those tweets had to be cleaned.

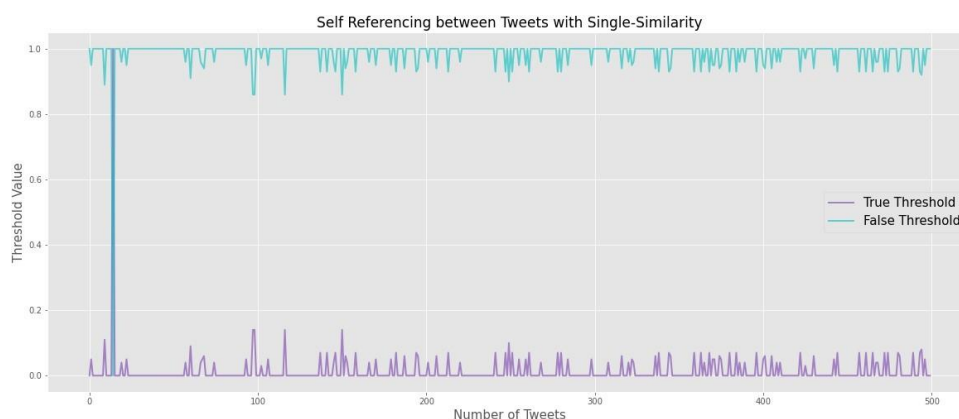
For each place in the tweet, if the words are comparable, it will return True or False declaring true, or else marking false if the phrases on the same spot are different. The most essential aspect to consider when there is a comparison in tweets is if the length of both tweets is comparable. For that reason, knowing that not every tweet would be comparable in length, we set the booleanar to disallow comparisons for the additional words and mark those false.

The reason we required these boolean values was to assess the percentage of how much the words were similar in two tweets. To assist in comprehending the list of true and false, we applied the Thresholder algorithm to our dataset, which computed the percentage of how much each tweet in the dataset was similar to each other. Thresholder operated by determining the amount of true and false values for a given set of compared tweets, determining which are unique, and setting an exclusion level for the set.

After the percentage is calculated from the Thresholder, it's passed on to a function called filterer, which uses it to filter tweets. This function analyzes the percentage and verifies that the false value threshold is less than 25% for false and also greater than 75% for true, thus implying that the filter has indicated that the tweet is similar. In turn, if the condition that holds are satisfied, then we understand that those two tweets aren't similar, and therefore they should remain unchanged in the dataset, Figures 4 and 5.

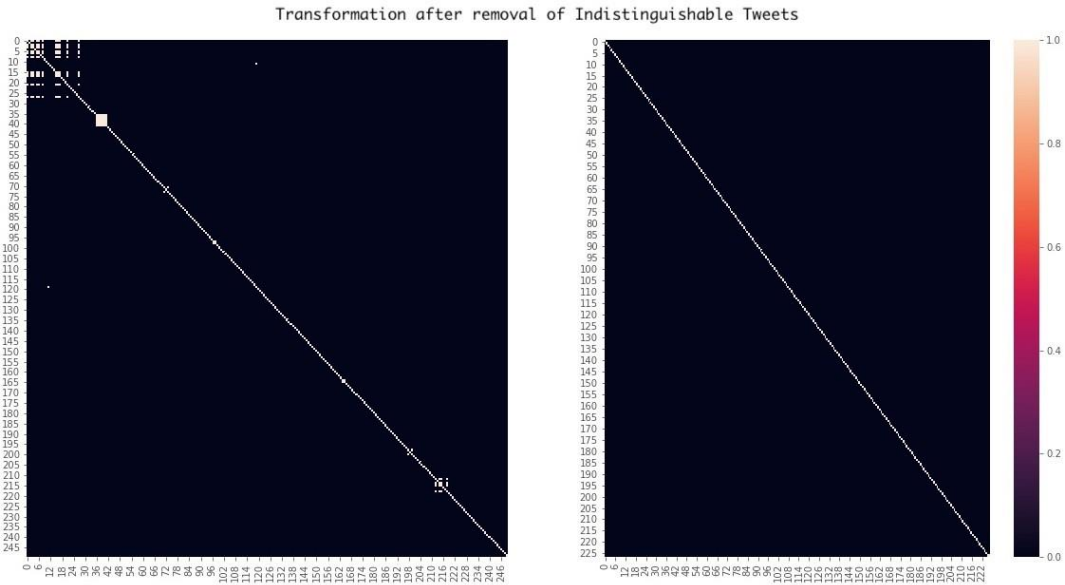
Our dataset was almost ready to be cleaned, but the elimination of duplicate and repeated tweets remained to be done last. But if we eliminate all these tweets that the filterer was looking for, then the dataset may become empty. Therefore, we construct an Extractor function that takes the tweets, creates an array that includes all the boolean values, namely true or false for each word of the tweet which each Tweet compared with the other Tweets existing in the dataset earlier. This function searches for the duplicity in the dataset after comparing all the tweets in the array. Where a set of similar tweets is found, just one tweet with the most words saved while the others, that are partially duplicates, are therefore eliminated Figures 4 and 5.

Once our tweets had been filtered, where all the above changes were made in our versions 1-4 Figure 7, We discovered an issue with our system. Due to several iterations and an inordinate amount of repetition in the reading and comparing of tweets, they required a lot of time to process. Tweets used to go through certain important phases while they were iterating and comparing with themselves; but rather just the extra time invested in self-comparison was useless. So instead of having 6 individual steps working step-by-step, we implemented an algorithm consisting of 2 phases that utilize time efficiently giving us better and optimum results, which we implemented for our version 5, Figure 7.



**Figure 5.** This illustration illustrates a comparison of two tweets after cleaning and here, no duplicates were found in either of them.



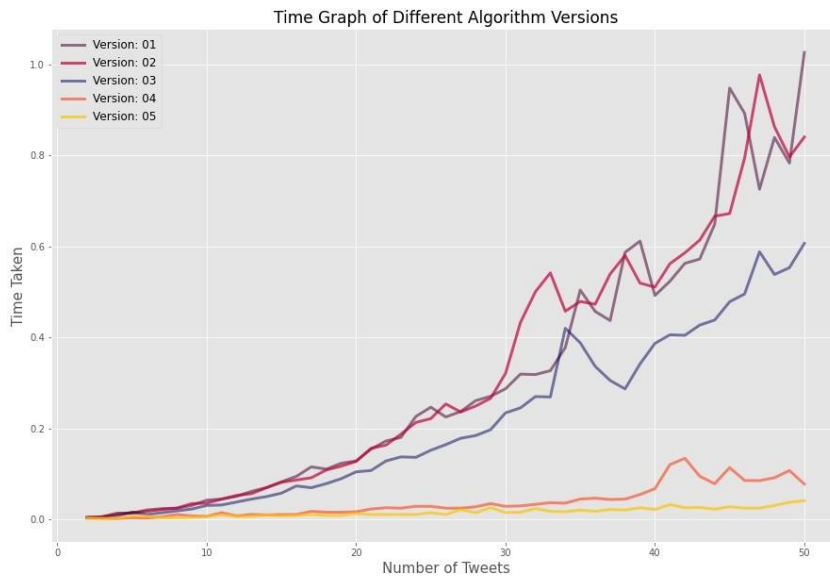


**Figure 6.** Our 1st algorithm is noticeably different from the 5th's. The diagonal line shows the self-referencing present in the graph on the left side with duplicates. Once we'd gone through the first step of our final algorithm, the distortions were no longer there which are shown in the graph at right side.

So, in Phase 1 of the algorithm's version 5, we start by taking a NumPy array of size  $n \times n$  but being only 1 dimensional, instead of using for a list, Figure 6. The iteration would have been slow if we'd have picked Cuda, Data-frame, List, etc., so we opted for one dimension, which resulted in fast and efficient computation. In the previous approach, we used the sum function over each tweet, one by one, to aggregate the data. We replaced it with an incrementation checker, this eliminated tweets that just weren't helpful and therefore only kept the one that we required for processing, decreasing the time necessary at a higher level.

We did some additional tweaking to the filter algorithm in Phase 2, Figure 6, Which was recalculating true and false for each word in the tweet differently, so when examined more closely, we saw that none of it was necessary, thus we only noted one of the two, i.e., the percentage of the 'true' in tweets and allowed false to be computed by subtracting the percentage of true from '1'. We also reduced the portion of the algorithm where tweets in the array were repeatedly compared with themselves, which increased the algorithm's performance exponentially. Additionally, we just had to write two lines of code to produce the same threshold value rather than the long code that we used in the prior approach.

This is how adjusting minute aspects of functions since our initial algorithm's version 1, all of it boosted and helped us achieve our goal of producing cleaner and more uniform tweets in the least amount of time, Figure 7. preprocessed with the version 5 of our algorithm.



**Figure 7.** This graph illustrates the relative performance of all the algorithm versions on a certain number of tweets. We have concluded that the version 5 algorithm is optimal and quick, when compared to each of the other different versions.



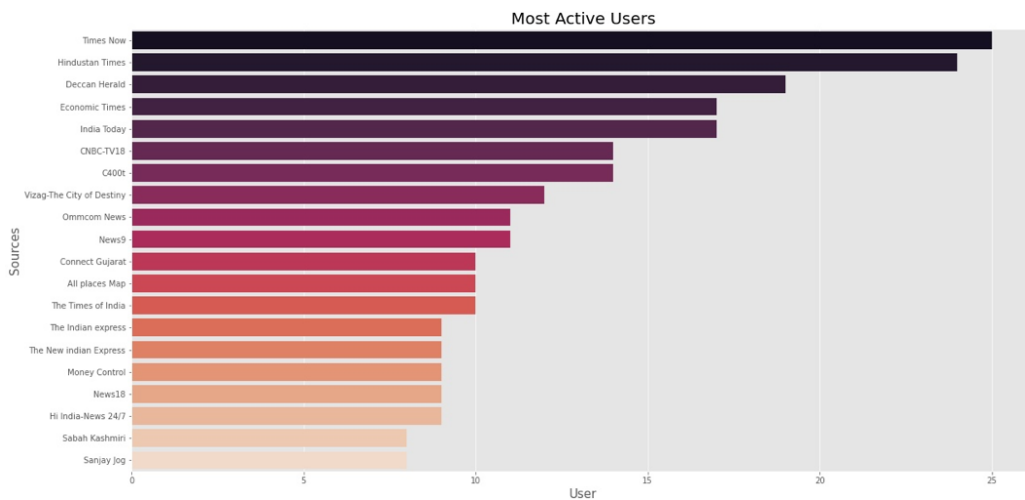
#### 4.2 Preprocessing of Data Frame

We ran our preprocessing tool and were left with about 11129 tweets, Figure 6. We tried to learn about the numbers of users, unique users in the dataset, the most frequent user, users who are verified, the number of retweets they've received, and their follower count in this processed tweet dataset. The preprocessing stage was so effective that the duplicates had already been eliminated.

Furthermore, we found that the tweets which were tweeted by lots of users were sourced, copied, framed, and added, etc. from various accounts, with verified ones being mostly sourced. Thus, being left out with no unique information in most of the non-verified tweets. There were many tweets that merely contained zero retweets, no location, etc. Also, the accounts themselves had almost no reach. Thus, only tweets with high reliability were included in the dataset; and to get these tweets, the dataset was filtered over the various filters given below: -

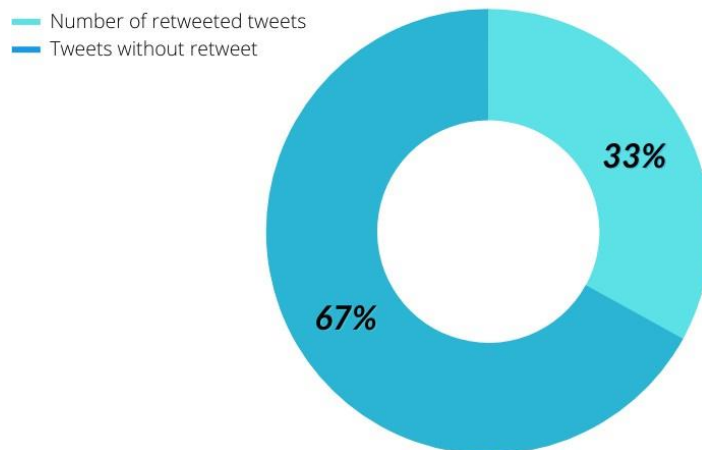
- User's account is verified, because they receive a higher reach and mostly are the highest active users, Figure 8.
- Tweets having retweets count greater than zero, as shown in Figure 9, they have higher reach and comprise around 33% of our dataset.
- Adding those tweets whose users, in particular, had following greater than 50 & also the location is null.
- We appended all the tweets with location not null, which is shown in Figure 10, due to their activeness in various regions.

These various filters allowed us to get relevant and particular tweets that had been narrowed down out of the rest of the Twitter flow in relation to our problem statement. Once the algorithms were completed, we had around 9525 Tweets remaining in the dataset on which we then run the algorithms [12].



**Figure 8.** This specific statistic displays the most active accounts on Twitter after their tweets have been cleaned.

#### Percentage of Retweets

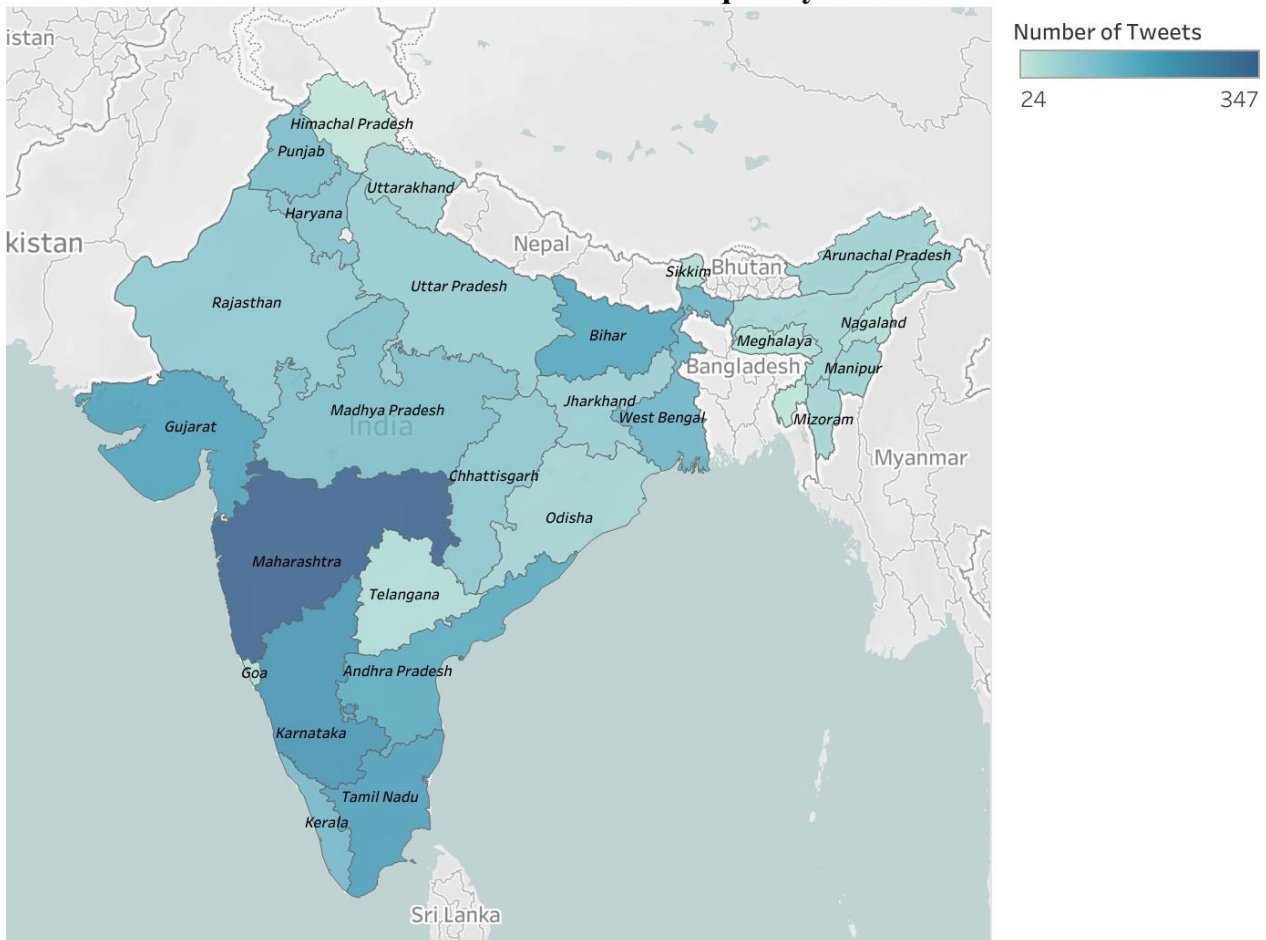


**Figure 9.** After the cleaning of dataset, we performed a differentiating tedious task over the tweets which gave us an overview that 33% of tweets in this donut graph had retweets from other users.

## V. CLUSTERING

Iterative methods are often used in the clustering process to group instances together based on shared features. Using these categories, we may examine the data, detect anomalous behavior, and then anticipate outcomes. When it comes to the visualization of clusters, the models of clustering may also assist you to discover connections that you might not be able to infer from simple browsing or just watching. Since some of these causes, clustering is often employed in machine learning projects in the early phases of the project to explore the data and find interesting results.

### State-wise division of frequency of Tweets

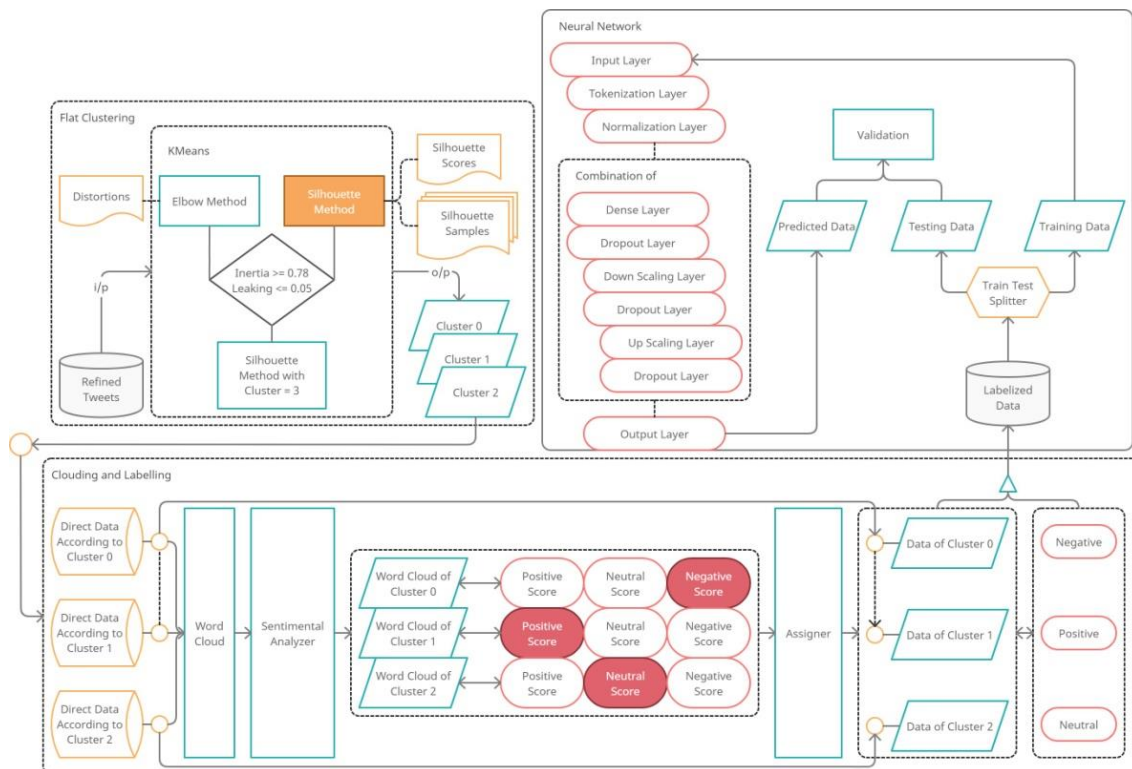


**Figure 10.** This is a location-based tweet frequency map that shows how often people tweeted from each state in India, except Jammu & Kashmir (Negligible Tweets from that area). Citizens in Maharashtra being the most active in relation to their activity over Twitter.

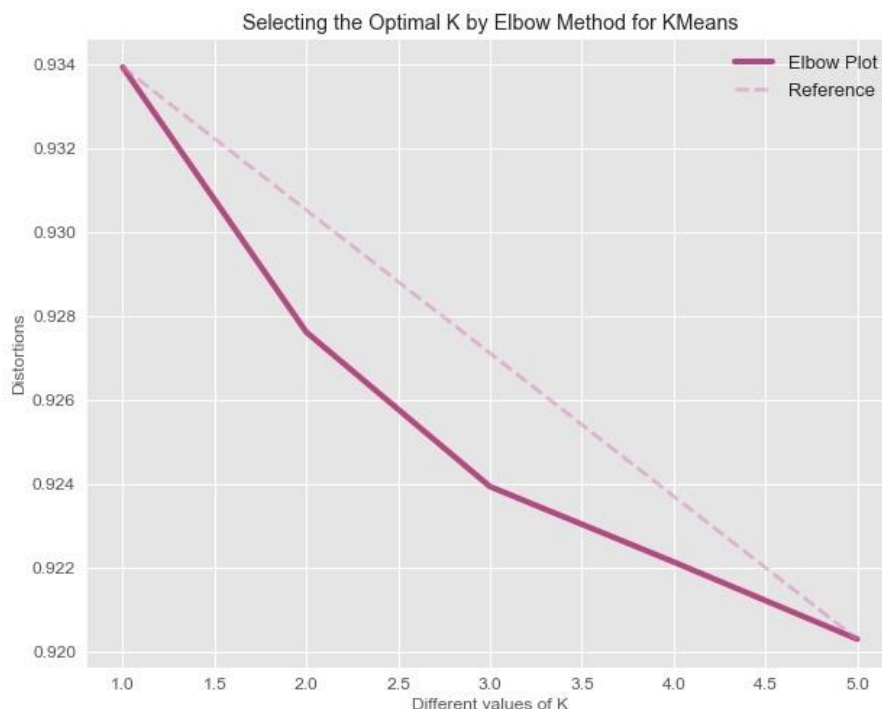
Having done all the preparation of tweets, we are now left with 9525 tweets on which we intend to apply to a cluster, Figure 11. In this stage of development, the data did say that we have prepared, we apply k-means clustering over it [18]. This method is aimed at partitioning  $n$  observations into  $k$  clusters, each of which has the closest mean value. Though I have to admit, working in  $n$ -dimensional areas is critical. If the effort fails, a part of the sample is shifted until each sample is closer to the cluster center.

#### *Elbow Method and Silhouette Method:*

To identify  $K$  clusters [18], we can see the 3 clusters established using the silhouette technique are shown. Based on these findings, we decreased the dimensions, and then assigned their domains as well as colors to distinguish them based on the cluster values, and setting  $K$  centroids to distinct values. The overall variance between each group decreases as we add additional clusters. These findings, therefore, do not demonstrate the ideal number of  $K$  clusters to seek. To achieve the correct  $K$  number of centroids for our prediction analysis, we must decrease the within-cluster sum of squares. To do so, we first utilized the Elbow technique to determine the ideal number of clusters [19], [20]. The graph and the illustration below are provided so that you can see the effect of varying the number of clusters using the Elbow technique, Figure 12. When taking the total of squares inside groups, one would expect to see 2 at the end of the elbow.

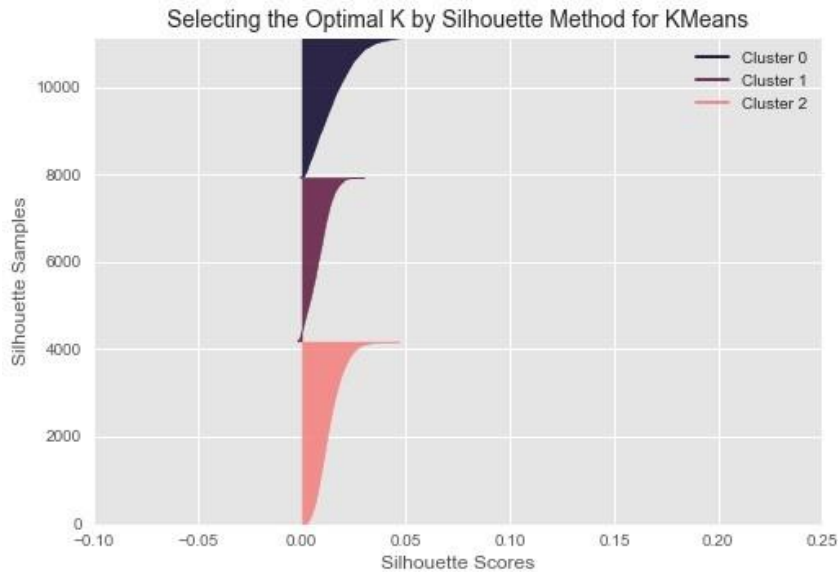


**Figure 11.** The diagram above depicts the transmission of data in flat clustering over three clusters which were predicted using the Silhouette Method. Following that, we create a word cloud and labelize various words into sentiments. All of this is then predicted and validated using a neural network.



**Figure 12.** The figure above illustrates the Elbow Method, which determines the optimum number of clusters using the elbows. The fact that this presents us with an elbow at 2 leaves us with some ambiguity

When you look at K equals 2, it seems to be an elbow. Adding the number of clusters to the within-group mean square would therefore result in a much worse answer [21], [22]. In this example, the best number of cluster centers is two. However, the narrative is open to interpretation, and although we aren't sure due to having the other point at 3, this further weakens our case. For that, we used the second strategy, the 'Silhouette technique,' which ensured that the correct amount of k clusters were created [23]. The figure below illustrates the Silhouette method's optimum number of clusters, Figure 13.



**Figure 13.** We choose Silhouette, which presents us with a close approximation of the number of clusters: three. In practice, since there is no negative cohesiveness, this is optimum cluster number.

To understand and validate consistency among data clusters, one may use the silhouette technique. It is determined by the closeness of an entity from its own (cohesion) cluster against the distance of that entity to all other (separation) clusters. A high score indicates that the item is quite well linked with its cluster [24], [25], Whereas a low score indicates that it is poorly matched to adjacent clusters. Setting up the clusters is proper if most items have a high value. However, if there are a significant number of points with a negative or negligible value, there may be an overabundance or deficit of clusters in the clustering arrangement. Observe in the following diagram that cluster number 2 has a negative value. When we examine the opposite side, we see that each point is bright and consistent in groups, and without disruption. Our prediction methodology relies on 3 clusters, which, thus, may be inferred to be the most optimum, Figure 14.

Simple counts provide a challenge because of how common some words like "the" are, and these high counts in the encoded vectors will be a small fraction of the total. There are several different ways to go about finding word frequencies. One of the most common is TF-IDF (Term Frequency–Inverse Document) [26]. A TF-IDF is assigned to each word to generate the score. Frequency indicates the appearance of a certain word in a given text. Words that often appear in several files TF-IDF uses word frequency scores to highlight more interesting terms, for example, how often a word occurs in a particular file, but not across other files. Tokenization, vocabulary learning, and document frequency weighting are all part of the Tfidf Vectorizer’s capabilities. A Tfidf Transformer is often used for monitoring the number of times text is reversed and the text encoding process is started. Normalized scores are typically in the range of 0 to 1, which allows for ready usage of encoded document vectors.

K-means clusters the documents, and TF-IDF always measures a non-negative value, thus every document in a cluster will include its centroid, which is equal to the mean of all documents in the cluster. So, in other words, the keywords with the greatest impact on the centroid are those that have the most impact throughout all of the papers in that cluster. Each word is included, but a lot of unimportant information is ignored [26]. In other words, words most important to the document vector have the greatest TF-IDF values Equation (1), while those most important to the cluster as a whole have the highest centroid values.

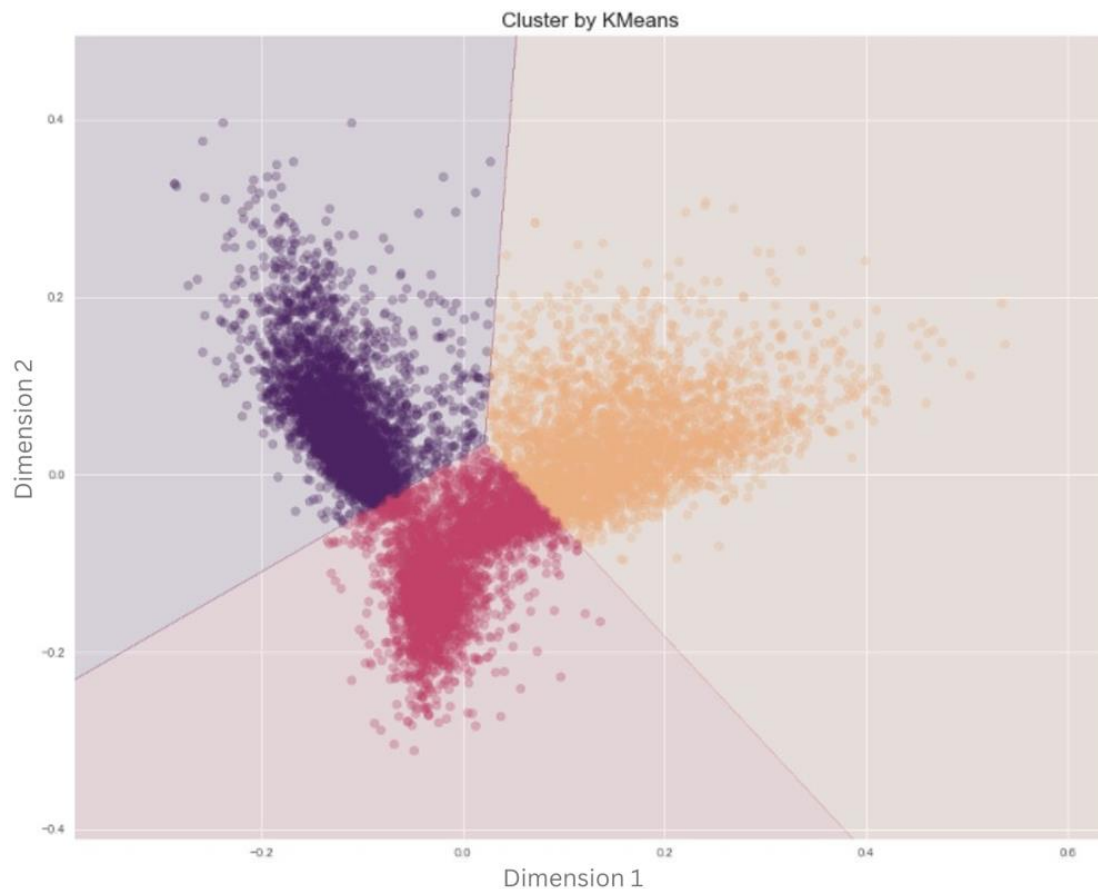
$$tf_{t,d} = \frac{count_{t,d}}{totalcount_d} \quad [1]$$

The total count (the entire number of all words in the text) is known as count (t, d). When IDF evaluates the degree to which a word is informative in a text for model training, it’s taking into consideration many different dimensions. One may calculate it as Equation (2).

$$idf = \frac{N}{Df_t} \quad [2]$$

The N-by-Df<sub>t</sub> matrix represents the total number of documents in the corpus that include the word t. When a word often appears in multiple texts, IDF assesses the weight of the term. An example of this is that stop words do have low IDF scores. As shown before, TF-IDF may be described as Equation (3).

$$tf - idf = tf_{t,d} \times \log(idf) \quad [3]$$



**Figure 14.** Here we can see the 3 clusters established using the silhouette technique are shown. Based on these findings, we decreased the dimensions, and then assigned their domains as well as colours to distinguish them based on the cluster values

When dealing with large datasets, it's always difficult to display the data, in our instance, tweets, and therefore very difficult to glean information from our tweets. As a result, it becomes our primary objective to decrease this greater dimensionality, which may be accomplished using methods such as PCA or TruncatedSVD [27], [28]. Either of the functions contributes to the reduction of the number of dimensions, which results in the elimination of outliers and the efficiency of calculation. It is usually suggested utilizing PCA or TruncatedSVD to decrease the number of dimensions to a manageable level when the number of features is extremely large [29]. This really does contribute significantly to noise reduction and acceleration in the calculation of pairwise distances between samples.

## VI. WORD CLOUD AND LABELING

A graphical depiction of tweets was produced by creating word clouds that help to show words as they arrive in tweets. Word clouds use the frequency of words to emphasize them. Throughout this study, the frequently performed word clouds in tweets linked to COVID-19 and its vaccination gave greater insight regarding the tweets about vaccine and COVID-19 over Twitter [30]. Based on the Figure 15, which is the preclustering word cloud with the information of all the words obtained from the preprocessed Twitter tweets, other very frequently present terms were associated with Vaccine, Vaccination, Age, and Group. These words correspond to the citizens' desire for vaccinations and concern about mandatory vaccination rules based on age group and the requirements being associated with COVID-19. In addition, the term people and death illustrate the distribution of the virus to people, along with their thoughts and fears, while the word 'Price' depicts the conditions that affect people's finances due to the continuing spread of the infection. There are eighteen, forty-five, state, and forty-four terms that are included in our word cloud to represent public perceptions of our research [31], [32], Figure 16.

We have two goals for this project: We want to analyze the impact of vaccines, and we want to learn more about people's emotions in general by looking at popular tweets which mention people ages 18-45 who need immunizations. Now we have three differentiating and distanced word clusters to look forward to. However, we did not know which cluster had positive, negative, and neutral terms. The following procedure shows how to assign labels to these clusters: SentimentIntensityAnalyzer, which is contained in the NLP package NLTK, is utilized [12].



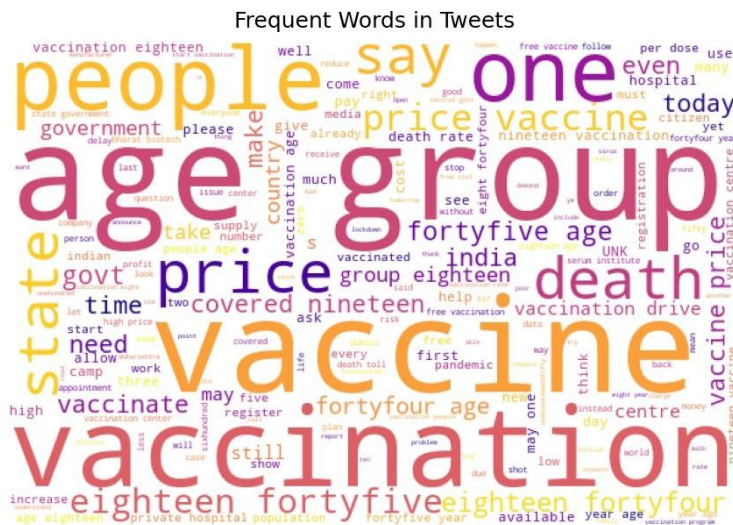


Figure 15. Here we see a word cloud of tweets that have not been clustered, which illustrates all the feelings and emotions of Twitter users.

This returns a dictionary having a collection of various scores there were 3 scores: negative, neutral, and positive. All of which added up to 1 and the sum couldn't be negative. To give an example, let's use the text "Wow, NLTK is extremely strong" with the algorithm pass. The algorithm produces 0.0, 0.295, 0.705, and 0.8012. The resultant of all three first measurements were negative, neutral, and positive and the last one was the compound in which the readings range from -1 to 1. This is used to identify and classify clusters based on which variable has the greatest level [25].

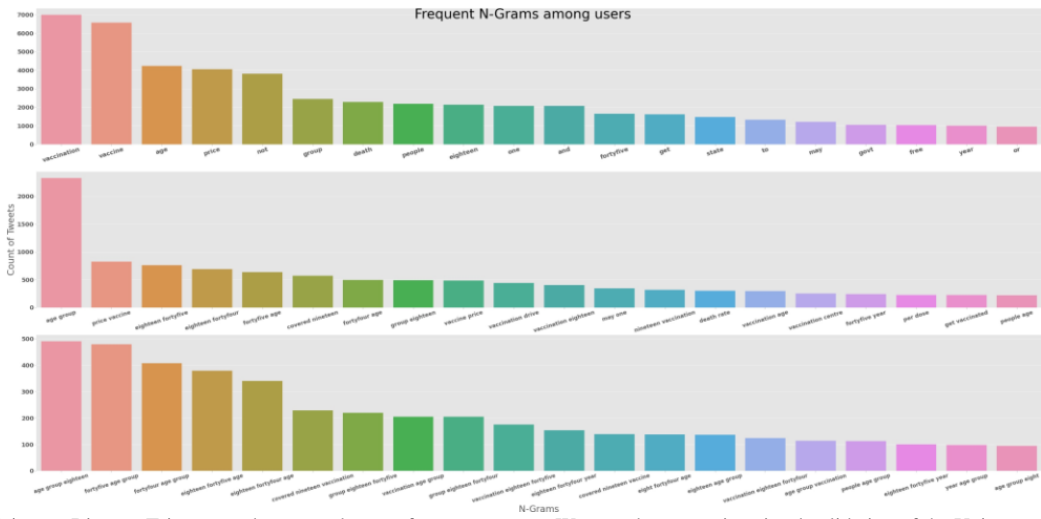


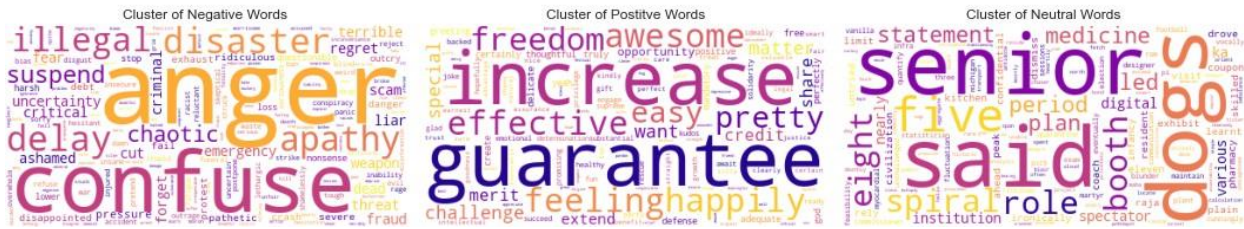
Figure 16. Unigram, Bigram, Trigram, and so on make up a frequent n-gram. We can observe a nice visual validation of the Unigram that pertains to vaccinations, age, pricing, and more, providing us confidence that the dataset is clean. In both cases, bigrams such as age group, price vaccine, and so on are listed as the concerns. Additionally, we see trigrams like "age group eighteen", "age group forty-five," and so on all reflecting the main problem as well.

The results show conflicting feelings when looking at Cluster 1 in the Word-cloud for all the Negative terms cluster [33], [34], [35]. Anger and confusion are very well expressed when it comes to this topic. Illegal is used to indicate that individuals in India have been receiving vaccine doses in different methods which aren't provided by the government and the citizens have been discussing it. We also look at the other words like disaster, apathy, delay, chaotic, suspend, terrible, scam, liar, etc. words used by people expressing their emotions towards this issue, Figure 17.

Cluster 2 for Positive words has been speaking more about Increase & Guarantee, which are national ideology-related to the government's ability to ensure everyone gets a vaccine [34]. We can note that citizens have been talking about hope and expectations arising from an increase in vaccination. We can observe the terms "freedom," "awesome," "effective," "happily," "special," "pretty," and other words related to happiness appear as people are overjoyed with this choice, which allows them to see hope and to express their enthusiasm. We get a glimpse of how this choice has made it simpler for the doses to be obtained via the term "easy", Figure 17.



## Topics Per Clusters



**Figure 17.** The image above is composed of 3 components, meaning essentially word cloud creation after being analyzed & clustered for the various emotions about the words in tweets. A) The leftmost being cluster over negative words, B) The one in center being cluster over positive words, C) Rightmost word cloud representing neutral words.

Neutral words tend to be hard to portray emotionally since they rely on neither positive nor negative [33], [35]. Despite these being matters of personal or public safety, we observe that there are discussions through tweets linked to senior citizens, dogs, institutions, booths, etc. Figure 17.

## VII. NEURAL NETWORK

A deep learning neural network is composed of multiple hidden nodes and therefore is inspired by the brain's neural networks, which are used to provide such a dependable output. This is useful for improving the precision of the Tweet user's feelings. This neural network is built using the TensorFlow framework [36]. The positive and negative keywords are first processed, and then the data from these processing stages is stored for comparisons. Tweets that have been analyzed before their publication are inspected for terms that are associated with either positive or negative sentiment, and then those tweets are assigned either to positive, negative, or neutral. Each tweet with a good, negative, or neutral message will have a corresponding score of  $\geq 1$ ,  $\leq -1$ , or  $= 0$  accordingly. So, using the Neural Network we had lots of variations in our Models. Starting from 1st to getting an accuracy of 97% in the 5th and our final Model of Neural Network.

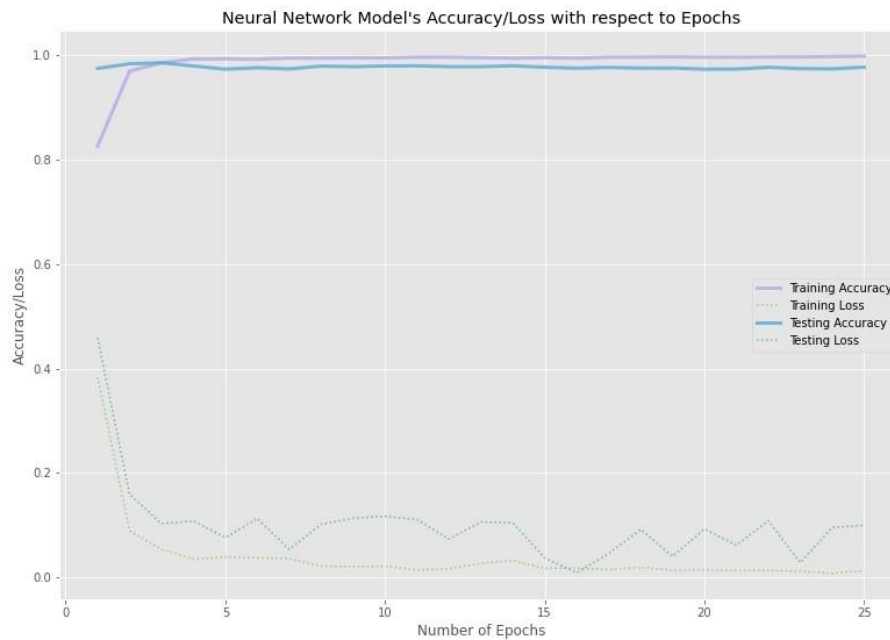
**1st Model:** At the beginning, we started by three dense layers connected to  $2^{14}$  hidden layers. The categorical cross-entropy loss function and SoftMax activation function are used in this research for multi-class and single-label predictions. We used tokenization, sequential embedding, and dense layer network building blocks to construct our network. This model is applied to the data that is remaining after training and is also validated on the validation set. We observe that the training loss decreases slowly while the validation loss continues to increase from epoch 5, Table 1. As both of them must increase together to get a good result. We conclude, that we had an overfitting problem in this model thus we went forward with another model.

**2nd Model:** To deal with these issues, we included dropout layers into the network model to help alleviate overfitting. Now we apply the dropout layer to the model to help with accuracy. LSTM is a kind of recurrent neural network, and like with RNN [37], it is a far more powerful choice when you require the network to remember information for a longer duration. We observe a slight variation in our accuracy but it's still not as good as we required because this well is having an overfitting problem, Table 1.

**3rd Model:** As a result, we go to the next model where we utilize various sampling schemes to improve the model's accuracy. After 10 epochs, the model begins to overfit and upscaling and downscaling begin again, Table 1. There was a rise in the rate of loss before models, though.

**4th Model:** We now use a bidirectional LSTM model to improve the model's accuracy [37]. Traditional LSTM help enhance model performance on sequence classification tasks, and these bidirectional LSTM are an extension of such models. Here we had better results when it came to training accuracy [38], [39], but when it came to real-world performance our F1 scores were poor and the model was unable to accurately predict the emotions of the tweets, Table 1.

**5th Model:** We ended up using dropouts, normalization, and several sampling approaches to enhance the accuracy of our model [36], [39]. This all led to excellent F1 ratings for each label, which also allowed us to pinpoint our issue statement, Figures 18 and 19.



**Figure 18.** In this figure, we took a model and fed the data to the Neural Network Architecture where it does forward and backward propagation, with respect to the epoch. We can see how the model reduces its losses and gains accuracy over the testing and training set.

**Table 1.** As can be seen from this table, there is a distinction between the matrices. We worked with a variety of models, but for the comparison, we chose five of the most optimal models. They are compared on their set levels over the matrix [Average Scores, F1 scores, Precision & Recall] in their respective classes.

Neural Network		Model 1	Model 2	Model 3	Model 4	Model 5
Average Scores	Training	0.99	0.98	0.97	0.95	0.95
	Testing	0.97	0.95	0.93	0.93	0.91
F1 scores	Class 0	0.98	0.96	0.97	0.95	0.91
	Class 1	0.96	0.93	0.95	0.93	0.88
	Class 2	0.96	0.95	0.95	0.92	0.89
Precision	Class 0	0.99	0.95	0.95	0.92	0.89
	Class 1	0.95	0.9	0.96	0.9	0.92
	Class 2	0.97	0.89	0.92	0.93	0.91
Recall	Class 0	0.97	0.9	0.89	0.9	0.91
	Class 1	0.95	0.91	0.88	0.91	0.92
	Class 2	0.97	0.89	0.93	0.9	0.92

## VIII. RESULT

The project's main aim was to analyze Vaccination tweets during the Covid-19 epidemic for emotion. We began taking a dataset of approximately 50,000 tweets from Twitter, and we worked our way through it to finally get a dataset of about 9525 tweets, Figure 1. In the process, we had found that only about 20% of the dataset were duplicates, and our tweets were especially relevant to the issue we had set out to study. A loss of data may be a concern, but it is much more critical to ensure that the cleanest possible clusters are created. Our final model, the Neural Network, trained with 99% accuracy and testing resulted in an overall rating of 97% Figures 18 and 19. A high level of accuracy concerning both class 1 and class 2 at 96% for both and a high level of precision for class 0 at 98%. The precision ranged from 95% and topped to 99% for the classes.

As a result, this became feasible, because we followed steps where we took the tweets, performed filtering over them during the data gathering process, where we filtered over our 724 keywords in hashtags out of a dataset aggregately having 2657 hashtags. As a result, we've received 30% of the total number of tweets as 15,174, i.e., 30% of the dataset. Removing URLs, Hashtags, utilizing Regular expressions, Tokenization and Stop words using NLTK [12]. 100% of the words were converted from their derived forms to their base forms because of the Lemmatization. Everything in this presentation was obtained in three different versions. As Version 3 was the most optimized, we calculated that our Version 1 for 25 tweets took about 30 seconds. While our most optimized version was 600% faster. This task, on average, took 0.2 seconds to do, Figure 3.

Our support count used to decrease due to the tweets by users, just being copied and appended with some extra words. Without removing this, we would have ended up with flawed clusters and irregularities, Figure 6. Ideating and creating an algorithm and having versions of it, we removed these distortions not just with accuracy but also with the least time taken. Our first comparison took 2 mins 20 secs to clean, while we beat the time scores with our accuracy being 100% and time is taken comparatively at  $3 \times 10^4$  times faster, only taking 0.5 secs for 500 tweets, Figure 7.

We found that using k-means clustering, our leaking factor was minimum, and validated the number of clusters using the Silhouette Method. It was observed that the silhouette method performed far better than the Elbow Method, which helped in forming profoundly distinguishable clusters, Figures 13 and 14. Thus, then all of this made it suitable for the categorization of the 9525, cleaned tweets which show that 33% are negatively classified, and 33% are positive, while the rest being neutral, Figure 17.

## IX. CONCLUSION

A systematic study of Covid-19 emotions stated in tweets by users is reported in this paper. Given both the worldwide escalation of the pandemic and the changing perceptions of the virus's consequences, such studies like we provide are becoming more important for researchers within the medical field in matters ranging from health issues to public awareness. This example illustrates the ability to summarize beliefs regarding experimentally validated disease preventive strategies via mining Twitter data.

We used several preprocessing methods and feature extraction techniques, beginning with the 49,345-row dataset, which got cleaned to 9525 Tweets, Figure 7, lastly used to train the Neural Network model. Immediate estimates of what people are saying and experiencing throughout the viral devastation may be produced with the assistance of abstract visualizations, such as the clustering methods shown. With using K-Means Clustering we could say that it aids classification well, given the tweets must be clean. We observe a success rate of 99% after a successful training and testing process. In researching F1 scoring, the literature has revealed scores of almost 98 percent, Figure 18. Using the approach, we conclude that with our preprocessing method, algorithm, and Neural network model we developed, we can be certain that our findings will continue to be optimal as the number of tweets about vaccines continues to grow day by day.

## X. FUTURE SCOPE

Our pipeline is designed to serve as a helpful and beneficial tool for healthcare professionals and public officials by presenting useful information about global health issues. In the analysis, one may determine which groups or subgroups have fallen short of being reached by current programs and make use of predictive modeling to help health organizations plan and optimize outreach initiatives.

Additionally, in the current research, the time of vaccination news and updates is limited to just a certain duration. Vaccine development stages, projection of vaccine availability, vaccination approvals, and early vaccine doses have taken the bulk of Twitter activity during this time period. Future studies may incorporate vaccine-related tweets as individuals in the age range over 18 are currently getting vaccinations.

## REFERENCES

- [1] Yang, J., Chen, X., Deng, X., Chen, Z., Gong, H., Yan, H., Wu, Q., Shi, H., Lai, S., Ajelli, M., Viboud, C., & Yu, P. H. (2020). Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China. *Nature Communications*, 1. <https://doi.org/10.1038/s41467-020-19238-2>.
- [2] Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., Duan, W., Tsoi, K. K., & Wang, F.-Y. (2020). Characterizing the Propagation of Situational Information in social media During COVID-19 Epidemic: A Case Study on Weibo. *IEEE Transactions on Computational Social Systems*, 2, 556–562. <https://doi.org/10.1109/tcss.2020.2980007>
- [3] Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. *IEEE Access*, 90225–90265. <https://doi.org/10.1109/access.2020.2992341>.
- [4] Long, S. W., Olsen, R. J., Christensen, P. A., Bernard, D. W., Davis, et al. (2020). Molecular Architecture of Early Dissemination and Massive Second Wave of the SARS-CoV-2 Virus in a Major Metropolitan Area. *MBio*, 6. <https://doi.org/10.1128/mbio.02707-20>.
- [5] Salzberger, B., Glück, T., & Ehrenstein, B. (2020). Successful containment of COVID-19: the WHO-Report on the COVID-19 outbreak in China. *Infection*, 2, 151–153. <https://doi.org/10.1007/s15010-020-01409-4>.
- [6] Vahidy, F. S., Drews, A. L., Masud, F. N., Schwartz, R. L., Askary, B. "Billy," Boom, M. L., & Phillips, R. A. (2020). Characteristics and Outcomes of COVID-19 Patients During Initial Peak and Resurgence in the Houston Metropolitan Area. *JAMA*, 10, 998. <https://doi.org/10.1001/jama.2020.15301>.
- [7] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 8, 727–733. <https://doi.org/10.1056/nejmoa2001017>.
- [8] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 2, 15–21. <https://doi.org/10.1109/mis.2013.30>.
- [9] Shen, K.-L., Yang, Y.-H., Jiang, R.-M., Wang, T.-Y., Zhao, D.-C., et al. (2020). Updated diagnosis, treatment and prevention of COVID-19 in children: experts' consensus statement (condensed version of the second edition). *World Journal of Pediatrics*, 3, 232–239. <https://doi.org/10.1007/s12519-020-00362-4>.
- [10] Cotfas, L.-A., Delcea, C., Roxin, I., Ioanas, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month Following the First Vaccine Announcement. *IEEE Access*, 33203–33223. <https://doi.org/10.1109/access.2021.3059821>.
- [11] Fan, G., Yang, Z., Lin, Q., Zhao, S., Yang, L., & He, D. (2020). Decreased Case Fatality Rate of COVID-19 in the Second Wave: A study in 53 countries or regions. *Transboundary and Emerging Diseases*, 2, 213–215. <https://doi.org/10.1111/tbed.13819>.
- [12] Jongeling, R., Datta, S., & Serebrenik, A. (2015). Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 531-535).
- [13] Liu B. (2011) *Opinion Mining and Sentiment Analysis*. In: *Web Data Mining. Data-Centric Systems and Applications*. Springer, Berlin, Heidelberg.
- [14] Liu B., Zhang L. (2012) *A Survey of Opinion Mining and Sentiment Analysis*. In: *Aggarwal C., Zhai C. (eds) Mining Text Data*. Springer, Boston, MA.

- [15] Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) (pp. 257-261).
- [16] Nelli F. (2018) Textual Data Analysis with NLTK. In: Python Data Analytics. Apress, Berkeley, CA.
- [17] Yogish D., Manjunath T.N., Hegadi R.S. (2019) Review on Natural Language Processing Trends and Techniques Using NLTK. In: Santosh K., Hegadi R. (eds) Recent Trends in Image Processing and Pattern Recognition. RTIP2R (2018). Communications in Computer and Information Science, vol 1037. Springer, Singapore.
- [18] Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 2, 451–461. [https://doi.org/10.1016/s0031-3203\(02\)00060-2](https://doi.org/10.1016/s0031-3203(02)00060-2).
- [19] Purnima Bholowalia, & Arvind Kumar (2014). Article: EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9), 17-24.
- [20] Yuan, Chunhui & Yang, Haitao. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J. 2*. 226-235. 10.3390/j2020016.
- [21] Aranganayagi, S., & Thangavel, K. (2007). Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)* (pp. 13-17).
- [22] Thinsungnoen, Tippaya & Kaoungku, Nuntawut & Durongdumronchai, Pongsakorn & Kerdprasop, Kittisak & Kerdprasop, Nittaya. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. 44-51. 10.12792/iciae2015.012.
- [23] Kodinariya, Trupti & Makwana, Prashant. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 1. 90-95.
- [24] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 8, 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- [25] Wagstaff, Kiri & Cardie, Claire & Rogers, Seth & Schrödl, Stefan. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of 18th International Conference on Machine Learning*. 577-584.
- [26] Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF\*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 3, 2758–2765. <https://doi.org/10.1016/j.eswa.2010.08.066>.
- [27] Sehgal, S., Singh, H., Agarwal, M., Bhasker, V., & Shantanu (2014). Data analysis using principal component analysis. In *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)* (pp. 45-48).
- [28] Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 228–233. <https://doi.org/10.1109/34.908974>.
- [29] Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., & Lin, S. (2007). Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 40–51, <https://doi.org/10.1109/tpami.2007.250598>.
- [30] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1833-1842).
- [31] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of LREC*. 10.
- [32] Saif H., He Y., Alani H. (2012). Semantic Sentiment Analysis of Twitter. In: Cudré-Mauroux P. et al. (eds) *The Semantic Web – ISWC 2012*. ISWC 2012. Lecture Notes in Computer Science, vol 7649. Springer, Berlin, Heidelberg.
- [33] Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733-2742.
- [34] Wang, X., Ma, X., & Grimson, E. (2007). Unsupervised Activity Perception by Hierarchical Bayesian Models. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8).
- [35] Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*. 3. 601-608.
- [36] Erik Wiener, Jan O. Pedersen, & Andreas S. Weigend. (1995). A Neural Network Approach to Topic Spotting.
- [37] Zaremba, Wojciech & Sutskever, Ilya & Vinyals, Oriol. (2014). Recurrent Neural Network Regularization.
- [38] Tsai, J.T., Chou, J.H., Liu, T.K (2006). Tuning the structure and parameters of a neural network by using hybrid Taguchi-genetic algorithm. *IEEE Transactions on Neural Networks*, 17(1), 69-80.
- [39] Leung, F., Lam, H., Ling, S., & Tam, P. (2003). Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Transactions on Neural Networks*, 14(1), 79-88.

# Chapter - 7

## An Intrusion Detection System Based on Data Analytics and Convolutional Neural Network in NSS-KDD dataset

Dr.D.Kalaivani

Associate Professor and Head, Department of Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, India

Email: dkalaivani77@gmail.com

*Abstract— Due to the internet's quick growth, intrusion attacks have been growing exponentially, making them a very important worry in the modern era. Cyber-attacks can target any of the millions of users of the internet, as well as international companies and government agencies. The creation of sophisticated algorithms to identify these network breaches is therefore one of the most important tasks in the field of cyber-security research. In order to recognise malicious traffic inputs, intrusion detection systems (IDS) are trained using data from internet traffic logs. Utilizing these techniques, malicious traffic inputs are detected. The most often used database for internet traffic record data is that maintained by the Network Security Laboratory's Knowledge Discovery and Data Mining (NSL-KDD) team. It also acts as the benchmark for present-day internet traffic. This framework seeks to discriminate between normal and abnormal (Denial of Service (DoS), Probe, User to Root (U2R), and Remote to Local (R2L)) categories in the NSL-KDD database with high detection precision and low false alarm rates. Several classifiers, including Naive Bayes (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), linear discriminant analysis (LDA), and Convolution Neural Network, will be used to achieve this (CNN). The unique and cutting-edge supervised detection techniques will be used in this study as the fundamental approaches to address the issue of the need for more labelled data during the IDS training process. The results of the trials show that, in terms of classification performance, the CNN classifier outperforms both recently presented approaches and other methods that are currently in use.*

*Keywords— Intrusion Detection System, Data Analytics, Convolutional Neural Network, NSS-KDD*

### I. INTRODUCTION

#### 1.1 OVERVIEW OF IDS

Services that are accessible over the internet have expanded fast in recent years. In fact, it is anticipated that there will be 50 billion devices online by 2020. Cyber dangers are a persistent threat to systems using information and communication technology, but there are benefits as well. (Wang, et.al 2020) suggests Malware attacks have become increasingly sophisticated and difficult to identify, and they can have detrimental impacts on the economy and society. A. Billions of dollars are lost each year as a result of IT service breaches, and in the coming years, this number is expected to increase. Consequently, the main concern in modern civilization is becoming cyber security.

---

© 2022 Technoarete Publishing

Dr.D.Kalaivani – “An Intrusion Detection System Based on Data Analytics and Convolutional Neural Network in NSS-KDD dataset” Pg no: 93 – 107.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch007>



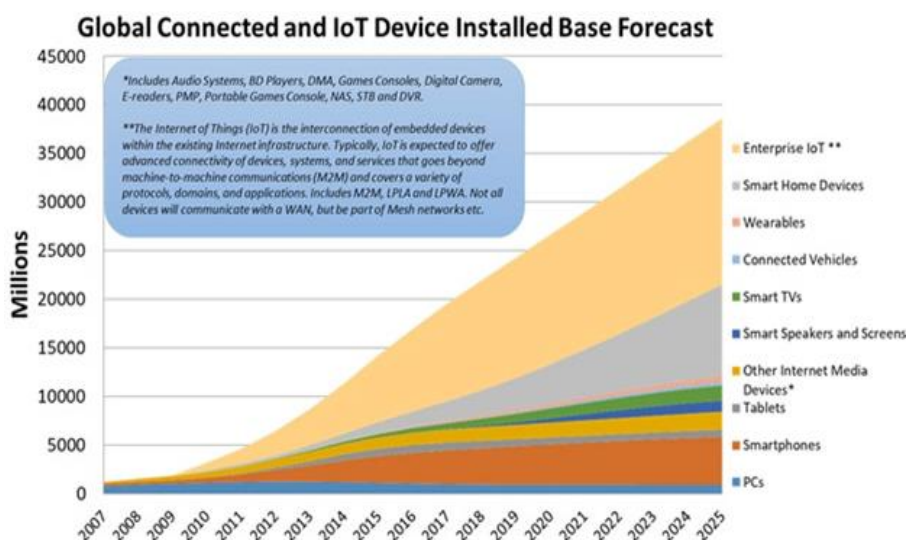


Figure 1. Number of connected devices in internet (www.businesswire.com)

According to the most current Strategy Analytics survey, there were 22 billion internet-connected devices worldwide as of the fourth quarter of 2018. (www.businesswire.com). By 2030, there will be 50 billion connected devices, and by 2025, there will be 38.6 billion. At this moment, network traffic data monitoring and analysis are essential to identifying potential attack trends. To address the dangers and security challenges, several academics are focusing on creating IDSs nowadays (A. Wang, 2020). IDSs' main responsibilities include keeping an eye on hosts and networks, analysing computer system behaviour, issuing warnings, reacting to suspicious behaviour, and generating alerts when it discovers suspicious acts and unknown risks.

An intrusion is any action that jeopardises an information system. Computer systems are monitored by intrusion detection systems, which check for odd behaviour that a conventional packet filter might miss (L. Nie et al, 2022). It's critical to achieve a high level of cyber resilience against harmful acts and to detect unauthorised access to an information system by scanning network packets for signs of malicious activity. To develop sophisticated, intelligent IDS capable of preventing attacks and guaranteeing improved cyber security in this situation. Many IDSs employ a number of techniques to initially find intrusions. Through the Signature-IDS (SIDS), also known as Knowledge-Based Detection, a signature identifier is produced for known malware so that malware can be identified in the future. If that exact signature is found again, the traffic can be identified as malicious. Especially for reported invasions, SIDS often offers excellent detection accuracy.

Because SIDS might be as simple as updating the signature-based information, three issues arise. First off, malware's polymorphic nature makes it easy to trick security mechanisms relying on signatures. This method fails the similarity test since it doesn't match any of the signatures in the IDS database, giving the attacker a way into the computer system. Second, when more signatures are stored in the database, it takes longer to analyse and practise the large amount of data. Anomaly-IDS (AIDS) systems, which are increasingly used to identify hostile attacks on computer systems, have overcome the limitation of SIDS. This approach is based on the notion that profiles of harmful activity deviate from those of typical user behaviour. A statistical model of the average user activity is created by AIDS, which detects any aberrant action that deviates from the standard model (A. Mishra et al., 2020). The goal of the AIDS technique is to keep track of actions, sketch up and display the predictable typical behaviour profile, and then classify anomalous events based on how much they depart from the expected behaviour. Anomaly detection systems can identify new assaults and have strong generalizability, but because cyberattacks are dynamic, they may have a high proportion of false alarms. The behaviours of unknown users are categorised as intrusions since they differ from customary behaviour. The training and testing phases are two of HIV/AIDS' phases.

The model is evaluated on a set of data that it was not exposed to during training after learning the typical traffic profile using data sets that replicate typical activity in the training stage. Initially, AIDS could be fought using machine learning-based methods. However, due to the enormous amount of heterogeneous data generated by multiple sources, machine learning (ML) techniques are ineffective and unsuitable for effectively addressing such security concerns (I. Kotenko et al., 2021). Even with its low computing cost, it is unable to comprehend the intricate non-linear relationships that are present in the sizable dataset. By utilising DL technology to create a more complicated IDS, the constraints mentioned above are eliminated, and intrusion detection performance is enhanced (B. Ge et al., 2020). In fact, DL can automatically generate extremely abstract levels from input representations using a hierarchical learning process.



### 1.2 EXISTING TECHNIQUES TO IDS

For the purpose of identifying intrusions, a variety of techniques are used, including host-based, network-based, rule-based, and signature-based systems. The aforementioned approaches have significant overheads and take a long time to compute. As of right now, there are just 2 types of IDSs: SIDS and network-related techniques, both of which can only detect known threats and have incredibly low false positive detection rates. In the attempt to create secure systems, ML and DL have proven to be some of the most successful technologies to date.

### 1.3 PROPOSED METHOD

The proposed study aims to provide a strong IDS framework that would boost network security. A system like that would be good at identifying attacks. To do this, we review, discuss, and assess a variety of cutting-edge ML and DL algorithms on the NSL-KDD dataset that can be applied to enhance network security. The proposed system, which includes feature extraction, pre-processing, and classification, is then described.

Stage 1: The basic data format is changed during the data pre-processing step, and data values are normalised. Normalized data must be transformed into image data format in order for the CNN model to function more effectively.

STEP 2: Feature extraction is one of the key stages in the development of ML/DL models. Relevant features in the data have an impact on the model's accuracy and lengthen the training process. An IDS must be created using the feature extraction method.

STEP 3: Training is done to improve the performance of the AI model by continuing parameter adjustments. To enhance the performance of the model, a number of parameters are also changed during the training process.

STEP 4: Test data should be utilised to evaluate the ML/DL model's accuracy following training (STEP 3). For instance, if the accuracy rate reached the required level, the training would be complete; otherwise, the model would go through the training process again.

STEP 5: It is evaluating step, which follows training, the model's performance is evaluated. Common evaluation measures include the accuracy rate, detection rate, and false alarm rate.

The following diagram uses the created data analytics algorithm and DL model to show the four steps of the IDS mode.

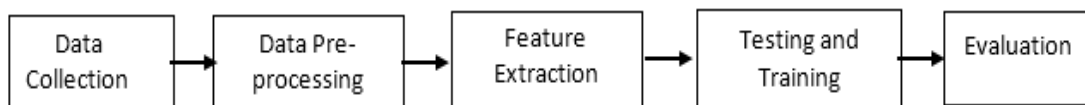


Figure.2 Intrusion Detection Model

The DL algorithm and DL model provided in this work are then used to demonstrate the 5 steps of the IDS in the accompanying image.

### 1.4 LIMITATION OF EXISTING TECHNIQUES

The NSL-KDD database has many redundant and null records, which is one of its key drawbacks. This concept forces learning algorithms to learn common recordings rather than unusual recordings, which are frequently more problematic for R2L and U2R networks. Due to their presence in the test set, these repeated records will also affect the evaluation findings. Here are a few more drawbacks:

- A high score was based on the supposition that the behaviour was acceptable in the system.
- The lengthy and challenging availability of a lot of data.
- Capable of being trained by assailants.
- Difficulty in setting metrics and parameters.

### 1.5 PROBLEM DEFINITION

The amount of network traffic in the web-enabled world has significantly increased as a result of the quick expansion of online users and their online communication. As the number of internet users increases, so do the security risks. The effectiveness of the network is significantly impacted by the magnitude and frequency of these attacks. Therefore, it is safe to say that integral protection against intrusions cannot be obtained despite the range of security measures that have lately been devised, such as peripheral protection mechanisms and numerous authentication and access control systems. Utilizing IDS is one solution to this issue. Recognizing and anticipating both normal and deviant behaviour is IDS's primary duty. In order to look for anomalous activities intended to illegally access, modify, and disable computer systems, an IDS collects and analyses data from numerous networks and computer sources. Any behaviour that deviates from the pre-defined (trained) data is recognised as an aberrant activity by IDSs, which display the typical system or network traffic behaviour. How these systems should be trained, or how to identify what constitutes typical behaviour of a system or network environment (which attributes are relevant), and how to represent this behaviour computationally, are significant challenges for anomaly-based IDSs.

This study was created utilising data analytics and AI (ML/DL) techniques to address these challenges. One of the IDS's

most popular tactics is the usage of AI algorithms because to their effectiveness. Machine and deep learning intrusion detection systems have increased their accuracy and capacity to identify new attacks in recent years.

### 1.6 CONTRIBUTION OF THIS WORK

In order to overcome the primary constraints like data recurrence and the presentation of null values, this research uses the feature extraction method in the NSL-KDD dataset to remove null and irrelevant features that interfere with the detection rate of the proposed work and increase the training time necessary to construct the IDS approach.

- Effective IDS is proposed with an architectural design.
- A thorough examination of the various assault types, attack datasets that are accessible, and attack detection techniques for research.
- The creation and application of data analytics-based AI techniques for IDS that improve classification accuracy and produce better outcomes.
- The development and application of supervised IDS to address the vast volume of labeled data.

## II. LITERATURE REVIEW

### 2.1 BACKGROUND RESEARCH

Machine learning, deep learning, and ensemble approaches are just a few of the cutting-edge technologies that various researchers have created in recent years to recognise invasions. Using methods for anomaly detection, the distribution of usual network data is determined; any data that differs from this distribution is referred to as an anomaly (U. S. Musa et al.,2020) For instance, the labelled dataset is used to build a classifier that uses ML methods like SVM, DT, and KNN, among others, to identify intrusions. According to R. Doshi, et.al (2019) Deep learning techniques like Auto Encoder, Deep Neural Network (DNN), Deep Belief Network (DBN), Recurrent Neural Network (RNN), and others make automatic feature extraction and categorization possible. The final category improves detection performance by using a range of ensemble and hybrid tactics, including bagging, boosting, stacking, and combination classifier algorithms. Using the KDD99 and NSL-KDD datasets, a number of IDSs have been published in the literature that assess the effectiveness and performance of our suggested models.

### 2.2 AI TECHNIQUES FOR IDSs

#### 2.2.1 ML METHODS

In 2017, Binhan Xu et al. proposed the incremental k-NN SVM technique, which combines the benefits of both SVM and k-NN. This method effectively increases the amount of training data and looks for k-NN using a R\*-tree. The expansion k-NN, SVM IDS methodology can train and update with new data in a reasonable period of time, and its forecasting time does not considerably increase throughout the expansion learning process, according to studies utilising the free database KDDCUP 99.

Anish Halimaa et al. compare the detection rate or classification accuracy and error rate or misclassification for the dataset after post-pre-processing (2019). The execution results show that for 19000 instances, the SVM's accuracy value is 93.85 percent, while the Naive Bayes' accuracy rate is 71.001 percent. For 19000 cases, Naive Bayes misclassifies more samples than SVM. For 19000 events, the Naive Bayes accuracy rate has fallen.

Extreme Learning Machines (ELM) and SVM are combined in a hybrid IDS proposed by Al-Yaseena et al. (2017) to improve the effectiveness of identifying both known and unknown assaults. On the KDD Cup 1999 dataset, the system had a false alert rate of 1.87 percent, which was good performance.

E. D. Alalade et al. (2020) created an IDS that combines ELM with an AI system to identify anomalies in the smart home system (AIS-ELM). The input parameters are evaluated by ELM for improved convergence in detecting aberrant activity after being optimised by AIS using the clonal technique.

The packet-level machine learning solution for DoS discovery proposed by Rohan Doshi et al. (2018) can successfully distinguish between genuine traffic and DoS attack traffic from consumer IoT initiatives. Using a database of routine then cyber-attack (DoS) movement created after a test network infrastructure for consumer IoT devices, they assessed five different machine learning classifiers, including KNN, LSVM, DT, RF, and NN. In this example, all algorithms achieve a 99.9% accuracy rate on the test set of data.

For the NSL-KDD database, Ingre et al. (2015) created a 3-layer MLP to detect attack classes. Objects are categorised into binary classes and five classes, with the following findings (type of attack). The analysis of the data using a variety of performance indicators revealed improved accuracy. Attack classification detection rates are 81.2 and 79.9 percent, respectively, while intrusion detection rates are 79.9 percent. The new method outperforms the existing system when compared to binary class and five class classification tasks, where a greater detection rate is attained.

According to Umar et al. (2020), feature selection failed to reduce computational time for models on the UNSW-NB15 dataset and, in the case of replicas built by NSL-KDD, decreased their working performance. In contrast, normalisation was more effective than feature selection in improving model performance on any dataset. The method also suggests that the

NSL-KDD dataset is less advanced and unsuitable for building reliable contemporary IDS when compared to the UNSW-NB15 database. Random Forest outperforms all other models on both datasets with accuracy on NSL-KDD data of 99.75 and 98.51 percent and UNSW-NB15 data of 98.51 percent, respectively.

Along with a comparison of those IDS datasets in terms of the types of computer attacks they contain and the features they choose, the investigation of several open datasets—including KDD-Cup'99, Center for Applied Internet Data Analysis (CAIDA), NSL-KDD, and then CICIDS2017—by Cremer.F et al. (2022) is presented. Finally, the authors published classification findings for the selected datasets based on their earlier research. Their model, which comprises of a payload classifier and an MLP neural network, had a 95.2% accuracy rate on CICIDS2017.

A K-means approach was proposed by Yonghao GU et al. (2019) to identify DDoS assaults at CICIDS2017. In order to avoid "the curse of feature dimensionality," they also created a fusion-based feature optimization strategy to remove arbitrary attributes from the model's input. The properties and features that are available are considered in this procedure. Data pre-processing, feature standing, and feature subset penetrating are the procedures that the features go through before being processed. A subset of the features that were processed are then output by the algorithm. Finally, they used their suggested optimization technique and got a 96.50 percent accuracy rate and a 30.5 percent error rate.

### 2.3 DEEP LEARNING METHODS FOR IDSs

Gurung Sandeep et al. (2019) introduced a 2-stage classification DL model system that divides traffic from the NSL-KDD Cup dataset into attack class and normal class using unsupervised feature learning. A sparse autoencoder is used to aid the deep learning model in learning features, and the sparsity penalty is then used for classification. A signature-based IDS strategy may be utilised to improve overall accuracy even while the model does a decent job of decreasing false positives.

Kehe Wu et al. (2018) used CNN to automatically detect traffic features from the raw data set and change the cost function weight coefficient of each class based on its numbers in order to address the issue of an unbalanced data set. This technique concurrently improves class with minuscule numbers accuracy while reducing the false alert rate (FAR). To further reduce computing expenses, they convert the raw traffic vector data into an image representation. Using the typical NSL-KDD data set, they evaluated the performance of the proposed CNN model. The experimental findings show that, in terms of accuracy, FAR, and computing cost, the suggested model performs better than traditional standard techniques. It is a well-known and reliable technique for network intrusion detection.

Samson Ho et al. (2021) proposed an IDS paradigm based on CNN to improve internet security. In order to identify network intrusions, the proposed IDS paradigm splits all network packet traffic into normal and up-normal categories. The proposed model was trained and validated using the CICIDS2017 dataset from the Canadian Institute for Cyber Security. Assessments have been made of the system's accuracy, attack true positive rate, false alarm rate, and training expense.

Jiyeon Kim et al. (2019) used DL methods to build a CNN model on the CICIDS2018 dataset, which had more samples but the same feature set as the CICIDS2017 dataset. The models for the study were created and assessed using portions of the CICIDS2018 database, which was a portion of the many different types of network traffic data. To simulate the models for multi-class classification, only some classes in the dataset—not all of them—were taken into account simultaneously. The RNN, a particular DL model that is generic when time sequence data is utilised as input data, may not perform as well as the CNN-based IDS, according to the experimental findings of this work. Using the subdataset from CICIDS2018 that was collected of the normal and DoS samples, the CNN method presented in this study was able to achieve a 96.77% accuracy rate. However, as compared to the CNN model and the RNN models examined in this study, only 82.84 percent of the dataset's predictions were accurate.

In order to assess the classification capabilities of a machine learning model with a DL model, K. Atefi et al. released a paper in 2019 employing anomaly analysis for intrusion detection using KNN and Deep Neural Networks. They used CICIDS2017 as the database to duplicate the model's performance from the study. They concluded that DNN performed significantly better than KNN. Their DNN in particular generates accuracy of 96.427 percent in contrast to the KNN's accuracy of 90.913 percent. The two models' running time overheads were also compared. The fact that DNN's 110-second CPU time is less than KNN's 130-second CPU time indicates that DNN has a shorter time than KNN.

2019 saw the introduction of the DL models for IoT network cyber security by Monika Roopak et al. The authors assessed the efficacy of the MLP model, Long Short Term Memory (LSTM), CNN, and a fusion technique of LSTM and CNN using the DDoS assault samples from CICIDS2017. The precision of the hybrid model was 97.41%, followed by that of the CNN model (98.14%) and the LSTM model (98.44%). Last but not least, the MLP model's simulation accuracy was 88.47%. Additionally, the authors compared their findings to a few ML models. After running the simulation, the authors found that, with the exception of MLP, all of the tested DL models outperformed ML models like SVM, Bayes, and RF.

DL-MAFID was created by Louati, Ktata, et al. (2020) to address the issues associated with multi-class cyber-attacks using multiagent systems and auto-encoder. The KDDCup'99 dataset is used to assess how well this approach uses an autoencoder to reduce dimensions. Shallow classifiers like MLP and KNN are also employed to distinguish between the five classes in the

dataset under study. The results of the testing demonstrated that DL-MAFID is capable of accelerating detection and achieving an accuracy of 99.95%.

A. Rashid et al. conducted a comprehensive analysis of the comparison between the industry-recognized databases NSL-KDD and CIDDS-001 in 2020. To achieve the best results, they applied fusion feature optimization procedures before utilising self-learning (ML / DL) classification algorithmic approaches as SVM, NB, k-NN, Neural Networks, DNN, and DAE. They evaluated the effectiveness of IDS using well-known performance pointer criteria. The experimental results demonstrate that k-NN, SVM, NN, and DNN classifiers perform nearly at 100% accuracy on the NSL-KDD dataset, but k-NN and NB techniques perform approximately at 99% accuracy rate on the CIDDS-001 dataset. The authors, suggested tactics, and the accuracy of their results are listed in the table below, which also summarises the examination of the literature.

**Table.1** Summary of literature review

Author Names (Years)	Proposed Method	Accuracy (%)
Halimaa et al. (2019)	SVM	93.85
Al-Yaseena et al. (2017)	SVM and Extreme Learning Machines	98.13
Rohan Doshi et al. (2018)	Packet-level Machine Learning	99.9
Ingre et al. (2015)	3-layers MLP	81.2
Umaret et al. (2020)	Random Forest	98.51
A. Khraisat et al(2018)	MLP	95.2
Yonghao Gu et al. (2019)	Semi-supervised K-means algorithm	96.50
K. Atefi et al. (2019)	DNN	90.913
Monika Roopak et al. (2019)	LSTM	98.44
Louati and Ktata et al. (2020)	Shallow classifiers (MLP and KNN)	99.95
A. Rashid et al. (2020)	DNN and DAE	95
Jiyeon Kim et al. (2019)	CNN model	96.77

This has assisted deep neural networks in dealing with intrusions in ways for intrusion detection, and hazardous actions are studied in-depth and categorised in this section of the literature study. For this, it first categorises the ML-based IDS schemes according to the features selection techniques they utilise, and it then specifies how each system aims to use ML techniques for identifying various types of intrusions. Furthermore, the shallow learning methods applied in conjunction with the ML methods in the aforementioned IDS systems are studied. To provide a complete knowledge of the analysed IDS frameworks, the primary contributions, advantages, and some shortcomings are also listed. Additionally, each area's datasets and evaluation measures are compared. Finally, it can be argued that DL is a fascinating method that presents the IDS with both opportunities and difficulties. According to the aforementioned literature research, our work focuses on DL approaches on the NSL-KDD database for intrusion detection framework.

### III. RESEARCH METHODOLOGY

#### 3.1 INTRODUCTION

Recently and with increasing the need to use the Internet in all applications and domains, the number of moving packets and the load over the network are increased. Therefore, the most important information is at risk despite the existence of multiple cyber protection systems such as the effective IDS, which is an effective system of protection and prevention. This section provides an overview of the NSL-KDD database, which was used for this investigation.

#### 3.2 CLASSIFICATION

##### 3.2.1 ML TECHNIQUES

In the interdisciplinary field of study known as machine learning (ML), the main emphasis has been on developing algorithms for machine learning. Learning is nothing more than learning from feature datasets. ML systems are frequently developed and applied in a way that enables the expert system to solve the diagnostic problem using past data. Reinforcement learning, unsupervised classification, and supervised classification are just a few of the various learning techniques that are now accessible for the classification task. Classification is one of the supervised learning procedures, and the target class is forecast using the classification tool (I. Abrar.et al, 2020).

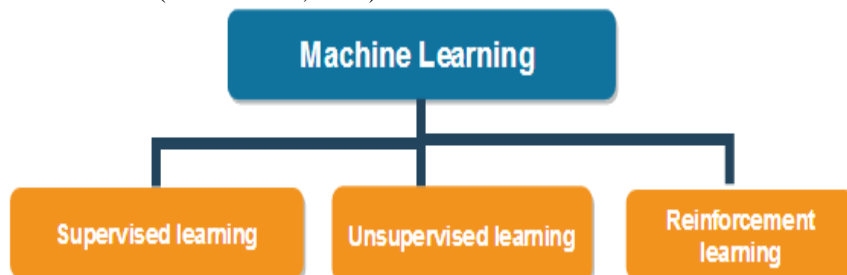


Figure.3 Machine Learning Algorithms

In the great majority of practical machine learning applications, supervised learning is used. Due to the fact that all of the data is labelled, the algorithms may learn to predict the results from the input data. A few examples include neural networks (Multilayer Perceptron), SVM, LR, NB, and KNN Algorithm. In unsupervised learning, all of the data is unlabeled, and the algorithms learn about the underlying structure from the incoming data. Examples include neural networks, fuzzy c-means, hidden markov models, and hierarchical, Gaussian, and c-means mixtures. The core of reinforcement learning is the sequential decision-making process. Simply put, the output is determined by the state of the current input, and the state of the next input is determined by the outcome of the previous input. Decision trees, linear regression, ensemble techniques, and neural networks are a few examples. An essential task in this system is choosing the right algorithm. Recently, a large number of supervised and unsupervised machine learning algorithms have been created, and each one uses a different learning strategy (S. Dwibedi, et al., 2020).

Using scientific data, the cyber-attack classification problem is a well-known classification problem that is addressed in this study using supervised machine learning techniques. We have a set of training records with the labels "A" and "B" for each of the different categories, namely  $D=X_1, X_2, \dots, X_n$ . The classification model associates one of the class labels with the features in the testing record. The model then forecasts a class label for a specific instance or set of features that belong to an undefined class. The basis for supervised learning approaches is the input image's annotated training features. The following image displays the supervised binary classification process (S. Dwibedi, et al., 2020).

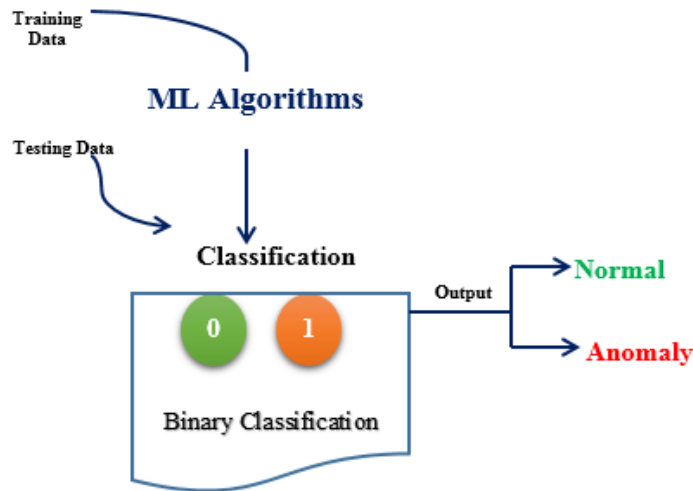


Figure.4 Functioning of supervised binary classification

The recommended method uses a variety of supervised learning techniques to categorise the NSL-KDD dataset into normal and malignant categories. The supervised learning is built upon the labelled instances in the training data set. It helps learning models be trained effectively, enabling them to provide high classification accuracy. It is essential to utilise learning algorithms with rigorous procedures as a result.

In supervised learning, a relationship is found between a set of input variables  $X$  and an output variable  $Y$ , and this relationship is then utilised to forecast the results for hypothetical data. The majority of practical machine learning algorithms are based on supervised learning. After all the data has been labelled, algorithms research the results based on the response data. This study uses several machine learning (ML) classification techniques, such as NB, KNN, SVM, LDA, and CNN, and shows how each algorithm performs given a set of features.

*Naïve Bayes (NB) Algorithm:*

The most crucial ML algorithms for classification are NB. The "NB" classifier supports this tactic by being built on the independence notion. It is employed for a variety of purposes, including text classification and spam filtering. When calculating the likelihood score of grades for an assumed feature subset, the combined prospects of attributes/features and grades are used. Using the straightforward probabilistic classifier, this classifier separated a set of data ( $d_r$ ) into classes ( $C$  ( $i=1$ )  $m = c_1, c_2, \text{ etc}$ ). (A. Ali et al, 2020). Maximum Posterior (MAP) class returns are the optimal class returns for the "NB" technique.

$$C_{\text{map}} = \underset{C_i \in C}{\text{argmax}} P(C_i) P(d_r | C_i)$$

In this situation, the class " $P(C_i)$ " can be identified by dividing the total number of features in class " $C_i$ " by the total number of features. The frequency of the characteristic in the data " $d_r$ " that corresponds to class " $C_i$ " was expressed by the expression  $P(d_r | C_i)$ . The probability " $P(C_i | d_r)$ " for each latent class will be calculated, but " $P(d_r)$ " stays the same. The denominator might be eliminated in light of this. By calculating the later likelihood of each class, it determines the most likely classes (called " $C_{\text{map}}$ ") given the data (called " $d$ ").

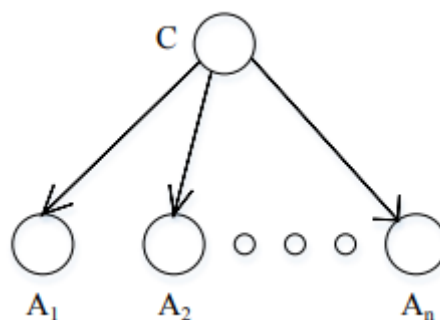


Figure.5 Functioning of NB



The following graphic depicts the simple structure of the NB classifier, where the classification node serves as the parent node for all other nodes. No further connections are allowed in a Naive-Bayes structure (A. Ali et al, 2020).

*K-Nearest Neighbour (KNN) Algorithm:*

Machine learning systems rely heavily on the KNN technique. It belongs to the area of supervised learning and is used extensively in a variety of fields, including pattern recognition and intrusion detection. These KNNs are applied in practical scenario when non-parametric approaches are required.

These techniques don't establish any patterns for the distribution of data. The KNN method splits the dataset's correlatives into clusters that can be recognised by a certain trait. This strategy's key advantage is that it generates comparable results for comparable training data. The most appropriate classification for all or portion of the samples is found for the input population (F. Z. Belgrana, et al 2021).

Examine the sample populations  $X_i = \{x_1, x_2, \dots, x_{iN}\}$  and  $X_j = \{x_1, x_2, \dots, x_{jN}\}$  in order to calculate the distance between them and determine how similar they are.

$$\text{Dist}(X_i, X_j) = \sqrt{\sum_{m=1}^N (x_{im} - x_{jm})^2}$$

The equation above states the Euclidean distance, which compares how similar two pixel locations are. The pixels are consequently assigned to the group to which a large majority of them frequently belong. The number of K nearest neighbours in the KNN is K. The number of neighbours is the main deciding factor. When there are two courses, K is frequently an odd number. The calculation is referred to as the closest neighbour computation when K=1. This is the circumstance that is easiest (F. Z. Belgrana, et al 2021).

*Support Vector Machines (SVM) Algorithm:*

SVM is a supervised machine learning (ML) model that can be built in both a linear and nonlinear manner. Because datasets are frequently nonlinearly inseparable, the main goal of the SVM approach is to capture the best surface that is available to discriminate between positive and negative training feature samples based on experimental threat (training set additionally test set error) reduction principal. This approach can try to represent a decision boundary in a high-dimensional feature space using hyper-planes. The hyper plane generates a decision based on this support vector by dividing the factorised input into two groups. You can use the example below to explain how the SVM works (W. Wang, et al, 2020).

We provide a feature vector "x" with "d" dimensions and a training set of "N" linearly separable objects. Two optimizations are required.

Here,  $\alpha \in \mathbb{R}^N$  then  $y \in \{1, -1\}$

The result of SVMs may then be summarized as below:

$$\vec{\alpha}^* = \text{argmin} \left\{ -\sum_{i=1}^n \alpha_i + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle \right\}$$

Where,

$$\sum_{i=1}^n \alpha_i y_i = 0; \quad 0 \leq \alpha_i \leq C$$

A linear dataset is split into two classes by an SVM classifier using a single hyperplane and a particular feature subset. In this case, kernel functions are used to lay out the data to a higher dimensional space that is linearly separable in order to manage nonlinear datasets with more than two classes.

The following diagram demonstrates the fundamental idea of SVM, which can be regarded of as having the objective of differentiating between the positive and negative classes in feature space. The main problem is to find a hyper plane that successfully divides those classes based on maximum margin.

*Linear Discriminant Analysis (LDA):*

Linear discriminant analysis can be carried out using a supervised classification method (LDA). In a general classification problem, an arbitrary variable named X comes from one of K classes, and some class-specific probability densities specified (x). A discriminant rule tries to partition the data space into K distinct chunks in order to represent all the classes (visualise the cases on a chessboard). If one of these regions contains x, discriminant analysis classification simply identifies x as being a member of class j. (S. Subbiah, et al., 2022). Two allocation criteria are used to logically locate the region of the data x:

Assuming that each class has an equal chance of happening, use the maximum likelihood rule and assign  $x$  to class  $j$  if

$$j = \arg \max_i f_i(x)$$

Bayesian rule: Allocate  $x$  to class  $j$  if we are aware of the class prior probabilities,

$$j = \arg \max_i \pi_i f_i(x)$$

Explicit variations of the aforementioned allocation rules can be made if it is assumed that the data is from a multivariate Gaussian distribution, where the distribution of  $X$  can be described by its mean ( $\mu$ ) and covariance ( $\Sigma$ ). According to the Bayesian rule, if data  $x$  has the highest likelihood among all  $K$  classes for  $I = 1$ , we allocate it to class  $j$  to  $K$ :

$$\delta_j(X) = \log f_i(x) + \log \pi_i$$

The phrase used to describe the aforementioned function is a discriminant role. Note the application of log-likelihood in this case. In other words, the discriminant function tells us of the probability that data  $x$  belongs to each class. The set of  $x$  where two discriminant functions have the same value is thus the decision boundary between any two classes,  $k$  and  $l$ . We were unable to decide, thus any data on the decision boundary is equally likely to come from either class.

LDA happens when we assume that the covariance in each of the  $K$  classes is equal. In other words, instead of having a covariance matrix for each class, all classes share a single one. Therefore, it is possible to obtain the discriminant function stated below:

$$\delta_k(X) = X^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Everyone can see that this is a linear function in  $x$ . Since each decision boundary between two classes is also a linear function in  $x$ , this technique is known as linear discriminant analysis. Without the equal covariance assumption, the likelihood's quadratic term does not cancel out, hence the resulting discriminant function is a quadratic function in  $x$ .

$$\delta_k(X) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k$$

In this case, the decision boundary is quadratic in  $x$ . This is known as linear discriminant analysis (LDA) (S. Subbiah, et al., 2022).

### 3.2.2 DEEP LEARNING (DL) TECHNIQUES

Deep learning is a crucial concept in machine learning. Data representation in DL is done using a layered hierarchy technique that mimics how the human brain functions. Deep learning may employ the analysis of enormous volumes of data to produce conclusions with accuracy rates that are higher than those of traditional categorization techniques. Many different fields are using DL techniques to address problems. Deep learning-related initiatives have been in the works at numerous prestigious companies, with a range of big data applications. However, there are a lot of challenges that come with these triumphs. Deep learning has a bright future because computer scientists are working together with specialists in a variety of industries to develop the algorithms that manage the current challenges (S. Subbiah, et al., 2022).

In both anomaly detection contexts, a wide range of machine learning techniques have been applied in numerous recent studies. Conventional machine learning (ML) methods struggle with a lack of labelled training datasets and mostly rely on manually extracted features, making it difficult to use on large platforms. DL is a cutting-edge paradigm in machine learning (ML) that was developed primarily using artificial neural networks (ANNs) and outperforms other conventional machine learning techniques (S. Subbiah, et al., 2022).

CNNs, Deep Belief Networks (DBNs), Restricted Boltzmann Machines (RBMs), Recurrent Neural Networks (RNNs), and Deep Neural Networks are only a few of the networks used in deep learning (DNN). These networks can learn in supervised, semi-supervised, or unsupervised ways. They also benefit from the use of hierarchical layers, which, as opposed to relying on manual features, work to extract appropriate high-level attributes from the raw input data. The literature offers a variety of DL-based anomaly detection models to handle various sorts of intrusions and security threats because DL has attracted a lot of interest in the IDS field.

#### Convolutional Neural Network (CNN)

CNNs can significantly reduce weight and computation while training with multidimensional data by utilising the convolutional feature. The back propagation technique is trained using a CNN, a multi-layer network, which is used to recognise two-dimensional figures and decrease the necessary parameters. The unlabeled data may also be used to extract various amounts of features (X. Zhang et al, 2019).

Xiao et al. introduced CNN-IDS, a network IDS built on a CNN paradigm. The network traffic features are extracted using CNN, and the data required for intrusion detection is collected using supervised learning. This method reduces the execution cost by turning the traffic vector into an image. The KDDCup99 performance evaluation tool was used by the authors to show that the CNN and IDS model works well based on a variety of parameters, including FAR, accuracy, and timeliness. However, the low detection rate of attacks like R2L and U2R is not resolved by this technique.

#### IV. MODEL IMPLEMENTATION

The practical applications of the model are covered in this section. The model's implementation is covered in this chapter, together with the software setup, hardware setup, and decision-making procedure. The entire test is done on a 64-bit macOS system with an Intel 2.6GHz 8-core i7 processor, 16GB of 2400MHz DDR4 RAM, and a Radeon Pro 560X 4GB GPU. The applications are all written in Python 3.8 and run in the Anaconda environment.

**Table.2** Important libraries

<b>Libraries</b>	<b>Version</b>
Keras	2.3.1
Scikit-learn	0.21.3
Numpy	1.21.5
Matplotlib	3.1.1
Pandas	0.25.1
Psutil	5.6.3
Joblib	0.13.2

##### 5.1 DATASET DESCRIPTION

Many intrusion detection systems typically use the NSL-KDD dataset as a benchmark. It is a subset of the original KDD99 dataset. Due to redundant and duplicate data in the test and train sets, the prior KDD99 was criticised for skewing the classifiers in favour of more frequent samples. These issues are actually addressed by NSL-KDD. The NSL-KDD dataset is available without charge from the Canadian Institute of Cybersecurity. KDDTrain+ and KDDTest+ are the names of the two data sets that it uses for testing and training. Particularly, the KDDTest+ dataset featured seventeen more attack kinds that weren't available in the KDDTrain+ for a fair classification, therefore the occurrences matching to such categories were deleted. The KDDTrain+ and KDDTest+ sets are detailed in Table 3 below.

**Table.3** Feature details of NSL-KDD dataset

<i>duration</i>	<i>destination bytes</i>	<i>num failed logins</i>	<i>num root</i>
<i>is guest login</i>	<i>error rate</i>	<i>dst host count</i>	<i>dst host srv diff host rate</i>
<i>protocol type</i>	<i>land</i>	<i>logged in</i>	<i>num file creations</i>
<i>count</i>	<i>srv error rate</i>	<i>dst host srv count cont</i>	<i>dst host error rate</i>
<i>service</i>	<i>wrong fragment</i>	<i>num compromised</i>	<i>num shells</i>
<i>srv count</i>	<i>same srv rate</i>	<i>dst host same srv rate cont</i>	<i>dst host srv error rate</i>
<i>flag</i>	<i>urgent</i>	<i>root shell</i>	<i>num access files</i>
<i>error rate</i>	<i>diff srv rate</i>	<i>dst host diff srv rate</i>	<i>dst host error rate</i>
<i>source bytes</i>	<i>hot</i>	<i>su attempted</i>	<i>Num outbound cmds</i>
<i>srv error rate</i>	<i>srv diff host rate</i>	<i>dst host same src port rate</i>	<i>dst-host srv error rate</i>
			<i>is host login</i>

##### 5.2 EXPERIMENTAL DESIGN AND IMPLEMENTATION

I'll describe my process for doing the implementation and the methods I employed in this part.

###### 5.2.1 DATA PREPROCESSING

The numeric features values  $z_{i2}$  were mapped using the min-max preprocessing approach into the numeric range [0-1], as shown in the following equation:

$$z_{i2} = \frac{z_{i1} - z_{min}}{z_{max} - z_{min}}$$

Where  $z_{i1}$  denotes the feature's initial data,  $z_{max}$  the feature's maximum data,  $z_{min}$  the feature's minimum data, and  $z_{i2}$  the data after  $z_{i1}$  was normalized. In contrast, the normalized feature value, [0-1], is represented by  $z_{i2}$ .

### 5.3 FEATURE EXTRACTION

The three categorical features, such as protocol type, service, and flag, were transformed into numerical values using the one-hot encoding technique. In particular, a binary value is used to represent each categorical property. For example, the protocol type has three attributes: tcp, udp, and icmp. Using the one-hot encoding method, each of them was independently converted into a binary vector, yielding the values [1,0,0], [0,1,0], and [0,0,1], respectively. In a similar fashion, service and flag were also turned to one-hot-encoding vectors.

A data frame with only categorical attributes is constructed after the categorical data attributes are selected. By Employing pandas, category attributes can be one-hot encoded using the obtain dummies () function. This processing module extracts the features that are highly correlated. The KDDTrain+ and KDDTest+ sets both look at each continuous feature's fraction of zeros. Figure 13 displays the distribution of null values for each numerical variable in the KDDTrain+ set.

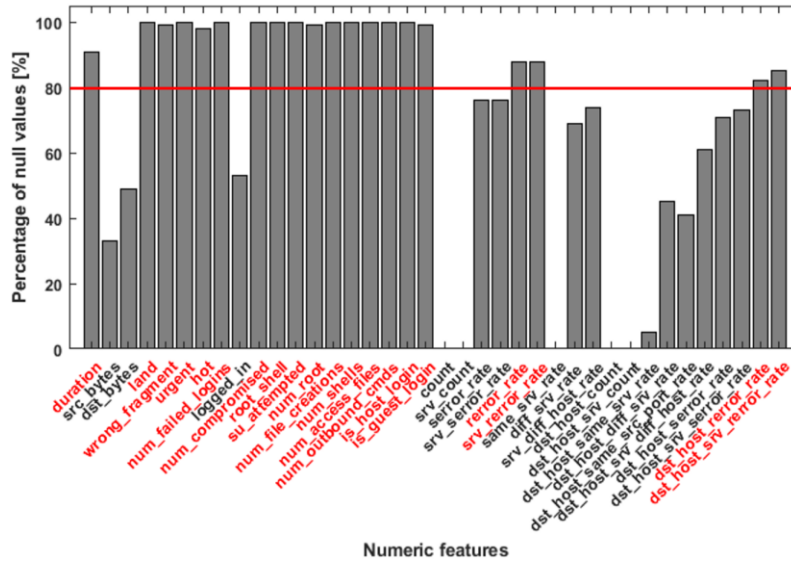


Figure.6 Graph of null values comprised in the 38 numeric attributes of the KDDTrain+ set

During the labelling process, attack labels are divided into "normal" and "abnormal" groups. A data frame is then created using only the binary class dataset's numeric properties and the encoded label attribute. Finding qualities having a correlation of at least 0.5 with the encoded attack label attribute is the next stage. When the encoding process is accomplished, attributes are selected using the pearson correlation coefficient method, and the combined encoded, one-hot-encoded, and original attack label attribute file is then saved to the disc.

### 5.4 CLASSIFICATION

The 20 characteristics that this study disapproves are underlined in red in Figure 13. While merging the remaining 18 continuous features with 84 one-hot-encoding vectors to get a 102-dimensional characteristics vector. To analyse IDS in comparison, different ML and DL algorithms, mentioned in the previous section, may be used. Only a few intrusion detection datasets are covered in the literature review. Since most researchers prefer to use the NSL-KDD dataset for IDS, we will compare various ML and DL IDS algorithms utilising it. Import various ML and DL models to determine each model's accuracy using the retrieved dataset.

5.4.1 PERFORMANCE EVALUATIONS METRICS

The IDS is often assessed using the confusion matrix measurement in the following ways:

		Predicted class	
		Normal	Intrusion
Actual class	Normal	True Negative (TN)	False Positive (FP)
	Intrusion	False Negative (FN)	True Positive (TP)

Figure.7 Confusion Matrix

- Accuracy or Classification Rate (CR): The CR is calculated as the IDS's overall classification accuracy for both routine and intrusion assaults:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (P): P is the ratio of the total number of true positive (TP) and false positive (FP) instances divided by the total number of true positive (TP) instances:

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

- Recall (R): Is the ratio of the total number of true positives (TP) and false negatives (FN) instances to the total number of correctly identified relevant outcomes (TR):

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

- F1-score: It represents the average of recall and precision. When only one precision metric is required as an evaluation measurement, F-Measure is preferred:

$$F - \text{measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

For the classification process, some effective classification algorithms as NB, KNN, SVM, LDA, and CNN are imported. The Adam optimizer can be used to iteratively alter network weights based on training data in a CNN model to increase accuracy. The dropout function is used in a CNN model to screen the data to prevent over fitting. The built-in functions were used in this work to train the data. 25% of the data were used for testing throughout evaluation, and the random state value was 42. The various machine learning models are trained using built-in libraries (X. Zhang et, al, 2019)..

Table.4 Accuracy values of ML classifier on KDDTest+ dataset

Algorithms	Accuracy
NB	82.6
KNN	85.6
SVM	88.57
LDA	89.7
CNN	92.23

Table.5 Shows how well the NB, KNN, LDA, and CNN classifiers

Algorithms	NB	KNN	SVM	LDA	CNN
Precision	82.5	80.5	82.3	86.3	92.7
Recall	85.67	85.89	85.61	87.37	89
F1-score	87.45	86.74	86.01	87.46	89.78

## 5.5 TECHNICAL CHALLENGES

Because ML/DL is a promising technology, its purpose needs to be developed. It is challenging to fight against cyber-attacks due to the sophisticated strategies that adversaries use. But if ML/DL is designed well from the start, it might be able to get past many challenges and be useful. Choosing the optimal strategy for cyber-security is an intriguing challenge.

## V. CONCLUSION AND RECOMMENDATIONS

### 5.1 CONCLUSION

Cybercriminals began utilising cutting-edge methods, such as polymorphism, to develop new attacks and frequently alter the signature to produce attacks in large quantities. IDSs should be able to reliably recognise both known and zero-day attacks in order to be considered effective. Through the use of techniques for both known and unidentified threat identification, this study suggests a novel paradigm for developing an intelligent IDS that outperforms current IDSs. The focus of the work is primarily on the latter. The two fundamental subcategories of intrusion detection systems are abuse detection strategies and anomaly detection techniques. The previously established signature of security threats and damaging behaviours is used by abuse detection algorithms to identify intrusions. In the security literature, dealing with intrusions is a topic that receives a lot of interest, and a variety of machine learning and deep learning techniques are examined. Recently, many IDS strategies have embraced the fascinating subject of deep learning to boost the efficiency and performance of misuse detection in a variety of scenarios. We have demonstrated that a CNN classifier can yield a better detection rate than other machine learning methods in terms of detection rate, accuracy, specificity, sensitivity, and F1 score. In this work, the ML and DL classifiers were used to categorise the intrusion. In terms of outputs, the CNN algorithm has significantly outperformed ML techniques.

ML and DL methods detect network intrusions by predicting the risk with the help of training the data. Various machine learning and deep learning methods have been proposed over the years which are shown to be more accurate when compared to other cyber security systems.

### 5.2 RECOMMENDATIONS

Future work will focus on developing IDSs based on DL that are very reliable in hostile, dynamic network environments.

The NSL-KDD dataset was used in this study as a benchmark to assess the IDS. The CNN intrusion detection method used in the study suggests an increase in training and recognition speed while maintaining accuracy rate. We intend to develop more precise deep architectures capable of controlling real-time data runs similar to NSL-KDD characteristics in order to deter harmful attacks in real-time systems. To benefit from long-term learning in real-time data, earlier decision-making criteria and less complicated processing are also required for big data analysis. Our upcoming research will focus on how to apply this method, which is also acceptable to boost the efficiency of IDSs in identifying diverse threats.

## REFERENCES

- [1] <https://www.businesswire.com/news/home/20190516005700/en/Strategy-Analytics-Internet-of-Things-Now-Numbers-22-Billion-Devices-But-Where-Is-The-Revenue>.
- [2] A. Wang, "Internet of Things Computer Network Security and Remote Control Technology Application," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 1814-1817.
- [3] L. Nie et al., "Intrusion Detection for Secure Social Internet of Things Based on Collaborative Edge Computing: A Generative Adversarial Network-Based Approach," in IEEE Transactions on Computational Social Systems, vol. 9, no. 1, pp. 134-145, Feb. 2022.
- [4] I. Kotenko, I. Saenko, O. Lauta and M. Karpov, "Situational Control of a Computer Network Security System in Conditions of Cyber Attacks," 2021 14th International Conference on Security of Information and Networks (SIN), 2021, pp. 1-8.
- [5] B. Ge and J. Xu, "Analysis of Computer Network Security Technology and Preventive Measures under the Information Environment," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 1978-1981.
- [6] U. S. Musa, M. Chhabra, A. Ali and M. Kaur, "Intrusion Detection System using Machine Learning Techniques: A Review," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 149-155.
- [7] B. Xu, S. Chen, H. Zhang and T. Wu, "Incremental k-NN SVM method in intrusion detection," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 712-717.
- [8] Anish Halimaa A, K. Sundarakantham: Machine Learning Based Intrusion Detection System. In: Proceedings of the Third International Conference on Trends in Electronics and Informatics, pp. 916-920. IEEE Xplore, Tirunelveli, India (2019).
- [9] E. D. Alalade, "Intrusion Detection System in Smart Home Network Using Artificial Immune System and Extreme Learning Machine Hybrid Approach," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-2.
- [10] R. Doshi, N. Apthorpe and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 29-35.
- [11] B. Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," 2015 International Conference on Signal Processing and Communication Engineering Systems, 2015, pp. 92-96.
- [12] Mubarak Albarka Umar, Chen Zhanfang Effects of Feature Selection and Normalization on Network Intrusion Detection, Communication, Networking and Broadcast Technologies, 2020, 10.36227/techrxiv.12480425.v2.
- [13] Cremer, F., Sheehan, B., Fortmann, M. et al. Cyber risk and cybersecurity: a systematic review of data availability. Geneva Pap Risk Insur Issues Pract 47, 698-736 (2022).
- [14] Y. Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised K-means DDoS detection method using hybrid feature selection algorithm," IEEE Access, vol. 7,



pp. 64351–64365, 2019.

- [15] I. Abrar, Z. Ayub, F. Masoodi and A. M. Bamhdi, "A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 919-924.
- [16] S. Dwibedi, M. Pujari and W. Sun, "A Comparative Study on Contemporary Intrusion Detection Datasets for Machine Learning Research," 2020 IEEE International Conference on Intelligence and Security Informatics (ISI), 2020, pp. 1-6.
- [17] A. Ali et al., "Network Intrusion Detection Leveraging Machine Learning and Feature Selection," 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), 2020, pp. 49-53.
- [18] F. Z. Belgrana, N. Benamrane, M. A. Hamaida, A. Mohamed Chaabani and A. Taleb-Ahmed, "Network Intrusion Detection System Using Neural Network and Condensed Nearest Neighbors with Selection of NSL-KDD Influencing Features," 2020 IEEE International Conference on Internet of Things and Intelligence System (IoT&IS), 2021, pp. 23-29.
- [19] W. Wang, X. Du, D. Shan, R. Qin and N. Wang, "Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine," in IEEE Transactions on Cloud Computing, 2020.
- [20] S. Subbiah, K. S. M. Anbananthen, S. Thangaraj, S. Kannan and D. Chelliah, "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," in Journal of Communications and Networks, vol. 24, no. 2, pp. 264-273, April 2022.
- [21] X. Zhang, J. Ran and J. Mi, "An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 456-460.

# Chapter - 8

## Disease Prediction using Deep Learning Algorithms in Healthcare Sector

Aman<sup>1</sup>, Rajender Singh Chhillar<sup>2</sup>

<sup>1</sup> Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India

<sup>2</sup> Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India

Email: <sup>1</sup> [sei@live.in](mailto:sei@live.in), <sup>2</sup> [chhillar02@gmail.com](mailto:chhillar02@gmail.com)

*Abstract— Deep Learning (DL) is a major focus of discussion in the healthcare sector. The healthcare sector in the United States generates around one trillion Gigabytes of clinical data annually. With limited resources, manually analyzing these massive amounts of data is a tremendously time-consuming. Finding useful patterns and acquiring knowledge from high-dimensional, poorly annotated, heterogeneous and complex clinical data continues to be a significant challenge in the health care sector. Latest advancements in DL have been shown as an efficacious approach to building end-to-end learning models for disease prognosis and diagnosis. In the past, discovering information from data has been accomplished through the use of conventional Machine Learning (ML) techniques. These techniques first require optimal features to be extracted from clinical data before building a disease predictive model on top of them. Problems with these techniques are that they do not scale properly with the increase in data due to a lack of domain knowledge. Firstly, this chapter explores popular DL algorithms for various types of clinical data. These algorithms can potentially prevent infectious disease, reducing operating costs, and efforts. Finally, the challenges while designing and implementing a holistic DL model have been discussed for disease prediction.*

*Keywords— Deep Learning (DL), Disease diagnosis, Deep Belief Networks (DBNs), Deep Convolutional Neural Networks (DCNNs), Recurrent Neural Networks (RNNs)*

### I. INTRODUCTION

With the latest advances in healthcare technology like Artificial intelligence (AI), Blockchain, and cloud computing, new methods for prognosis and diagnosis of diseases enter our daily practice. According to a report [1], from 2019 to 2021, nearly 86% of the health institutions have implemented at least an elementary Electronic Health Record (EHRs) framework. These frameworks store medical reports, medical imaging, prescription written by doctors/specialists, genomics, etc. To summarize, acquiring all important data from healthcare has become critical.

Deep learning (DL) is a subclass of ML, and its structure consists of multiple layers used to extract high-level information from input. These layers convert the incoming data into visuals and output the result by detecting the disease. These layers change incoming data using non-linear functions before sending it on. First and final layers are input and output, while intermediary layers are concealed. Three layers are considered DL (including input and output). Each node level in a DL network trains different resource sets based on the previous level's output. The neural network perceives more complex resources as it traverses as the nodes merge and recombine the upper-level resources. DL layers include input, convolution, fully connected, sequence, activation, normalization dropout, cropping, pooling and un-pooling, combination, object identification, GAN, and output. These images are processed by DL layers, which are used to detect various diseases. DL, in some ways, mimics AI and how the human brain works in data processing to create a decision-making pattern. Neural networks, Convolutional Neural Networks (CNN), and Artificial Neural Networks (ANN) are examples of DL models that are considered to be the first steps in the process of system automation. CNNs are the foundation of DL; research in this field began in the late 1970s, with the first application in medical informatics occurring in 1995. However, these were considered the initial achievements of CNNs, and CNNs did not mature until powerful new techniques for training deep networks were developed. ImageNet, also known as Alex Net in 2012, was a watershed moment in the history of CNNs. Several improvements were made in the following years to reach the maturity level. DL applications have grown in popularity due to their rapid growth. It is time to search for each disease's diagnosis separately in the vast field of medical analysis. This void can be filled by combining diagnostic techniques for common diseases in a single study. In recent years, much work has been presented as a clear summary. The survey classifies healthcare techniques according to human body systems and focuses on AI methods that have been implemented. Computer-aided design (CAD) systems are classified according to diseases that affect the three systems of the human body.

---

© 2022 Technoarete Publishing

Aman – “Disease Prediction using Deep Learning Algorithms in Healthcare Sector” Pg no: 108 – 115.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch008>

DL techniques are increasingly being used to assist and comprehend healthcare-related issues. The "vulnerabilities" of the American Recovery and Reinvestment Act (ARRA) have resulted in a steady increase in public health data. Lung cancer is common in men and women. Lung cancer causes 27% of cancer fatalities. If Lung nodules are examined early, it helps in the patients' survival rates. Parkinson's disease is the usual form of dementia, a neurological disorder and deteriorating brain disease, which disturbs problem-solving capabilities, memory issues, and physical activities and affects other necessary daily life activities. According to [2], 47 million persons have dementia. 2030 and 2050 will see population growth. Nearly 8% of women develop breast cancer in their lifetime. Digital mammography detects breast cancer. Dense imaging has some limitations; ultrasound imaging (US) is an alternative. Depression is on the rise worldwide; a thorough survey was conducted to determine the prevalence and risk factors for depression. Depression affects one out of every 13 people worldwide, according to the World Health Organization (WHO), DL methods can be used to determine the causes of depression and its symptoms. Myocardial ischemia, a well-known heart disease, is caused by a decrease in blood supply to the myocardium, which alters the morphology of electrocardiogram (ECG) signals. Cardiac arrhythmia causes chest discomfort, cardiac arrest, and abrupt death. Physiological signals identify and cure illness. Heart-related diseases and their diagnosis using ECG are critical in medical applications. Every heartbeat in the ECG waveform represents a time series of the electrical activity of the heart. Any ambiguity in the heartbeat rate or variation in the morphological pattern is a symptom of an arrhythmia, which can be detected by examining an ECG waveform. Speech is the primary mode of communication; Gaussian Mixture Models are the DL methods used for speech recognition. The sound waves were represented by Gaussian models in the Hidden Markov Models. For non-linear functions, these models are amateurish. These models are more effective for short-term signals and produce better results in deep neural networks.

## II. HEALTHCARE SECTOR

The Healthcare sector has grown as a leading sector in India for generating both employment and revenue. It includes outsourcing, health insurance, telecare, medical equipment, and health institutions. The Indian medical system is booming due to improved coverage, services, and governmental and private spending. The burden of data collection and the difficulties of calculation have historically limited predictive patient analysis. ML eliminates the limitations of human data collecting and calculation. Using these strong algorithms, one may imagine individualized treatment decisions and improved outcomes[3], [4].

According to a report of (the Ministry of Finance, 2022) on the Economic survey of India for 2022, In 2021-22, government spending on healthcare was 2.1 percent of Gross Domestic Product (GDP), up from 1.8 percent in 2020-21 and 1.3 percent in 2019-20. The e-health market is expected to be worth \$10.6 billion by 2025 [6]. As per the information provided by the health & family welfare ministry, the doctor to patient ratio is 1:854 [6]. The Indian healthcare industry employs 4.7 million people and creates more than 500,000 jobs each year, making it one of India's major employers. India is on the verge of an exciting digital healthcare transformation. ML and other innovative technologies, such as automation and other AI techniques such as Natural Language Processing (NLP), are rapidly gaining traction—especially with 5G on the horizon. With a growing population to fill new roles, the country has a thriving startup ecosystem and established health-tech companies. Healthcare providers are becoming more aware of technologically enabled ways to accomplish more with less manual effort. The government has increased spending on evolving healthcare delivery, and the public is supportive.

The healthcare system is being inundated with previously unheard-of amounts of complex data derived from physicians' notes, medical devices, labs, and other sources. Remote patient wearables are adding to the deluge. EHRs help with information digitization, but their role is not to reduce administrative workload or to provide decision support at a glance [7]–[9].

All the data that comes in is only as valuable as the insights that can be gleaned from it and applied correctly and quickly to improve healthcare delivery. ML can aid this process, particularly for digital data sets with distinct patterns. Data from various sources is collected and combined by ML. It can perform the complex computations required by doctors, nurses, and other healthcare team members to quickly interpret raw physiological, behavioral, and imaging data. By using algorithms to gather insights, ML reduces the workload of physicians, radiologists, pathologists, and other providers. Automated workflows based on how healthcare teams work frequently used in tandem for easy information sharing and collaboration.

Due to the Covid-19 pandemic, the government has prioritized investing in India's healthcare infrastructure since 2020. This has also allowed technology companies to enter the healthcare sector and innovate to contribute to improving healthcare facilities in the country. Part of the Digital India Initiative, the government launched Ayushman Bharat Digital Mission (ABDM) [10]. The idea creates digital health records that people and families may exchange online. Under this objective, citizens will be issued a randomly generated 14-digit number to uniquely identify, authenticate, and link their health information across numerous systems and stakeholders. Furthermore, inclusion is a key principle of ABDM. ABDM's e-health system fosters primary, secondary, and tertiary care continuity. Telemedicine promotes access to health care services in distant and rural locations. India's digital health startups provide solutions for the government's push to strengthen digital healthcare

infrastructure. The Indian healthcare startup landscape goes beyond a single disease, therapy, geography, product, service, or business model. In a country where affordable healthcare is a problem, the public benefits from Digital Health's growth. ABDM unifies India's healthcare system and promotes industry innovation. It's uncertain how the legal system will perceive Digital Health, with the government and entrepreneurs worried about public interest. AI and ML have established a grip in India in the previous year and have a promising future.

### III. IMPORTANCE OF DEEP LEARNING IN HEALTHCARE SECTOR

It is critical for health issues that both models describe performance and interpretability. Doctors are not likely to employ a system that is difficult to understand. Outstanding performance is one of the hallmarks of DL models. In order to describe a stable system using this model, however, it is necessary to explain and understand the conclusions acquired using this model. For this problem, the algorithm is required to describe the DL models and methodologies used to assist the existing tools in explaining a particular data-driven system. DL models' computer power has streamlined hospital processes. DL networks transform patient care and health systems' clinical practice. Computer vision, NLP, and reinforcement learning are widely employed in healthcare.

*Medical imaging.* Picture recognition and object detection are used in MR and CT for image segmentation, disease diagnosis, and prediction. DL models can effectively interpret imaging data by combining tissue size, volume, and shape. These models can highlight critical areas in images. DL systems detect diabetic retinopathy, Parkinson's, and breast nodules. DL improvements will allow future analysis of most pathology and radiology pictures. DL techniques simplify complex data analysis, allowing irregularities to be discovered. CNNs let doctors detect health concerns more quickly and correctly. CNNs identified melanoma in dermatology photos more accurately than professionals, according to a 2018 study.

*Healthcare data analytics.* DL models can assess structured and unstructured EHR data, such as clinical notes, lab test results, diagnoses, and prescriptions, quickly and accurately. Smartphones and wearables offer lifestyle data [11]. DL mobile apps can alter medical data by monitoring risk variables. FDA authorized Current Health's AI wearable gadget for home use in 2019.

*Mental health chatbots.* Happify, Moodkit, Woebot, and Wysa are AI mental health apps (including chatbots). Some chatbots can employ DL for more realistic patient conversations. According to a Stanford University study, an intelligent conversational agent can reduce depression and anxiety in students and is an entertaining approach to deliver mental health care [12], [13].

*Personalized medical treatments.* By analyzing patients' medical histories, symptoms, and tests, DL solutions enable healthcare organizations to provide personalized patient care. NLP extracts patterns from raw medical information to recommend the most appropriate medical treatments [14].

*Prescription audit.* DL models can compare prescriptions to patient health records to detect and correct potential diagnostic or prescription errors [12].

*Underwriting.* Insurance businesses utilize DL models to make customer offerings. Learn how AI is improving underwriting in our article.

*Fraud detection.* DL systems detect medical insurance fraud by studying deceitful actions from EHRs like claims history, hospital-related information, and patient features [15].

*Drug discovery.* Technological advances have increased DL models' contributions to drug discovery and interaction prediction. DL systems evaluate genomic, clinical, and population data to find feasible drug combinations. Pharmaceutical researchers employ DL toolkits to analyze big data sets [16].

*Genomics analysis.* DL models increase biodata interpretation. DL models' complicated data analysis helps scientists comprehend genetic variation and build genome-based therapies. CNNs are commonly used to extract DNA sequence characteristics [17], [18].

*Mental health research.* DL models improve mental health clinical practice. Deep neural networks are used to study mental illness and other brain problems. DL models beat ordinary ML models, say researchers. DL can discover brain biomarkers.

*Covid-19.* With COVID-19, DL models are more important. Researchers are researching DL applications for early diagnosis of Covid-19 by evaluating Chest X-ray (CXR) Chest CT images to predict intensive care unit admission, identify high-risk patients, and estimate mechanical ventilation needs [19].

### IV. VARIOUS TYPES OF CLINICAL DATA

*Electronic Health Records (EHRs).* Real-time patient-centered records are made possible by EHRs. DL has recently been utilized to handle aggregated EHRs, including structured and unstructured data (e.g., clinical notes). Most of this material focused on processing EHRs for a specific, supervised, predictive clinical task. DL surpasses typical ML models in ROC curve, accuracy, and F-score. EHRs include patients' medical and treatment histories, but they go beyond traditional clinical data gathered in a provider's office to provide a wider perspective of a patient's care. It helps schedule and manage clinician

workflow. These systems help with clinical, financial, and administrative coding. Service requests and refund claims are supported [20].

*Clinical imaging.* First uses of DL to clinical data focused on image processing, namely brain MRI scans to predict Parkinson's disease and its variants. CNNs have been utilized in other medical domains to segment cartilage and predict osteoarthritis risk from low-field knee MRI data. This strategy outperformed one that used 3D multi-scale features manually selected. DL was also utilized to segment MS lesions in 3D MRI and identify benign from malignant breast lumps in ultrasound pictures [21], [22].

*Genomics.* DL captures data set structure in high-throughput biology (e.g., DNA sequencing, RNA measurements). Deep models allow the discovery of high-level characteristics, which improves performance, interpretability, and understanding into biological data structure. In the literature, various works have been proposed [23]. Here we review the broad concepts and refer the reader for more in-depth reviews. First neural network applications in genomics superseded traditional ML with deep architectures. Sparse AEs can categorize cancer cases based on gene expression profiles or predict protein backbones. Deep neural networks also improved the genomic drug discovery pipeline.

*Mobile.* Sensor-enabled cellphones and smartwatches are redefining mobile applications like patient monitoring. These devices might allow patients immediate access to personal data to improve health, promote preventative treatment, and manage illness. DL is crucial for interpreting this new data. Few recent health care sensing works have used deep models due to technology limitations. Running a deep architecture on a mobile device to handle noisy and sophisticated sensor data is difficult and likely to exhaust device resources [24]–[27].

## V. DEEP LEARNING (DL)

ML is a general-purpose AI method that can learn relationships from data without having to define them beforehand. The capacity to develop predictive models without making significant assumptions about the underlying systems is appealing. ML pipeline includes data cleansing, reinforcement learning, model fitting, and assessment [28]. Over years, constructing a ML model needed meticulous engineering and domain knowledge to translate raw data into a schema from which a classifier could find patterns in data. Conventional approaches can't process raw data since they require a single, frequently linear, input space transformation. DL teaches representations from raw data differently than typical ML. DL allows neural network-based computational models to learn multi-level data representations [29]. DL differs from standard ANNs in its hidden layers, connections, and capacity to learn meaningful abstractions of inputs. Traditional ANNs have three layers and are trained to provide supervised representations optimal for the particular task at hand and not generalizable. Each layer of a DL system optimizes a local unsupervised criterion to represent observable patterns based on data from the layer below. DL layers are learnt from data, not developed by engineers. Deep neural networks analyze inputs layer-by-layer to pre-train (initialize) hidden nodes to learn generalizable 'deep structures and representations. These representations are supplied into a supervised layer, which utilizes the backpropagation technique to fine-tune the entire network. Deep models have become state-of-the-art owing to breakthroughs in unsupervised pre-training, innovative ways to minimize overfitting, the use of GPUs to speed up calculations, and the creation of high-level modules to quickly assemble neural networks (e.g., Theano, Caffe, TensorFlow). DL has shown useful for uncovering detailed patterns in high-dimensional data and for object identification in photos, audio recognition, and NLP. Clinical findings successes in health care (e.g., detecting retinopathy in retinal images, classifying breast cancer, and predicting DNA and RNA-binding protein sequence specificities) open the way for a new concept of smart DL tools for real clinical care.

### 5.1 ARTIFICIAL NEURAL NETWORKS (ANNs)

ANNs are brain-based electrical models. Brains learn through experience. Small energy-efficient packages can handle challenges beyond the capabilities of today's computers. Brain modelling could make machine solution development less complex. Computers excel at routine jobs like maintaining ledgers and doing calculations. ANNs can analyze data in parallel, thus they can tackle several tasks at once. Resistant. Losing one or more cells or neural networks affects ANN performance. It's being progressively dismantled, so they won't abruptly cease working. Computers have trouble understanding basic patterns, much alone generalizing prior behaviors. Biological advances offer an understanding of natural thought. This study demonstrates that brains store information as patterns. Some of these patterns are extremely complex, allowing us to recognize individual faces from a variety of perspectives. Using patterns to solve issues is a new discipline of computing [30]. ANNs rely on parallel processing, hence they require processors that support it. Since it's related to human brain functions, we may not know the correct ANNs network topology. ANNs and statistical models can be trained with only numeric data, so it's hard for them to understand the problem statement. When an ANN gives a solution to a problem statement for which we don't know the foundation, the ANN is not dependable. ANN are computers with architectures that are modeled after the brain. They have hundreds of basic processing units connected by a complicated network. Each node is a schematic version of a biological



neuron that activates when it receives a strong signal from other nodes (see Figure 1 **Error! Reference source not found.**). Equation 1 shows output  $\hat{z}$  of a node in ANNs which includes bias  $\hat{b}$  and summation of input  $p_i$  and their respective weight  $q_i$ . When discussing neural networks, we should more commonly refer to them as ANNs. ANNs are computers with architectures inspired by the human brain. They are typically composed of hundreds of simple processing units linked together by a complex communication network. NLP, image recognition, and optimization techniques employ ANNs. These apps are produced by adjusting neurons' synaptic connections, analogous to biological learning.

$$\hat{z} = f\left(\hat{b} + \sum_{i=1}^n p_i q_i\right) \quad (1)$$

### 5.2 DEEP BELIEF NETWORKS (DBNs)

DBNs can learn high-dimensional data manifolds. DBNs are multilayer neural networks with both directed and undirected connections. The top two layers' connections are undirected, while all other layers' connections are directed. DBNs can be thought of as a stack of greedily trained Restricted Boltzman Machines (RBMs). RBM layers communicate with each other and with previous and subsequent layers (See Figure 2 **Error! Reference source not found.**). A feed-forward network and several layers of RBM serve as feature extractors in this model. An RBM has only two layers: a hidden layer and a visible layer [31]. Figure 6 depicts the architecture of the DBN methodology that has been adopted from, where (v) is the deep belief model's stochastic visible variable. DBNs is generative models which use a sophisticated network of Restricted Boltzmann machines (RBM) [32]. Each RBM model transforms its input vectors nonlinearly (similar to a neural network) and outputs vectors for the next model in the sequence. This gives DBNs a lot of flexibility and makes them easier to scale. Because DBNs are generative models, they can be used in both unsupervised and supervised settings. DBNs can discover and classify features, which is important in many applications. In particular, in feature learning, we perform unsupervised layer-by-layer pre-training on the various RBMs that comprise a DBN, and We fine-tune classification and other tasks using backpropagation (gradient descent) on a small labelled dataset.

### 5.3 DEEP CONVOLUTIONAL NEURAL NETWORKS (DCNNs)

DCNNs recognize patterns in pictures and video. DCNNs use a three-dimensional neural pattern inspired by animal visual brain [33], [34]. Multidimensionally correlated data have not yet been used to train neural network models. DCNNs layer well. It processes red, green, and blue aspects of a picture simultaneously. It trains classifiers using picture inputs. Instead of matrix multiplication, the network uses "convolution." It consists of convolution, pooling, activation layers (See Figure 3 **Error! Reference source not found.**). The ReLU function ( $R$ ) outputs the input straight if it's positive and zero otherwise (See Equation 2). Softmax activation function ( $S$ ) converts  $n$  real values into  $n$  real values that total to 1. Softmax turns positive, negative, zero, and greater-than-one input values into probabilities between 0 and 1 (See Equation 3).

$$R(p) = \max(0, p) \quad (2)$$

$$S(p_i) = \frac{e^{p_i}}{\sum_{i=1}^n e^{p_i}} \quad (3)$$

### 5.4 RECURRENT NEURAL NETWORKS (RNNs)

It is possible to use a RNNs to recognize patterns in streamed or sequential input such as speech, handwriting, and text. With respect to its internal architecture [35], RNNs has cyclic connections **Error! Reference source not found.**. Using these cyclic connections of hidden units, the input data is processed sequentially. A state vector is maintained in hidden units for each of the preceding input data, and this state vector is used to compute the outputs. As a result, RNNs compute a new output based on the current input and the prior input. Despite RNNs' promising performance, the approach is doomed by the vanishing gradient problem during data training. LSTM networks and Gated recurrent unit (GRUs) are two viable alternatives that holds data for a long time. Figure 4 depicts the RNNs' design. Despite encouraging results, the usefulness of GRU in addressing the vanishing gradient problem is heavily reliant on the input data and the complexity of the challenge.

## VI. KEY CHALLENGES FACED WHILE DESIGNING AND IMPLEMENTING THE DEEP LEARNING MODEL

DL in healthcare has both obstacles and potential, as we'll explore in this section. Despite strong results, several difficulties remain. Check Figure 5 for step included in workflow while designing and implementing the deep learning model.

### 6.1 BAD DATA

ML relies on data. ML experts often lack high-quality data. Unclean, noisy data may be tiring. Data collection is the first step in the process. DL necessitates a set of computational models that are exceedingly demanding. The classic example is a fully



linked multilayer neural network. A huge number of network parameters must be accurately estimated by the network. The proposed objective can only be achieved if a large amount of data is available. In general, there aren't any rules dictating how many data sets should be used. DL is more successful when there is a large amount of data accessible, as a general rule is to have at least 10x the number of parameters in the network. Despite the fact that healthcare is a different field, there is no standard method for collecting data. It is also difficult to obtain high-quality data and annotate a large number of subsequent publications using the old dataset. China physiological signal challenge 12 lead 2018 used high-quality data, while physio Net computing in cardiology challenge 2017 used high-quality data. Short-term ECG recordings are still the emphasis. High-quality ECG datasets with observations are available in the long term, which could spur new studies. Deleting outliers, filtering missing results, and deleting unnecessary characteristics must be done perfectly.

### 6.2 UNDERFITTING AND OVERFITTING ON LEARNING DATA

Overfitting is when a ML is trained with too much data, reducing its accuracy. As compared to other domains, healthcare, and biomedical concerns are more complex. Underfitting happens when input and output variables aren't accurately correlated. There are a wide variety of diseases, yet we still know very little about their causes, how they spread, and how to heal them. As a result of a shortage of patients, we are unable to collect sufficient datasets in some cases. However, in the future, this problem can be overcome. Every patient's possible information should be gathered and processed in a novel way so that these data sources can be included in our analysis.

### 6.3 SLOW ADOPTION

ML practitioners often confront this difficulty. ML models are accurate yet time-consuming. Slow programmers, data overload, and excessive requirements delay proper outcomes. The best production demands continual monitoring and maintenance. The best production demands continual monitoring and maintenance.

### 6.4 TUNING OR OPTIMIZING DL MODEL

Black-box models describe DL algorithms. Models with multiple parameters or complex architectures are common. The results of the generation of such models are difficult to grasp. The quantitative algorithm's performance in the medical and medical Area is critical, as is the algorithm's function and how it works. The dilemma is made considerably worse in the medical community by the fact that medical professionals will not accept diagnoses that cannot be explained.

## VII. CONCLUSIONS

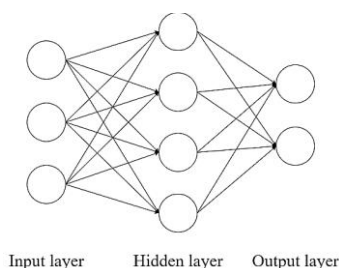
This chapter covers machine learning and deep learning in healthcare. Later, we discussed how healthcare may use deep learning to find patterns in EHRs, Imaging data, Genomics, and Mobile data. We have covered ANNs, DBNs, DCNNs, and RNNs. We have covered the issues programmers/researchers encounter while creating and deploying a deep learning model in healthcare.

## REFERENCES

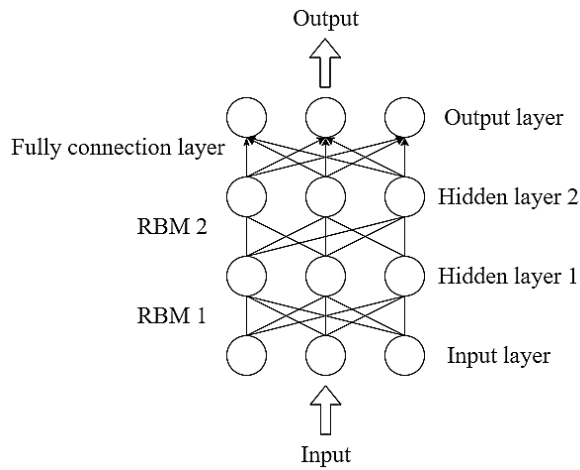
- [1] Office of the National Coordinator for Health Information Technology., "Adoption of Electronic Health Records by Hospital Service Type 2019-2021 | HealthIT.gov," Apr. 2022. <https://www.healthit.gov/data/quickstats/adoption-electronic-health-records-hospital-service-type-2019-2021> (accessed Jul. 12, 2022).
- [2] "World Alzheimer Report 2016 | Alzheimer's Disease International (ADI)," 2016. <https://www.alzint.org/resource/world-alzheimer-report-2016/> (accessed Jul. 20, 2022).
- [3] Aman and R. S. Chhillar, "Disease predictive models for healthcare by using data mining techniques: State of the art," *SSRG Int. J. Eng. Trends Technol.*, vol. 68, no. 10, Art. no. 10, 2020, doi: 10.14445/22315381/IJETT-V68I10P209.
- [4] Aman and R. S. Chhillar, "Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease using WEKA Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, Art. no. 8, Oct. 2021.
- [5] Ministry of Finance, "Economic Survey," 2022. <https://www.indiabudget.gov.in/economicsurvey/doc/eschapter/echap09.pdf> (accessed Jul. 14, 2022).
- [6] "Indian Healthcare Industry Analysis | IBEF," *India Brand Equity Foundation*, 2022. <https://www.ibef.org/industry/healthcare-presentation> (accessed Jul. 14, 2022).
- [7] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
- [8] Y. Kumar, S. Gupta, and A. Gupta, "Study of Machine and Deep Learning Classifications for IOT Enabled Healthcare Devices," in *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, Nov. 2021, pp. 212–217. doi: 10.1109/ICTAI53825.2021.9673437.
- [9] F. H. Masmali, S. J. Miah, and N. Y. Mathkoo, "Internet of Things-based innovations in Saudi healthcare sector: A methodological approach for investigating adoption issues," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Dec. 2020, pp. 1–5. doi: 10.1109/CSDE50874.2020.9411588.
- [10] "Update on Ayushman Bharat Digital Mission," 2022. <https://pib.gov.in/pib.gov.in/Pressreleaseshare.aspx?PRID=1813660> (accessed Jul. 20, 2022).
- [11] K. S. P. A. M. V., A. N. S. A. B., and K. K. R., "Smart Health Monitoring System Using ANN Algorithm," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Jul. 2021, pp. 1–5. doi: 10.1109/ICCES51350.2021.9489239.
- [12] S. Revathy, Niranjani. R., and R. Kanushya. J., "Health Care Counselling Via Voicebot Using Multinomial Naive Bayes Algorithm," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Jun. 2020, pp. 1063–1067. doi: 10.1109/ICCES48766.2020.9137948.

- [13] N. M. Sharef, M. A. A. Murad, E. I. Mansor, N. A. Nasharuddin, M. K. Omar, and F. Z. Rokhani, "Personalized Learning Based on Learning Analytics and Chatbot," in *2021 1st Conference on Online Teaching for Mobile Education (OT4ME)*, Nov. 2021, pp. 35–41. doi: 10.1109/OT4ME53559.2021.9638893.
- [14] L. B. Rebelo dos Santos, M. dos Santos Silvério, C. de Castro Mario, C. Guellner Ghedini, and R. J. Soares, "A system to Support the Physiotherapeutic Treatment of Chronic Pain in the Spine," in *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, Jun. 2021, pp. 1–7. doi: 10.23919/CISTI52073.2021.9476549.
- [15] W. Yang, W. Hu, Y. Liu, Y. Huang, X. Liu, and S. Zhang, "Research on Bootstrapping Algorithm for Health Insurance Data Fraud Detection Based on Decision Tree," in *2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, May 2021, pp. 57–62. doi: 10.1109/BigDataSecurityHPSCIDS52275.2021.00021.
- [16] R. Biswas, A. Basu, A. Nandy, A. Deb, K. Haque, and D. Chanda, "Drug Discovery and Drug Identification using AI," in *2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, Feb. 2020, pp. 49–51. doi: 10.1109/Indo-TaiwanICAN48429.2020.9181309.
- [17] P. Thareja and R. S. Chhillar, "Comparative Analysis of Data Mining Algorithms for Cancer Gene Expression Data," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 12, no. 10, Art. no. 10, 54/31 2021, doi: 10.14569/IJACSA.2021.0121035.
- [18] P. Thareja and R. S. Chhillar, "A Detailed Survey on Data Mining Based Optimization Schemes for Bioinformatics Applications," *ECS Trans.*, vol. 107, no. 1, pp. 4689–4696, Apr. 2022, doi: 10.1149/10701.4689ecst.
- [19] A. Mehla and S. Singh, "Deep Learning based Diagnosis of Chest X-rays for Multiple Diseases of Lungs: A Review," *J. Orient. Res. Madras*, pp. 35–46, 2021.
- [20] P. Thareja and R. S. Chhillar, "A review of data mining optimization techniques for bioinformatics applications," *Int. J. Eng. Trends Technol.*, vol. 68, no. 10, Art. no. 10, Oct. 2020, doi: 10.14445/22315381/IJETT-V68I10P210.
- [21] N. Li *et al.*, "Simultaneous Head and Spine MR Imaging in Children Using a Dedicated Multichannel Receiver System at 3T," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3659–3670, Dec. 2021, doi: 10.1109/TBME.2021.3082149.
- [22] V. Solanki, "Brain MRI Image Classification using Image Mining Algorithms," *2018 Second Int. Conf. Comput. Methodol. Commun. ICCMC*, no. Iccmc, Art. no. Iccmc, 2018, doi: 10.1109/ICCMC.2018.8487690.
- [23] A. Darolia and R. S. Chhillar, "Analyzing Three Predictive Algorithms for Diabetes Mellitus Against the Pima Indians Dataset," *ECS Trans*, vol. 107, no. 1, p. 2697, 2022, doi: <https://doi.org/10.1149/10701.2697ecst>.
- [24] T. Javid, M. Faris, H. Beenish, and M. Fahad, "Cybersecurity and Data Privacy in the Cloudlet for Preliminary Healthcare Big Data Analytics," in *2020 International Conference on Computing and Information Technology (ICCI-1441)*, Sep. 2020, pp. 1–4. doi: 10.1109/ICCI-144147971.2020.9213712.
- [25] H. Saini and G. Singh, "Analysis and Comparison of Various Privacy Preservation Mechanisms for Cloud Network," in *Multidisciplinary Subjects for Research*, vol. 2, Redshine Publisher, 2021, pp. 120–125. Accessed: Jul. 20, 2022. [Online]. Available: <https://redshine.co.in/product/979-8-7188-7982-7/>
- [26] H. Saini and G. Singh, "A Study of Challenges and Methods of Privacy Preservation in Social Networking Data," *EPRA Int. J. Multidiscip. Res.*, vol. 7, no. 1, pp. 411–414, 2021.
- [27] F. M. J. M. Shamrat, P. Ghosh, M. Hasan, S. Shomi, M. Fradulent, and A. Deteciton, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," no. December, Art. no. December, 2020, doi: 10.1109/INOCON50539.2020.9298026.
- [28] D. Bordoloi, V. Singh, S. Sanober, S. M. Buhari, J. A. Ujjan, and R. Boddu, "Deep Learning in Healthcare System for Quality of Service," *J. Healthc. Eng.*, vol. 2022, p. e8169203, Mar. 2022, doi: 10.1155/2022/8169203.
- [29] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi: 10.1093/bib/bbx044.
- [30] M. Akgül, Ö. E. Sönmez, and T. Özcan, "Diagnosis of heart disease using an intelligent method: A hybrid ANN – GA approach," in *Advances in Intelligent Systems and Computing*, Jul. 2020, vol. 1029, pp. 1250–1257. doi: 10.1007/978-3-030-23756-1\_147.
- [31] C. Gong, "Mathematical Evaluation Model and Intelligent Prediction Research about Health Status Based on SSA-DBN," in *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Apr. 2022, pp. 610–613. doi: 10.1109/IPEC54454.2022.9777356.
- [32] I. A. Sattar, R. S. Alhamdani, and M. N. Abdulah, "Utilizing latent Features for Building Recommender system Based on RBM Neural Network," in *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, Apr. 2021, pp. 281–286. doi: 10.1109/BICITS51482.2021.9509886.
- [33] J. C. Alcaraz, S. Moghaddammia, M. Penner, and J. Peissig, "Monitoring the Rehabilitation Progress Using a DCNN and Kinematic Data for Digital Healthcare," in *2020 28th European Signal Processing Conference (EUSIPCO)*, Jan. 2021, pp. 1333–1337. doi: 10.23919/Eusipco47968.2020.9287324.
- [34] P. Pal and M. Mahadevappa, "Adaptive Multi-Dimensional dual attentive DCNN for detecting Cardiac Morbidities Using Fused ECG-PPG Signals," *IEEE Trans. Artif. Intell.*, pp. 1–10, 2022, doi: 10.1109/TAI.2022.3184656.
- [35] N.-Y. Tung *et al.*, "Numerical prediction for Systolic Blood Pressure in Intradialytic Hypotension Using Time-relevant RNN Models," in *2021 IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, May 2021, pp. 57–59. doi: 10.1109/ECBIOS51820.2021.9510228.

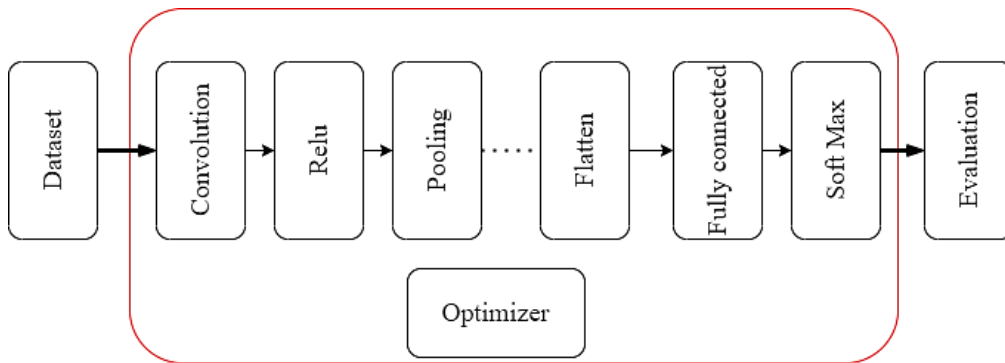
## APPENDIX



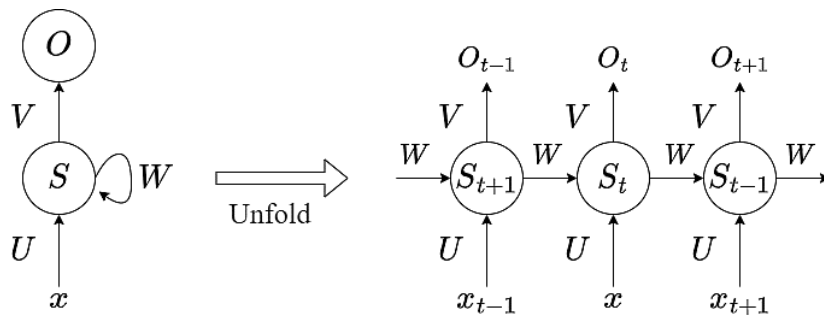
**Figure 1:** Interconnection between different layers of ANNs



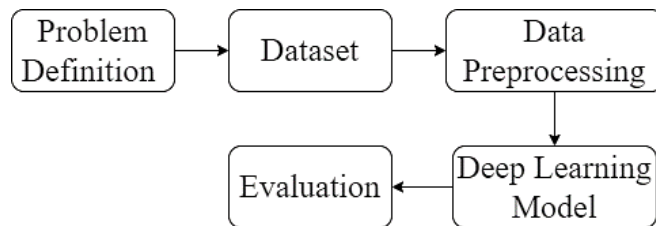
**Figure 2:** Stacked RBMs forming DBN



**Figure 3:** Schematic diagram of DCNNs



**Figure 4:** Unfolding of a state of RNNs



**Figure 5:** Workflow in designing and implementing a deep learning model

## Applications of Deep Learning Models in Bioinformatics

Preeti Thareja<sup>1</sup>, Rajender Singh Chhillar<sup>2</sup>

<sup>1</sup> Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India

<sup>2</sup> Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India

Email: <sup>1</sup> [preetithareja10@gmail.com](mailto:preetithareja10@gmail.com), <sup>2</sup> [chhillar02@gmail.com](mailto:chhillar02@gmail.com)

*Abstract—Deep learning (DL) models have had an influence on machine learning-based in bioinformatics applications since they allow for the learning of complicated non-linear interactions between functionalities. Deep learning models also enable information utilized from large unlabeled data that is unrelated to the problem under investigation. Protein-protein interactions (PPIs) are important in a variety of biological functions, including cell signaling, immune function, and cellular organization. PPIs analysis is thus vital, as it may spotlight the detection of targeted proteins and their role in the disease and thus help in designing treatments for it. PPIs play critical roles in life processes, and abnormal interactions are linked to a variety of disorders. Identification of interaction sites is critical for understanding disease mechanisms and designing new drugs. Because of the overall cost of experimental methods, effective and efficient computational methods for PPI prediction are extremely valuable. Machine learning and deep learning techniques have produced remarkable results, but their efficacy is highly reliant on protein interpretation and feature extraction. This chapter will explain various deep learning models that can be used in Bioinformatics as well as the challenges they face.*

*Keywords—Deep Learning, Bioinformatics, PPIs, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Deep Belief Network (DBN)*

### I. INTRODUCTION

Deep learning has recently become one of the most effective machine learning algorithms thanks to the considerable advancements in big data and computer capacity. It has been consistently improving the performance of many DL jobs at the cutting edge and promoting the growth of numerous fields. Convolutional neural network-based techniques, for instance, already predominate in the three main areas of computer vision, such as image identification, object detection, and picture inpainting and super-resolution. Recurrent neural network-based techniques often offer the most advanced performance in the field of natural language processing for a variety of applications, including text categorization, speech recognition, and machine translation. Additionally, DL has proven to be very successful in accelerating the field of bioinformatics, primarily in the fields of protein sequences, hierarchy prediction and restructuring, biomedical property and feature prediction, biomedicine image analysis and prognosis, and biomaterials interaction forecasting and cell biology [1]. Use DL to predict protein contact maps and the structure of membrane proteins; use DL to simulate the secondary structure of proteins when they interact with other molecules, and Bayesian inference using DL to speed up fluorescence microscopy super-resolution. DL is used to predict protein Gene Ontology (GO), protein subcellular location, and enzyme detailed function concerning the biomolecular property and function prediction. Peptide intracellular localization can also be predicted using DL [2].

The primary factor in deep learning's success in bioinformatics, in addition to rising processing power and enhanced methods, is the data [3]. DL is particularly well suited for biological analysis due to the massive amount of data being generated in the biological area, which was originally regarded to be a huge challenge [4]. DL has particularly demonstrated its advantage when handling the following biological data types. First off, sequence data such as DNA, RNA, protein, and Nanopore signal have all been successfully handled by deep learning. DL is an expert at finding hidden motifs, patterns, and domains in sequence data because it is trained through backpropagation and stochastic gradient descent. RNNs and CNNs with 1-dimensional filters are capable of handling this type of data. CNNs are typically the best option for biological sequence data if one wishes to figure out the hidden patterns revealed by the neural network, however, it is difficult to explain and display the trend revealed by recurrent neural networks. Second, DL is particularly effective in processing 2D and tensor-like data, such as gene expression profiles and biological pictures [5].

When handling biomedical data, standard convolutional neural networks and their variations, such as residual networks, densely connected networks, and dual route networks, have demonstrated excellent performance. These networks can autonomously convert the initial input to a chosen concealed domain, where the high-level description is highly instructive and suitable for supervised methods, and then methodically investigate the patterns concealed in the initial map at various scales. Thirdly, DL may be utilized to handle graph data, including symptom-disease networks, gene co-expression networks, protein-protein interaction networks, and cell system hierarchy, and it can be used to get cutting-edge results [6].

In this chapter, we will present a comprehensive overview of DL models, varying from simplistic neural nets to deep learning models and their above-mentioned variations, which are appropriate for biological data analysis. This is because DL has a meaningful ability to promote bioinformatics research. Those examples will include deep belief networks (DBN), recurrent neural networks (RNN), and conventional convolutional neural networks (CNN). Before examining how effectively these techniques have been applied in bioinformatics, biomedical imaging, biomedicine, and drug discovery, we particularly explore the fundamentals of DL models. The significance of DL in bioinformatics is then examined in a variety of domains, including microarray, text mining, system biology, genomics, and proteomics. Although there are many different and varied ways that ensemble DL is used in bioinformatics, we highlight and explore the typical problems and opportunities in the context of bioinformatics research. We categorize studies both by the bioinformatics area and DL framework to present a useful and comprehensive perspective, and we include concise descriptions of each work. We also explore theoretical and practical challenges related to DL in bioinformatics and make recommendations for future research. We are certain that this study will offer insightful information and act as a springboard for scholars wishing to use DL techniques in their bioinformatics research.

## II. MODELS IN DEEP LEARNING

### 2.1 Convolutional Neural Networks

The connectivity structure of convolutional neural networks, sometimes referred to as CNNs or ConvNets, a type of feed-forward artificial neural network, is modeled after how the visual brain of animals is organized. Only certain edge orientations trigger a response or firing from a particular neuronal cell in the brain. While some neurons fire when shown edges in a vertical direction, others fire when shown edges in a horizontal or diagonal direction. Neural nets used in DL to assess visual data are called CNN [7].

Figure 1 illustrates how CNNs can learn complex objects and patterns thanks to their input layer, an output layer, several hidden layers, and millions of parameters. Before applying an activation function, it subsamples the input using convolution and pooling techniques. All of the hidden layers are partially connected in the beginning, and the output layer is the final fully connected layer. The input and the output are similar.

**Convolutional Layer:** Convolutional layers make up the majority of CNN construction components. This layer frequently includes output vectors like a feature map, filters like a feature detector, and input vectors like an image. Following the convolution operation, the input is condensed to a feature map, also termed an activation layer, as shown in equation 1.

$$feature\ map = input\ image \times feature\ detector \dots\dots\dots(1)$$

Only the receptive field to which each convolutional neuron has been assigned receives data from the neuron. These convolutional operations are used in CNN's convolutional layers to discover information and classify input. And over source pixel is where the kernel's core is located. The source pixel is then replaced by a weighting factor of the pixels around it. Local connection and parameter sharing are two aspects employed by CNN. Parameter sharing explains how all cells in a feature receive weight. Each neuron is thought to be connected to just a portion of the incoming image by a local connection. By doing so, the system's parameter count is decreased and the calculation is completed more quickly.

**Pooling:** Pooling aids in decreasing the number of variables and computations in the network by significantly lowering the dimension of the representation and minimizing overfitting. The result would have the same quality as the source if no pooling is applied. The pooling layer handles each feature map uniquely. The methods for pooling as depicted in figure 2. The feature map's most important piece is selected via max-pooling. The resulting max-pooled layer stores the significant features of the feature map. It is the most widely used approach since it yields the best results. As opposed to this, average pooling involves averaging over each area of the feature map.

**ReLU:** The rectified linear activation function (ReLU) is a simple linear function that, whenever the input is favorable, outputs the input directly, and when the input is not in favor, it outputs zero. It has evolved as the default activation function for many varieties of neural networks since a model using it trains more quickly and typically performs better.

**SoftMax Activation Function:** The final layer of the network, which acts as a classifier, is often given the activation layer called soft-max. The input that is presented is classified into many categories by this layer. The softmax process turns a network's non-normalized results into a probability distribution. The classification layer employs the softmax function, which has good multiclass performance. The layer consists of N units, where N represents the size of units. Each layer uses equation



2 to determine the accuracy of the classifier on N and is fully related to the previous layer.

$$\text{Softmax} = \frac{e^{w_{Nx} + b_N}}{\sum_{m=1}^N e^{w_{mx} + b_m}} \dots \dots \dots (2)$$

$W_m$  is the weight vector connecting the  $m^{\text{th}}$  unit to the former layer,  $x$  is the previous layer's output, and  $b_m$  is the  $m^{\text{th}}$  unit's bias.

*Role of CNN in bioinformatics.* Given CNN's abilities to analyze spatial data, it is clear that the majority of its bioinformatics research so far has been concentrated on biomedical imaging. It stands to reason that CNN is not the default for DL architecture in genomics and biomedical data processing since the common data in the area doesn't seem to include spatial features [8]. However, 2D information, such as relationships involving gene sequences and the time-frequency vector of a biomedical stimulus, can still be regarded as spatial information. As a result, we think CNN has a lot of promise in these areas and is well-positioned to have a big effect in the future.

### 2.2 Recurrent Neural Networks

RNNs are a DL method for modeling sequential information. Before the development of attentive models, RNNs was the recommended solution for managing sequential information. Each component of the sequence may require a different set of parameters for the deep feedforward model. Furthermore, it might not apply to sequences of different lengths.

Figure 3 illustrates how RNNs generalize to sequences of various lengths by using the same weights for each element in the sequence, reducing the number of parameters. Because of how they are constructed, RNNs can be applied to various types of structured data besides sequential data, like geographic or graphical data [9]. The behavior of RNNs, which are built from feedforward networks, is comparable to that of human brains. Simply put, recurrent neural networks are better than other algorithms at anticipating sequential data as can be seen in figure 4.

All of the endpoints in conventional neural nets operate independently of each other. It's crucial to recall the past words because there are instances when past words are necessary, such as when guessing the following word in a sentence. RNN was developed as a result, and it used a Hidden Layer to solve the issue. The most important RNN component is the Hidden state, which preserves precise details about a sequence. RNNs have a memory where they keep track of all the calculations data. Since it reaches a similar outcome by carrying out a similar operation on all intake or hidden nodes, so it utilizes similar parameters for each intake.

For each timestep  $t$ , the activation  $a^{<t>}$  and the output  $y^{<t>}$  is expressed by equation 3 and 4.

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \dots \dots \dots (3)$$

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y) \dots \dots \dots (4)$$

Where,  $W_{ax}$ ,  $W_{aa}$ ,  $W_{ya}$ ,  $b_a$ ,  $b_y$  are coefficients that are shared temporally and  $g_1$ ,  $g_2$  activation functions.

RNNs have a loop where the input travels through before reaching the middle-hidden layer. Before sending the input to the middle layer of the neural network, the input layer  $x$  receives and analyses it. A handful of hiding layers, with their collection of activation functions, weights, and biases, are included in the intermediate layer  $h$ . If the individual hidden layer parameters are not affected by the hidden layer before it, or if the neural network has no memory, then RNN can be used. To ensure that each layer hiding has similar characteristics, the RNN will unify the various activation functions, weights, and biases. Instead of generating numerous hidden layers, it will just generate one and cycle over it as much as is required.

*Function of RNN in bioinformatics.* In comparison to DNN and CNN, RNN in bioinformatics is still in its infancy, but its capacity to analyze sequential data raises expectations to an extremely high level. There are several domains where RNN has enormous potential, one of the areas is the analysis of dynamic CT and MRI, which consists of several consecutive pictures [10]. With RNN, biomedical text analysis, such as that of electronic medical records and research articles, will be able to advance significantly.

### 2.3 Deep Belief Network

To solve problems with traditional neural networks in deep layered networks, DBNs are developed. A few examples include the necessity for a large number of training data, slow training, and being stuck in a locally optimal solution as a result of poor parameter selection [11]. Numerous layers of unpredictable latent variables make up a DBN. Binary latent variables also referred to as feature detectors, are variables with a binary representation. A hybrid generative graphic model is DBN. There is no direction in the top two layers. Direct links to lower layers are present in the layers above. The DBN technique uses probabilistic DL without supervision. Although they are not the same, DBNs are a type of DL method that resembles DNNs. These neural networks are feedforward and have a deep architecture or numerous hidden layers as shown in figure 5.

*DBN structure:* A DBN is built by sequentially linking a set of restricted Boltzmann machines. The outcome of the Boltzmann machine's "output" layer is added as an input to the following Boltzmann machine successively. Then, as illustrated



in figure 6, it is trained until it converges and then uses the same strategy to complete the entire network. Associative memory is formed by the symmetric and undirected connections that connect the top two levels of the DBN. The relationships between the lower layers are indicated by arrows pointing in the direction of the layer closest to the data. The input Binary or real data is delivered to the lowest layer of visible units. There are no intra-connections between layers, similar to the RBM. The correlations in the data are represented by the hidden units as features. A  $W$  vector of symmetrical values links two layers together. Each layer's units will be interconnected with those in the one above them.

*Working of DBN.* Figure 7 depicts the operating flow for Deep Belief Network. To pre-train DBN, the Greedy learning method is used. These weights regulate the connection between variables in a single layer and variables in the layer over that. DBN can do countable steps of Gibbs sampling on the top two layers that are hidden. Since the RBM's top two hidden layers define it, this stage essentially consists of taking a sample from them. Then, a sample from the visible units is taken using an ancestor sampling run across the rest of the model. To deduce the values of the hidden variables in each layer, a solo bottom-up method is used. Greedy pretraining starts with an observed data vector in the lowest layer. It then adjusts the generative weights in the other direction. Using the greedy learning method, the entire DBN is trained. Till all the RBMs are taught, the greedy learning algorithm teaches one RBM at a time.

*Role of DBN in bioinformatics.* The investigation of internal correlations in high dimensional data is a popular application for DBN. Considering that bioinformatics information is frequently complicated and has high dimensional data, several studies are leveraging DBN over genomics, proteomics, biomedical imaging, and biomedical signal processing [12]. So, DBN's potential has not yet been completely realized. Although the primary idea behind DBN is that hierarchical features can be learned from data, raw data forms were frequently substituted for human-designed features as input. Researching appropriate methods to encode the raw data forms and learning appropriate features from the raw form is what is anticipated and will lead to future advancements in DBN in bioinformatics.

### III. DEEP LEARNING IN BIOINFORMATICS

The enormous increase in the amount of biological information available creates two issues: the effective management and storage of the data, and the extraction of information that may be used. One of the primary difficulties in computational biology is the second issue, which calls for the creation of tools and techniques that can convert all of these diverse data sets into biological knowledge about the underlying mechanism [13]. With the aid of these techniques and tools, we ought to be able to deliver information in the form of verifiable models rather than just a simple description of the data. We can derive system predictions by using this abstract abstraction that serves as a model.

Neural network models are used for information extraction from data in several biological disciplines. An outline of the primary biological issues being addressed by computational techniques is shown in Figure 8. These issues have been categorized into six different fields: text mining, microarrays, systems biology, evolution, genomics, and proteomics [14]. The remaining issues are gathered under the heading "other applications." Particularly genomics and proteomics, which are viewed in this review as the study of nucleotide chains and proteins, respectively, should be interpreted in a very general sense.

#### 3.1 Genomics

Genomic tracking, inheritance, and editing are the main topics of research in the crucial area of bioinformatics known as genomics [15]. An organism's genome is its whole collection of genetic material. There are three major divisions within genomics as follows:

1. Regulatory genomics: It involves an analysis of genomic expression regulation. Producing RNA-binding proteins and transcription factors, as well as predicting and categorizing gene expression, are examples of neural network models in this area of genomics.
2. Structural genomics: It uses computational and experimental methods to characterize genomic structures. This section uses machine learning in bioinformatics to categorize the primary, secondary, and tertiary protein structures.
3. Functional genomics: Experts try to explain the relationships and roles of genes in this area. Classifying mutations and protein subcellular localization in biology can be aided by deep learning.

Large volumes of biological data related to genomics have been analyzed by researchers using DL techniques and natural language processing. By doing this, they can quickly find solutions to problems like relation extraction and named entity recognition. According to DL, the business now gives a comprehensive choice of goods and facilities to the community. The industry is projected to increase to an astounding 54.4 billion USD by 2025 [16]. Applications of machine learning in genomics include:

1. Genome Sequencing: It is essential for diagnostic purposes. Researchers can now sequence human genomes in a day thanks to machine learning-enabled DNA sequencing techniques like next-generation sequencing, as opposed to the traditional Sanger Sequencing Technology, which took over ten years to read a human genome.
2. Gene Editing: The act of replacing, removing, or inserting DNA sequences to alter an organism's genetic makeup is

known as gene editing. It employs a technology called CRISPR, which is a quicker and less expensive way to carry out the procedure. However, there is still work to be done by researchers to choose the proper DNA sequence, which can be a time-consuming and error-prone process. Machine learning has saved the day by making it simpler to pinpoint the right target market, drastically cutting the cost and time needed to carry out gene editing.

3. Clinical Workflow: The clinical workflow process has been greatly changed by deep learning. For instance, it has always been difficult for healthcare professionals to access patient information that is stored in electronic records, paper charts, and other sources [17]. However, healthcare institutions may now fully utilize patient data thanks to the introduction of ML-enabled solutions like Intel's Analytics Toolkit.

### 3.2 Proteomics

The study of protein elements, their connections with one another, and their functions in an organism are known as proteomics. Numerous human proteins have been examined thanks to mass spectrometry-enabled proteomics [18]. Its development has been hampered by computational and experimental issues, necessitating informatics methods like machine learning to evaluate and understand vast biological data sets. Due to its high throughput operations, mass spectrometry is a method of analysis utilized in omics investigations to analyze biological samples. Proteins are not directly measured by mass spectrometry in their typical form. Instead, it divides them into smaller units made up of around 30 building block amino acid sequences. The amino acids are then assigned to particular proteins after being compared to the database. The consequences are not correct since roughly a few proteins have not been fully identified.

A variety of proteins can be identified from a given sample using machine learning techniques. They may be applied to:

1. The mass spectral peaks: Without knowing which proteins and peptides are present, samples are evaluated. Instead, potential biomarkers are compiled from peaks with high signal intensities.
2. Proteins recognized by sequence database searching- The studied sample is searched for peptide masses, and those masses are then used to determine which proteins they correspond to.

These technologies have a clear advantage over more conventional ones like enzyme-linked immunosorbent assays (ELISAs), protein arrays, affinity separation, and 2D gel electrophoresis in the detection of many disorders. The creation of a program named Prosit is one of the most recent developments in the application of machine learning in proteomics. It was used by scientists at the Technical University of Munich (TUM) to swiftly and precisely identify protein patterns.

### 3.3 Microarrays

Microarrays are lab instruments used to simultaneously detect several different gene expressions. This method is useful for researching genome organization, gene expression, and chromatin structures, which are all fields of genetic research that are becoming more and more popular in animals, plants, and microbes [19]. A microarray is a collection of several probes (DNA, RNA, tissues, proteins, and peptides) that are arranged in a certain pattern, typically on a silicon microchip or glass slide, and correlate to distinct gene segments. This technique is based on the idea that under the correct circumstances, complementary sequences will bond to one another, whereas non-complementary ones won't. Fluorescence represents the level of hybridization between modern probes.

Microarray data sets are becoming increasingly complicated at a rapid rate. The simultaneous monitoring of thousands of probes is necessary for large-scale studies. It is now simpler to identify important interactions in complex studies thanks to machine learning. Gene classification and clustering are two of the most frequently recognized applications in microarray analysis where it has been extensively employed. For instance, using machine learning techniques, Neural Designer has enabled researchers to find deep linkages and recognize complex patterns in microarray data. Additionally, public databases like Array Express store all the details of a microarray experiment, making it simple for the academic community to reuse the data.

A few examples of how DL techniques have been used with microarrays include:

1. Gene Analysis: It determines if changes in gene patterns are brought on by a specific disease or whether they are a result of normal aging.
2. Differentiate gene stages: It determines the conditions that cause genes to mutate from a healthy state to a diseased state.
3. Predict future gene stages: It creates models utilizing previous biological data that can forecast gene alterations in the future.
4. Prevents diseases: Through the application of predictive modeling for early diagnosis and preventative medicine, it aids in the finding of connections between genes and diseases.

### 3.4 Text Mining

Text analytics is another name for text mining. It is a deep learning-based technique that analyses massive amounts of texts using natural language processing to find fresh information that contributes to the resolution of research topics. It is now more challenging for researchers to go through various sources and assemble pertinent data on a given issue as a result of the rise of

biological publications [20]. Machine learning can handle and evaluate data by using various human-generated report kinds in databases, which lowers labor costs and expedites research without sacrificing quality. Bioinformatics uses DL for:

1. large-scale investigation of PPI
2. Content transformation into various languages
3. identifying innovative pharmacologic goals (since it requires the mining of information stored in biological journals and data sets)
4. Functions of genes and proteins are automatically annotated
5. DNA expression arrays analysis

### 3.5 Systems Biology

Systems biology is the computational and statistical analysis of how biological elements including molecules, cells, organs, and organisms interact and behave. In this field, computational modeling is a useful tool [14]. It simulates the behavior of the entire system and employs mathematical modeling to capture the interconnections between biological parts. However, given the complexity and incomplete knowledge of the underlying mechanisms, it is challenging to construct a stable mathematical model.

However, with the development of data-driven machine learning approaches, modeling complex linkages in domains like signaling pathways networks, genetic connections, and metabolic functions have become easier. When there is enough biological data but not enough biological understanding to create theory-based models, machine learning is useful in biological systems. The exposure to the relationship between the *S. cerevisiae* genotype and phenotype is a remarkable example.

The lack of theory-based models that show how the variation in genotypes determines the strain phenotypes persists despite the abundance of strains with documented phenomes and genomes. Given this, DL is used to establish the bond amid phenotypes and genotypes by making an administered model with input in form of a genome and output in the form of phenomes. The interpretation of the final model provides cues about the essential genetic makeup of the organism. It aids in determining the most important aspects that influence the model's ability to forecast the future.

One of the DL procedures that are operated the most in systems biology is the probabilistic graphical model. It models genetic networks and determines the relationship between various variables. Genetic algorithms are another often-used method. It has also been used to model regulatory systems and genetic networks based on the natural process of evolution. Additionally, DL is employed in systems biology to address issues like finding transcriptional binding sites using the Markov chain optimization method. A Markov Chain is a stochastic model that uses information gathered from prior occurrences to explain a probable sequence of events.

## IV. DEEP LEARNING FOR PROTEIN-PROTEIN INTERACTIONS (PPIS)

Because it plays a crucial role in predicting the protein function of the target protein and the drug-like properties of compounds, protein-protein interactions (PPIs) are important in system biology with a variety of biological processes. Because they offer an alternative to laboratory tests and a cost-effective technique to forecast the most likely collection of interactions at the complete proteome scale, many studies have been published to predict PPIs computationally. DL has recently gained popularity in computational approaches thanks to various scientific studies [21].

[22] created a DL model that is attention-based and was motivated by techniques for assigning captions to images when designing peptide or protein sequences. These protein-protein interface core elements, give rise to these interaction fragments. The trained model enables the one-sided creation of a specific protein fragment, which is useful for redesigning protein interfaces or creating brand-new interaction pieces from scratch. Simulating naturally occurring protein-protein interactions, such as antibody-antigen complexes, show its potential. The created interfaces exhibit native-like binding affinities and successfully predict key native interactions. Additionally, it does not require a precise backbone position, making it a desirable tool for working with protein-protein interaction de novo design.

To model and forecast PPIs solely based on sequence information, [23] introduces DPPI, a revolutionary DL framework. By using pre-existing, high-quality experimental PPI data and historical data about a protein pair under prediction, the approach successfully predicts PPIs. It achieves this by using a deep CNN with Siamese-like architecture, random projection, and data augmentation. The test results show that DPPI outperforms state-of-the-art methods on a variety of metrics in terms of the area under the precision-recall curve (auPR), and is computationally more efficient. Moreover, it shows how effective DPPI is in a wide range of applications, including estimating cytokine-receptor binding affinities and correctly predicting homodimeric interactions in cases when other methods fall short.

[24] presented a sequence-based method for PPI prediction utilizing many parallel convolutional neural networks called DeepTrio. Experimental analyses reveal that DeepTrio outperforms several cutting-edge approaches in terms of different quality criteria. Additionally, DeepTrio is expanded to offer more details on how each input neuron contributes to the outcome of the prediction.

[25] introduced a technique for forecasting PPI sites dubbed DeepPPISPXGB, which is based on DL and XGBoost. The feature extractor function of the DL model was used to exclude extraneous data from protein sequences. Using Extreme Gradient Boosting, a method for detecting PPI sites was developed. The DeepPPISP-XGB achieved the following findings to compete with cutting-edge methods: an area under the receiver operating characteristic curve of 0.681, a recall of 0.624, and an area under the precision-recall curve of 0.339. Additionally, we confirmed that global traits play a beneficial impact in predicting protein-protein interaction locations.

[26] provided a better method for learning graph representation to address the problem of representing graph information. Our model is capable of investigating PPI prediction using both sequence data and graph structure. In addition, our study utilizes a representation learning model and a graph-based DL approach for PPI prediction, which outperforms current sequence-based approaches. Based on statistics, our method produces the best results on the Human protein reference database (HPRD), *Drosophila*, *Escherichia coli* (*E. coli*), and *Caenorhabditis Elegans* (*C. Elegans*) datasets of the Database of Interacting Protein (DIP).

## V. DEEP LEARNING IN BIOINFORMATICS: CHALLENGES AND LIMITATIONS

We thoroughly outlined the fundamental yet crucial DL theories, techniques, and contemporary applications in various biomedical studies in the study. Reviewing common DL models like RNN, CNN, and DBN enables us to emphasize how important the application scenario or context is when designing a suitable DL method to extract knowledge from data. As a result, characterizing and interpreting data features is still a challenging task in DL workflow. Numerous variations on traditional network models, such as the network models shown above, have been used in recent DL experiments to demonstrate how model choice affects how well DL applications function.

Second, we should reconsider the method's nature to understand its limitations and potential directions for improvement. DL is mainly a continual manifold transformation between different vector spaces, but due to the complex geometric transformation, many problems cannot be turned into a DL model or a learnable technique. DL differs from traditional statistical learning or Bayesian approaches in that it is typically a big-data-driven technique. As a result, integrating or embedding DL with other traditional methods to handle such challenging jobs is a new route for the technology.

Third, in terms of expansion in computational tools and methods, DL is a big data-driven inference technique that requires high-performance parallel computing infrastructure, further algorithmic innovations, and the rapid collection of diverse perceptual data. It is having a lot of success across many different applications and sectors. It has observed striking changes in its research methodologies, particularly in the data-oriented field of bioinformatics and computational biology.

Finally, given the remarkable creativity and triumphs that DL has achieved in a variety of subfields, some have even suggested that DL could usher in a new wave similar to the internet. Long-term, DL technology will have a significant impact on how our lives and civilizations develop. However, DL has many technical issues to resolve because of its nature and should not be misunderstood or overvalued in either academia or the AI sector.

The conversion of enormous datasets produced by recently developed technology into informative data is one of the most urgent problems in bioinformatics and biology as a whole. However, DL in bioinformatics is playing a crucial part in bringing about this revolution as we move into the era of artificial intelligence and big data.

Some of the fundamental ideas of DL and its most current applications in biology and bioinformatics have been covered in this article. It has been seen that DL may be employed to create intricate models and algorithms that aid in the forecasting of trends across several biological fields. Ultimately, for these models to function, good data must be provided in terms of statistical power and sample sizes.

In general, the following summarizes certain difficulties and restrictions:

1. Large amounts of data are needed for DL processing. Health information is not always publicly available (public). A tiny set of tuples in the database for some rare conditions can cause misclassification.
2. It could take some time to run out of computational resources.
3. The architecture of DL is sometimes viewed as a mystery. Sometimes, researchers are unable to correct misclassifications, which results in poor accuracy.
4. One of the drawbacks of DL is overfitting.
5. Sometimes a gradient will vanish. The DL model's accuracy will suffer as a result.
6. The cost will go up since high computational processing processors are needed to increase complexity.
7. Scaling uncertainty is another difficult task.
8. Catastrophic forgetting: DL models are unable to pick up new information without influencing the old.
9. Because the biological data is unbalanced, using unbalanced data to train a neural network could produce unfavorable results.

## VI. CONCLUSIONS

The field of DL in bioinformatics is new. Informatics research is being done intensively to address the issues in this field. An in-depth discussion of the various DL applications in bioinformatics was provided in this chapter. We have multiple methods to use for researchers to implement various DL techniques. Although DL in bioinformatics has shown promising results, there are still significant difficulties with its application, including misclassification, overfitting, imbalanced data, and results in interpretation. In this chapter, difficulties with DL are also explored. We think that the scientific community will benefit from this thorough review's perspective and assist it to move closer to using DL architectures in bioinformatics.

## REFERENCES

- [1] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, vol. 166, no. April 2019, pp. 4–21, 2019, doi: 10.1016/j.ymeth.2019.04.008.
- [2] Z. Liao, G. Pan, C. Sun, and J. Tang, "Predicting subcellular location of protein with evolution information and sequence - based deep learning," pp. 1–22, 2021.
- [3] P. Thareja and R. S. Chhillar, "A review of data mining optimization techniques for bioinformatics applications," *Int. J. Eng. Trends Technol.*, vol. 68, no. 10, pp. 58–62, 2020, doi: 10.14445/22315381/IJETT-V68I10P210.
- [4] Aman and R. S. Chhillar, "Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease using WEKA Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, p. 2021, Oct. 2021.
- [5] P. Thareja and R. S. Chhillar, "Comparative Analysis of Data Mining Algorithms for Cancer Gene Expression Data," vol. 12, no. 10, pp. 322–328, 2021, doi: <http://dx.doi.org/10.14569/IJACSA.2021.0121035>.
- [6] N. Sapoval *et al.*, "deep learning across the biosciences," pp. 1–12, 2022, doi: 10.1038/s41467-022-29268-7.
- [7] Aman and R. S. Chhillar, "Disease predictive models for healthcare by using data mining techniques: State of the art," *Int. J. Eng. Trends Technol.*, vol. 68, no. 10, pp. 52–57, 2020, doi: 10.14445/22315381/IJETT-V68I10P209.
- [8] C. Zhang, Y. Lu, and T. Zang, "CNN - DDI : a learning - based method for predicting drug – drug interactions using convolution neural networks," pp. 1–11, 2022.
- [9] E. Elbasani, S. N. Njimbouom, T. J. Oh, E. H. Kim, H. Lee, and J. D. Kim, "GCRNN : graph convolutional recurrent neural network for compound – protein interaction prediction," pp. 1–13, 2021.
- [10] D. Griffith and A. S. Holehouse, "PARROT is a flexible recurrent neural network framework for analysis of large protein datasets," pp. 1–17, 2021.
- [11] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion : a review," vol. 23, pp. 1–15, 2022.
- [12] P. Supriya, B. Marudamuthu, S. K. Soam, and C. S. Rao, "Trends and Application of Data Science in Bioinformatics BT - Trends of Data Science and Applications: Theory and Practices," S. S. Rautaray, P. Pemmaraju, and H. Mohanty, Eds. Singapore: Springer Singapore, 2021, pp. 227–244.
- [13] P. Thareja and R. S. Chhillar, "A Detailed Survey on Data Mining based Optimization Schemes for Bioinformatics Applications," 2021.
- [14] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Front. Genet.*, vol. 10, no. MAR, pp. 1–10, 2019, doi: 10.3389/fgene.2019.00214.
- [15] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep learning in the biomedical applications: Recent and future status," *Appl. Sci.*, vol. 9, no. 8, Apr. 2019, doi: 10.3390/APP9081526.
- [16] G. Market, "Genomics Market by Product & Service (System & Software, Consumables, Services), Technology (Sequencing, PCR), Application (Drug Discovery & Development, Diagnostic, Agriculture), End User (Hospital & Clinics, Research Centers) – Global Forecast to 2025." <https://www.marketsandmarkets.com/Market-Reports/genomics-market-613.html>.
- [17] A. Darolia and R. S. Chhillar, "Analyzing Three Predictive Algorithms for Diabetes Mellitus Against the Pima Indians Dataset," *ECS Trans.*, vol. 107, no. 1, pp. 2697–2704, 2022, doi: 10.1149/10701.2697ecst.
- [18] B. Wen *et al.*, "Deep Learning in Proteomics," *Proteomics*, vol. 20, no. 21–22, Nov. 2020, doi: 10.1002/PMIC.201900335.
- [19] H. S. Basavegowda and G. Dagnev, "Deep learning approach for microarray cancer data classification," vol. 5, pp. 22–33, 2020, doi: 10.1049/trit.2019.0028.
- [20] C. Simon, K. Davidsen, C. Hansen, E. Seymour, M. B. Barnkob, and L. R. Olsen, "BioReader : a text mining tool for performing classification of biomedical literature," vol. 19, no. Suppl 13, 2019.
- [21] A. R. Jamasb, B. Day, ~ Ta ~ Lina Cangea, P. Liò, and T. L. Blundell, "Chapter 16 Deep Learning for Protein-Protein Interaction Site Prediction," doi: 10.1007/978-1-0716-1641-3\_16.
- [22] R. Syrlybaeva and E.-M. Strauch, "Deep learning of Protein Sequence Design of Protein-protein Interactions," doi: 10.1101/2022.01.28.478262.
- [23] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein-protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018, doi: 10.1093/bioinformatics/bty573.
- [24] X. Hu, C. Feng, Y. Zhou, A. Harrison, and M. Chen, "DeepTrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks," *Bioinformatics*, vol. 38, no. 3, pp. 694–702, 2022, doi: 10.1093/bioinformatics/btab737.
- [25] P. Wang, G. Zhang, Z. G. Yu, and G. Huang, "A Deep Learning and XGBoost-Based Method for Predicting Protein-Protein Interaction Sites," *Front. Genet.*, vol. 12, no. October, pp. 1–11, 2021, doi: 10.3389/fgene.2021.752732.
- [26] J. Yang, N. Li, S. Fang, K. Yu, and Y. Chen, "Semantic Features Prediction for Pulmonary Nodule Diagnosis Based on Online Streaming Feature Selection," *IEEE Access*, vol. 7, pp. 61121–61135, 2019, doi: 10.1109/ACCESS.2019.2903682.



APPENDIX

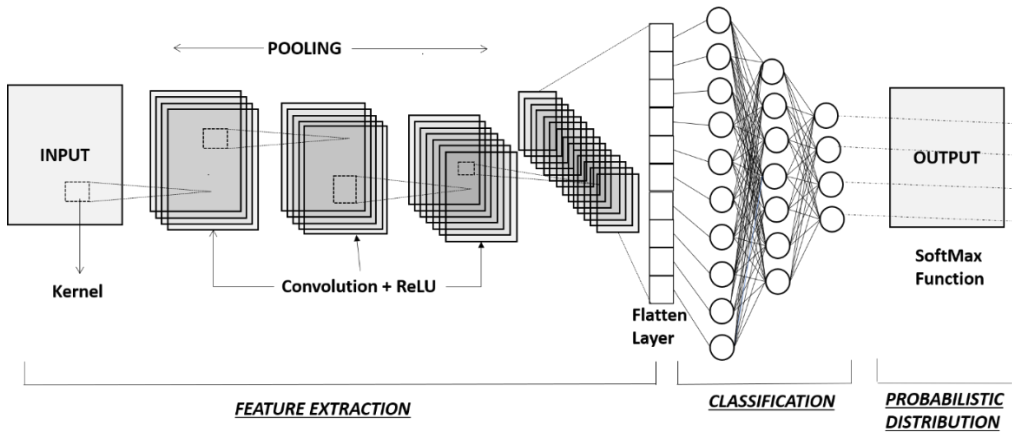


Figure 1 CNN Architecture

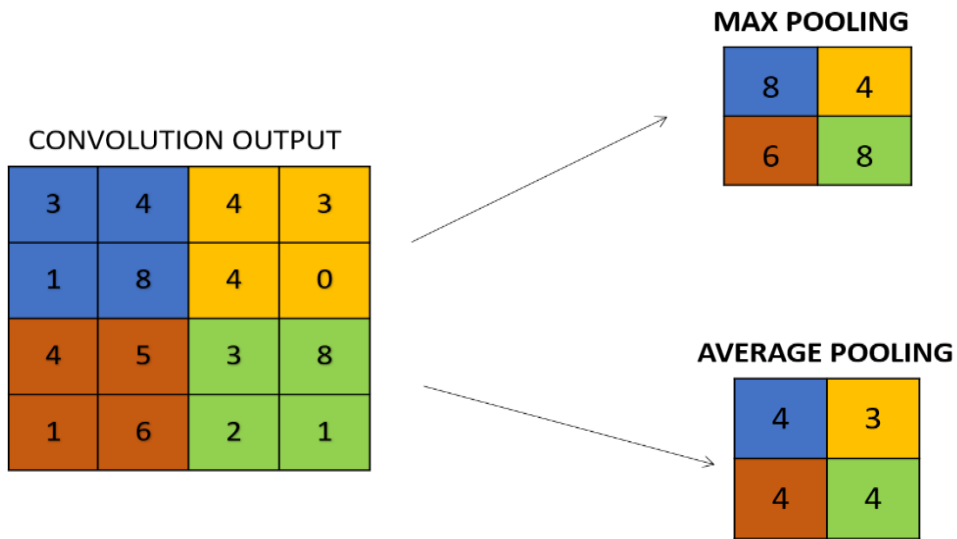


Figure 2 Pooling Method in CNN

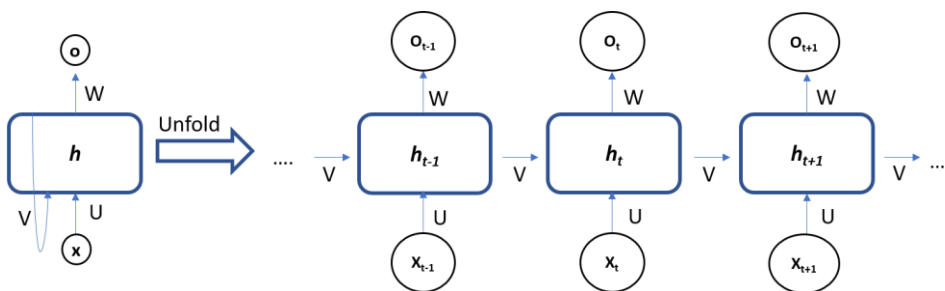
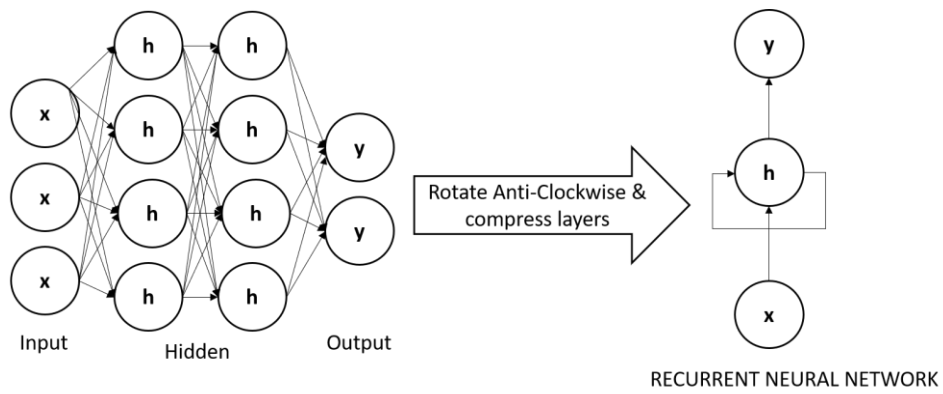
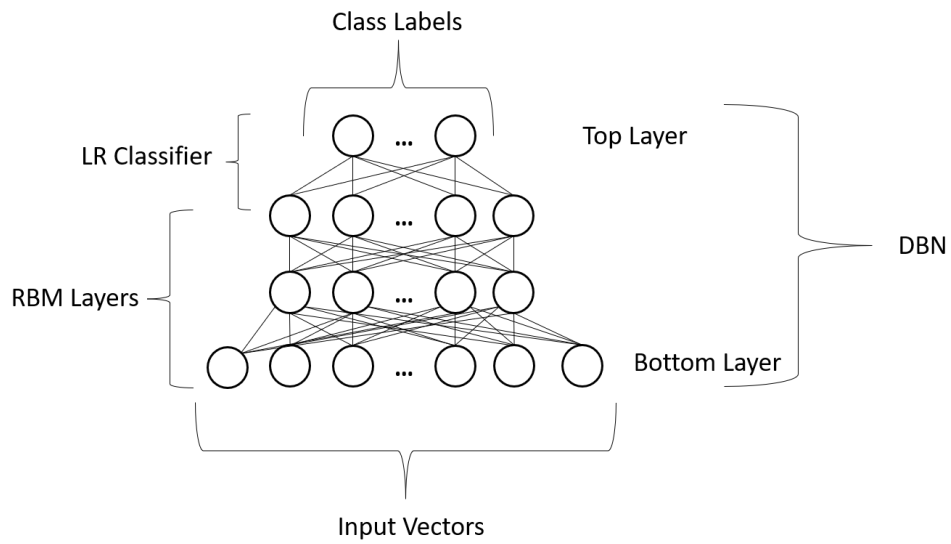


Figure 3 Sequences Unfold in RNN

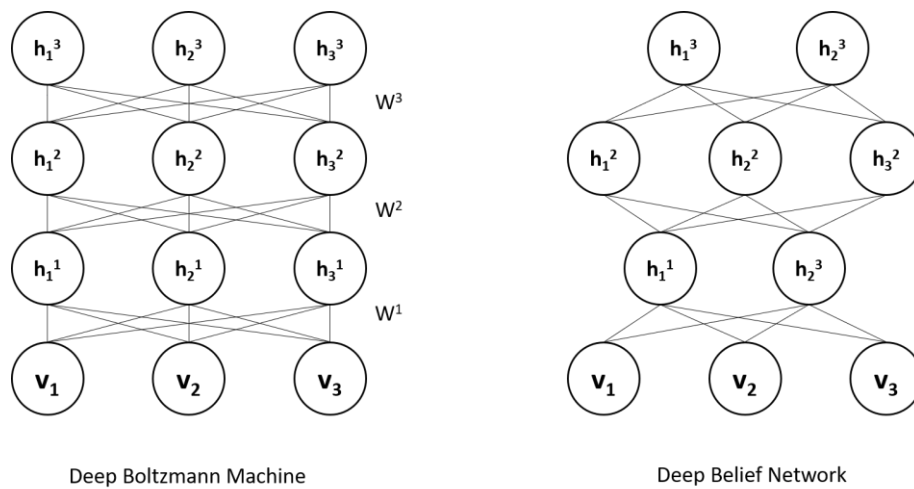




**Figure 4** RNN Architecture



**Figure 5** DBN Technique



**Figure 6** DBN Architecture



# Chapter - 10

## Machine Learning Approach for Early Detection of Plant and Fish Diseases

Dewi Syahidah<sup>1\*</sup>, Bernadetta Rina Hastilestari<sup>2</sup>

<sup>1</sup> Research Centre for Veterinary Science, National Research and Innovation Agency of Indonesia (BRIN), Indonesia

<sup>2</sup> Research Centre for Genetic Engineering, BRIN, Indonesia

Email: <sup>1</sup> [dewi050@brin.go.id](mailto:dewi050@brin.go.id) / [dewi.syahidah@my.jcu.edu.au](mailto:dewi.syahidah@my.jcu.edu.au)

*Abstract— The information technologies currently used in plant and fish farming are largely based on equipment and mechanism, image processing, and pattern acknowledgement, computerized modelling, geographical information systems, expert systems (Pakar), data supervision, artificial intelligence (AI), decision maker devices, and care centres or links. The use of advanced technologies eases the prediction and prevention of parasite infestation and other disease outbreaks. The food productivity of the food sources, including plants and fish, is limited by diseases. The early detection of the disease's infection by naked eyes is somehow difficult. Therefore, early detection through different image processing tools has been introduced widely. Due to the increasing number of reported paper on the potential use of data quarrying and types of machine learning (ML) for plant and fish disease prediction, this chapter consolidates and presents scientific information on the application of data mining and ML in both types of diseases and discussed how imaging technology can be applied to study the diseases and the method in the detection, with comprehensions on the different encounters and prospects. In addition, the potential application of ML in terms of plant and fish disease discoveries in Indonesia are put forward.*

*Keywords— Agriculture, Aquaculture, Data mining, Fish diseases, Image Processing, Machine Learning, Pakar, Plant diseases.*

### I. INTRODUCTION

The age of "big data" began in the 1980s, have directed most society to living in the era of "data mining", where most people have the capacity and freedom to explore vast quantities of specific and complex information. The application of Machine Learning (ML) as a part of artificial intelligence (AI), in this case is enough to make a big impact in almost all industries such as technology, banking, marketing, agriculture, animal husbandry, fisheries, forestry, and entertainment. Although algorithms haven't developed abundant, big data, and considerable calculating support AI to acquire through [1]. The use of computer technology such as ML and computer visions methods for early detection various diseases, including plant and fish diseases has become even more important during pandemic like the COVID-19, which have been occurred since 2019, forcing everyone to depend and work on advanced technology based on digital platforms [2] and the technologies give better solution to farmers to fight against diseases [3-4].

The use of previous diagnostic techniques or macro detection of plant and fish diseases is important but considered less sensitive. Collecting history, reading water quality, visual examination, microscopic observation, laboratory sampling for testing bacteriology, haematology and serum biochemistry, necropsy, and histopathology sampling are part of standard routine for fish diagnostics to investigate diseases in fish [6]. The early detection of the disease's infection by naked eyes, especially for plant diseases is somehow difficult. Therefore, the introduction of different computer processing tools to overcome the issue is important.

Lots of work have been approached in order to detect the disease in early stage by either single ML or its 'combination with different computer processing. Four ML approach, including Support Vector Machine (SVM) as image classification, K-nearest Neighbour (K-NN) as image segmentation, Multi-Layer Perceptron (MLP) as non-linear decision representative, and Convolutional Neural Network (CNN) as image differentiator and identifier, were examined in distinguishing plant and fish diseases. In different reports, each of the approach has shown different function and validation accuracy.

ML can be defined as a part of computer science, by which the program can identify the input automatically [7]. The work mechanisms of ML is merely based on observations or previous experiences and studies, i.e examples or coaching, for identifying data pattern using examples, delivering methods to improve choices. The main purpose of ML is to operate computer automatically without human intervention [8].

---

© 2022 Technoarete Publishing

Dewi Syahidah – "Machine Learning Approach for Early Detection of Plant and Fish Diseases" Pg no: 127 – 136.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch010>

Concerning the growing study and attention of using ML for predicting both plant and fish diseases, this chapter consolidates and presents scientific information on the application of data quarrying and ML for both type of diseases and discussed how the technology can be applied to study disease detection and the related method, with some overviews on the different encounters and prospects. For these reasons, this chapter consists of three parts. Initially, both plant and fish diseases are discussed. A broaden description of the application of ML for the detection of plant and fish diseases are presented in the subsequent part. The emerging uses of ML to enhance plant and fish diseases detection in Indonesia are also put forward at the end of this chapter.

## II. DISEASE THREAT IN FOOD SYSTEM

### II.A. Disturbing organism in staple food

Major staple foods of the world population are wheat, rice, corn, cassava and potatoes [9]. Plant disturbing organisms are posing a threat to the crop yield [10]. Disease-infected plants will indicate sign in the form of certain pattern and discoloration on the plant parts such as leaves, panicles, and stems. Disease symptoms on rice leaves are the easiest part to identify, because usually they have wider surface compared to other plant's organs. Therefore, discoloration and shape changes are easily detected and be used as an initial step for disease detection in rice (Fig. 1).

In Indonesia, yellow rice stem borer (*Scirpophaga incertulas* Walker) causes high amount of yield losses [11]. The borer infects rice plant stage of seedling to maturity and destroys tillers leading to dead hearts formation or dryness of central tiller at vegetative stage and causes empty grains or whiteheads at reproductive stage [12]. (Fig. 1A).

Another disease called Tungro, which is caused by a double infection of two viruses, including a dependent virus namely rice tungro baciliform virus (RTBV) and a helper virus namely rice tungro spherical virus (RTSV). The viruses are transmitted by the green leafhopper *Nephotettix virescens* [13]. Symptoms of the disease could be detected through yellowish leaf discoloration, reduced tillers, and sterile panicle [14; 15] (Fig. 1B), whereas rice blast disease, which is caused by fungi, *Pyricularia oryzae* Cav [synonym of *Magnaporthe oryzae* (Hebert) Barr], has spread in rice crops worldwide [16]. The symptoms of this disease is lesion on leaves, panicles, seeds, stems and also roots. Symptom of this disease is white lesion on the leaves [16; 17] (Fig. 1C). Equally important, bacterial blight is caused by *Xanthomonas oryzae* pv. *Oryzae*. The disease can attack in every phase of plant growth with symptom of leaves discoloration [18; 19].

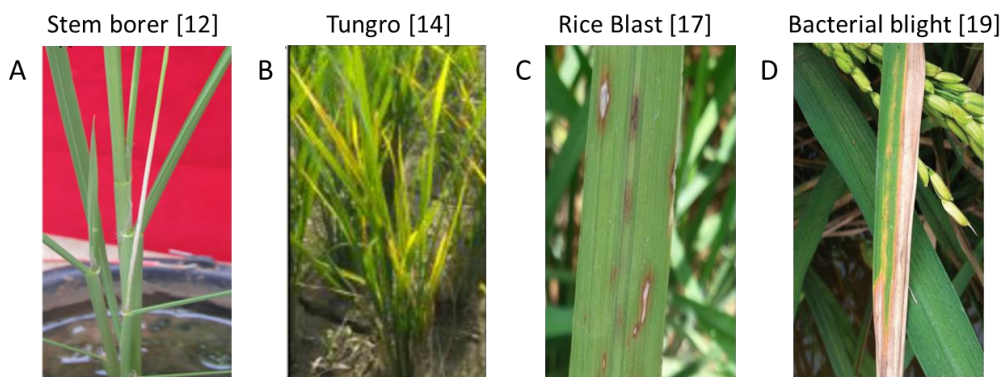


Figure 1. Symptoms of main rice disturbing organisms.

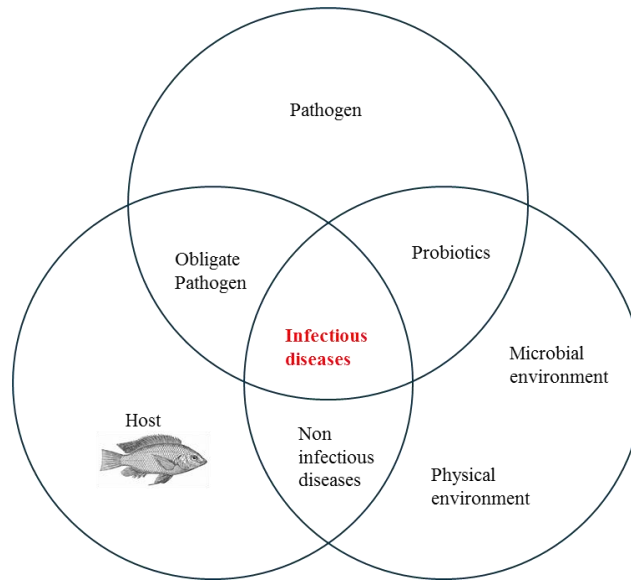
Those disease symptoms are usually detected by direct observation and farmer's experiences. However, if the cultivation area is big, it is difficult to do direct observation because it should be carried out by experts or agricultural workers. In Central Java, Indonesia, agriculture has tried to deal with problem such as decline number of young farmers [10] and decrease of qualified agricultural officers [20]. Therefore, with such limited human resources, as it is hard to monitor each of the crop plants, it is necessary to develop an application based on ML for disease detection which can help farmers to find out early symptoms.

With ML implementation, the risk of yield losses can be reduced at early stage so that the food security could be achieved. Therefore, in this chapter, plant disease detection using ML is presented in sub section III.A. and the potential application of Image technology to detect plant disease is explored in sub section IV.A.

### II.B. Fish Diseases

Fish diseases are substantial sources of economic loss to the aquaculture industry. Disease infection and the occurrence significantly surge production costs because of the investment lost in mass mortalities, cost of treatment, and poor growth during convalescence. A diagram called the Sneizko three-ring Venn (Fig. 2) illustrates the interactions between fish as the

host, pathogen, and the environment, in which the fact, that, to incident, most contagious disease is a triangle interaction, covering all components of pathogen, host, and environment. Several variation of the model have been made to illustrate specific point. Non-infectious disease is an interaction merely between the fish and environment. The overlapping area between pathogen and host represents obligate pathogen, the most threatening group as they do not need environment stress to cause clinical disease [21].



**Figure 2.** A Sneizko three-ring venn (modified from [21]).

All fish carry pathogens and parasites, making them susceptible to various types of diseases. If the impact of a pathogenic infection incurs a very high cost, it can be categorized as a disease. Diseases in fish are mostly not well understood. The most frequently discussed are diseases of aquarium fish, and more recently, with cultured fish. Two main factors cause fish disease, namely biotic, including parasites, fungi, bacteria, and viruses. While the quality of feed and poor environmental conditions are abiotic factors [22]. Some of fish bacteria and parasites are pathogenic and can cause zoonosis [23] (Table 1). Zoonosis are diseases and infectious agents that are naturally transmitted from vertebrate animals, including fish to human [23]. In direct zoonosis, the agent requires only one host to complete its entire life cycle, without significant changes during transmission [23].

**Table 1.** Some common finfish diseases that can cause zoonosis

No	Types of diseases	Caussative-agents	Name of diseases
1	Bacterial	<i>Aeromonas</i> sp. <i>Pseudomonas</i> sp. <i>Aeromonas salmonicida</i> <i>Mycobacterium tuberculosis</i> <i>Chondrococcus columnaris</i> <i>Dactylogyrus</i> spp. <i>Aeromonas hydrophylla</i> * <i>Mycobacterium marinum</i> * <i>Vibrio</i> sp.* <i>Streptococcus iniae</i> * <i>Edwardsiella</i> spp.*	Fin and tail rot Skin ulcer Furunculosis Fish Tuberculosis Cotton mouth disease Flukes Dropsy Akuarium granuloma Vibriosis Meningoencephalitis Scepticaemia
2	Fungal	<i>Saprolegnia</i> <i>Branchiomyces demigrans</i> <i>Aphanomyces invadans</i> <i>Oomycetes</i> sp.	Cotton wool disease or Saprolegniasis Branchiomycosis or gill rot Epizootic ulcerative syndrome (EUS) or red spot Dermatomycosis

3	Parasitic	<i>Ichthyophthirius multifiliis</i> <i>Argulus sp.</i> <i>Lernaea sp.</i> <i>Myxosoma cerebralis</i> <i>Ichthyobodo necator</i> <i>Oodinium Pilularis</i> <i>Anisakis sp.*</i> <i>Diphyllbothrium latum*</i> <i>Gnathostoma sp.*</i> <i>Spirometra erinacei-europaei *</i> <i>Angiostrongylus cantonensis*</i> <i>Clonorchis sinensis*</i>	White Spot disease Argulosis Anchor worm Whirling disease Slime disease Velvet or rust Anisakiasis Diphyllbothriasis Gnathostomiasis Sparganosis Angiostrongyliasis hepatobiliary
4	Protozoans	<i>Costia necatrix</i> <i>Myxobolus cerebralis</i> <i>Ichthyophthirius multifiliis</i> <i>Diplostomulum</i> <i>Eubothrium</i>	Costiasis Whirling Disease Ichthyophthirius Diplostomis Gut blocking
5	Viral	<i>RNA viruses</i> <i>Infectious pancreatic necrosis virus</i> <i>Rhabdovirus carpio</i> A herpes virus	Viral haemorrhagic septicaemia (VHS) Infectious pancreatic necrosis (IPN) Spring viremia of carp Channel vatfish virus disease

\*Can cause zoonosis

Conventionally, agar plates and then phenotypic and serological properties of pathogens or histological analysis were used to diagnose fish diseases [24], [25]. However, conventional diagnostic procedures cannot distinguish some bacteria from phenotypically comparable bacteria of one genera [26]. Several efforts using biochemical assessment, DNA homology, and protease variability have been conducted [27] – [29]. However the methods were ineffective because the need for previous pathogen isolates and less sensitive to detect low-level pathogens.

During the last decades, great improvements have been conducted in studying the molecular biology of fish pathogens and their hosts. Molecular biology has become a routine instrument in the examination for enhanced techniques of fish disease control, the epidemiology of bacterial, viral, and parasitological diseases [30]. Molecular detection of nucleic acid has confirmed its practicality for stressing scarcely cultivable, non-cultivable, and even dead microbes, creating proper novel or additional machineries [30].

Considering the important of combating fish diseases and couple with the development of the use of digital platforms in aquaculture sector nowadays, the use of ML as part of AI to detect fish diseases is indispensable. In this chapter, the assessment of ML for detecting diseases in fish is presented in subsection III.B, while studies using expert system called Pakar to predict fish diseases in Indonesia is presented in subsection IV.B.

### III. DISEASES DETECTION FOR PLANT AND FISH

#### III. A. Plant Disease detection

After being infected by pathogen, plants send signal through intracellular responses through Mitogen-activated protein kinase (MAPK) cascades [32]. These signals induce morphological - physiological and omics changes of the plants. Physiological changes can be noticed by the discoloration, closure of stomata, biosynthesis of hormones, cell wall strengthening, production of reactive oxygen species (ROS) and cell death [33]. Omics changes can be observed in the upregulation and / or down gene expression, changes in metabolite profiling and protein [34].

Plant – pathogen interaction can be detected by destructive and non-destructive techniques. Destructive techniques can be done by a laboratory detection using immunological and serological as well as molecular methods [35]. Those methods are conducted by destructive method for extracting genetic material or certain metabolite, while non-destructive one can be applied through observation of morphological changes and ML based plant disease detection.

A touch of technology and accurate statistical data on agriculture are able to boost the yield. Accurate data are the first step in digitalization of agricultural process from seed selection to post harvest activities. These data are collected from many growth phases, disease symptoms and other sectors for further detection and decision making [36]. In this 4.0 industry era, data collection has been conducted in a fast and comprehensive way. Big data serves an opportunity to identify crop disease for which disease control can be taken and predicted in advance [37], [38]. Data can be collected from laboratory data, Geographical Information System (GIS), Global Positioning System (GPS), remote sensing technology, and virtual satellite



imaging by integrating soil, climate and environment information [39].

Immunological technique relates to experimental strategy by using antibodies for detecting particular proteins in the targeted samples; while serological technique is used by antigen detection assays [35]. Immunological and serological diagnosis can be applied to detect plant viruses, plant bacterial infection and fungi infection [40], [41]. Some techniques belong to this detection are such as fluorescence in-situ hybridization (FISH), enzyme-linked immunosorbent assay (ELISA). FISH is applied to investigate pathogen infections using hybridization of DNA probes and infected plant sample [42]. This technique can be used to detect pathogen, but it may lead to false positive due to auto-fluorescence [40], [44] – [42].

ELISA can be applied to detect plant disease based on colour changes due to binding of the antigen of pathogen and the monoclonal or recombinant antibodies [44], [45]. This detection method can be used to confirm the disease after appearing visible symptoms, but unfortunately is not favourable for early detection without visible symptoms [45].

Owing to advanced technologies of genomic identification, cost for disease detection using nucleic acid based technologies are getting lower [46]. Genomic based disease detection using technology such as polymerase chain reaction (PCR) around the nineties [47] and this technology has been further developed into quantitative PCR (qPCR) and digital PCR (ddPCR) [48], [49]. These methods are quite good for plant disease detection but they require High cost on equipment and reagents [50], [51].

Disease detection as a part of AI is developed by data production without repeat human intervention [52]. Some research on ML to detect plant disease has been conducted for several plants, for instance, corn using Naive Bayes method [53], tomato using Raspberry with 84.22% - 100% detection accuracy [54], tea using SVM [55], [56], MLP [55] - [57], and CNN [56], [57], potato using CNN with 94% validation accuracy [58]. This CNN method had high error rate of around 16% for images clustering into possible categories [59], but then the error could be fixed until 3.5% with some advances in this method [60], [61]. Besides, plant disease has been detected using conventional methods [52] and deep learning [61], [62]. This CNN method has been used as a robust approach to identify plant diseases in tomato [64], while a model in detecting plant diseases in leaves using deep convolutional neural network have been promoted [36].

These ML approaches have been quite efficient for saving cost, labour and times to classify and identify pathogen. What is important in plant detection using ML methods is feature optimization to differentiate the target. It will be more difficult if the symptoms is more complex. However, with the advances of technology and innovation, these limitation can be improved for optimising disease detection and preventing yield losses.

### III. B. Prediction of Fish Diseases

Diseases diagnostics in fish by experienced farmers' bare eyes are sometimes not error free. Despite being laborious since some lab works are required in determining the relevant pathogens, the classical technique most frequently creates an erroneous and ambiguous result. Aqua-culturant cannot provide exact handling for unexpected fish infections. Consequently, it is difficult to dealing appropriate and efficient procedures. Furthermore, that fish disease spreads extensively and cause mass mortalities of fish and huge loss to the farmers [65]. Therefore, a technology such as ML emerges as an advanced detection platform as an on time fish diseases alarm for farmers.

The algorithms used in ML are reliable for different purposes with fast and efficient. The different regressions are applied for different infection and efficiently forecast disease [66]. Efforts to detect fish diseases using ML in previous works have been published. A configuration named the grouping of the image substances that obliged diverse segmented image engagements based on a dimension of some groups was used [67]. Here, selected indicators for definite objects and objects were used and faced with a certain indicator. Lastly, the proportion of an object in the image and the proportion of diseased part to the fish body was calculated to differentiate fish disease. However, a particular marking of an object is ineffective and time wasting [67].

Four infected Epizootic Ulcerative Syndrome (EUS) species such as catfish (*Clarias batrachus*), swamp barb (*Puntius chola*), bata fish (*Labeo bata*), and kuria labeo (*Labeo gonius*) from different sites of the Barak Valley, Assam were identified [68]. Two types of algorithms including Principal Component Analysis (PCA) and K-means clustering were used with a flow chart for the research. The outcomes showed that the algorithm has more than 90% precision for PCA. Therefore, for skin colour identification and texture feature extraction, HSV is a worthy option whereas morphological procedure should be used for fish pathogen identification.

A specific fish disease detection methodology of EUS was proposed by [69]. Combination among the PCA and Histogram of Oriented Gradients (HOG) was used with Features from Accelerated Segment Test (FAST) feature detector and then categorize over ML algorithm (neural network or NN). The sequence of FAST-PCANN gave 86 % precision through the classifier, and HOG-PCA-NN gave 65.8 % precision [69].

Equally important, by using combination among behavioural observing, with the surveillance of noticeable disease syndromes, farm data systems, and smart processes, it is potential to detect precisely, forecasting, and avoid disease outbreak. Fish Doc- tor, SALMEX, Fish-Expert and AquaSDS, were developed with some of them based on ambiguous logic and inference system [70]. In these methods, rule-based data quarrying processes was used to quarry evidence from miscellaneous

image and acquaintance database works (e.g., farmer and human expert assessments) to consistently analyse infections. For instance, Fish-Expert, a web-based smart disease diagnosis system by Chinese researchers could be used to identify 126 types of nine major freshwater fish diseases [71].

Alternatively, Aquaconnect was developed for predicting diseases, based on water condition on farm, nursing, and growing data was developed [72]. Whereas, under the NCE (Norwegian Centers of Expertise) Sea food Innovation project advances the AquaCloud platform, using a workflow that comprise adjustment of data pool of farm level fish health (Fishtalk), enablement of numerical records interchange (Mercatus) and smart processing of data for reinforce the decision sustenance related to fish health and welfare. Another commercial data processing, oiFarm by BioSort and tested by Cermaq, Norway to monitor the health and growth of up to 150,000 specific fishes for targeted health interventions is nowadays being improved.

EUS ulcer can be predicted appropriately and efficiently using a Probabilistic Neural Network (PNN). Input from different sources and database images from numerous internet resources were used. Then images were put through the pre-processing to avert annoying biases or to augment several appearances, which are valuable for additional processing in which Red Green Blue (RGB) to grey conversion has been used. Several extraction approaches have been used Curvelet Wavelet Transform (CWT) for the effectiveness of recognition and lastly the target diseases in fish, including ammonia poisoning, camallanus worm and dropsy were categorized. Others which were not infected are documented and divided [65].

Recently, an important ML-based SVM to detect diseases in infected cultured salmon was introduced by [73]. Two groups of dataset, from A (163 infected and 68 fresh) and B (785 infected and 320 fresh) are used to train their model is novel. Fishes were classified into two individual classes, including fresh and infected fish and appraised the model with different metrics and confirmed the classified result with graphical interface from those grouping outcomes. The work contributed to conveying out an advanced programmed fish detection method than the previous systems based on image processing or lesser precision. The analyses were not only be determined by the contemporary image processing procedure but also attach demandable controlled learning procedures. The classifier is formed to envisage diseased fish with the preeminent precision level than other systems for the authors' actual-world novel dataset [73].

#### IV. POTENTIAL APPLICATION OF ML TO DETECT PLANT AND FISH DISEASES IN INDONESIA

##### IV. A. Image technology for plant diseases detection

Whenever plant has been infected by pathogen, there will be symptoms in several parts of the plants, for instance leaves, stems, panicles and fruits, and roots. Direct observation is possible if the number of plants is small, but if the number is quite big, technology such as image based ML are able to classify the plant disease quite accurately. ML is considered a powerful model to perform image segmentation and classification, even if the number of data is huge [74]. Some models of pre-training in CNN approach can be applied through for instance Alexnet, Googlenet and ResNet model. The last model could identify 13 diseases in plants as well as differentiating healthy leaves using stochastic gradient descent (SGD). Leaves discoloration in the citrus can be detected using hybridization method by CNN methods [75], in which discoloration spot in the fruit was segmented and classified using multiple SVM. The characteristics of data images such as colour and texture are transferred into a codebook. In this system, pertaining approach is important to get good accuracy, some models for instance AlexNet, GoogleNet, ResNet and VGGNet can be applied [76]. Then, data images are adjusted using sensors and transferred to the segments to identify the region of interest (ROI). From this ROI, features are extracted and classification approaches are implemented the performance [76], [77].

The selected image are segmented using the algorithm, producing binary set as 1 and 0, which are classified as infected and noninfected plants respectively. This classification can be done manually that is indicated by (p) and the automatically classified set (t). The algorithm accuracy needs to be tested within the parameters used in the manual and automatic classification. These classification is further compared mathematically from two approaches, i.e. pixels in the image (z) and pixel as part of infected area (d) (Eq. (1)) [78].

$$z = \sum_{i=1}^m \sum_{j=1}^n I(i, j); \quad d = \sum_{i=1}^m \sum_{j=1}^n = 1 \quad (1)$$

The success of final result discrimination is color transformation. This approach goal is to transform the image for discriminating infected and non-infected parts. This color transformation provides information about disease area and the background. HSV (Hue, Saturation and Value) is applied in order to distinguish colors associated with the wavelength of light, to indicate how much white color is given and to indicate the amount of light received by the eye regardless of color [79].

##### IV.B. Pakar System for the Detection of Fish Diseases

Barriers to fish farming business are mainly caused by disease disorders, both infectious and non-infectious. In addition with the progress of AI technology, the function of information system technology in overcoming fish diseases is very important. One of these technologies is called an expert system or so-called "Pakar". Experts try to implement anthropological acquaintance to processors that are intended to address issues like professionals [80] based on a knowledge base that can be

obtained from textbooks or make diagnostic reasoning based on rules [81] and can replace the role of humans (experts) in carrying out their duties [82], [83].

As a part of AI, Pakar is used to solve complex problems and only an expert can solve them, but Pakar is not made to get rid of an expert but only as a guide [84], [85]. The most commonly used Pakar system is the rule-based expert system. Instruction-based knowledgeable schemes have many benefits including lowering costs, being permanent, increasing effectiveness because it reduces errors experienced by humans, and if designed by many experts can increase trust [86]. Pakar systems based on rule processing (working memory) are widely applied in diagnosing fish diseases and the methods that are often applied to experts are certainty factor and forward chaining. Certainty Factor (cf) which is a clinical parameter value to show the amount of confidence, which is accompanied by a level of confidence in the percentage that can be used to the need for diagnosis of fish diseases [83]. The certainty factor is also a way of combining belief and disbelief in a single field [87].

Forward chaining (fc) is a forward tracking method that is driven by existing data and combining rules that are used to create a conclusion or goal into a decision. Starting with a series of tracing the input of known facts and applying rules to generate new facts where the basis of the conclusion matches the acknowledged evidences, and lasts this procedure until it ranges the programmed objective, or until no extra evidences occur from other premise, matches the recognized evidences. Fc confirms evidences compared to a predetermined probe or objective, and shows that implication transfers onward from evidences to objectives [88] – [90] and can help provide output regarding disease diagnosis [91].

Expert system testing (Pakar) using the cf and fc methods for diagnosing fish diseases in Indonesia is growing. A new software on a website-based expert system to identify diseases in freshwater fish consumption (Table 2). In the process of searching the information, it is supported by the Bayes Theorem to support its certainty [92]. The resulting software is able to identify 14 diseases in Indonesian freshwater fish based on the symptoms entered and provide solutions like an expert. fc was found suitable for the detection of infectious and non-infectious diseases of freshwater ornamental fish, including *Symphysodon* sp., *Cyprinus rubrofusculus*, *Carassius auratus auratus*, *Pterophyllum scalare*, *Scleropages formosus*, *Poecilia reticulata*, and *Paracheirodon innesi*, along with their symptoms and treatment [92].

Fc was successfully used in detecting the attack of protozoa, *Aeromonas hydrophila* bacteria, *Trichodina* sp., and itching disease in *Clarias gariepinus* [94]. Meanwhile, a 100% validity with an average user acceptance of 83.6% was achieved after using the same method to diagnose *C. batrachus* diseases caused by 11 pathogens, including *Pseudomonas hydrophilla*, *Aeromonas hydrophilla*, *Aeromonas punctate*, *Columnaris*, *Edwardsiella* sp, *Mycobacterium tuberculosis*, White fungus, White spot, Itch, Trematoda, and *Lernea* sp. [95]. Cf was used to diagnose 9 *Cyprinus carpio* including Argulus, white spot disease, Lerniasis disease, itching disease, Aeromoniasis disease, Columnaris disease, Koi Herves disease, Virus (KHV), EUS and Saprolegnia with 30 physical symptoms in goldfish caused by parasites, bacteria, and viruses [96].

A 100% accuracy rate was obtained after testing the cf to diagnose 11 diseases in *Pangasius pangasius*, including *Aeromonas hydrophilia*, *Dactylogyrus* sp., *Edwardsiella ictaluri*, *Edwardsiella tarda*, *Epistylis* sp., *Flexibacter* sp. *Ichthyophthirius multifiliis*, *Oodinium* sp., *Saprolegnia* sp., *Trichodina* sp., and *Vorlicella* sp., by providing symptom value data, obtained from experts [97]. Meanwhile, cf was applied to help to overcome *Oreochromis mossambicus* diseases by designing an expert system to find out the type of disease and how to overcome it [98].

The combination between fc and cf resulted a builded up application for detecting *Betta splendens* disease. Fc method is able to provide a large amount of information from only a small amount of data. While cf technique is suitable for use in Pakar systems to measure whether something is certain or uncertain in diagnosing disease as one example [99].

The application of cf for the detection of *Osphronemus goramy* diseases is considered as an alternative solution for detecting good quality gouramy disease and maintaining fish health. The process of determining the diagnosis can be done more accurately and precisely compared to just checking and estimating yourself [100]. Meanwhile, a level of confidence above 95% was achieved after testing cf to diagnose parasitic diseases in fish, such as White Spot, Trichodina, Myxobolosis, Heneguya, Epistylis, Oodinium, Kutilan, Stye, Lernaeciasis, Ergasilusis, Kuan, Saprolegnesis and Achlyasis [101].

**Table 2.** The use of “Pakar” for the Detection of Fish Diseases in Indonesia (2013-2021)

NO	Methods	Fish	References
	fc & cf	human consumption freshwater fish	[92]
	fc	Discus ( <i>Symphysodon</i> sp.) Koi ( <i>Cyprinus rubrofusculus</i> ) Goldfish ( <i>Carassius auratus auratus</i> ) Manfish ( <i>Pterophyllum scalare</i> ) Arwana ( <i>Scleropages formosus</i> ) Guppy ( <i>Poecilia reticulata</i> ) Neon tetra ( <i>Paracheirodon innesi</i> )	[93]
	fc	Lele dumbo ( <i>C. gariepinus</i> )	[94]
	fc & cf	Betta fish ( <i>Betta</i> sp.)	[99]

cf	Tilapia ( <i>Oreochromis niloticus</i> )	[98]
cf	Patin ( <i>Pangasius pangasius</i> )	[97]
cf	Gouramy ( <i>Osphronemus goramy</i> )	[100]
cf	Carp ( <i>Cyprinus carpio</i> )	[96]

\*cf: Certainty Factor; fc: Forward Chaini

## V. CONCLUSION

The use of digital platform for timely recognition of vegetal and fish disease infection is important and is considered indispensable nowadays. The assessment of ML for the detection of fish diseases in Indonesia mostly focused on Pakar system with Forward Chaining (fc) and Certainty Factor (cf) as the main methods. Therefore, the opportunity to develop Pakar systems in Indonesia is widely open. Meanwhile, detection of plant disease can be designed using the image as there is usually changes of colour of plant parts and spots. ML can be applied using CNN approach. This network can be done for detecting the data in the form of images. CNN can extract image feature and classify the layers which are potential for further development. This ML approach is quite new in agriculture, and it is not easy in the beginning. However, with good investment on the ML approaches, pathogen detection can be applied in an easy and better way. Therefore, we propose that the development of a user-friendly and reliable approaches to detect diseases in plant and fish for not only improving the yield but also preventing the spread of the diseases.

## REFERENCES

- [1] R. Anyoha, "The history of artificial intelligence (AI). Blog, special edition of artificial intelligence," 2017. [Online]. Available: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence>.
- [2] C. Puttamadappa and B.D. Parameshachari, "Demand side management of small scale loads in a smart grid using glow-worm swarm optimization technique," *Microprocessors Microsystems*, vol 71, pp. 102886, 2019.
- [3] R.P., Shaikh, and S.A. Dhole, "Citrus Leaf Unhealthy Region Detection by using Image Processing Technique, in: IEEE International Conference on Electronics," *Communication and Aerospace Technology*, pp. 420–423, 2017.
- [4] D.L. Vu, T.K. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, and P.H. Phung, "HIT4Mal: hybrid image transformation for malware classification," *Transportation Emerging Telecommunication Technology*, vol. 31, no. 11, pp. e3789, 2020.
- [5] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Trans. Intelligence & Transportation Systems*, vol, 22, no. 7, pp. 4337–4347, 2020.
- [6] R. Loh, 2013. *Fish Pathology*, 4th edn Edited by Ronald J Roberts. Wiley Blackwell, Oxford. 597pp.
- [7] E. Gavin, "Discusses about machine learning: an introduction," 2018. [Online]. Available: <https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0>.
- [8] D. Gupta, A. Julka, S. Jain, T. Aggarwal, A. Khanna, N. Arunkumar, and N.V.C. De Albuquerque, "Optimized cuttlefish algorithm for diagnosis of Parkinson's disease," *Cognition System Research*, 52: 36e48, 2018, doi : 10.11648/j.sjac.20190704.12.
- [9] A. Bayata. "Review on nutritional value of cassava for use as a staple food". *Sci J Anal Chem*, Vol.7, No. 4, pp.83-91, Sep. 2019. doi : 10.11648/j.sjac.20190704.12.
- [10] S.H. Susilowati. "Fenomena penuaan petani dan berkurangnya tenaga kerja muda serta implikasinya bagi kebijakan pembangunan pertanian". *Forum Penelitian Agro Ekonomi*, Vol.34, No.1, pp.35-55, Jul. 2016.
- [11] I.S. Dewi, I.H. Somantri, D. Damayanti, A. Apriana and T.J. Santoso. Evaluasi tanaman padi transgenik Balitbio terhadap hama penggerek batang. *Laporan Hasil Penelitian Balitbio, Bogor*, pp. 141 – 150, Nov. 2002, <http://repository.pertanian.go.id/handle/123456789/12199>.
- [12] B.R. Hastilestari, C.F. Pantouw, S. Nugroho and A. Estiati. " Uji ketahanan padi transgenik mengandung gen Cry 1B dibawah kontrol promoter terinduksi pelukaan Mpi terhadap hama penggerek batang kuning (Scirpophaga Incertula WK.) pada fase vegetatif". *Prosiding Seminar Nasional 2013 : Inovasi Teknologi Padi Adaptif Perubahan Iklim Global Mendukung Surplus 10 Juta Ton Beras 2014*. Balai Penelitian dan Pengembangan Pertanian Kementerian Pertanian, pp. 215 – 223, Jul. 2014.
- [13] H.Tyagi, S. Rajasubramaniam, M.V. Rajam, and I. Dasgupta, "RNA-interference in rice against Rice tungro bacilliform virus results in its decreased accumulation in inoculated rice plants". *Transgenic Res.*, Vol. 17, No.5, pp.897-904, Feb. 2008, doi: 10.1007/s11248-008-9174-7.
- [14] O.Azzam, and T.C. Chancellor, "The biology, epidemiology, and management of rice tungro disease in Asia". *Plant Dis.*, Vol. 86, No.2, pp. 88-100, Feb.2007, doi : 10.1094/PDIS.2002.86.2.88.
- [15] B.R.Hastilestari, D. Astuti, A. Estiati and S. Nugroho, "Sequence analysis of ORF IV RTBV isolated from tungro infected *Oryza sativa* L. cv Ciherang". *AIP Conference Proceedings*, Vol. 1677, No. 1, p, 090013, Sep, 2015. <https://doi.org/10.1063/1.4930758>.
- [16] X. Wang, et al., "Current advances on genetic resistance to rice blast disease". In *Rice-Germplasm, genetics and improvement*, 2014, pp.195-217. InTech, Rijeka, Croatia.
- [17] S. Zahrah, R. Saptono, and E. Suryani, "Identifikasi Gejala Penyakit Padi Menggunakan Operasi Morfologi Citra". In *Seminar Nasional Ilmu Komputer (SNIK 2016)-Semarang* , Vol. 10, Oct, 2016 .
- [18] R. Olivia, et al. "Broad-spectrum resistance to bacterial blight in rice using genome editing". *Nat biotechnol.* Vol. 37, No. 11, pp.1344-1350, Oct, 2019.
- [19] M.M.Faizal Azizi and H.Y. Lau, "Advanced diagnostic approaches developed for the global menace of rice diseases: a review." *Canadian Journal of Plant Pathology*, Vol. 44, No.5, pp 627 – 651, Mar 2022, doi: 10.1080/07060661.2022.2053588.
- [20] R. Moordiani, A. Wildani and S. Widayani, S. "Analisis Kebutuhan Penyuluh Pertanian Mendukung Jawa Tengah Menjadi Lumbung Pangan Nasional". In *Prosiding Seminar Nasional Fakultas Pertanian UNS* Vol. 2, No. 1, pp. C53 – C60, 2018.
- [21] L. Owens, "Diseases," in *Aquaculture. Farming Aquatic Animals and Plants*. J.S. Lucas, J.S. and P.C. Southgate, Eds., Blackwell Publishing, 2015, pp. 199-214.
- [22] M. Sharma, A.B. Shrivastav, Y.P. Sahni, Y.P., and G. Pandey, "Overviews of the treatment and control of common fish diseases. International Research," *J. of Pharmacy*, vol. 3, no. 7, pp. 123-127. 2012 [Online] Available: [www.irjponline.com](http://www.irjponline.com).
- [23] World Health Organization (WHO). 2004. *Waterborne Zoonosis: Identification, Causes and Control*. World Health Organization, Geneva.
- [24] J.F. Bernardetn, A.C. Campbell, J.A. Buswell, "Flexibacter maritimus is the agent of 'black patch necrosis' in Dover sole in Scotland," *Dis. in Aquat. Org.*, vol. 8, pp. 233-237. 1990.



- [25] F. Pazos, Y. Santos, A.R. Macías, S. Núñez, and A.E. Toranzo, "Evaluation of media for the successful culture of *Flexibacter maritimus*," *J. of Fish Dis.*, vol.19, pp. 193-197. 1996.
- [26] J.M. Shewan and T.A. McMeekin, "Taxonomy and ecology of the Flavobacterium and related genera," *Ann. Rev. in Micr.* vol. 37, pp. 233-252. 1983
- [27] S.W. Pyle and E.SuppB. Shotts, "A new approach for differentiating flexibacteria isolated from cold water and warm water fish," *Can. Jour. of Fish and Aquat. Sci.*, vol. 37, pp. 1040-1042.1980.
- [28] J.M. Bertolini and J.S. Rohovec, "Electrophoretic detection of proteases from different *Flavobacterium columnare* strains and assessment of their variability," *Dis. in Aquat. Org.*, vol. 12, pp. 121-128. 1992.
- [29] M.F. Chen, D. Henry-Ford, and J.M. Groff, "Isolation and characterization of *Flexibacter maritimus* from marine fishes of California," *J. of Aquat. Anim. Health*, vol. 7, pp. 318- 326. 1995.
- [30] J.A. Plumb, "Health maintenance and principle microbial diseases of cultured fishes," Iowa State University Press. Ames, Iowa. 344 pp. 1999.
- [31] I. Altinok and I. Kurt, "Molecular Diagnosis of Fish Diseases: a Review," *Turkish J. of Fish. and Aquat. Sci.*, vol. 3, pp. 131-138. 2003.
- [32] S.Bartels, et al., "MAP Kinase phosphatase1 and protein tyrosine phosphatase1 are repressors of salicylic acid synthesis and SNC1-mediated responses in Arabidopsis," *The Plant Cell*, Vol. 21, No.9, pp. 2884-2897, Sep. 2009. doi : 10.1105/tpc.109.067678.
- [33] T.Boller and G. Felix, "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors," *Ann Rev Plant Biol.* Vol.60, pp.379 – 406. 2009, doi : 10.1146/annurev.arplant.57.032905.105346.
- [34] X.Meng and S. Zhang, "MAPK cascades in plant disease resistance signaling," *Annu Rev Phytopathol*, Vol.51, No.1, pp. 245-266, May. 2013, doi : 10.1146/annurev-phyto-082712-102314.
- [35] Y.Fang, and R.P. Ramasamy, "Current and prospective methods for plant disease detection", *Biosensors*, Vol. 5, No.3, pp. 537-561, Aug. 2015, doi : 103390/bios5030537.
- [36] A. Zhang, E. Jakku, R. Llewellyn, and E.A. Bake, "Surveying the needs and drivers for digital agriculture in Australia," *Farm Policy J*, Vol.15, No.1, pp. 25-39, 2018.
- [37] K.Thongboonnak, and S. Sarapirome, "Integration of Artificial Neural Network And Geographic Information System For Agricultural Yield Prediction," *Suranaree J. Sci.Technol*, Vol.18, No.1, pp. 71-80, Jan, 2011.
- [38] A.K.Rumpf, et al., "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance," *Comput Electron Agric*, Vol.74, No.1, pp. 91–99. Oct.2010, doi : 10.1016/j.compag.2010.06.009
- [39] J. Schuster. "Big data ethics and the digital age of agriculture," *Resource Magazine*, Vol.24, No.1, pp. 20-21, 2017.
- [40] M. Hijri. "The use of Fluorescent in situ hybridisation in plant fungal identification and genotyping," In *Plant Pathology*, pp. 131-145. Humana Press, Totowa, NJ.
- [41] A. Kliot, et al., "Fluorescence in situ hybridizations (FISH) for the localization of viruses and endosymbiotic bacteria in plant and insect tissues," *J. Vis. Exp.* Vol. 84, p. e51030, Feb. 2014, doi : 10.3791/51030.
- [42] V.A.J. Kempf, K. Trebesius and I.B. Autenrieth 2000. "Fluorescent in situ hybridization allows rapid identification of microorganisms in blood cultures," *Am Soc Microbiol*, Vol. 38, No. 2, pp. 830–838, Feb. 2000, doi : 10.1128/JCM.38.2.830-838.2000.
- [43] E.F. DeLong, G.S. Wickham, and N.R. Pace. "Phylogenetic stains: Ribosomal RNA-based probes for the identification of single cells," *Science*, Vol. 243, No. 4896, pp. 1360–1363, Mar. 1989, doi: 10.1126/science.2466341.
- [44] M.F. Clark, and A.N. Adams. "Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses," *J Gen Virol*, Vol. 34, pp. 475–483, Mar.1977, doi : 10.1099/0022-1317-34-3-475.
- [45] M.M. López, et al., "Strategies for improving serological and molecular detection of plant pathogenic bacteria," In : De Boer, S.H. (eds) *Plant Pathogenic Bacteria*, Springer, Dordrecht, pp. 83–86, 2001, doi : 10.1007/978-94-010-0003-1\_15.
- [46] P. Baldi, and N. La Porta. "Molecular approaches for low-cost point-of-care pathogen detection in agriculture and forestry," *Front Plant Sci*, Vol. 11, p.570862, Oct. 2020, doi : 10.3389/fpls.2020.570862.
- [47] F. Martinelli, et al., "Advanced methods of plant disease detection. A review," *Agron Sustain Dev*, Vol. 35, pp. 1-25, Sep.2014, doi: 10.1007/s13593-014-0246-1.
- [48] I. S. Fotiou, P.G. Pappi, K.E. Efthimiou, N.I. Katis, and V.I. Maliogka, "Development of one-tube real-time RT-qPCR for the universal detection and quantification of Plum pox virus (PPV)," *J Virol. Methods*, Vol.263, pp. 10–13, Oct.2018, doi : 10.1016/j.viromet.2018.10.006.
- [49] X. Zhong, L. Xue-lu, L. Bing-hai, Z. Chang-yong, and W. Xue-feng, "Development of a sensitive and reliable droplet digital PCR assay for the detection of *Candidatus Liberibacter asiaticus*," *J. Integr. Agric.*, Vol.17, No.2, pp. 483–487, 2018, doi : 10.1016/S2095-3119(17)61815-X.
- [50] J.A.Tomlinson, et al. "On-site DNA extraction and real-time PCR for detection of *Phytophthora ramorum* in the field," *Appl. Environ. Microbiol.*, Vol. 71, pp. 6702–6710, Nov. 2005, doi : 10.1128/AEM.71.11.6702-6710.2005.
- [51] T.M. Voegel, and L.M. Nelson, "Quantification of *Agrobacterium vitis* from grapevine nursery stock and vineyard soil using droplet digital PCR," *Plant Dis.*, Vol. 102, No.11, pp. 2136-2141, Sep. 2018, doi : 10.1094/PDIS-02-18-0342-RE.
- [52] S. Ramesh et al. "Plant disease detection using machine learning," *2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C)*, pp. 41-45, IEEE, Apr. 2018.
- [53] H.T. Sihotang, "Sistem pakar untuk mendiagnosa penyakit pada tanaman jagung dengan metode bayes," *Journal of Informatic Pelita Nusantara*, Vol. 3, No. 1, pp. 17-22, 2018.
- [54] A. Muchtar, D. Nur, E. Tungadi, and M.N.Y Utomo, "Perancangan Back-End Server Menggunakan Arsitektur Rest dan Platform Node. JS (Studi Kasus : Sistem Pendaftaran Ujian Masuk Politeknik Negeri Ujung Pandang)," *Seminar Nasional Teknik Elektro dan Informatika (SNTEI)*, pp. 72-77, Oct. 2020.
- [55] S. Hossain, et al. "Recognition and detection of tea leaf's diseases using support vector machine," In *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 150-154, IEEE, Mar. 2018.
- [56] J. Chen, Q. Liu, and L. Gao, L., "Visual tea leaf disease recognition using a convolutional neural network model," *Symmetry*, Vol. 11, No.3, p. 343, Mar.2019, doi : 10.3390/sym11030343.
- [57] X.Sun, S. Mu, Y. Xu, Z. Cao, and T.Su. "Image recognition of tea leaf diseases based on convolutional neural network," *2018 International Conference on Security, Pattern, Analysis, and Cybernetics (SPAC)*, 2018, pp. 304-309, doi : 10.1109/SPAC46244.2018.8965555.
- [58] A.J. Rozaqi, A. Sunyoto, and M.R. Arief, "Deteksi Penyakit Pada Daun Kentang Menggunakan Pengolahan Citra dengan Metode Convolutional Neural Network," *Creat. Inf. Technol. J.*, Volume 8, No.1, pp. 22-31, Mar. 2021, doi : 10.24076/citec.2021v8il.263.
- [59] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, *arXiv preprint arXiv:1404.5997*. [Online]. Available <https://arxiv.org/abs/1404.5997>.
- [60] Y. He, G. Kang, X. Dong, Y. Fu, and Y. Yang, "Soft filter pruning for accelerating deep convolutional neural networks," 2018, arXiv:1808.06866. [Online]. Available: <http://arxiv.org/abs/1808.06866>.
- [61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [62] G. Wang, Y. Sun, and J. Wang, "Automatic image-based plant disease severity estimation using deep learning". *Comput. Intell. Neurosci.*, Vol.2017, pp. 1–8, Jul. 2017, doi : 10.1155/2017/2917536.

- [63] K.P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput Electron Agric*, Vol. 145, pp. 311-318, Feb. 2018, doi: 10.1016/j.compag. 2018.01.009.
- [64] A. Fuentes, D.H. Im, S. Yoon and D.S. Park "Spectral analysis of CNN for tomato disease identification," In *International Conference on Artificial Intelligence and Soft Computing*, pp. 40 – 51, Springer, Cham, 2017.
- [65] S.J. Divinely, K Sivakami, and V. Jayaraj,"Fish diseases identification and classification using machine learning," *Intl. J. Adv. Res. Bas. Eng. Sci.Tech. (IJARBEST)*, vol. 5, pp. 46–51. 2019.
- [66] P. Jayanthi," Machine learning and deep learning algorithms in disease prediction: future trends for the healthcare system,"*In Deep Learning for Medical Application with Unique Data*, pp. 123-152. 2022.
- [67] V. Lyubchenko, R. Matarmeh, O. Kobylin, and V. Lyashenko,"Digital image processing techniques for detection and diagnosis of fish diseases," *Intl. J. Of Adv. Res. in Com. Sci. and Soft. Eng.*, vol. 6, pp. 79–83. 2016.
- [68] H. Chakravorty, P. Rituraj P., and P. Das," Image Processing Technique to Detect Fish Disease," *Intl. J. of Com. Sci. and Sec. (IJCSS)*, vol. 9, no. 2, pp. 121-131. 2015.
- [69] T.K. Malik, Shaveta, and A.K. Sahoo,"A novel approach to fish disease diagnostic system based on machine learning," *Adv. in Im. and Vid. Proc.*, vol. 5, no. 1, pp. 49–49. 2017.
- [70] M. Alagappan and M. Kumaran, "Application of expert systems in fisheries sector – a review," *Res. J. Anim. Vet. Fish. Sci.*, vol. 1, no. 8, pp. 19–30. 2013.
- [71] D. Li, Z. Fu, and Y. & Duan," Fish-Expert: a web-based expert system for fish disease diagnosis," *Expert System Application*, vol. 23, no. 3, pp. 311–320. 2022.
- [72] M. Føre, K. Frank, T. Norton, E. Svendsen, J.A. Alfredsen, T. Dempster, H. Eguiraun, W. Watson, A. Stahl, L.M. Sunde, C. Schellewald, K.R. Skøien, M.O. Alver, and D. Berckmans," Precision fish farming: a new framework to improve production in aquaculture," *Bios. Eng.*, vol. 173, pp. 176–193. 2018.
- [73] M.S. Ahmed, T.T. Aurpa, and M.A.K. Azad," Fish Disease Detection Using Image Based Machine Learning Technique in Aquaculture," *J. of King Saud Univ. – Com. and Inf. Sci.* 2021. doi: <https://doi.org/10.1016/j.jksuci.2021.05.003>.
- [74] D.Klauser "Challenges in monitoring and managing plant diseases in developing countries," *J Plant Dis Prot*, Vol. 125, No.3, pp. 235-237, Jan. 2018, .doi :/10.1007/s41348-018-0145-9.
- [75] M.Sharif,et al., "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *J Exp. Theor. Artif. Intell*, Vol. 33, No.4, pp.577-599, Feb, 2019, doi : 10.1080/0952813X.2019.1572657.
- [76] K.Thenmozhi, And U.S. Reddy, "Crop pest classification based on deep convolutional neural network and transfer learning," *Comput. Electron. Agric*, Vol.164, p.104906, Aug. 2019, doi : 10.1016/j.compag.2019.104906.
- [77] S. Iqbal, M.U. Ghani, T.Saba, and A. Rehman, "Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN)," *Microsc. Res. Tech*, Vol.81, No.4, pp. 419-427., Jan. 2018, doi : 10.1002/jemt.22994.
- [78] A. Camargo, and J.S. Smith, "An image-processing based algorithm to automatically identify plant disease visual symptoms," *Biosyst. Eng*, Vol. 102, No.1, pp.9-21. Jan.2009, doi : 10.1016/j.biosystemseng.2008.09.030.
- [79] S.D.Khirade and A.B. Patil"Plant Disease Detection Using Image Processing," 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 768 -771, doi: 10.1109/ICCUBEA.2015.153.
- [80] S. Kusrini,"Sistem pakar teori dan aplikasi," Andi offset, Yogyakarta. 2006.
- [81] W. Wang, M. Yang, M., and P.H. Seong,"Development of a rule-based diagnostic platform on an object-oriented expert system shell," *Annals of Nuc. En.* vol. 88, pp. 252-264. 2016.
- [82] R.R. Al Hakim, "Pencegahan Penularan Covid-19 Berbasis Aplikasi Android Sebagai Implementasi Kegiatan KKN Tematik Covid-19 di Sokanegara Purwokerto Banyumas," *Commun. Eng. and Emerg. J. (CEEJ)*, vol. 2, no.1, pp. 7–13. 2020.
- [83] R.R. Al Hakim, E. Rusdi, E., and M.A. Setiawan,"Android based expert system application for diagnose covid-19 disease: cases study of banyumas regency," *J. of Intell. Com. & Health Inf.*, vol. 1. No.2, pp.1–13. 2020.
- [84] S. Kusumadewi,"Artificial Intelegence (Teknik dan Aplikasinya)," Yogyakarta: Graha Ilmu. 2003.
- [85] T.S. Saptadi and V.S. Sebukita,"Pengambilan keputusan dalam penerimaan karyawan bank dengan pendekatan terstruktur berbasis sistem pakar," *J. Tek. Kom. dan Inf.*, p. 81. 2012.
- [86] G. Engin, B. Aksoyer, M. Avdagic, D. Bozanli, U. Hanay, d. Maden, and G. Ertek,"Rule-based expert systems for supporting university students," *Proc. Com. Sci.*, vol. 31, pp. 22-31. 2014. DOI: 10.1016/j.procs.2014.05.241.
- [87] M. Arhami,"Konsep dasar sistem pakar," Penerbit Andi. Yogyakarta. 205 p. 2004.
- [88] I. Akil,"Analisa efektifitas metode forward chaining dan backward chaining pada sistem pakar," *J. Pilar Nusa Man.*, p. 13. 2017.
- [89] A. Al-Ajlan, A., "The comparison between forward and backward chaining. international journal of machine learning and computing, "5, 2nd ser. 2015
- [90] D. Novaliendry, and C.H.Y. Yang," The expert system application for diagnosing human vitamin deficiency through forward chaining method," *Inf. and Comm. Tech. Conv. (ICTC)*, pp. 53-58. 2015. DOI: 10.1109/ICTC.2015.7354493.
- [91] I. M. Shofi, L.K. Wardhani, and G. Anisa, " Android Application for Diagnosing General Symptoms of Disease Using Forward Chaining Method," *Cyber and IT Service Management*, Bandung, Indonesia, 25-27 April. 2016. DOI: 10.1109/CITSM.2016.7577588.
- [92] Elfani and A. Pujiyanta," Sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website. sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website," vol. 1, no. 1, pp. 42–50. 2013.
- [93] T.H. Yunianto, "Sistem pakar diagnosa penyakit pada ikan hias," pp. 17. 2013.
- [94] David," Sistem pakar diagnosa penyakit ikan lele dumbo. konferensi nasional sistem & informatika," STMIK STIKOM Bali, 9 – 10 Oktober. 2015, pp. 107-112.
- [95] M.N. Rachmatullah and I. Supriana,"Low Resolution Image Fish Classification Using Convolutional Neural Network 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA) pp 78-83. 2018.
- [96] Z. Hakim and R. Rizky,"Sistem pakar diagnosis penyakit ikan mas menggunakan metode certainty factor di upt balai budidaya ikan air tawar dan hias kabupaten pandeglang banten," *J. Tek. Inf. Unis*, vol. 7, no.2, pp.164–169. 2020.
- [97] T.S. Dewi and R. Arnie," Sistem Pakar Diagnosa Penyakit Ikan Patin Dengan Metode Certainty Factor Berbasis Web," *J. TIMES*, vol. 6, no. 1, pp. 1311–1448. 2017.
- [98] P.I. Hidayati," Penerapan metode cf (certainty factor) pada diagnosa penyakit ikan nila," *Teknologi Informasi*, vol 8, no.2, pp. 127–134. 2017.
- [99] S. Budi," Kombinasi metode forward chaining dan certainty factor untuk mendiagnosa penyakit pada ikan cupang," *Tek-Sis. Inf. Unus. PGRI Kediri*, vol. 1, no.1, pp. 1–6. 2017.
- [100] Lestari," Penerapan Metode Certainty Factor Pada Sistem Pakar Diagnosa Penyakit Ikan Gourami Berbasis Website (Studi kasus UPTD Balai Benih Kota Binjai)," Thesis. Universitas Pembangunan Panca Budi Medan, pp 77. 2019.
- [101] R.R. Al Hakim, A. Pangestu, and A. Jaenul., A. 2021. Penerapan metode certainty factor dengan tingkat kepercayaan pada sistem pakar dalam mendiagnosis parasit pada ikan," *J. of Inf. Tech. Res.*, vol.2 no.1, pp. 27-37. 2021.



# Chapter - 11

## State-of-the-Art Analysis and Research Direction towards Secure Mobile Edge Computing in Transport System

Atul Anil Kumar Kumbhar<sup>1</sup>, Dr G Manjula<sup>2</sup>, Dr Roopa R Kulkarni<sup>3</sup>, Dr. Prashant P. Patavardhan<sup>4</sup>

<sup>1</sup> Research Scholar, DSATM, Research Centre, Karnataka, India

<sup>2</sup> Associate Professor, Dept. of ISE, DSATM, Karnataka, India

<sup>3</sup> Associate Professor, Dept. of ECE, SATM, Kerala, India

<sup>4</sup> Professor, Dept. of ECE, DRVITM, Karnataka

Email: <sup>1</sup> [atulkumbhar.edu@gmail.com](mailto:atulkumbhar.edu@gmail.com), <sup>2</sup> [manjula-ise@dsatm.edu.in](mailto:manjula-ise@dsatm.edu.in), <sup>3</sup> [roopakulkarni-ece@dsatm.edu.in](mailto:roopakulkarni-ece@dsatm.edu.in),  
<sup>4</sup> [prashantpp.rvitm@rvei.edu.in](mailto:prashantpp.rvitm@rvei.edu.in)

**Abstract** - Mobile Edge Computing is the current paradigm in transportation systems (MEC). In order to demonstrate certain significant paradigm capabilities to visit nearby destination sites, this computing approach is simulated. With the use of network apps and services, this strategy exchanges information with the least amount of delay possible while displaying real-time capabilities that are immediately available. Researchers have developed an intelligent framework that makes use of cutting-edge applications for deploying strategies, constructing architecture, and creating communication methodologies by merging the efficient transport system with MEC. In order to satisfy this need, this chapter covers the several traditional ways for integrating MEC in vehicle networks. This criterion suggests that an intelligent transportation system should be developed in the context of future smart cities.

In order to provide the best performance, security is a major concern in MEC-based automotive systems. Security precautions are managed via a Cyber-Physical Transportation System (CPTS), which combines a large number of sensors and wireless mobile devices. Sensing, communications, and traffic control are all things it is capable of. Maintaining the heterogeneous nature of variables as traffic sensors in Vehicular Ad Hoc Networks (VANET) while taking into account a diversity of abilities necessitates the use of the CPTS of MEC technique. Furthermore, the connected cars are used in the real-time application execution and application with respect to the edge node as the network edge for computational reasons. Due to its secure service deployment for autonomous vehicles, the Internet of Things (IoT), and Internet of vehicles, researchers use MEC for a variety of applications. The edge of networks have been deployed using MEC because to its properties and without the use of terminal servers.

However, only a few research studies have been systematically used for MEC deployment. Secure service design settings with MEC are also somewhat uncommon. Therefore, using intelligent ways to analyse numerous secured MC research is necessary. Here, we cover some of the challenging and unresolved issues surrounding the secure design of mobile services in edge computing. The first problem is a significant security restriction on secure access control systems. Second, as information and new services continue to proliferate and develop online, security difficulties with data transfer have arisen. These challenges include high traffic volumes, scalability issues, and other issues. Thirdly, the widespread use of in-car MEC could lead to improper exploitation of vehicle position data. Additionally, a subsequent study will investigate the development of MEC in more challenging contexts and use the acquired information in a variety of application areas. Additionally, in a vehicular edge-based environment, it is necessary to supply a wide range of comprehensive middleware solutions to support a variety of classes of communications between the cloud server and the sensing layer. Last but not least, a number of unresolved concerns are not included in the current studies that examine potential future study directions.

**Keywords** - Mobile Edge Computing (MEC), Vehicular Ad Hoc Networks (VANETs), Vehicular Cloud Computing (VCC), Edge Computing Vehicles (ECVs), Edge Cloud Computing (ECC), Vehicles-to-Everything (V2X) communication system, Edge Content Delivery and Update (ECDU)

### I. INTRODUCTION

In today's date, Mobile Edge Computing (MEC) has become the emerging paradigm in transport systems. This computing technique is simulated to bring out some significant paradigm capabilities to access the nearby destination points. Due to this process, it has the key benefits of exhibiting the easily accessible resource of real-time traits and achieving less latency measure for sharing the information via network applications and services. Several scholars have developed the intelligent framework by superimposing the effective transport system and MEC, where it exploits the novel application for deploying the techniques, constructing the architecture and communication methodologies.

---

© 2022 Technoarete Publishing

Atul Anil Kumar Kumbhar – “State-of-the-Art Analysis and Research Direction towards Secure Mobile Edge Computing in Transport System”

Pg no: 137 – 144.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch011>

In order to meet this requirement, this chapter explores the different traditional approaches to incorporate the MEC in vehicular networks, which implies to design an intelligent transport system in smart cities in future perspective.

Smart cities in the name itself it defines the connection of mobile devices for vehicle communication, to perform the data communication over the smart network, etc. The major intention of intelligent transport model is to improve the experiences in terms of driving, transportation and measures of traffic safety. In this way of context, Vehicular Ad Hoc Networks (VANETs) is one of the prime techniques in research world. This type of network is constructed with the enormous number of vehicles with large scale of data traffic. Nevertheless, it subsists with remarkable constraints due to the problems arises by flexibility, connectivity, intelligence and scalability. Over the past recent years, Vehicular Cloud Computing (VCC) is eminently involved to grasp the advantages of cloud computing that joints with vehicle services of VANET. Thus, the combination of vehicular network infrastructure and cloud computing has an objective of providing the dynamic real-time applications by forecasting traffic incidents and adapting the cloud environmental changes. To exhibit efficient performance, MEC analyses the results in terms of supporting the real time access, less latency and more bandwidth, which can be achieved by giving the network services. On the contrary, MEC is interconnected with number of vehicles to explore the rapid response once the request has been sent that can be performed like a smart mobile device, which comes under the future generation of computing techniques. Albeit, VANETs suffers with the inclusion of traffic data, it leads to degrade the performance like low scalability measure and intelligence, less reliable and poor connectivity. Also, while VANETs are distributing the services with the cloud computing environment, it also struggles with far apart places of mobile vehicular nodes and cloud servers. It is caused since the occurrence of delay in the transmission process and inadequate capacity. To mitigate this problem, this work explores several MEC techniques for intelligent system models of transport systems in smart world.

## II. SECURE MOBILE EDGE COMPUTING IN TRANSPORT SYSTEM

In MEC based vehicular system, security plays a major concern to provide the effective performance. By governing the security measure, a Cyber-Physical Transportation System (CPTS) is employed, where enormous wireless mobile devices and large set of sensors are involved. It has the potential to attain communications, traffic control and sensing. Consideration of various abilities, the CPTS of MEC approach is entailed for ensuring the heterogeneous nature of factors as traffic sensors in VANET. Moreover, the connected vehicles are involved to execute the application in a real-time manner and applied with respect to the edge node as network edge for computing purpose.

Over the past one decade, the security issue has given the more attention for research developers, standard organization, academic institutions and industry sector. Nevertheless, some former implemented methodologies are in need of achieving more power for computing and so on. Yet, such models are confined with the performance enhancement due to insufficient computing energy and less number of vehicle nodes or sensors. To alleviate these challenges, some inventor has developed the model of physical-layer security for MEC in transport system. Therefore, the security becomes the major issue while developing the intelligent transportation systems that should be taken into the account of addressing when conducting various intelligent models for MEC related network.

## III. RESEARCH CHALLENGES

Over the past few years, owing to provide the secure services, MEC can be applicable for distinct applications of deploying the techniques with Internet of Things (IoT), autonomous vehicles and Internet of vehicles. Also, due to the technical features of MEC along with destined servers, MEC has been employed with respect to the edge of the networks and edge node of vehicle network. Some literature work explains that the MEC has been used for real-time applications, but still it exists with open challenging issues. Additionally, various approaches use different kind of algorithms and validate the effectiveness regarding distinct measures. Consideration of multiple research papers of MEC faces such complications and challenging issues, which is to be resolved and provoked a better design in future direction. Some of the common issues of MEC in transport system are sequenced as below.

- The MEC-based system model comprises with myriads nodes and large scale of cloud resources, it causes computational issue and structural complexity.
- Another issue is maintaining the security level when transmitting the data from source to destination. Without any authorization or authentication, the malicious node can be easily making an entry in the network and act as a legitimate user to send or receive the information.
- Since the entailment of malware activities, the system performance gets degraded in terms of reliability, scalability, predictive results when attack occurs, etc. Also, traffic data is one of the reasons to create an impact of inefficiency of the model.
- In VANET network, many vehicles are participated to transmit the information, Due to the congestion of vehicles,

MEC tends to offer some misclassification results, and any vehicle can misuse the information.

- While cloud computing is merged with MEC of VANET, both becomes fragile by having some naked limitations occur over the cloud servers and edge related information of vehicular network.
- These are the issues to be taken into steps for resolving the future development of MEC in transport system and also integration of cloud computing with MEC.

#### IV. EXISTING RESEARCH WORKS ON MEC IN TRANSPORT SYSTEMS

In 2020, Cui *et al.* [1] have developed the model for data preserving and energy efficient with the concept inference of edge computing in VANET network. Here, Road Side Unit (RSU) was used to monitor the encrypted requests from the vehicle nodes with the help of private key information. Subsequently, the RSU has received the data from neighbor vehicles, termed as Edge Computing Vehicles (ECVs). When the vehicles were in need of getting the data, it could be able to download from the ECVs, where the model has improved the downloading efficiency. Thus, the performance was validated and its results were assessed regarding security level. On the contrary, the suggested method has outperformed the efficient performance of network.

In 2019, Ismail *et al.* [2] have instigated the “Active Queue management-based green Cloud model for Mobile edge computing (AGCM)”, in which the mobile users were involved with less latency and energy consumption. It was also used to eliminate the congestion problem in the cloud network by storing the data packets securely that was then distributed over the nodes. The proposed work was simulated using NS2 and its experimental results have ensured to provide a higher performance. Hence, the proposed work has attained the lower latency value of 65% and more throughput of 42% was increased rather than traditional approaches.

In 2021, Vaiyapuri *et al.* [3] have implemented the CBR-ICWSN by incorporation of Cluster Based Routing (CBR) protocol with Information Centric Wireless Sensor Networks (ICWSN) in IoT sector. In order to provide optimal value, Black Widow Optimization (BWO) was enhanced to choose the optimal Cluster Heads (CHs). In addition to this, an Oppositional Artificial Bee Colony (OABC)-aided routing was infused to select the optimal path for data transmission. Therefore, the CBR-ICWSN has obtained the desired outcome regarding energy efficiency and network lifetime against conventional methodologies.

In 2020, Li *et al.* [4] have developed the novel criterion methods for retaining the energy efficiency among Electric Vehicles (EVs). Moreover, an "effective charging information dissemination algorithm" was utilized to resolve the multi-objective issue. Then, the technique of local relying was used to increase the efficiency and lessen the overhead. Thus, the proposed model has improved the efficiency level in VANET.

In 2018, Al-Badarnah *et al.* [5] have presented the Software-Defined Edge Computing for VANET, where the services were done between the vehicles. These services were made with low latency of network for V2I and V2V communications. Hence, the simulated results have demonstrated that the novel technique has exploited the impressive results.

In 2020, Noorani and Seno [6] have processed the inter-vehicle communications to share the data packets. To meet this, fog-based computing and Software-Defined Networks (SDN) were deployed. An improved routing algorithm was mainly used to securely transmit the data packets. In this way, SDN and Fog computing-based Switchable Routing (SFSR) has offered the best path for data transmission. When the data packet was not supposed to transmit through VANET, it was transmitted via internet. Hence, the SFSR method has exhibited the higher results regarding delay, routing failure rate, packet loss and packet delivery ratio and routing overhead.

In 2020, Huang *et al.* [7] have recommended the task offloading scheme for performing the uplink transmission with respect to packet drop value and energy consumption. Further, the "computation resource allocation scheme" was employed to allocate the resource for MEC. By combining computation resource allocation and dynamic task offloading, a new model was developed as Lyapunov-based dynamic offloading decision algorithm to reduce the utility function of network. Finally, the findings have revealed that it has provided the better performance.

In 2015, Fathian *et al.* [8] have implemented the Improved Ant System-based Clustering algorithm (IASC1 and IASC2) in VANET. It was simulated by using the most commonly used clustering techniques. Thus, the findings have proved that the proposed work has achieved higher stability of the network model and low runtime to improve the computing performance in VANET.

In 2020, Lie *et al.* [9] have introduced the novel blockchain techniques of Named Data Networking (NDN) with Vehicular Edge Computing (VEC). The most effective consensus algorithm was developed in blockchain model to carry out the experimentation and validate the results regarding cache poisoning defense schemes, key management protocols and access control strategies. Thus, the suggested blockchain has improved the security performance when compared to classical methods.

In 2021, Sun and Samaan [10] have included the novel techniques of optimizing the signal control to traffic intersections while constructing the VCC infrastructure. The model of diffusion approximation was deployed to analyse the traffic flow in road map. Thus, by the optimal selection of control parameters to monitor the traffic flows, the suggested approach has

provided the expected outcome to improve the performance.

In 2019, Ding *et al.* [11] have developed the new framework of Edge Content Delivery and Update (ECDU) for MEC approach. Here, it has combined both the "Edge Content Delivery (ECD) and Edge Content Update (ECU) schemes". In ECDU, ECD was used to provide the content to cache pool that related to cloud data whereas ECU has reduced the content with respect to the frequency and ranking of cache pool. Hence, the extensive results have ensured that the model has exhibited the efficient network performance.

In 2015, Huang *et al.* [12] have resolved the problem of V2V2I VANET by using k-hop-limited offloading mechanism. Since the vehicles were in the form multi-hop communication, MEC has garnered the context of vehicles and given to offloading scheme. Thus, the novel MEC-assisted offloading has outperformed the performance rather than other existing works.

In 2020, Zhang *et al.* [13] have considered the bandwidth of V2V and V2I link to achieve the less average delay in the network. In addition to this, edge cooperative cache algorithm was included. The developed cache algorithm was used to minimize the delay by increasing the convergence rate. This was accomplished by sequential allocation of bandwidth in space. Finally, the experimental outcome has delivered the efficient performance regarding delay and convergence.

In 2019, Chen *et al.* [14] have presented the adaptive framework of MEC in VANET. Initially, the offloading scheme was used to define the mobile edges and devices in cloud environment. Due to the adaptive nature of model, it was also applicable for real-time implications. The performance was analysed with measures like response time and energy consumption. Thus, the novel framework has exploited the better results regarding less response time of 8 to 50% and 9 to 51% of energy consumption.

In 2019, Wang *et al.* [15] have reduced the network overhead by implementing the Vehicular User (VU), in which the optimization was carried out for allocating the communication and computing resources. Initially, the problem was decomposed into equivalent level of two problems and also using the low-complexity method to render the optimal best results. Finally, the simulated results have provided the effective results over existing approaches.

In 2019, Haitao *et al.* [16] have developed the optimization-assisted multipath transmission to handle the issue of complexity of edge node as the Virtual Machine (VM) migration was happened at the edge nodes. In the first hand, it was used to select the node that has closest to vehicles, where the response time was high when it became more than capacity. On the second hand, depends on the response time, the performance was improved with cloud edge nodes. Since the resources have various sizes, it was placed according to the VM via convex optimization. Hence, it has ensured that it has effectively reduced the average response time.

In 2022, Wu *et al.* [17] have included the flight trajectory algorithm for Unmanned Aerial Vehicles (UAVs) that relied on traffic awareness. The Cloud Computing Center (CCC) has determined the real-time traffic situations in the consideration of various mission regions. The optimization algorithm was mainly used to reduce the UAVs cost. Additionally, Deep Reinforcement Learning (DRL) was employed to raise the hovering position of UAV that was optimally happened. Finally, the performance results have proved that it has achieved the less consumption of energy.

In 2021, Liu *et al.* [18] have suggested the Real-time Distributed Strategy (RtDS) for preventing the issue of Multi period Offloading Problem (MOP), which was done by integration of vehicles and its edge nodes. Thus, the simulation outputs have shown the impressive results over other former techniques.

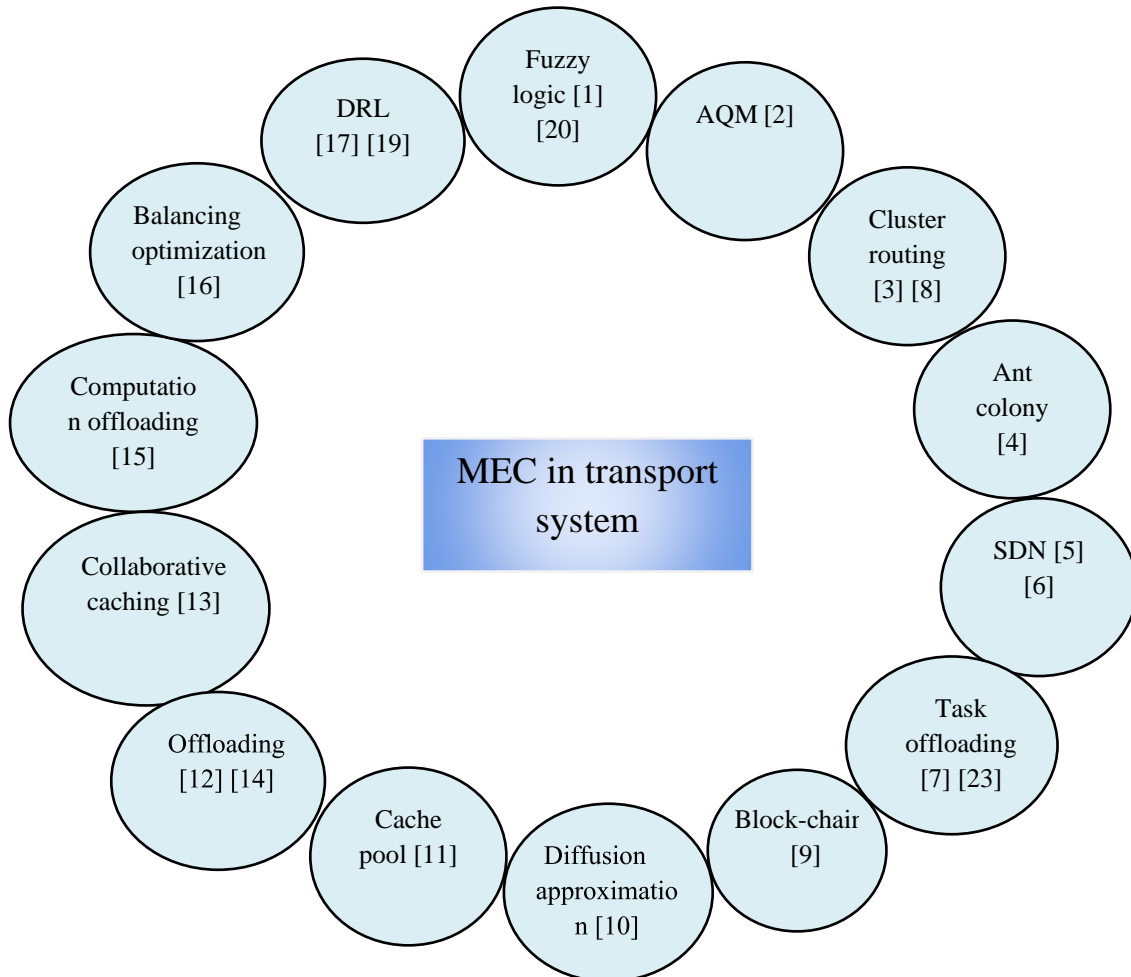
In 2020, Li *et al.* [19] have investigated the "collaborative edge computing framework" in MEC applications. Initially, Task Partition and Scheduling Algorithm (TPSA) was employed to schedule the tasks depends on the computation criteria. Further, an artificial intelligence (AI) was deployed to estimate the "task offloading, computing and result delivery policy for vehicles". This formulation was modeled by the Markov decision process and to determine the optimal solution with the aid of deep deterministic policy gradient. Due to the optimal results, the latency and penalty was reduced over the network, thereby, proposed work has delivered the superior performance.

In 2020, Zhang *et al.* [20] have recommended the privacy-preserving authentication framework by integrating the fifth Generation communication technology (5G). The suggested method was developed that depends on the inter-vehicle communication, where the transmission was happened between the vehicles in the name of device-to-device technology. Here, the most challenging factor was maintained the security level for the 5G-enabled technology. An authentication process was segmented into two divisions such as (a) fuzzy logic model was implemented to select the edge related vehicles for verifying the authentication and (b) mutual authentication was processed between normal and edge vehicles. Therefore, the information was shared among the vehicles in a secure way. Thus, the validation results have elucidated that the suggested model has effectively reduced the communication and computation overhead in contrast with other approaches.

## V. CLASSICAL ALGORITHMS USED FOR DESIGNING MEC IN TRANSPORT SYSTEMS

Different traditional algorithms or techniques have been implemented to validate the performance of MEC in transport system. Recently, various approaches are used for our literature works that is represented in Figure. 1. Such developed methodologies for MEC-aided transport systems are fuzzy logic [1] [20], active queue management [2], cluster-based routing

[3] [8], ant colony optimization [4], SDN [5] [6], task offloading [7] [23], blockchain [9], Diffusion approximation [10], Cache pool mechanism [11], offloading [12] [14], Collaborative caching [13], computation offloading [15], Balancing optimization [16] and deep reinforcement learning [17] [19].



**Figure 1:** Algorithm diagram of existing works of MEC in transport system

Thus, all the aforementioned algorithms are executed, and it tries to offer the effective network performance and also it ensures to improve the security level for MEC in transport system, where the data can be shared efficiently.

## VI. DISCUSSION ON SIMULATION PARAMETERS OF EXITING MEC IN TRANSPORT SYSTEMS

This section illustrates the network parameter analysis of MEC using “simulation area, number of vehicles and simulation time”. In [1], the area is fixed as  $2500 \times 2500(m^2)$ , vehicles as 50 and simulation time as 200s. In [2], the parameters like packets size, number of servers and packets sent are discussed. In [3], number of nodes is taken for performance. In [4], the network area is set as  $10 \times 9$  km and total vehicles as 20. The request content identifier and threshold for MEC nodes is taken in [5]. In [6], the simulation area is defined as  $3000 \text{ m} \times 2000 \text{ m}$  and its time as 300s. In [7], it includes the transmission power of vehicles, channel bandwidth and packet length. In [8], transmission range is considered. In [9], the simulation time as 10s and also contains the link bandwidth and delay. In [10], the number of vehicles as 50 is used. In [11], it has taken the five numbers of servers. In [12], simulation time is 200s. In [13], the factors like number of hidden units and its learning rate is fixed. In [14], the different locations are taken for performing the offloading schemes. In [15], the latency, coefficient of energy consumption and vehicle speed are utilized. The number of VM is used to set for simulation in [16]. The flying height, transmission power and frequency of UAV are considered in [17]. The metrics like communication radius, transmission power and task generated for each vehicle is given in [18]. In [19], the activation function and neuron count is taken for simulating the experimentation. In [20] the area is fixed as  $2500 \times 2500\text{m}$  and simulation time as 200s. These are all the simulated parameters that are used for analyzing the performance of MEC techniques, which will help the future researchers towards developing a



new model.

### VII. PERFORMANCE MEASURES CONSIDERED IN DEVELOPING MEC IN TRANSPORT SYSTEMS

The performance measure is mainly used to validate the system in terms of different kinds of metrics, which is given in Table. 1. Here, various techniques have been developed for MEC in transport system. Each and every system is analyzed with the measure like packet loss ratio, delay, overhead complexity, latency, energy consumption, throughput, energy efficiency and so on. These are all the metrics are involved to prove that the implemented work has provided the effective performance.

Citation	Packet loss ratio	Delay	Latency	Energy Efficiency	Packet delivery rate	Miscellaneous measures
[1]	✓	✓	-	-	-	Busy Time
[2]	-	-	✓	-	-	Throughput, Routing Load, Energy Consumption
[3]	✓	✓	-	✓	✓	Network lifetime
[4]	-	✓	-	-	✓	Overhead, Waiting time
[5]	-	-	✓	-	-	-
[6]	✓	✓	-	-	✓	Routing overhead, routing failure rate
[7]	-	-	-	✓	-	Packet drop rate, queue length
[8]	-	-	-	-	-	Run time, Average CH change
[9]	-	✓	✓	-	-	Throughput, storage overhead, calculation overhead
[10]	-	-	-	-	-	waiting time
[11]	-	-	-	-	-	Cost estimation
[12]	-	-	-	-	-	Offloading fraction, path life time, and average session time
[13]	-	✓	-	-	-	System overhead
[14]	-	-	-	-	-	Response time, energy consumption
[15]	-	-	✓	-	-	Computation overhead
[16]	-	✓	-	-	-	Response time,
[17]	-	-	-	-	-	RMSE, energy consumption, reward
[18]	-	✓	-	-	-	Task completion ratio, Reward
[19]	-	✓	-	-	-	Run time, service time
[20]	-	-	-	-	-	Computation overhead, verification delay

### VIII. FUTURE RESEARCH DIRECTION TOWARDS MEC

This section explores the future research direction towards MEC in transport systems and also integration of cloud computing with MEC. In generally, the MEC contains heterogeneous nature of stationary and mobile nodes like “sensors, actuators, home appliances, smart phones, and personal computers”. Being an edge node network, it has the leverage applications for various technologies such as “3G, 4G, 5G, Wi-Fi and private networks”, thus, the aspect here considers for analyzing the challenging issue, which stimulated to develop the future research work of MEC operations. Some of the challenging factors of MEC in transport system are as given below.

- The evolution of MEC brings some key advantages to perform the process in transport systems. But, owing to less awareness of network integration with reputed techniques, it fails to implement with mainly accounting of grasping the information related to edge nodes for computing. Thus, it leads to imprecise results that degrade the robustness of the system.
- Cost effective becomes another challenge. MEC model consumes with various nodes or sensors to process the system, but some industries cannot be able to afford it. Thus, the reduction of cost and applicable for real-time implications come under the category of challenging factor.
- MEC in transport system in the sense it belongs to the VANET network. Also, in the name itself, edge computing is often suffering by the edge nodes of the vehicles. Since such works exploit higher results, still it faces some issues like complexity of structural representation, computation and time complexity.



- Some other algorithms of MEC confuses with some parameters like delay, transmission time, security and privacy of data.
- The major challenging concern is security issue. Over the vast network, several attacks can be easily entered into the network to perform the data transmission.
- When VANET is combined with cloud computing model, it also faces some challenges like utilization of cloud resources in form malware activities, inadequate energy saving and privacy preserving data.

Accounting of all the aforementioned challenges, the major prerequisites of future direction of MEC in transport system is summarized as below.

- ✓ **Intermittent connectivity:** In MEC, the management and control of network connection between the edge nodes of vehicles are considered as future research development. Thus, the lack of intermittent connections is causing the impact of vehicle mobility and frequently packet loss occurred in vehicular environment, which is to be resolved in future.
- ✓ **Location awareness and high mobility:** In future work of MEC-related VANETs, it requires the information related to mobility nature and “location awareness” of participated vehicle nodes for data communication. Every vehicular node should occupy their respective position in the network to evade the inclusion of security threats.
- ✓ **Heterogeneous vehicle management:** Since the smart vehicles are differing with its own characteristics. The vehicle management is one of the future scopes of developing the MEC in transport systems.
- ✓ **Security:** When more number of nodes and cloud resources are involved, it will eventually rises the challenging problem of security issue. Always, there is a risk to preserve and secure the data. Thus, in future development, approaches like security protocols, encryption algorithms, and attack detection process are suggested to include. Also, in a vast environment, the vehicles can communicate with any other node without the awareness of it belongs to normal or attack node. Thus, security management is primarily focuses in future research trends.
- ✓ **Support of network intelligence:** The future work of MEC in VANET requires network intelligence to support the model. The network is to be supported in the form of more nodes relied on vehicles and collects the edge cloud data, which is in need of pre-processing the gathered data before transmission happens.
- ✓ **Real-world implications and low latency:** Some of the algorithms of MEC techniques fail to offer the low latency measure, which is the basic prerequisite for future VANETs with respect to real-world implications. Therefore, the future techniques should support the real-time applications with less latency level to achieve the effective performance.
- ✓ **High bandwidth:** In future, big traffic data leads to reduced bandwidth among the vehicles. This will be considered as future scope for MEC in VANET.
- ✓ **Connectivity:** Seamless connection is required to avoid the attack intrusion in the network, thus, future VANETs are in need of connecting all the participating vehicles and also in VCC, the connectivity is established among the cloud related vehicle nodes. Due to the reliable connection between the fog nodes of cloud networks and vehicles, it has the capacity to prevent transmission failures in the data communication process.

Hence, in future research consideration, MEC technique can be integrated with smart device applications of IoT and establish a new model like Vehicles-to-Everything (V2X) communication system and computing the techniques of Edge Cloud Computing (ECC) and so on. On the second, nowadays, the technologies are focusing on the approaches of machine learning and deep learning model. In general, the learning model is categorized into two ways: supervised and unsupervised learning. When the MEC is intruded with several attacks, several inventors have used the machine or deep learning technique to detect the attack and mitigate it from the network in an effective way. It can also reduce the limitation that is addressed from the existing works of MEC. Some learning models or classifiers as Recurrent Neural Network (RNN), Artificial Neural Network (ANN), Deep Neural Network (DNN), Naive Bayes (NB) classifier, Long-Short Term Memory (LSTM) and so on. Moreover, these learning techniques are also facing some restrictions to give the precise outcome. Recently, this will be handled by employing some heuristic algorithms to find the optimal results. Mainly, it is used to identify the optimal solution for parameter utilization related to features and machine or deep learning methods. Some heuristic algorithms are classified as swarm intelligence, evolutionary based algorithms, human-based and animal-based optimization. These are the scopes that this chapter suggests to improve the performance towards the future direction of MEC in transport systems.

## REFERENCES

- [1] J. Cui, L. Wei, H. Zhong, J. Zhang, Y. Xu and L. Liu, "Edge Computing in VANETs-An Efficient and Privacy-Preserving Cooperative Downloading Scheme," in *IEEE Journal on Selected Areas in Communications*, Vol. 38, No. 6, pp. 1191-1204, 2020.
- [2] Alshimaa H. Ismail, Nirmeen A. El-Bahnasawy and Hesham F. A. Hamed, "AGCM: Active Queue Management-Based Green Cloud Model for Mobile Edge Computing", *Wireless Personal Communication*, Vol. 105, pp. 765-785, 2019.
- [3] Thavavel Vaiyapuri, Velmurugan Subbiah Parvathy, V. Manikandan, N. Krishnaraj, Deepak Gupta and K. Shankar, "A Novel Hybrid Optimization for Cluster-Based Routing Protocol in Information-Centric Wireless Sensor Networks for IoT Based Mobile Edge Computing", *Wireless Personal Communications*, 2021.

- [4] G. Li, X. Li, Q. Sun, L. Boukhatem and J. Wu, "An Effective MEC Sustained Charging Data Transmission Algorithm in VANET-Based Smart Grids," in *IEEE Access*, vol. 8, pp. 101946-101962, 2020.
- [5] Jafar Al-Badarnah, Yaser Jararweh, Mahmoud Al-Ayyoub, Ramon Fontes Mohammad Al-Smadia and Christian Rothenberg, "Cooperative mobile edge computing system for VANET-based software-defined content delivery", *Computers & Electrical Engineering*, Vol. 71, pp. 388-397, 2018.
- [6] Naserali Noorani and Seyed Amin Hosseini Seno, "SDN- and fog computing-based switchable routing using path stability estimation for vehicular ad hoc networks", *Peer-to-Peer Networking and Applications*, Vol. 13, pp. 948-964, 2020.
- [7] Xiaoge Huang, Ke Xu, Chenbin Lai, Qianbin Chen and Jie Zhang, "Energy-efficient offloading decision-making for mobile edge computing in vehicular networks", *EURASIP Journal on Wireless Communications and Networking*, No. 35, 2020.
- [8] Mohammad Fathian, Gholam Reza Shiran and Ahmad Reza Jafarian-Moghaddam, "Two New Clustering Algorithms for Vehicular Ad-Hoc Network Based on Ant Colony System", *Wireless Personal Communications*, Vol. 83, pp. 473-491, 2015.
- [9] Kai Lei, Junjie Fang, Qichao Zhang, Junjun Lou, Maoyu Du, Jiyue Huang, Jianping Wang and Kuai Xu, "Blockchain-Based Cache Poisoning Security Protection and Privacy-Aware Access Control in NDN Vehicular Edge Computing Networks", *Journal of Grid Computing*, Vol. 18, pp. 593-613, 2020.
- [10] P. Sun and N. Samaan, "A Novel VANET-Assisted Traffic Control for Supporting Vehicular Cloud Computing," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6726-6736, Nov. 2021.
- [11] Chuntao Ding, Ao Zhou, Jie Huang, Ying Liu and Shangguang Wang, "ECDU: an edge content delivery and update framework in Mobile edge computing", *EURASIP Journal on Wireless Communications and Networking*, No. 268, 2019.
- [12] Chung-Ming Huang, Shih-Yang Lin and Zhong-You Wu, "The k-hop-limited V2V2I VANET data offloading using the Mobile Edge Computing (MEC) mechanism", *Vehicular Communications*, Vol. 26, 2020.
- [13] Mu Zhang, Song Wang and Qing Gao, "A joint optimization scheme of content caching and resource allocation for internet of vehicles in mobile edge computing", *Journal of Cloud Computing*, Vol. 9, No. 33, 2020.
- [14] Xing Chen, Shihong Chen, Yun Ma, Bichun Liu, Ying Zhang and Gang Huang, "An adaptive offloading framework for Android applications in mobile edge computing", *Science China Information Sciences*, Vol. 62, No. 82102, 2019.
- [15] J. Wang, D. Feng, S. Zhang, J. Tang and T. Q. S. Quek, "Computation Offloading for Mobile Edge Computing Enabled Vehicular Networks," in *IEEE Access*, vol. 7, pp. 62624-62632, 2019.
- [16] Z. Haitao, D. Yi, Z. Mengkang, W. Qin, S. Xinyue and Z. Hongbo, "Multipath Transmission Workload Balancing Optimization Scheme Based on Mobile Edge Computing in Vehicular Heterogeneous Network," in *IEEE Access*, vol. 7, pp. 116047-116055, 2019.
- [17] Z. Wu, Z. Yang, C. Yang, J. Lin, Y. Liu and X. Chen, "Joint deployment and trajectory optimization in UAV-assisted vehicular edge computing networks," in *Journal of Communications and Networks*, vol. 24, no. 1, pp. 47-58, Feb. 2022.
- [18] Chunhui Liu, Kai Liu, Hualing Ren, Xincuo Xu, Ruitao Xie and Jingjing Cao, "RtDS: real-time distributed strategy for multi-period task offloading in vehicular edge computing environment", *Neural Computing and Applications*, 2021.
- [19] M. Li, J. Gao, L. Zhao and X. Shen, "Deep Reinforcement Learning for Collaborative Edge Computing in Vehicular Networks," in *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 4, pp. 1122-1135, Dec. 2020.
- [20] J. Zhang, H. Zhong, J. Cui, M. Tian, Y. Xu and L. Liu, "Edge Computing-Based Privacy-Preserving Authentication Framework and Protocol for 5G-Enabled Vehicular Networks," in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7940-7954, July 2020.

# Chapter - 12

## Knowledge Discovery and Intelligent Data Mining

Sasi Kumar M<sup>1</sup>, Sasi Kumar V<sup>2</sup>, Samyukthaa LK<sup>3</sup>, Gokul Karthik S<sup>4</sup>, Abirami A<sup>5</sup>, Lakshmanaprakash S<sup>6</sup>

<sup>1,2,3,4,5,6</sup> Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India

E-mail: <sup>1</sup> [sasikumarmurugan02@gmail.com](mailto:sasikumarmurugan02@gmail.com), <sup>2</sup> [sasikumarskvs@gmail.com](mailto:sasikumarskvs@gmail.com), <sup>3</sup> [samyusamyukthaa@gmail.com](mailto:samyusamyukthaa@gmail.com),  
<sup>4</sup> [gokulkarthik48@gmail.com](mailto:gokulkarthik48@gmail.com), <sup>5</sup> [abirarmia@bitsathy.ac.in](mailto:abirarmia@bitsathy.ac.in), <sup>6</sup> [lakshmanaprakashs@bitsathy.ac.in](mailto:lakshmanaprakashs@bitsathy.ac.in)

**Abstract**— Knowledge Discovery in Databases (KDD) is a programmed, exploratory examination and demonstrating of enormous information storehouses. KDD is the coordinated course of recognizing legitimate, novel, valuable, and justifiable examples from enormous and complex informational collections. Data Mining (DM) is the core of the KDD interaction. The model is utilized for figuring out peculiarities from the information, investigation and expectation. The bringing together objective of the KDD cycle is to extricate information from information with regards to huge data sets. KDD is a fantastic tool for keeping organizations and sectors up to date on consumer demands, behaviours, and actions. There are several obvious benefits to employing the KDD technique, as well as some drawbacks. The Intelligent Data Mining and Analysis is a trend setting innovation in data handling to remove rules and information from huge data sets deliberately and break down the nonlinear connection among info and result factors in complex issues or peculiarities. Data mining is a full-grown application in different regions like marketing. The proposed model comprises of extricating Twitter client information utilizing programmable bookkeeping sheet apparatuses like Google Docs. This content purposes the Twitter API alongside passing a bunch of boundaries, similar to client handle or hashtags to bring client metadata. In any case, straightforwardly utilizing Twitter API is additionally conceivable yet it builds the intricacy of the system.

**Keywords**— KDD Process, Data Mining, Extract Knowledge, Non-linear Connection, Growing technique, twitter client information

### I. INTRODUCTION

Data mining most well-known approach to sorting out huge educational assortments to recognize models and associations that can help with putting everything in order issues through data assessment. Data mining methodologies and instruments engage tries to expect future examples and seek after more-instructed business decisions.

Information mining is a significant piece of in general information examination and is one of the centre areas of information science that utilizations progressed logical procedures to track down helpful data in datasets. At an additional definite level, information mining is a stage in the information disclosure process in a data set (KDD), an information science technique that gathers, processes, and breaks down information [1]. Information mining and KDD are now and again alluded to reciprocally, yet they are frequently thought to appear as something else.

#### 1.1 Why Data Mining Is Necessary

Information mining is an urgent part of effective examination drives in associations. The data it produces can be utilized in business knowledge (BI) and progressed examination applications that include investigation of verifiable information, as well as constant investigation applications that look at streaming information as it's made or gathered.

Successful information mining supports different parts of arranging business methodologies and overseeing activities. That incorporates client resisting capabilities, for example, showcasing, publicizing, deals and client assistance, in addition to assembling, production network the board, money and HR (Human Resource). Information mining upholds misrepresentation location, risk the executives, online protection arranging and numerous other basic business use cases. It additionally assumes a significant part in medical services, government, logical exploration, science, sports and that's only the tip of the iceberg

The Intelligent Data Mining and Analysis (IDMA) is a trend setting innovation in data handling to extricate rules and information from huge data sets efficiently and examine the nonlinear connection among information and result factors in complex issues or peculiarities. Data mining is a full-grown application in different regions like promoting. It is anyway generally ongoing that the power business has shown an interest in these methods. Profound learning, one more current hotly debated issue is tended to in this Working Group.

### 1.2 How Does Data Mining Works

Information mining is ordinarily finished by information researchers and other talented BI (Business Intelligence) and investigation experts. Yet, it can likewise be performed by information keen business experts, chiefs and laborers who capability as resident information researchers in an association. Its centre components incorporate AI (Artificial Intelligence) and measurable investigation, alongside information the executive errands done to plan information for examination. The utilization of AI calculations and man-made consciousness (AI) devices has mechanized a greater amount of the interaction and made it more straightforward to mine gigantic informational indexes, for example, client data sets, exchange records and log documents from web servers, versatile applications and sensors [1].

1. **Data Gathering:** Relevant information for an investigation application is recognized and gathered. The information might be situated in various source frameworks, a data distribution centre or data lake, an undeniably normal store in enormous information conditions that contain a blend of organized and unstructured information. Outer data sources may likewise be utilized.
2. **Information Processing:** This stage incorporates a bunch of moves towards the information to be mined. It begins with information investigation, profiling and pre-handling, trailed by information purifying work to fix mistakes and different information quality issues. Information change is likewise finished to make informational collections predictable, except if an information researcher is hoping to examine unfiltered crude information for a specific application
3. **Mining the data:** When the data is ready, an information researcher picks the fitting data mining method and afterward carries out at least one calculation to do the mining. In AI applications, the calculations ordinarily should be prepared on example informational collections to search for the data being looked for before they're gone against the full arrangement of information.
4. **Data Investigation and Understanding:** The data mining results are utilized to make scientific models that can assist with driving navigation and other business activities. The information researcher or one more individual from an information science group likewise should convey the discoveries to business leaders and clients, frequently through information representation and the utilization of information narrating strategies.

### FOUR STAGES OF DATA MINING

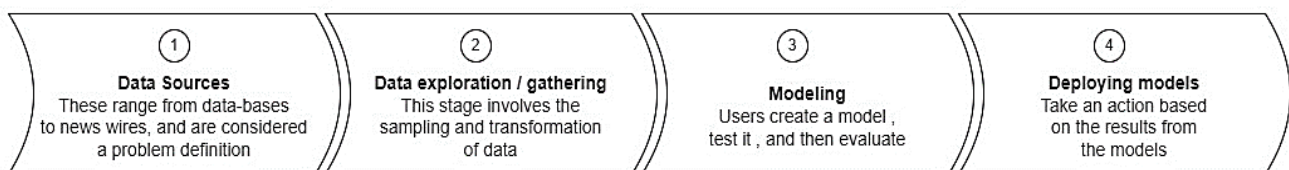


Figure 1. Four Stages of Data Mining

In this chapter we will discuss about the overview of knowledge discovery followed the architecture of the KDD and it's various aspects along with their uses and role in the discovery. Continuing with the various steps involved in the knowledge discovery such as data pre-processing, transformation, pattern evaluation etc., After this we will discuss about the intelligent data mining with a problem statement on Amber-Heard twitter and provide the result based on the sentimental analysis through SVM (Support Vector Machine) data mining & concluded by extracting the data using the bot sentinel tool that detect the spam accounts or bot accounts.

### 1.3 Data Mining methods:

The distinction is helpful for understanding the overall discovery aim even when the lines between prediction and description are not clearly defined (some of the predictive models can be descriptive, to the extent that they are intelligible, and vice versa). For different data mining applications, the relative weights of prediction and description can vary greatly. However, in KDD, description frequently takes precedence over prediction. In contrast, prediction is frequently the main objective in many machine learning and pattern recognition applications [1]. The following are the main data mining techniques that are used to accomplish prediction and description goals. Learning a function for classification involves mapping input items into one of

several predefined classes.

- **Regression:** Regression is the process of learning a function that converts a data point into a real-valued prediction variable and finding functional connections between variables.
- **Clustering:** Clustering is the process of selecting a small number of categories or clusters to explain the data. A methodology for estimating from data the joint multi-variate probability density function of all the variables/fields in the database, probability density estimation is closely related to clustering.
- **Summarization:** Finding a succinct description for a small set of data is known as summarization. Examples include using multivariate visualisation techniques and deriving summary or association rules.
- **Dependency Modelling:** Finding a model that captures important interdependencies between variables is known as "dependency modelling" (e.g., learning of belief networks).
- **Change and Deviation Detection:** Finding the biggest differences between the data and previously measured or observed values is known as "change and deviation detection."

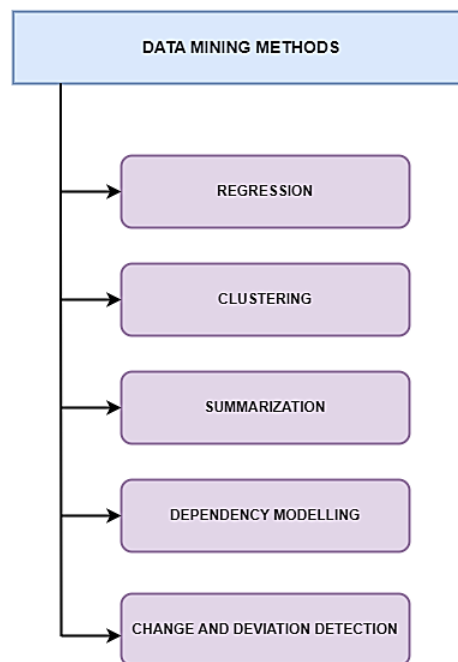


Figure 2. Data mining methods

#### 1.4 Components of Data Mining Algorithm

After outlining the fundamental data mining techniques, it is time to create the individual algorithms that will carry out these techniques. Any data mining technique may be broken down into three main parts: model representation, model evaluation, and search. This reductionist viewpoint isn't necessarily comprehensive or all-inclusive; rather, it's a practical technique to explain the essential ideas behind data mining algorithms in a style that's largely uniform and condensed.

- **Model Representation:** Model Representation is the terminology used to describe patterns that can be found. No amount of training time or instances will be able to create an accurate model for the data if the representation is too constrained. A data analyst must completely understand any representational presumptions that may be included in a given approach. Equally crucial is that an algorithm designer explicitly states the representational presumptions a certain algorithm makes. It is important to keep in mind that stronger representational power for models increases the risk of overfitting the training data, which lowers prediction accuracy on unobserved data.
- **Model Evaluation:** Model evaluation criteria are numerical claims (or "fit functions") that describe how well a specific pattern (a model and its parameters) satisfies the KDD process' objectives. For instance, the empirical prediction accuracy on a test set is frequently used to evaluate predictive models. The predictive accuracy, novelty, utility, and understandability of the fitted model for descriptive models can all be assessed.
- **Search Method:** The data mining problem has been reduced to a simple optimization work after the model representation (or family of representations) and the model evaluation criteria have been fixed: discover the parameters/models from the selected family that optimise the tile evaluation criteria. It consists of two components:



- 1) **Parameter Search:** In parameter search, the algorithm must look for the parameters that, given observed data and a defined model representation, optimise the model assessment criteria.
- 2) **Model Search:** The parameter search method is looped over to consider a family of models as the model representation is modified.

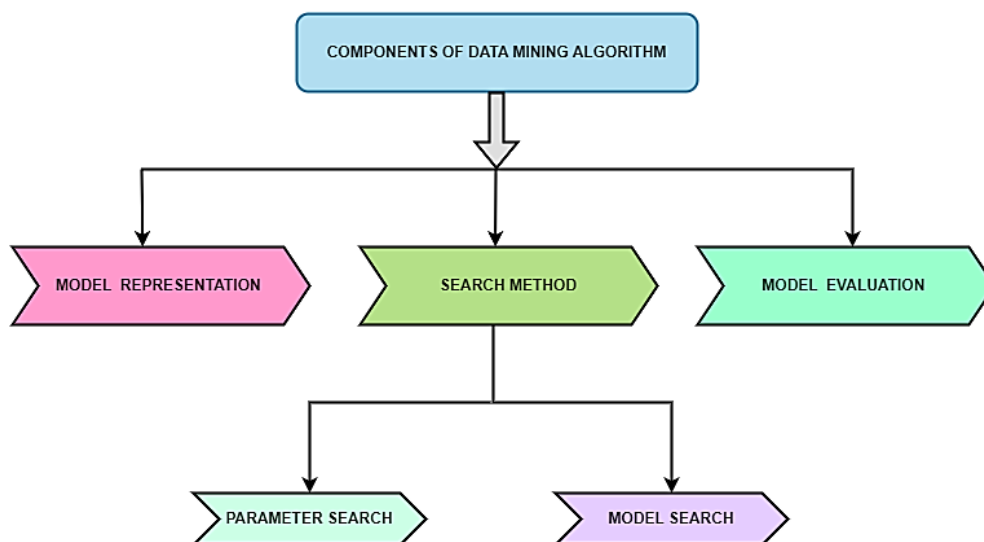


Figure 3. Components of Data Mining Algorithm

## II. LITERATURE SURVEY

Keeney, in his assessment communicated that affirmation of business and information mining (specific) targets is a basic piece of the KDDM cycle [2]. This section addresses the early phase of the KDDM interaction. Considering this reality, it is direct that shameful arrangement of objections can incite gambling with the entire KDDM project. Information mining writing and handle models see the vitality of this part, yet give no ways of managing completing it. We perceive a couple of systems proposed in the writing that can be used. In any case, we discuss regard focused thinking or VFT proposed as strategies for specifying objections and targets. Second, we discuss Shrewd methodology for figuring objections that is consistently proposed in the expert writing. [3] Berry and Linoff communicated that Programmed group location is used for seeing as huge designs in information. Grouping gives a way to deal with learn about the design of complicated information. When the right groups have been described, typically possible to find fundamental examples inside each bunch, as analyse the going with characteristics of programmed bunch recognition, in bunching, there is no pre characterized information and no refinement among independent and subordinate variables. In a greater sense, in any case, bunching can be an organized development since groups are searched for some business reason. In advancing, groups moulded for a business justification behind existing are commonly called "portions", and client division is a notable use of bunching. Programmed group identification is an information mining technique that is rarely used as a piece of withdrawal since finding bunches isn't consistently an end in itself. [4] Osei Bryson suggests a multi measures dynamic method for managing direct decision of the best choice tree from a broad game plan of choice trees. The suggested approach portrays such standards that could be used for evaluating the execution of choice trees and using them in a multi-rules dynamic construction to help assurance of the best mode and analyse that Delphi may be depicted as a methodology for coordinating a get-together correspondence process so the cycles practical in allowing a get-together of individuals, in general, to deal with a marvellous issue. [5] It is turned out to be a renowned gadget in IS look at in recognizing and sorting out issues for regulatory navigation. Delphi is in like manner relevant to the evaluation adventure of the KDDM cycle, as picked appraisal standards ought to be coordinated before information mining models can be picked. Information digging issue makes are generally organized into gathering, assessment, expectation, alliance standards, bunching and discernment. A fairly novel arrangement and organize issues in perspective on the showing assumption as (a) showing to understand, (b) exhibiting to request, and (c) showing to expect. Schenkerman investigated that AHP recommends breaking down an issue into a game plan of components, giving out mathematical loads or needs to those components, and taking a gander at changed decisions concurring in light of their scores on the picked set of components. These various choices would then have the option to be rank arranged to make an assurance. One of the primary characteristics of AHP (Analytical Hierarchy Process) is that it can get both emotional and furthermore target appraisal models. While AHP has been over a wide grouping of choice conditions, it isn't without criticism. Intellectuals of AHP have shown irregularity of results inferable from usage of abstract scales, rank reversals, and Instigation of Non-existent Request, etc. Verbal showdowns



between the critics and backers have similarly been displayed in the writing. Not with standing, AHP continues to be used as a common dynamic instrument by trained professionals and academicians, it has furthermore been participated in sort of the business programming Master Decision. [6] Inmon communicated in an investigation that an information Distribution Centre (DW) is a social occasion of composed, subject-arranged data sets expected to help the DSS (choice support) work, where each unit of information is non-volatile and pertinent to some moment in time. The Information Distribution centre contains functional information stores and information shops. The functional information store is the most generally perceived section of the DW condition. Its fundamental ordinary limit is to store the information for a singular, specific plan of functional applications. [7] The information shop is consistently viewed as a way to deal with get section into the space of information distribution centres and to commit all blunders on a more diminutive scale.

As the greater part of current philosophies of information mining investigate information in single information table. [8] However, as of late a large portion of these techniques are extended to social cases. Social data mining incorporates applying data mining approach on different table data for abstracting the data in it discuss appropriate Techniques and ways of thinking are expected in future to give food the necessities of data mining field as it is examining an always expanding number of amazing fields with the objective that we can research the such muddled conditions where data is massive yet is stacked with hidden away information.[9] This review paper noticed the usages of data mining techniques that fostered extra opportunity to help data the board communication as it is generally broadly used across various fields and each field is being maintained by discrete data mining methodology, it was shown that data mining can be integrated into data the board framework and work on that collaboration with dominating information and presented data mining as commonly vigorous and connecting with research locale which is procuring appeal in clinical region. Data mining gives a couple of benefits in clinical benefits space. It overhauls the clinical advantages in viable manner [10]. Anand. V Saurkar portrayed data mining as "interdisciplinary field which involves consolidated data bases, man-made thinking, AI, estimations thus forth.". They described data mining as multi-step process which contains availability of data for mining, mining computations, assessment of results and comprehension of results [11]. A few applications, tasks and issues associated with it have moreover been framed. Jawline AngWu present a model multifaceted connection mining structure, which with shrewd assistance through the assistance of the ontologies, can help clients with building significant data mining models, thwart inadequate model age, find thought extended oversees, and give a working data re-finding system [12]. Anastasia Giachanou examined about the key responsibilities of this survey consolidate the presentation of the proposed approaches for feeling assessment in Twitter, their request according to the system they use, and the discussion of late investigation examples of the point and its associated fields. [7] Apoorv Agarwal examined about the new components (connected with as of late proposed features) and the tree segment perform generally at a comparable level both beating the bleeding edge standard [13].

### III. OVERVIEW OF KDD PROCESS

The term KDD represents Knowledge Discovery in Databases. It implies to the wide system of finding knowledge discovery in information and stresses the significant level utilizations of explicit Data Mining methods. It is a field important to specialists in different fields, including man-made consciousness, AI, design acknowledgment, data sets, measurements, information obtaining for master frameworks, and information representation [14].

#### 3.1 KDD, Data mining and related to other fields

Generally, the thought of finding valuable examples in information has been given various names including information mining, information extraction, data disclosure, data collecting, information palaeontology, and information design handling. In our view KDD alludes to the general course of finding helpful information from information while information mining alludes to a specific move toward this cycle. Information mining is the utilization of explicit calculations for removing designs from information. The qualification between the KDD cycle and the information mining step (inside the interaction) is a main issue of this paper. The extra strides in the KDD cycle, for example, information readiness, information determination, information cleaning, consolidating fitting earlier information, and legitimate translation of the consequences of mining, are fundamental to guarantee that helpful information is gotten from the information. Blind use of information mining strategies (properly scrutinized as "information digging" in the factual writing) can be a perilous action effectively prompting disclosure of futile examples.

KDD has developed, and keeps on advancing, from the convergence of exploration fields, for example, AI, design acknowledgment, information bases, insights, man-made reasoning, information securing for master frameworks, information perception, and superior execution figuring.

The binding together objective is removing significant level information from low-level information with regards to enormous informational collections. KDD covers with AI and example acknowledgment in the investigation of specific information mining speculations and calculations: implies for demonstrating information and removing designs. KDD centres around parts of finding justifiable examples that can be deciphered as valuable or fascinating information, and puts areas of

strength for an on working with enormous arrangements of certifiable information. Consequently, scaling properties of calculations to enormous informational collections are of major interest.

KDD likewise shares a lot of practically speaking with insights, especially exploratory information examination techniques. The factual methodology offers exact techniques for evaluating the inborn vulnerability which results when one attempts to deduce general examples from a specific example of a general populace. KDD programming frameworks frequently insert specific factual strategies for testing and displaying information, assessing theories, and taking care of commotion inside a general information disclosure structure. As opposed to customary methodologies in measurements, KDD approaches commonly utilize more pursuit in model extraction and work with regards to bigger informational collections with more extravagant information structures.

Notwithstanding its solid connection to the data set field (the second 'D' in KDD), one more related region information warehousing, which alludes to the famous business pattern for gathering and cleaning value-based information

to make them accessible for on-line investigation and choice help. A famous methodology for examination of information stockrooms has been called OLAP (on-line scientific handling), after a bunch of standards proposed by Codd (1993). OLAP apparatuses centre around giving complex information examination, which is better than SQL in figuring synopses and breakdowns along many aspects. OLAP apparatuses are designated towards rearranging and supporting intuitive information investigation, while the KDD's device will likely computerize however much of the interaction as could be expected.

### 3.2 Definitions Related to The Knowledge Discovery Process

Knowledge revelation in data sets is the non-minor course of recognizing substantial, novel, possibly valuable, and eventually justifiable examples in information. Intriguing quality is a general proportion of example esteem, consolidating legitimacy, oddity, value, and straightforwardness.

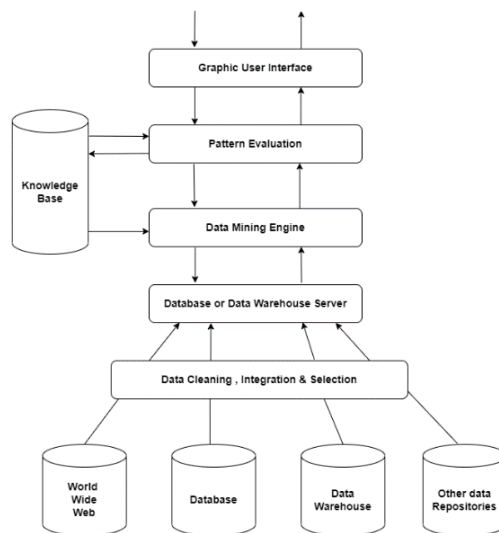
**Table 1.** Definition related to the Knowledge Discovery

<b>Data</b>	<b>A set of facts, F.</b>
<b>Pattern</b>	An articulation E in a language L depicting realities in a subset FE of F.
<b>Process</b>	KDD is a multi-step process including information readiness, design looking, information assessment, refinement with cycle after change.
<b>Valid</b>	Discovered patterns should be true on new data with some degree of certainty. Generalise to the feature (other data).
<b>Novel</b>	Pattern should be previously known.
<b>Useful</b>	Noteworthy, example ought to possibly prompt a few valuable activities.
<b>Understandable</b>	The cycle ought to prompt human knowledge, designs should be made justifiable to work with better comprehension of the basic information.

The primary target of the KDD cycle is to remove data from information with regards to enormous data sets. It does this by utilizing Data Mining calculations to distinguish what is considered information. The Knowledge Discovery in Databases is considered as a modified, exploratory examination and displaying of immense information repositories. KDD is the coordinated methodology of perceiving substantial, helpful, and reasonable examples from colossal and complex informational indexes.

### 3.3 Architecture of KDD Process

Data mining is a critical strategy where beforehand obscure and possibly helpful data is extricated from the tremendous measure of information. The data mining process incorporates a couple of parts, and these parts involve a data mining structure engineering [15].



**Figure 4.** Architecture of KDD Process

### 1- Data Sources

The genuine foundation of data is the Database, data dispersion focus, World Wide Web (WWW), text records, and various reports. You truly need a huge proportion of irrefutable data for data mining to make progress. Affiliations consistently store data in informational collections or data stockrooms. Data conveyance focuses could include something like one informational index, text records accounting sheets, or various chronicles of data. Sometimes, even plain text reports or estimation sheets could contain information. Another fundamental wellspring of data is the World Wide Web or the web.

### 2- Prior Process to Done

Preceding passing the data to the informational collection or data conveyance focus server, the data ought to be cleaned, consolidated, and picked. As the information comes from various sources and in different associations, it can't be used clearly for the data mining strategy in light of the fact that the data may not be done and exact. Subsequently, the primary data hopes to be cleaned and bound together.

### 3- Database and Warehouse Server

The data set or information distribution centre server comprises of the first information that is fit to be handled. Thus, the server is cause for recovering the applicable information that depends on information mining according to client demand.

### 4- Data Mining Engine

The data mining engine is a critical piece of any data mining structure. It contains a couple of modules for working data mining tasks, including connection, depiction, request, gathering, assumption, time-series assessment, etc. With everything taken into account, we can say data mining is the foundation of our data mining plan. It contains instruments and programming used to gain encounters and data from data accumulated from various data sources and set aside inside the data stockroom.

### 5- Pattern Evaluation

The Pattern evaluation module is basically obligated for the extent of assessment of the model by using a cut off regard. It collaborates with the data mining engine to focus in the pursuit on astounding models. This piece routinely uses stake assesses that assist with outing the data mining modules to focus the chase towards captivating models. It could utilize a stake cut off to filter through tracked down plans.

### 6- Graphical User Interface

Graphical User Interface (GUI) module confers between the data mining system and the client. This module helps the client to actually and capably use the system without knowing the complexity of the cycle. This module assists the data mining system when the client demonstrates a request or a task and introductions the results.

### 7- Knowledge Base

The information base is valuable in the entire course of data mining. It might be valuable to coordinate the chase or evaluate the stake of the result plans. The information base could attempt to contain client points of view and data from client experiences that might be helpful in the data mining process. The model evaluation module reliably teams up with the information base to get inputs, and moreover update it.

#### 3.4 Steps Involved in Knowledge Discovery

KDD in information mining is an iterative cycle that examines designs in light of three variables

- Significance
- Convenience
- Understandability

KDD includes a bunch of characterized stages for the treatment of the information prior to applying the various information mining procedures in the quest for buried designs in them to at last make the examination of the examples found lastly give a helpful result [16].

The reason for KDD is the translation of examples, models, and a profound examination of the data that an association has assembled to go with better choices. This incorporates conduct, needs, customs, client question, look by clients, and so forth. The KDD includes 9 stages and their arrangement is significant for acquiring the normal outcomes. At times, returning after the distinguishing proof of a chance for development in the handling of the data might be vital [17].

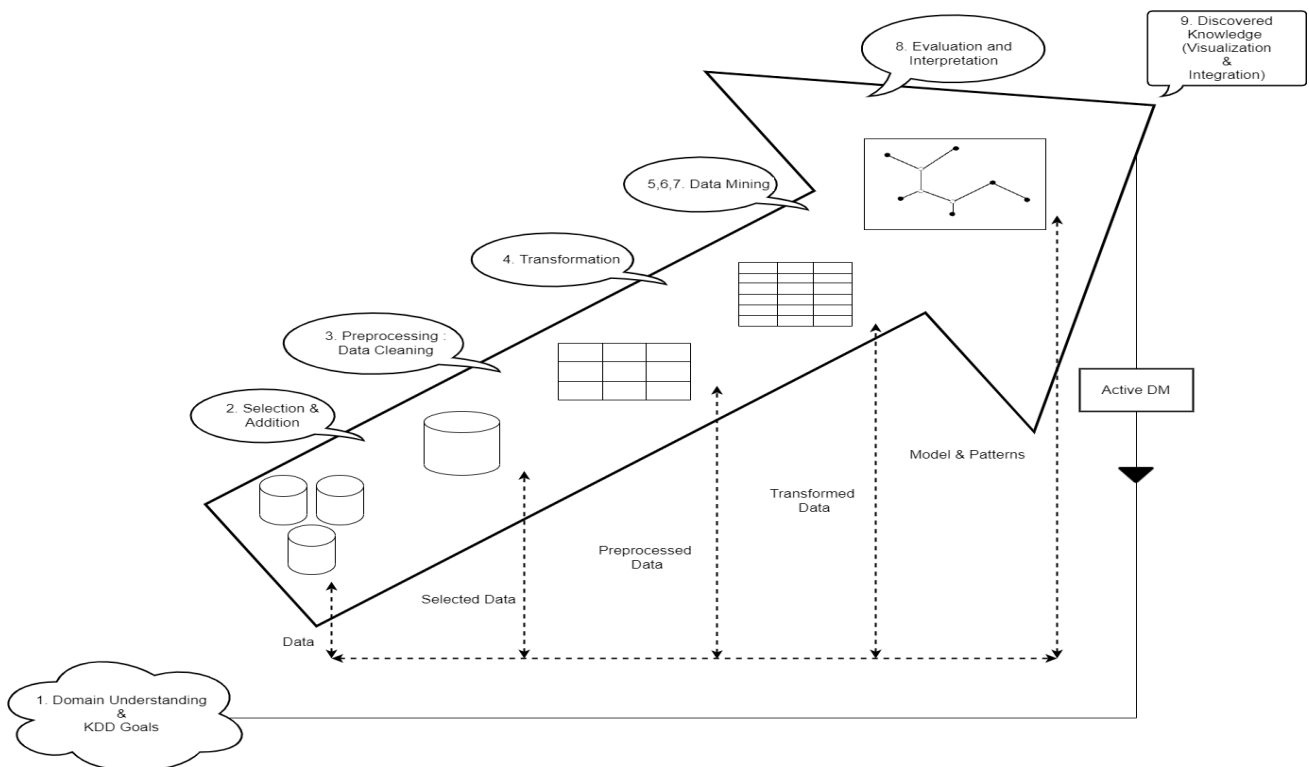


Figure 5. Steps involved in KDD Process

## 1 - Understanding the Data Set

Not everything is science and estimations, but understanding the issues we will stand up to and having setting to propose sensible and certifiable game plans is. It is basic to know the properties, obstructions, and rules of the data or information understudy, and portray the goals to be achieved.

## 2 - Data Selection

From the arrangement of data accumulated and the objectives to be achieved as of now described, open data ought to be concluded to do the survey and integrate them into a singular one that can help with showing up at the objectives of the assessment. Generally, this information can be found in a comparable source or can in like manner be conveyed.

## 3 - Cleaning and Pre-handling

At this stage, the trustworthiness of the not altogether settled, or possibly, doing tasks that guarantee the handiness of the data. For this, the data cleaning is done (treatment of lost data or killing special cases). This recommends clearing out variables or qualities with missing data or killing information not significant for this sort of task like text, pictures, and others.

## 4 - Data Transformation

At this stage, the idea of the data is improved with changes that incorporate either dimensionality decline (reducing the number of variables in the enlightening list) or changes for instance, exchanging the characteristics that are numbers over totally to straight out (discretization).

## 5 - Select the Appropriate Data Mining Task

In this stage, the right data mining cycle can be picked - be it portrayal, backslide, or assembling, according to the objectives that have been set for the cooperation.

## 6 - Choice of Data Mining Algorithms

In this manner, we keep on picking the procedure or estimation or both, to search for the model and get data. The meta-learning revolves around figuring out the inspiration driving why a computation ends up being better for explicit issues, and for each method, there are different possible results of how to pick them. Each computation has its own encapsulation, its own specific way of working and getting the results, so it is judicious to know the properties of those likelihood to use and see which one best fits the data.

## 7 - Application of Data Mining Algorithms

Finally, when the techniques have been picked, the resulting stage is to apply them to the data recently picked, cleaned, and dealt with. It is possible that the execution of the computations in a couple endeavouring to change the limits that smooth out the results. These limits shift according to the picked system.

## 8 - Evaluation

At the point when the estimations have been applied to the educational assortment, we keep on evaluating the models that were delivered and the show that was procured to make sure that it meets the targets set in the essential stages. To finish this evaluation there is a procedure called Cross-Validation, which performs data fragment, parcelling it into planning (which will be used to make the model) and test (which will be used to see that the computation really works and deals with its business skilfully).

## 9 - Interpretation

If all of the means are followed precisely and the results of the appraisal are satisfied, the last stage is simply to apply the data found to the interesting circumstance and begin to handle its interests. If regardless, the results are not tasteful it is vital to return to the past stages to roll out a couple of improvements, separating from the assurance of the data to the evaluation stage. Results ought to be presented in a reasonable plan. Consequently, data portrayal procedures are huge for the results to be significant since mathematical models or portrayals in message arrangement can be hard for end-clients to unravel.

### 3.5 Alternatives to KDD Process

There are various cycles that endeavour to get information from rapidly collecting information, then, at that point, circle that information to additionally refine the method involved with determining quality information and streamline activities [18]. Three normal cycles are the

- Knowledge Discovery in Databases (KDD) Process
- Sample, Explore, Modify, Model, and Access (SEMMA)
- the CRoss Industry Standard Process in Data Mining (CRISP-DM).

### 3.6 Predictive Analytic Modern Technology

The most broadly involved Cross Industry Standard Process for Data Mining strategy is utilized to foster prescient scientific models [19]. It incorporates 6 stages:

**Best Understanding:** The understanding of business stage incorporates appreciate and portray the use case or business issue, the business target and the business question that hope to be answered. It furthermore consolidates describing accomplishment rules. Then the standard endeavour related action hope to be process. These endeavours incorporate portraying resource needs like describing any goals, development, people, cash, making an endeavour arrangement, essentials, assessing bets and making a crisis strategy.

**Data Preparation:** The perception of data stage consolidates data needs, for instance, inside and external data sources, starting and data credits (component and quality) including 3Vs data volumes, collection, speed, plans, and so on, similarly whether the data is in a social informational collection, level records, a Hadoop Distributed File System (HDFS) then again if it is live, streaming data. This stage in like manner consolidates data examination and assessment using real assessment to look at embrace data, besides, a data quality assessment integrates understanding how much data is missing, has bumbles, is duplicated, and is clashing.

**Information readiness:** The objective of the data arranging stage is to make a lot of information that can be dealt with into AI algos. This communication requires different endeavours including isolating and cleaning; data change; data change; data progression; and variable unmistakable confirmation, which is generally called dimensionality decline or part assurance. Variable's ID will likely make an educational file of the main elements to be used as model commitment to stop by ideal results. The point is similarly to wipe out factors from an enlightening record that are not useful as model commitment without compromising the model's accuracy for portrayal, the precision of the assumptions it makes.

**Model development:** The model improvement stage is about the improvement of an AI model. Models can be developed to foresee, gauge or dissect data to find examples like sets, gatherings and affiliations

Two kinds of AI can be utilized in model turn of events:

1. Supervised learning

## 2. Unsupervised learning

Normally, perceptive models are created using coordinated learning. For portrayal, expecting we hope to cultivate a model for equipment disillumination assumption, we can use data that depicts gear that has truly failed. We can use that data to set up the new model to perceive the profile of a piece of stuff that is colourable going to crash and burn. To fulfil this profile affirmation, we parcel the data segments which exhaustive bombarded gear data records into a test educational file and a readiness instructive assortment. Then, we train the model by fill the readiness educational assortment and segments into an estimation, different of which can be used for assumption. Then, we test the model by test educational assortment.

Solo learning is a method for analysing data to endeavour to glance through disguised plans in the data that show thing connection and groupings-for frame, client division. Gathering relies upon restricting or intensifying comparability. The K-exhibits bundling computation is a most extensively elaborate estimation for this philosophy. Perceptive and illustrative astute models can be created using advanced Developed data mining contraptions, assessment fogs, data science instinctive activity manuals with procedural or logical programming vernaculars and robotized model headway instruments.

**Modern Evaluation:** A brief time frame later Model made, the accompanying stage is to evaluate the precision and faultlessness of assumptions. For assumptions, this assessment infers understanding what number of estimates were correct and wrong? Different cooperation can achieve this evaluation. Key appraisals in model evaluation are the number of authentic up-sides, certified negatives, fake up-sides and misdirecting negatives. The surface line is that we need to make certainly that the model is accurate; differently, it could make embrace deceiving up-sides that could achieve wrong exercises and decisions.

**Modern Deployment:** At the point when we are satisfied with the model we've made, the last stage remembers sending models to run for various different environment. These circumstances consolidate bookkeeping sheets, examination servers, data base organization structures (DBMSs), applications, keen social informational collection organization systems, Apache Hadoop, Apache Spark and streaming assessment stages.

## IV. INTELLIGENT DATA MINING

INTELLIGENT DATA MINING AND ANALYSIS (IDA) attempts to change crude information into data using broad foundation information on unambiguous areas from an earlier time and uses it to take care of the issues of present. It deals with an outcome driven approach as it underscores finding an answer for the issue utilizing a shut circle of social event information, making a thorough information base and far-reaching examination of it [20]. In the interim, information mining is just a piece of it that utilizes a pre-processed and changed information to figure out unambiguous examples inside the information. While data mining has different applications in the field of medical services, training, market-based examination, client relationship the board, extortion recognition, and so forth, it turns out to be more critical when it is utilized alongside intelligent data mining as its subpart [21].

# INTELLIGENT DATA MINING

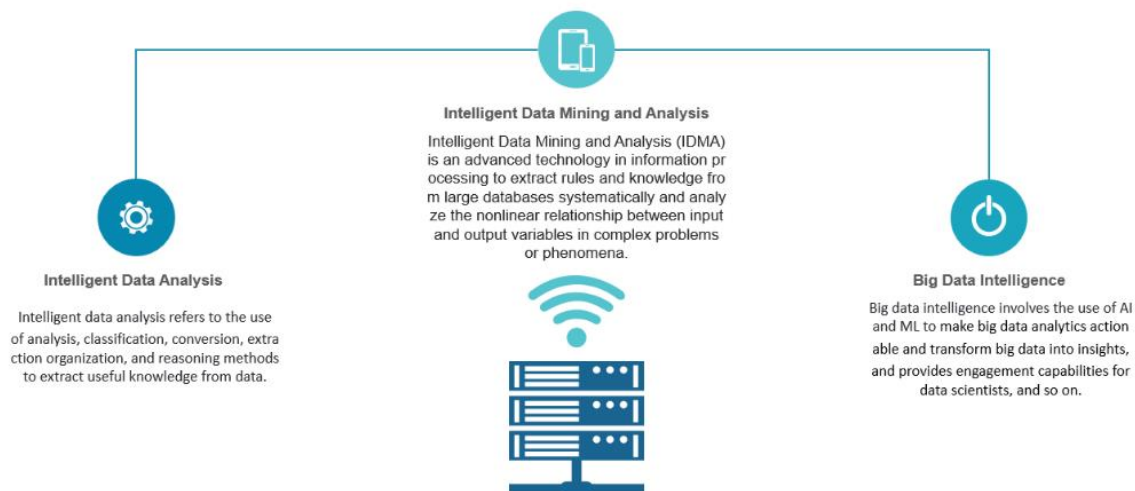


Figure 6. Intelligent Data Mining



#### 4.1 Intelligent Data Mining Software and Tools

Information mining devices are accessible from countless sellers, regularly as a feature of programming stages that likewise incorporate different kinds of information science and progressed examination instruments. Key highlights given by information mining programming incorporate information planning capacities, worked in calculations, prescient displaying support, a GUI-based improvement climate, and devices for conveying models and scoring how they perform [22].

Merchants that deal apparatuses for information mining incorporate Alteryx, AWS (Amazon Web Services), Databricks, Dataiku, Data Robot, Google, H2O.ai, IBM, Knime, Microsoft, Oracle, RapidMiner, SAP, SAS Institute and Tibco Software, among others. Different free open-source advances can likewise be utilized to mine information, including Data Melt, Elki, Orange, Rattle, scikit-learn and Weka. Some product sellers give open-source choices, as well.

#### DATA MINING SOFTWARE AND TOOLS

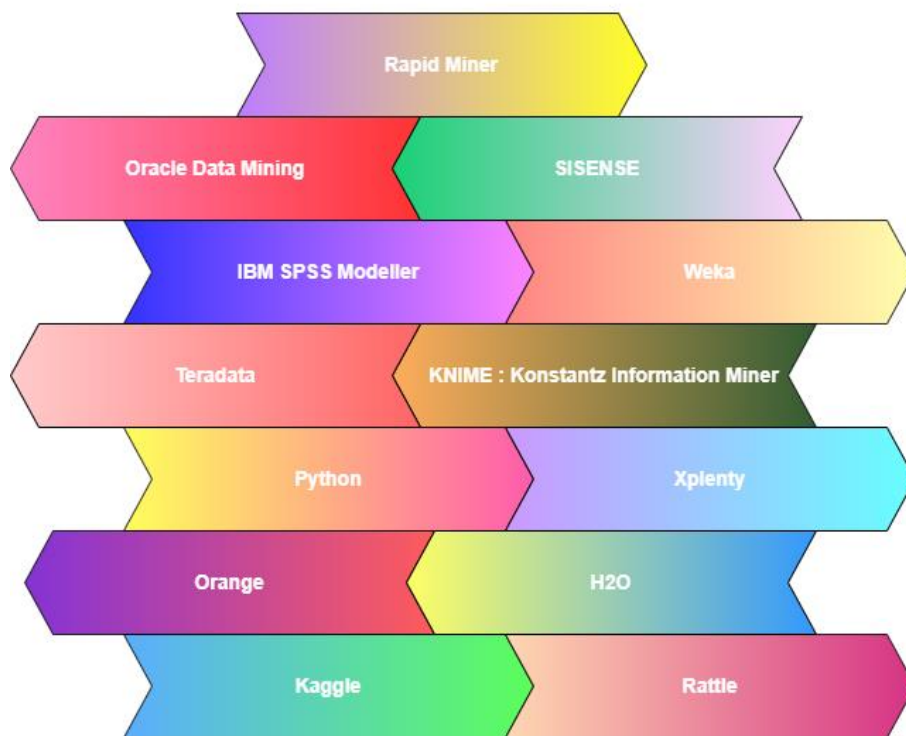


Figure 7. Data Mining Software and tools

**Rapid Miner:** A data science programming stage giving an integrated environment to various periods of data showing including data status, data cleaning, exploratory data examination, portrayal and that is just a hint of something larger. The strategies that the item helps with are AI, significant learning, text mining and judicious examination. Easy to use GUI mechanical assemblies that take you through the showing framework. This device made absolutely in Java is an open-source framework and is amazingly popular in the data mining world [23].

**Oracle Data Mining:** Prophet, the world boss it informational collection programming, solidifies its capacity in informational collection progresses with Analytical gadgets and presents to you the Oracle Advanced Analytics Database part of the Oracle Enterprise Edition. It incorporates a couple of data digging estimations for gathering, backsliding, assumption, peculiarity revelation and that is only the start. This is prohibitive programming and is maintained by Oracle particular staff in helping your business with building a strong data mining establishment at the endeavour scale.

**IBM SPSS Modeller:** IBM (International Business Machine) is again a significant name in the data space concerning colossal undertakings. It gets well together with driving developments to do an enthusiastic endeavour wide game plan. IBM SPSS Modeler is a visual data science and AI plan, helping in shortening a potential chance to regard by speeding up practical endeavours for data scientists. IBM SPSS Modeler will deal with you from improved on data examination to AI [24].

**KNIME:** Konstanz Information Miner is an open-source data assessment stage, that helps you with gather, association and scale rapidly. The mechanical assembly hopes to help with making judicious information accessible to natural clients. The thing features itself as an End-to-End Data Science thing, that makes and produce data science using its single straightforward and normal climate [25].

**Python:** Python is an unreservedly accessible and open-source language that is known to have a speedy expectation to learn and adapt. Joined with its capacity as a broadly useful language and huge library of bundles assist with building a framework for making information models from the scratch, Python makes for an extraordinary device for associations who need the product they use to be exceptionally worked to their determinations. One of the highlights Python is known for in this field is strong on the fly perception highlights it offers.

**Orange:** Orange is an AI and information science suite, utilizing python prearranging and visual programming highlighting intelligent information examination and part based get together of information mining frameworks. Orange offers a more extensive scope of elements than most other Python-based information mining and AI instruments. Programming has north of 15 years of dynamic turn of events and use. Orange likewise offers a visual programming stage with GUI for intuitive information representation.

**Kaggle:** The biggest local area of information researchers and AI experts. Kaggle in spite of the fact that began as a stage for AI rivalries, is currently expanding its impression into the public cloud-based information science stage field. Kaggle presently offers code and information that you really want for your information science executions. There are over 50k public datasets and 400k public journals that you can use to increase your information mining endeavours. The enormous web-based local area that Kaggle appreciates is your wellbeing net for execution explicit difficulties.

**Rattle:** The clatter is a R language-based GUI apparatus for information mining necessities. The instrument is free and open-source and can be utilized to get factual and visual rundowns of information, the change of information for information models, construct directed and solo AI models and look at model execution graphically.

**Weka:** Waikato Environment for Knowledge Analysis (Weka) is a set-up of AI gadgets written in Java. A grouping of portrayal gadgets for judicious showing in a GUI show, helping you with building your data models and test them, seeing the model displays graphically [26].

**Teradata:** A cloud information examination stage promoting its no code required devices in an extensive bundle offering venture scale arrangements. With Vantage Analyst, you needn't bother with to be a software engineer to code complex AI calculations. A straightforward GUI-based framework for speedy undertaking wide reception.

**H2O:** H2O is an open-source ML stage that plans to make computerized reasoning (AI) innovation accessible to everybody. It upholds the most widely recognized ML calculations to help clients in rapidly and effectively fabricating and conveying ML models, regardless of whether they are not specialists.

**Apache Spark:** A strong examination motor, Apache Spark accompanies a large number of APIs that urge Data Scientists to over and over get to information for Machine Learning, SQL Storage, and different purposes. You can construct equal applications with Apache since it's an almighty, iterative, publicly released, and in-memory appropriated examination motor.

**SISENSE:** This information mining programming can be utilized to examine enormous, different datasets and produce explicit business designs with regards to making reports for an association. For a non-specialized crowd, you can likewise make reports with rich visuals in light of the information you've refined.

**Xplenty:** Xplenty offers a stage with information reconciliation, handling, and planning capacities for examination. With the assistance of Xplenty, organizations can make the most of the potential outcomes introduced by large information, all without spending any cash on staff, equipment, or programming. An across the board, answer for making information pipelines. With a rich articulation language, you can perform complex information planning undertakings. It has a simple to-utilize ETL, ELT, or replication arrangement execution interface. A work process motor will allow you to coordinate and timetable pipelines.

## V. PROBLEM STATEMENT

Nowadays many people utilize informal communities to speak with one another and it has a ton of impact in the day-to-day existence. Since it is developing quickly, a lot of examination is happening about informal organization's development over the long haul and the impact of interpersonal organizations on youthful age. One of the most famous interpersonal organizations is Twitter [27]. Twitter permits individuals to share their reasoning and the subtleties of their existence with one another while famous people impart their way of life and way of life to their devotees to grow their own image. Since, in informal community it is not difficult to control an individual, a developing number of associations are utilizing bot to spread their items [28]. Returning a couple of years, there's been a great deal of conversation of "bots" on the web. Over the long haul, in any case, it's turned into a stacked and frequently got term wrong. The term is utilized to misrepresent accounts with mathematical usernames that are auto-created when your inclination is taken, and all the more worryingly, as a device by those in places of political ability to discolour, the perspectives on individuals who might contradict them or online popular assessment that is not positive. There are likewise numerous business benefits that indicate to offer experiences on bots and their action on the web, and oftentimes their emphasis is altogether on Twitter because of the free information we give through our public APIs. Twitter clients into three gatherings, telecasters, colleagues, reprobates and evangelists and finished up this depended on the quantity of devotees, the quantity of following or the supporter following proportion separately. Mis love, investigated the

design of a few internet based virtual entertainment organizations. He examined that it is normal in web-based interpersonal organizations to show commonality. In one more paper they likewise researched on spammers on twitter and reasoned that a high supporter to following proportion expands the possibilities of the record being a spam [29]. This empowers the clients to stop on the time and execution cycle of extricating additional highlights [30].

## VI. RESULT AND DISCUSSION

For data variety, the Twitter API is used here. Considering what we've portrayed over, there's a great deal of justifiable disarray and we truly need to improve at representing ourselves. Altogether, a bot is a motorized record. Returning several years, automated accounts were an issue for us. We zeroed in on it, made the endeavours, and have seen immense augmentations in taking care of them across all surfaces of Twitter. That doesn't mean our work is done. Our proactive work is revolved around control in many designs and that consolidates the malicious use of automation. It's basic to note, not a wide range of motorization are basically encroachment of the Twitter Rules. We've seen creative and inventive reasons for computerization to propel the Twitter understanding - for example, accounts like @pentametrone and @tinycarebot. Motorization can similarly be an astounding resource in client support joint efforts, where a conversational bot can help with finding information about orders or travel reservations normally.

### The thing may be said about devices like Botometer and Bot Sentinel?

Bot Sentinel is looking for accounts that participate in harmful trolling and works in collaboration with Botometer to identify automated or partially automated accounts. You can recognise these accounts when you see them, according to Bouzy. They may be automated or operated by people, and they breach Twitter's rules of service by harassing or spreading false information.



Figure 8. Botometer and Bot Sentinel

These devices start with a human taking a gander at a record - much of the time using a comparable freely available report information you can see on Twitter - and recognizing characteristics that make it a bot. So generally, the record name, the level of Tweeting, the region in the bio, the hashtags used, etc. This is an unimaginably limited approach. As referred to, a record with an impossible to miss handle is commonly someone who was thus proposed that username because their veritable name was taken at join. A record with no photo or region may be someone who has individual opinions on electronic security or whose use of Twitter could open them to risk, like a radical or dissident [28].

### 6.1 Investigation of Twitter Over Concerns of Potentially False

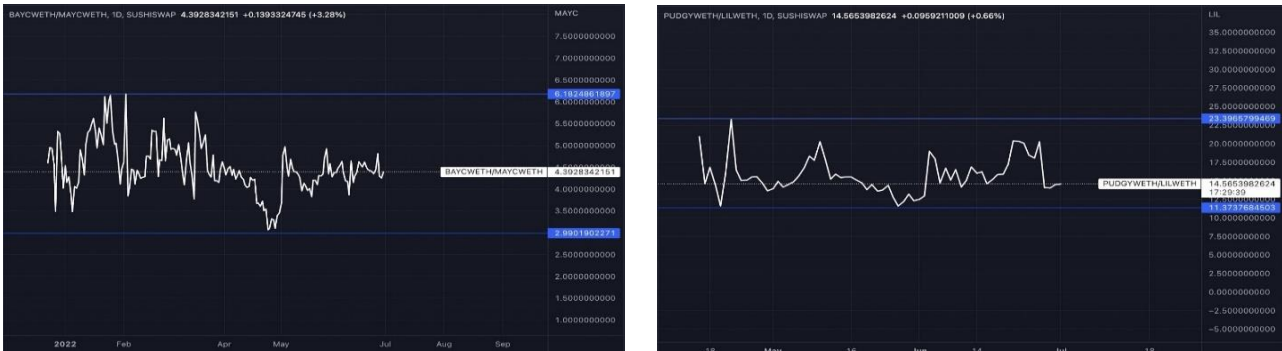
Texas' Attorney General, Ken Paxton, has sent off an examination of Twitter over worries of "possibly misleading" reports connected with the quantity of bots and other phony records on the informal community. In an official statement Monday, Paxton claims inauthentic records might assist with swelling "the worth" of Twitter — consequently he plans to seek after the examination under the state's Deceptive Trade Practices Act, which safeguards against deceiving promoters, organizations and ordinary clients. Paxton's office is chasing after the case similarly as Tesla CEO Elon Musk is apparently endeavours to abandon his own bid to buy Twitter. Musk has, for quite a long time, been recommending the stage's bot numbers might be far more noteworthy than its ongoing initiative are detailing. It's fascinating timing for Musk and Paxton's inclinations to adjust: Tesla just opened a Gigafactory in Texas, and is moving its base camp to the locale. That is a ton of possible business, and it comes as the state has offered tax reductions to organizations building neighbourhood offices [31].

Envision a Twitter bot where in the event that you label it on a string it would quantify the proportion between statement tweets/answers to likes with a rate like "75% Replies 15% Qrt 10% Likes". @Elocremarc twitter bot @theflipwars posts the

proportion between the Bigs and the Lils for Pudgy contrasting it with the Apes and the Mutants. Because of all being accessible on @NFTX\_ I can likewise graph the proportions utilizing Trading view to picture a similar data over the long run [32].

**Bored Ape floor 90.49 ₿**  
**Mutant Ape floor 17.50 ₿**  
**Ratio 5.17**  
**Total combined market cap 1,244,803 ₿**  
**OG dominance 72.69%**

**Pudgy Penguins floor 1.25 ₿**  
**Lil Pudgys floor 0.092 ₿**  
**Ratio 13.60**  
**Total combined market cap 12,943 ₿**  
**OG dominance 85.84%**



**Figure 9.** Ratio between the Bigs and Lils for Pudgy

## VII. AMBER HEARD - BOT SENTINEL INCORPORATED

### 7.1 Key Findings

1. We recognized 627 Twitter accounts zeroed in transcendently on tweeting adversely about Amber Heard and her female allies.
2. 3,288 records were tweeting #AmberHeardIsAnAbuser, #AmberHeardLsAnAbuser, #AmberHeardIsALiar, and #AmberHeardLsALiar, and 19% of those records were devoted to spamming the hashtags.
4. Someone utilized a photograph of a lady's departed kid to make a phony record and savage the lady since she tweeted on the side of Amber Heard.
5. Toxic savages utilized hashtag spamming to drift against Amber Heard hashtags falsely.
6. Over 24% of the records tweeting against Amber Heard hashtags were made inside the beyond seven months.

In 2020, Amber Heard's legitimate group reached Bot Sentinel after we distributed our discoveries on the organized assault focusing on Lisa Page. Golden Heard's lawful group recruited us to decide if the web-based entertainment action against Ms. Heard was natural or on the other hand assuming there was another clarification. They established that a huge piece of the action wasn't natural and placed our discoveries in a report. In June 2022, we started reconsidering the action after the Depp v. Heard decision. Neither Amber Heard nor anybody from her group recruited Bot Sentinel to survey the movement. Nobody employed Bot Sentinel to accumulate and distribute this report [33]. During the Depp versus Heard preliminary, they noticed enemy of Amber Heard hashtags routinely moving on Twitter.

# INVESTIGATION OF TWITTER OVER POTENTIALLY FALSE

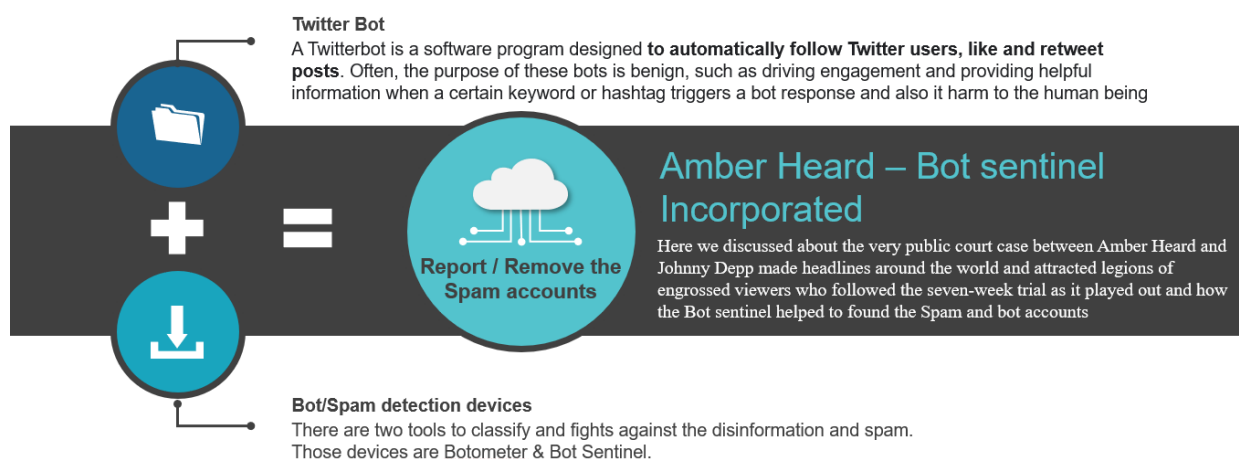


Figure 10. Amber Heard -Bot Sentinel Incorporated

## 7.2 Hashtag Spamming

There are 14,292 tweets with the hashtags #AmberHeardIsAnAbuser, #AmberHeardLsAnAbuser, #AmberHeardIsALiar, and #AmberHeardLsALiar. We recognized 627 Twitter accounts zeroed in overwhelmingly on tweeting adversely about Amber Heard and her female allies. Roughly 24.4% of the Twitter accounts tweeting negative Amber Heard hashtags were made inside the beyond seven months. The quantity of new records tweeting about Amber Heard was altogether higher than accounts tweeting about different subjects. Hashtag spamming was fundamental to intensifying enemy of Amber Heard content and moving enemy of Amber Heard hashtags misleadingly. Hashtag control strategies sent the mixed signal of overpowering resistance to Amber Heard [34].

## 7.3 Tweet Data





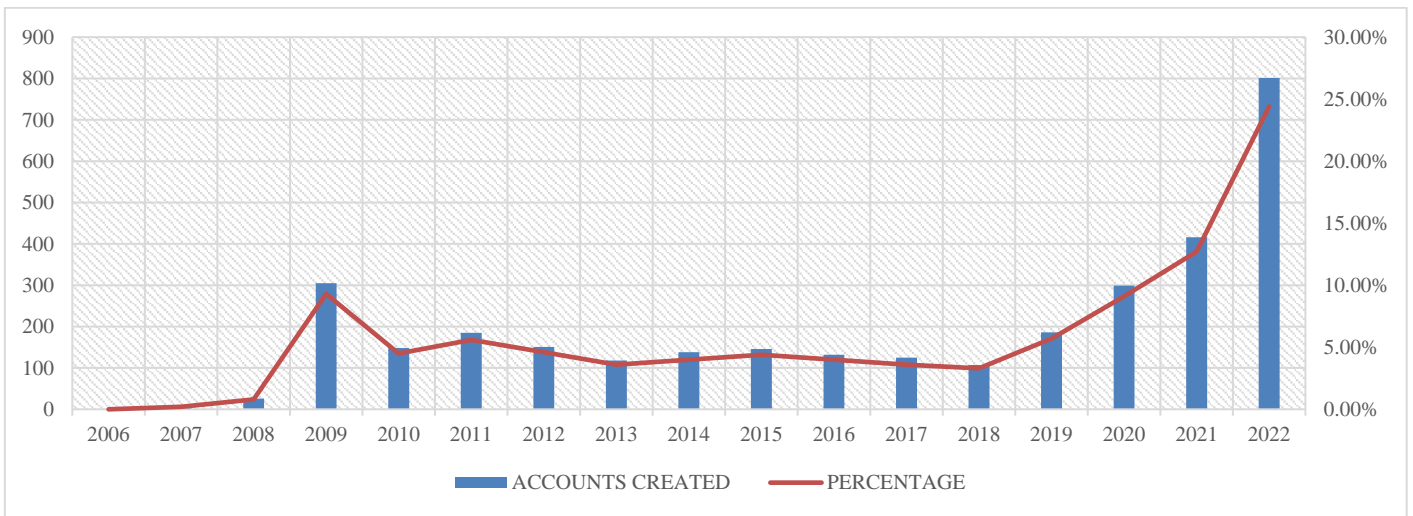
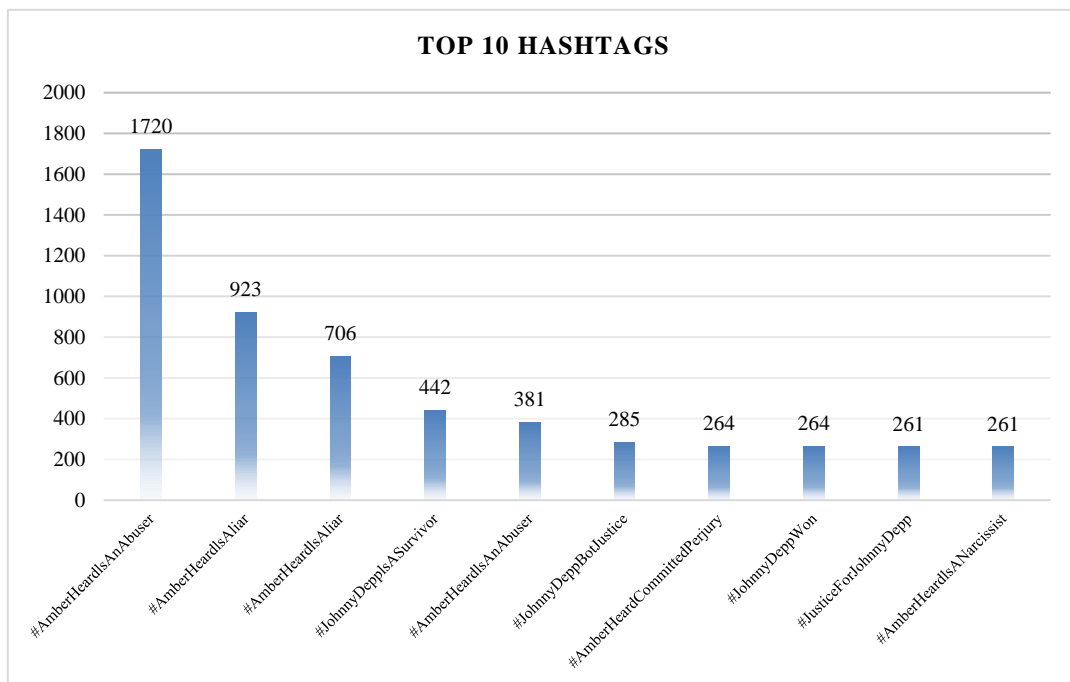


Figure 11. Spam Account created in a year 2006 – 2022

Hashtags	Tweets
#AmberHeardIsAnAbuser	1720
#AmberHeardIsAliar	923
#AmberHeardIsAliar	706
#JohnnyDeppIsASurvivor	442
#AmberHeardIsAnAbuser	381
#JohnnyDeppBotJustice	285
#AmberHeardCommittedPerjury	264
#JohnnyDeppWon	264
#JusticeForJohnnyDepp	261
#AmberHeardIsANarcissist	261

SOURCE	TWEETS
twitter for iphone	1190
twitter For android	1168
twitter web app	808
twitterr for ipad	112
tweetdeck	5
IFTTT	2
fenixapp	1
twitter for mac	1
even newer newerr dank bot	1





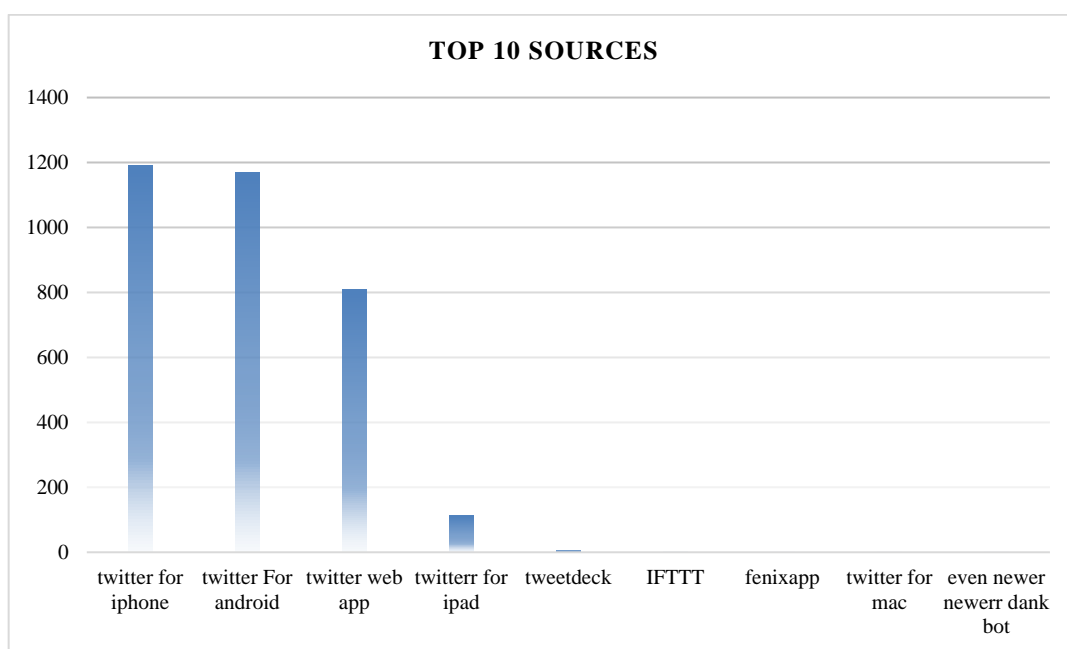


Figure 12. Top 10 Hashtags and Sources

#### 7.4 List of Accounts Violating Twitter Rules

There are 628 spamming accounts violating the rules of twitter and out of those 628 accounts, we mentioned only 50 accounts here.

Table 2. List of accounts violating twitter rules

ID	HANDLE	JOINED	TWEETS	FOLLOWING	FOLLOWERS
2.41E+09	bassgodess2	Thu Mar 27 21:01:46 +0000 2014	1428	73	21
1.36E+18	kiko73990379	Thu Feb 18 01:04:33 +0000 2021	1911	836	101
4.02E+08	MarciaModenese8	Mon Oct 31 12:03:35 +0000 2011	87168	1821	1961
1.31E+18	GladysG61310464	Fri Oct 02 23:11:18 +0000 2020	71	1	0
1.55E+18	NoTimeForEffery	Fri Jul 08 19:19:55 +0000 2022	16	14	3
1.54E+18	RealitySyndrome	Thu Jun 09 22:42:20 +0000 2022	13	27	3
1.52E+18	Isabel44482167	Tue May 10 00:09:26 +0000 2022	84	9	7
1.52E+18	AliceLElliott1	Fri May 06 15:23:29 +0000 2022	14	28	2
1.39E+18	Christi61096812	Sun May 02 21:51:19 +0000 2021	133	67	3
1.54E+18	Lizabethmunoz09	Sat Jul 02 01:35:29 +0000 2022	26	23	1
1.55E+18	factsact15	Thu Jul 07 12:57:44 +0000 2022	24	1	0
1.11E+18	Francis41189221	Wed Mar 27 17:31:39 +0000 2019	16	73	9
1.54E+18	TheShip101	Wed Jun 22 18:05:00 +0000 2022	22	122	7
1.13E+18	Mastiddle	Sat May 11 23:53:22 +0000 2019	14	24	29
1.53E+18	SukhpreetK9	Mon May 30 15:59:26 +0000 2022	17	32	1
1.52E+18	Tracey87160403	Thu May 12 22:58:12 +0000 2022	20	30	2
1.39E+18	gg56890	Tue May 11 01:39:07 +0000 2021	20	15	0
1.53E+18	LivinMyLife_23	Fri May 20 00:40:17 +0000 2022	28	20	0
1.5E+18	nadamfarr	Mon Mar 14 03:01:28 +0000 2022	78	72	14
1.52E+18	AliceChi1357	Mon May 09 06:50:39 +0000 2022	46	5	8
1.53E+18	xule_x	Thu May 19 18:16:07 +0000 2022	223	61	8
1.53E+18	Kelsbells3691	Sun May 29 12:27:13 +0000 2022	31	9	2
1.52E+18	melody96723428	Fri Apr 29 15:00:08 +0000 2022	35	2	0
1.53E+18	Penny0295	Sun May 15 20:24:37 +0000 2022	30	44	1
1.45E+18	laptitenad	Fri Oct 22 21:09:13 +0000 2021	36	0	0
1.53E+18	jameskasey8684	Tue May 31 14:23:25 +0000 2022	38	23	1
2.93E+08	1974pp	Wed May 04 15:19:21 +0000 2011	89	101	106
9.39E+17	iamfaraaz21	Wed Dec 06 22:45:48 +0000 2017	143	10	15

1.54E+18	maharaj_roniel	Thu Jun 23 17:42:19 +0000 2022	1411	93	27
36788704	bvdbilt	Thu Apr 30 22:11:17 +0000 2009	3139	24	34
1.3E+18	bazarette74	Wed Aug 19 19:23:20 +0000 2020	6564	20	61
1.54E+18	ImSuperSpying	Wed Jul 06 22:43:00 +0000 2022	1308	64	29
1.46E+18	BrianLe83870828	Fri Nov 26 00:48:12 +0000 2021	616	0	2
8.32E+17	LittlePeeps1963	Tue Feb 14 20:14:17 +0000 2017	17493	156	399
1.32E+18	AngelaMClark7	Fri Oct 16 04:18:34 +0000 2020	1841	7	23
1.53E+18	Ferns7Amanda	Tue May 24 21:45:17 +0000 2022	169	83	13
1.52E+18	Gisland08	Wed Apr 27 10:47:08 +0000 2022	667	16	9
1.53E+18	Nora79317370	Thu May 19 01:54:29 +0000 2022	161	15	3
1.49E+18	CraigWhyte110	Mon Jan 24 10:23:15 +0000 2022	277	4	10
1.53E+18	nienn88	Fri May 13 19:09:46 +0000 2022	176	44	6
1.44E+18	BerthaMiehle	Wed Sep 08 11:36:26 +0000 2021	605	16	3
1.54E+18	BlackSwanAlpha	Mon Jun 20 16:15:26 +0000 2022	554	3	3
3.44E+08	AsiaSays1	Fri Jul 29 01:06:42 +0000 2011	840	17	31
1.63E+08	tsukineko_96	Mon Jul 05 05:31:22 +0000 2010	7120	237	193
5.96E+08	envienvienvi	Fri Jun 01 10:00:02 +0000 2012	6907	1804	321
7.12E+08	Queeniesthought	Mon Jul 23 08:04:04 +0000 2012	2955	357	163
1.53E+18	LindaBa92057241	Thu May 26 14:33:54 +0000 2022	52	2	3
1.52E+18	Kim32741392	Thu Apr 28 02:18:52 +0000 2022	437	62	7
1.54E+18	Natasha00198878	Fri Jun 24 16:28:03 +0000 2022	131	5	3

### 7.5 Experimental Setup and Result Analysis

The proposed model comprises of extricating Twitter client information utilizing programmable bookkeeping sheet apparatuses like Google Docs. This content purposes the Twitter API alongside passing a bunch of boundaries, similar to client handle or hashtags to bring client metadata [35]. In any case, straightforwardly utilizing Twitter API is additionally conceivable yet it builds the intricacy of the system. To limit the query output, geo-directions and sweep can be determined in the boundaries to adjust and explicitly focus on any actual area. For this examination, various strategies were utilized which included the information recovery utilizing client handle and an assortment of hashtags. Without utilizing the geo-arranges, the information being gotten is of exceptionally wide\ area.

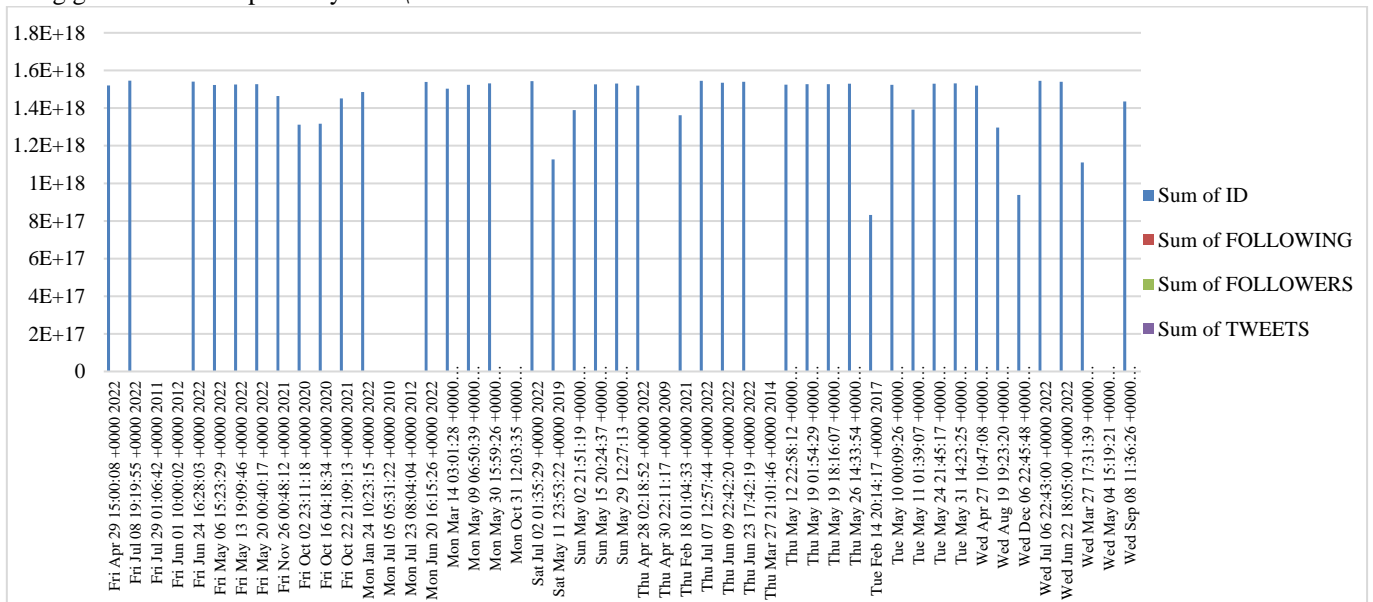


Figure 13. Graphical representation of Spamming accounts

The bots are known to utilize loads of hashtags and client makes reference to contrasted with a human client, which expands the length of the tweet. Subsequently, this can be an element to distinguish likely bots. Then, the level of retweets from the tweets is determined as wistful bots are known to advance opinions. Retweeting tweets for a positive or negative opinion can be simple and proficient method for helping its advancements.

The quantity of hashtags and client referenced in the tweet is separated from the recovered information and utilized among the other discovery highlights. Bots had been known to control a gathering of individuals or advance their substance on the

web-based entertainment. A regular wistful bot conduct is advancing positive items for a certain measure of time. This advancement is focused on to a general arrangement of individuals to control their view on the theme [36]. Then, the bot unexpectedly changes its way of behaving by advancing negative items on the point to one more arrangement of individuals. Making a contention of opinions is chiefly finished among the general individuals [37].

## VIII. CONCLUSION

We have provided some definitions of fundamental KDD terms. Clarifying the connection between knowledge discovery and data mining is a fundamental goal. We gave an overview of the KDD procedure and the fundamentals of data mining. Our succinct summary is obviously constrained by the wide range of data mining methods and algorithms available; there are several data mining approaches, particularly specialised strategies for particular types of data and domains. Although distinct methods and applications could seem very different at first glance, it is typical to discover that they have a lot in common. The task of any data mining method is clarified by comprehending data mining and model induction at this component level. This facilitates the user's comprehension of its overall significance and relevance to the KDD process. Here, we analysed and provide an information on the growth of the bot past few years.

Currently, social media such as Facebook, Twitter are of growing concern everyone around the world and for many reasons. Among all the reasons autonomous entities or BOTs are one of the major concerns. Every social media platform is now working on a way to eliminate all these autonomous entities by implementing different AI tools that can identify the autonomous activities as these bots can generate tweets, retweet, follow, like and spread information rapidly and cause social unrest. These autonomous entities or bots are already making impact on different occasions on different countries around the world. A classification between a human and a Bot was performed by using syntax analysis and user behaviour along with the sentiment analysis of random tweets, user specific tweets and extracted features by aggregating the tweets by their senders. This report will delineate one of the most pessimistic scenarios of stage control and egregious maltreatment from a gathering of Twitter accounts. It will show how savages utilized strategies to control discussions and patterns across Twitter while focusing on and mishandling ladies to stifle any sure tweets supporting Amber Heard. Here, we discussed about the Amber Heard case and how this bot plays a major role in dominating the Amber and listed an account violating the twitter rules and analysed the top hashtags and sources spread against the Amber Heard case. Botometer or Bot sentinel played a major role in controlling the twitter bot.

## REFERENCES

- [1] Model, Four Element, Graham J. Williams, and Zhexue Huang. "Modelling the KDD Process." (1996).
- [2] Keeney, R. L. (2006). "Value-Focused Thinking: Identifying Decision Opportunities and Creating Alternatives." *European Journal of Operations Research* 92: 537- 549.
- [3] Berry, M. and G. Linoff (2007). *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley and Sons.
- [4] Osei-Bryson, K.-M. (2004). "Evaluation of Decision Trees." *Computers and Operations Research* 31: 1933-1945
- [5] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. & Kobler, A. (2007). Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, 40, 438-447. doi:10.1016/j.jbi.2006.10.003
- [6] Schenkman, S. (2007). "Inducement of nonexistent order by the analytic hierarchy process." *Decision Sciences* 28(2): 475-482. 6.
- [6] Inmon, W. H. (2012). *Building the Data Warehouse*. New York, Wiley
- [7] Mirza S, Mittal S, Zaman M. A review of data mining literature. *International Journal of Computer Science and Information Security (IJCSIS)*. 2016 Nov;14(11):437-42.
- [8] Reddy DL. A Review on Data mining from Past to the Future. *International Journal of Computer Applications*. 2011;975(2011):8887.
- [9] Silwattananusarn T, Tuamsuk K. Data mining and its applications for knowledge management: a literature review from 2007 to 2012. *arXiv preprint arXiv:1210.2872*. 2012 Oct 10.
- [10] Parvathi I, Rautaray S. Survey on data mining techniques for the diagnosis of diseases in medical domain. *International Journal of Computer Science and Information Technologies*. 2014;5(1):838-46.
- [11] Solieman OK. Data mining in sports: A research overview. *Dept. of Management Information Systems*. 2006 Aug.
- [12] Giachanou A, Crestani F. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*. 2016 Jun 30;49(2):1-41.
- [13] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* 2011 Jun (pp. 30-38).
- [14] Fayyad U. Knowledge discovery in databases: An overview. *Relational data mining*. 2001:28-47.
- [15] Zhong N, Liu C, Kakemoto Y, Ohsuga S. KDD Process Planning. In *KDD 1997* Aug 14 (pp. 291-294).
- [16] Verma IS. Knowledge Data Discovery and Its Issues. Expansion, Impact and Challenges of IT & CS. 2015 Sep 21:88.
- [17] Shafique U, Qaiser H. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*. 2014 Nov 20;12(1):217-22.
- [18] Düntsch I, Gediga G, Nguyen HS. Rough set data analysis in the KDD process. In *Proc. of IPMU 2000 (Vol. 1)*, pp. 220-226).
- [19] Wagner C, Saalmann P, Hellingrath B. Machine condition monitoring and fault diagnostics with imbalanced data sets based on the KDD process. *IFAC-PapersOnLine*. 2016 Jan 1;49(30):296-301.
- [20] Chen Z. *Intelligent Data Warehousing: From data preparation to data mining*. CRC press; 2001 Dec 13.
- [21] Ruan D, Chen G, Kerre EE, Wets G, editors. *Intelligent data mining: techniques and applications*. Springer Science & Business Media; 2005 Aug 24.
- [22] Liu X. Intelligent data analysis. In *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications 2008* (pp. 308-314). IGI Global.
- [23] Ma S, Chowdhury SK. Application of LC-high-resolution MS with 'intelligent' data mining tools for screening reactive drug metabolites. *Bioanalysis*. 2012 Mar;4(5):501-10.

- [24] McCormick K, Salcedo J. IBM SPSS Modeler essentials: Effective techniques for building powerful data mining and predictive analytics solutions. Packt Publishing Ltd; 2017 Dec 26.
- [25] Dietz C, Berthold MR. KNIME for open-source bioimage analysis: a tutorial. *Focus on Bio-Image Informatics*. 2016;179-97.
- [26] Mohammad MN, Sulaiman N, Muhsin OA. A novel intrusion detection system by using intelligent data mining in weka environment. *Procedia Computer Science*. 2011 Jan 1;3:1237-42.
- [27] Sarlan A, Nadam C, Basri S. Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia 2014* Nov 18 (pp. 212-216). IEEE.
- [28] Wang L, Niu J, Yu S. SentiDiff: combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*. 2019 Apr 26;32(10):2026-39.
- [29] Pawar KK, Shrishrimal PP, Deshmukh RR. Twitter sentiment analysis: A review. *International Journal of Scientific & Engineering Research*. 2015 Apr 4;6(4):957-64.
- [30] Philander K, Zhong Y. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*. 2016 May 1;55(2016):16-24.
- [31] Rauchfleisch A, Kaiser J. The false positive problem of automatic bot detection in social science research. *PloS one*. 2020 Oct 22;15(10):e0241045.
- [32] Geeng C, Yee S, Roesner F. Fake news on Facebook and Twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems 2020* Apr 21 (pp. 1-14).
- [33] Maryn AG, Dover TL. Who gets canceled? Twitter responses to gender-based violence allegations. *Psychology of Violence*. 2022 Jun 6.
- [34] Jeffares S. *Interpreting hashtag politics: Policy ideas in an era of social media*. Springer; 2014 May 15.
- [35] Jain AP, Katkar VD. Sentiments analysis of Twitter data using data mining. In *2015 International Conference on Information Processing (ICIP) 2015* Dec 16 (pp. 807-810). IEEE.
- [36] Foysal A, Islam S, Rahaman T. Classification of AI powered social bots on Twitter by sentiment analysis and data mining through SVM. *International Journal of Computer Applications*. 2019;117:13-9

# Chapter - 13

## Data Visualization Techniques

Sasi Kumar V<sup>1</sup>, Sasi Kumar M<sup>2</sup>, Samyukthaa R<sup>3</sup>, Vinothraja R<sup>4</sup>, Abirami A<sup>5</sup>, Lakshmanaprakash S<sup>6</sup>

<sup>1,2,3,4,5,6</sup>Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Alathukombai, Sathyamangalam, Erode, Tamilnadu, India

E-mail: <sup>1</sup>[sasikumarskvs@gmail.com](mailto:sasikumarskvs@gmail.com), <sup>2</sup>[sasikumarmurugan02@gmail.com](mailto:sasikumarmurugan02@gmail.com), <sup>3</sup>[samyuktharavisamyuktha@gmail.com](mailto:samyuktharavisamyuktha@gmail.com),  
<sup>4</sup>[vinothrajar049@gmail.com](mailto:vinothrajar049@gmail.com), <sup>5</sup>[abirarmia@bitsathy.ac.in](mailto:abirarmia@bitsathy.ac.in), <sup>6</sup>[lakshmanaprakashs@bitsathy.ac.in](mailto:lakshmanaprakashs@bitsathy.ac.in)

**Abstract**— Data visualisation is the representation of the information using standard images like charts, plots, infographics, and even animations. These data visualisations convey complicated data linkages and data-driven insights in an easy-to-understand manner. It aids in the explanation of facts and the selection of courses of action. It will assist any field of research that demands novel approaches to presenting massive amounts of complicated data. Modern visualisation has been shaped by the introduction of computer graphics. A taxonomy of visualisation approaches is also offered, based on the number of variables that may be shown. Novel trends in user interface design are examined, as well as a range of new visualisation approaches and their applicability. In the topic of software visualisation, there are several novel visualisation approaches and tools for studying the datasets. However, finding the correct technology to meet user needs for displaying huge datasets remains a challenge. It gives a quick rundown of a few of the most popular visualisation tools and examines their suitability for supporting research with big volumes of environmental data and also provides significant opportunities for technical communication researchers to expand the field's knowledge of environmental data visualizations and their function in environmental communication. Here, we imported the dataset in data mining tool i.e., Rapid Miner and by using the dataset, we designed various types of data visualization chart.

**Keywords**— Data Visualization, Environmental Communication, Geographical data, User interface design, Rapid Miner

### I. INTRODUCTION

Data visualization is the portrayal of information through utilization of normal designs, like outlines, plots, infographics, and even activities. These visual presentations of data impart complex information connections and information driven experiences in a manner that is straightforward. By using visual parts like layouts, graphs, and guides, information perception contraptions give an accessible technique for seeing and sort out examples, special cases, and models in data. In the domain of Big Data, information portrayal mechanical assemblies and headways are crucial to separate colossal proportions of information and make data driven decisions [1].

Our eyes are attracted to varieties and examples. We can rapidly distinguish red from blue, square from circle. Our way of life is visual, including everything from workmanship and promotions to TV and motion pictures. Data visualization is one more type of visual workmanship that snatches our advantage and keeps our eyes on the message. At the point when we see a graph, we rapidly see patterns and exceptions. On the off chance that we can see something, we assimilate it rapidly. It's narrating with a reason. On the off chance that you've at any point gazed at a huge calculation sheet of information and couldn't see a pattern, you know the amount more successful a representation can be.

Visualization is an inexorably key apparatus to get a handle on the trillions of lines of information produced consistently. Data representation assists with recounting stories by organizing information into a structure more obvious, featuring the patterns and exceptions. A decent visualization recounts a story, eliminating the commotion from information and featuring the helpful data. In any case, it's not just as simple as sprucing up a diagram to cause it to seem overall more appealing or slapping on the "data" part of an infographic. Powerful data representation is a fragile difficult exercise among structure and capability. The plainest chart could really wear out get any notification or it make tell a strong point; the most shocking representation could totally fall flat at passing on the right message or it could say a lot. The information and the visuals need to cooperate, and there's a craftsmanship to consolidating incredible examination with extraordinary narrating [1].

In this chapter we will discuss about the importance of data visualization in career followed the various surveys by the researchers in the field of data visualization. Next, we will see about the features, advantages and usage of the visualization technique along with big data sets. Continued with the major tools we use for visualization and we will see one tool RAPID MINER in detail. Later we will analyse the types of data visualization like pie chart, bar graph, scatter plot etc., for a specific data set and will produce the result and conclusion based on it.

### **Is data visualization important for career?**

It's difficult to consider an expert industry that doesn't profit from making information more justifiable. Each STEM (Science, Technology, Engineering and Mathematics) field benefits from grasping information — thus do fields in government, finance, advertising, history, purchaser products, administration ventures, schooling, sports, etc. While we'll continuously wax idyllically about data visualization (you're on the Tableau site, all things considered) there are useful, genuine applications that are indisputable. Furthermore, since visualization is so productive, it's additionally one of the most valuable expert abilities to create. The better you can convey your focuses outwardly, whether in a dashboard or a slide deck, the better you can use that data. Data visualization is critical for essentially every business. It will in general be used by teachers to show student test results, by PC scientists researching movements in man-made awareness or by pioneers wanting to give information to accomplices.

Data visualisation also provides the following benefits:

1. A greater knowledge of the next steps that must be taken to develop the organisation; the capacity to take in information rapidly, gain better insights, and make quicker judgments; an improved capacity to hold an audience's interest with information they can understand;
2. A simple information flow that increases the chance of insight sharing among all parties;
3. Since data is easier to access and interpret, there will be less need for data scientists. There will also be a better ability to move rapidly on findings and, as a result, achieve success more quickly and with fewer errors.

#### *1.1 Advantages of Data Visualization*

Colours and patterns catch our attention. Red and blue may be immediately distinguished, as can squares and circles. Everything in our culture is visual, from TV and movies to ads and art. Another sort of visual art that captures our attention and keeps it fixed on the message is data visualisation. We can immediately see trends and outliers when we look at a chart. We easily assimilate something if we can see it. It's narrative with a goal. If you've ever tried to discern a trend in a huge spreadsheet of data, you know how much more impactful a visualisation can be.

The following are additional benefits of data visualisation:

1. sharing information is simple.
2. Investigate possibilities in conversation.
3. Visualize relationships and patterns.

#### *1.2 Disadvantages of Data Visualization*

Even while there are many benefits, some of the drawbacks might not be as clear. For instance, it's simple to get the wrong conclusion while looking at a visualisation containing numerous different datapoints. Or, perhaps, the visualisation is simply poorly conceived, leading to bias or confusion.

Additional drawbacks include:

1. Inaccurate or biased information.
2. Not all correlations indicate cause and effect.
3. Translation errors might obscure important points.

## **II. LITERATURE SURVEY**

To comprehend information by graphs and guides representation utilized in China as soon as 1137. In all fields there has been immense advancement in perception methods. To inspect data and information representation help to picture and express thoughts in engineering. With the approaching of PC reproduction perception congruity has been encourage fortified. A wide collection of PC based device in developing plan into computer aided design (PC Helped Plan) configuration is given. The utilization of computer aided design for seeing plan has been taken on incredibly rapidly by, Proficient all through the world. Data perception used to give plan information the guide of drawings and graphs and information is normally applied or unique, we require logical perception methods like outlines and chart and so forth. An unlined change from manual to computerized method in the erect plan representation has followed the as of now ruling standards of representation. There is a coextensive thought process to additionally ask the staying alive and new techniques for perception that practically present complex information. The representation plan ought to acquire from manual techniques where potential to help fashioners make a transformation from their practice. Perception ought to have ability to introduce complex information and it should be synergistic and grant adequate correspondence. Utilizing variety coding and layering site examination information is given on



the drawing the power of controlling the perceivability of layers as pined for by the originators. The space of perception develops, the instrument are endeavouring clients begin in our examination research centres [2].

In direct to face accomplishable manifest of assessable addition that will advance more remote of perception in which the functionality found out trials and studies reports are valuable only there is an emerging need as substitute strategy of rating. Data perception ordinarily part of some originate activity that needs client to develop theories, look through examples and avoidance, and the clean their theory [3]. Client as often as possible expect to see the comparable information according to unique direct point of view or more a year. They could require a sort of instruments to achieve their points, stubbornly bringing in and exportation of information. Scientists portray egressing research which shows up effortlessly obliged to inspect the originate activities that client of data perception seek after in. To help the objectives of perception of data the ethnographic strategies envisioned. To decide the benefits and burdens of their new representation of data instrument the designer or examiner are sharp. The representation local area has recently found disserve above the actions and impact of unreasonable chart beautification and documentation. Perception specialists like Stephen Few and Edward Tufte energized the customary view, contains that the perception ought to introduce the information obviously with next to no irritate and should exclude graph garbage. Brain research lab considered has additionally been upheld this view, which present that basic and clear representation are not difficult to decipher [4]. Memorability tries result shows that representation is as such critical with consistency over individuals. Representation is less extraordinary than natural scenes however prefer to pictures of faces, which could hint at general nonfigurative, qualities of human maintenance. Not astoundingly, credits like understanding and shade of a human unmistakable point increment memory power. Making a perception remarkable expects making the representation —stickl in the observer minds. We require the main material features of information the essayist is endeavouring to send to stick [5]. The endeavour of large information band is huge however hard difficulty. Perception of data method could help to sort out the difficulty. Undertaking of visual information has bunches of utilizations like information mining and misrepresentation identification use perception of data method for repaired information investigation. Undertaking of visual information ordinarily allows a speedier information endeavour habitually supplies more helpful outcome, especially in examples where reflex calculations fails. In the judgments of the investigation visual information endeavour methods outfit a frequently more elevated level of affirmation. This data reaches out to a famous necessity for visual undertaking methods and builds them fundamental in colligation with programmed campaign methods. Data representation concentrates on informational indexes insufficient fundamental Two-Layered (2D) or Three-Layered (3D) substance and consequently as well as insufficient an action portrayal of the nonfigurative information onto the effective screen. For representation there a ton of long recognizable techniques for informational indexes are x-y plots, histograms, and line plots. These strategies are utile for information undertaking however are limited to relatively minor and little layered information groups [6].

Perception procedure develop huge and complex data justifiable. Data perception is a visual point of interaction that take into account understanding of data to the exploiter. To develop things simple to interpret and understand what's more, simple to utilize. To achieving the visual portrayal all functionality, arise are influential for think about [7]. The entrance of creating significantly representation the system can be characterized into unique advances, for example, determination, show, planning, ease of use, assessment and intuitiveness, which portrays primary activities with respect to representation to coordinate definite and amazing plan. Perception procedures are arranged in an unexpected way, there is three classes of perception for example logical representation, data representation and programming perception. There are a large number formal information perception procedures, for example, table, pie diagram, bar graph, histograms, bubble outline, region diagram and line diagram. To begin the essential of representation for the future, remembering the significance of perception [8]. Individuals' collaboration with perception device has emphatically impact on the comprehension of information and framework capabilities. In this manner human connection contribute essentially job in the valuation and plan of perception apparatus [9].

Prior to starting our information base inquiries, we had put two limitations on our ventures. To start with, we limited our search of the writing to post-2000, with most of the examination we analysed intently being in the last 10 years. The better instruments for information perception creation also, more information promptly accessible implies that strategies utilized quite a while back have minimal bearing on current rehearses [10]. For model, some examination put together suggestions with respect to perception types on their own exploration of "best rehearses" that were not obviously made sense of nor exactly upheld. Along these lines, this sort of exploration article was not remembered for our investigation. We likewise restricted our pursuit to concentrates on that crossed with conveying patient centred data in wellbeing what's more, clinical settings. The kind of data remembered for these materials should show restraint focused furthermore, they need to track down ways of talking about subjects, for example, commonness of illness in a populace, likelihood that a positive experimental outcome shows a "genuine positive", recurrence of different results from various treatment strategies in ways that patients can comprehend. With wellbeing education levels in the US floating at around 12% as per the US government, scientists in wellbeing and medication are leading the most developed and different around approaches to impart complex information and data in both visual furthermore, printed ways [11]. We played out our writing look through on three significant information bases: Scopus, PubMed, and Web of Science. These three data sets cover most of studies directed in medication, science,

nursing, and the united wellbeing science fields. It likewise incorporates significant trains, for example, brain science. We likewise looked through IEEE Explore and the ACM data sets. We physically checked on the major diaries in TPC: IEEE Transactions on Professional Correspondence, Journal of Business and Technical Correspondence, Journal of Technical Writing and Correspondence, Technical Communication, and Specialized Communication Quarterly [12].

### III. DATA VISUALIZATION – THE ESSENTIAL

Data visualization gives a fast and successful method for imparting data in a widespread way utilizing visual data. The training can likewise assist organizations with distinguishing which variables influence client conduct; pinpoint regions that should be improved or need more consideration; make information more essential for partners; comprehend when and where to put explicit items; and anticipate deals volumes.

#### **Different advantages of information perception incorporate the accompanying:**

- the capacity to retain data rapidly, further develop experiences and settle on quicker choices;
- an expanded comprehension of the following stages that should be taken to work on the association;
- a superior capacity to keep up with the crowd's advantage with data they can comprehend;
- a simple dissemination of data that expands the chance to impart bits of knowledge to all interested parties;
- take out the requirement for information researchers since information is more available and reasonable; and
- an expanded capacity to follow up on discoveries rapidly and, in this manner, make progress with more noteworthy speed and less mix-ups.

#### *3.1 Data visualization and Big Data*

The extended standing of tremendous information and data examination projects have made portrayal huger than any time in late memory. Associations are logically using AI (Artificial Intelligence) to gather immense proportions of data that can be inconvenient and slow to sort out, understand and figure out. Portrayal offers a method for speeding this up and acquaint information with business visionaries and accomplices in habits they can understand [13].

Enormous data portrayal often goes past the normal methods used in commonplace discernment, for instance, pie frameworks, histograms and corporate graphs. It rather uses more mind-boggling depictions, for instance, heat guides and fever charts. Enormous information representation requires solid PC structures to assemble unrefined data, process it and change it into graphical depictions that individuals can use to quickly draw pieces of information. While colossal information perception can be profitable, it can address a couple of downsides to affiliations. They are according to the accompanying: To capitalize on huge data visualization devices, a perception expert should be recruited. This expert should have the option to recognize the best informational indexes and representation styles to ensure associations are streamlining the utilization of their information.

- Huge data representation projects frequently require contribution from IT, as well as the executives, since the perception of large information requires strong PC equipment, productive capacity frameworks and, surprisingly, a transition to the cloud.
- The bits of knowledge given by enormous data visualization may be basically as precise as the data being envisioned.

#### **The three Vs of Big Data**

Accordingly, it is fundamental to have individuals and cycles set up to administer and control the nature of corporate information, metadata and information sources.

**Volume:** The measure of information matters. With enormous information, you'll need to handle high volumes of low-thickness, unstructured information. This can be information of obscure worth, for example, Twitter information channels, clickstreams on a site page or a portable application, or sensor-empowered hardware. For certain associations, this may be many terabytes of information. For other people, it could be many petabytes.

**Velocity:** Velocity is the quick rate at which information is gotten and (maybe) followed up on. Ordinarily, the most noteworthy speed of information streams straightforwardly into memory as opposed to being composed to plate. Some web empowered shrewd items work progressively or close to ongoing and will demand continuous assessment and activity.

**Variety:** Variety alludes to the many sorts of information that are accessible. Conventional information types were organized and fit conveniently in a social data set. With the ascent of enormous information, information comes in new unstructured information types. Unstructured and semi structured information types, like text, sound, and video, require extra pre-processing to determine significance and backing metadata.

#### *3.2 Workflow for Creating Visualizations:*

Big Data is a large volume, complex dataset. So, such data cannot visualize with the traditional method as the traditional data visualization method has many limitations.

- **Perceptual Scalability:** Human eyes cannot extract all relevant information from a large volume of data. Even sometimes desktop screen has its limitations if the dataset is large. Too many visualizations are not always possible to fit on a single screen.
- **Real-time Scalability:** It is always expected that all information should be real-time information, but it is hardly possible as processing the dataset needs time.
- **Interactive scalability:** Interactive data visualization help to understand what is inside the datasets, but as big data volume increases exponentially, visualizing the datasets take a long time. But the challenge is that sometimes the system may freeze or crash while trying to visualize the datasets.

### 3.3 Use Cases of Big Data Visualization tools:

- **Sports Examination:** In light of past datasets with the assistance of representation devices, a triumphant rate forecast is conceivable. Diagram plotting for the two groups or players is conceivable, and investigation can be performed.
- **Misrepresentation Identification:** Extortion recognition is a popular use instance of huge information. With the assistance of perception devices in the wake of breaking down information, a message can be created to other people, and they will be cautious about such extortion episodes.
- **Value Improvement:** In any business item, cost set is a huge issue with picturing devices and every one of the parts utilized; cost can be examined lastly contrasted and market cost, and afterward an important cost can be set. Security Insight: Picturing hoodlums' records can foresee how much danger they are to society. Every nation has its security knowledge, and its errand is to imagine data and illuminate others about a security danger.

## IV. TOOLS USED IN DATA VISUALIZATION

### 1. Rapid Miner:

A data science programming stage giving an integrated environment to various periods of data showing including data status, data cleaning, exploratory data examination, portrayal and that is just a hint of something larger. Rapid miner is an extensive information science stage with visual work process plan and full mechanization. It implies that we don't need to do the coding for information mining assignments. Rapid miner is perhaps of the most well-known datum science apparatuses.

This is the graphical UI of the clear cycle in rapid miner. It has the store that holds our dataset. We can import our own datasets. It likewise offers numerous public datasets that we can attempt. We can likewise work with an information base association. The strategies that the item helps with are AI, significant learning, text mining and judicious examination. Easy to use GUI (Graphical User Interface) mechanical assemblies that take you through the showing framework. This device made absolutely in Java is an open-source framework and is amazingly popular in the data mining world [14].

### 2. Tableau

TABLEAU is a great data representation and business insight instrument utilized for revealing and dissecting huge volumes of information. An American organization began in 2003 — in June 2019, Salesforce obtained Tableau. It assists clients with making various diagrams, charts, guides, dashboards, and stories for imagining and breaking down information, to help in pursuing business choices

Tableau has a ton of extraordinary, energizing elements that make it perhaps of the most well-known device in business knowledge (BI). We should find out about a portion of the fundamental Tableau Desktop highlights. Since it has become so undeniably obvious what is scene precisely, let us see a portion of its notable elements [15].

#### TABLEAU FEATURES

- Tableau upholds strong information disclosure and investigation that empowers clients to address significant inquiries in a flash
- No earlier programming information is required; clients without significant experience can begin promptly with making representations utilizing Tableau
- It can associate with a few information sources that other BI devices don't uphold. Scene empowers clients to make reports by joining and mixing different datasets
- Tableau Server upholds a concentrated area to deal with all distributed information sources inside an association

### 3. Jupyter

Jupyter Notebooks give an information representation structure called Qviz that empowers you to envision data-frames with improved outlining choices and Python plots on the Spark driver. Qviz gives a presentation capability that empowers you to plot graphs, for example, table diagram, pie outline, line graph, and region graph for the accompanying information types:

- Spark data-frames
- pandas data-frames
- SQL (%%sql) sorcery

You can likewise make representation for custom plots straightforwardly on the Spark driver by utilizing the upheld Python libraries [16]. The **Quiz** framework provides various options to visualize data and customize the charts.

- [Visualizing Spark Data frames](#)
- [Visualizing Pandas data frames](#)
- [Visualizing SQL](#)
- [Visualising Using Python Plotting Libraries](#)
- [Using Quiz Options](#)

#### 4. Google Chart

One of the central parts in the data visualization market space, Google Charts, coded with SVG and HTML5, is popular for its capacity to deliver graphical and pictorial information perceptions. Google Charts offers zoom usefulness, and it furnishes clients with unequalled cross-stage similarity with iOS, Android, and, surprisingly, the prior renditions of the Internet Explorer program.

The Pros of Google Charts:

- Easy to use stage
- Simple to incorporate information
- Outwardly appealing information charts
- Similarity with Google items.

The Cons of Google Charts:

- The product highlight needs tweaking
- Lacking demos on apparatuses
- Needs customization capacities
- Network availability expected for representation

#### 5. Sisense

Sisense is a data visualization tool that permits you to make intuitive perceptions from your information without any problem. With Sisense, you can rapidly and effectively make broad, instructive dashboards that will assist you with understanding your information better. It has an exceptionally strong yet basic and natural point of interaction that permits you to move your information onto the material and make representations with a couple of snaps of a mouse.

It is additionally completely coordinated with a few BI devices like Microsoft Excel, BIRT, Pentaho, Qlikview and Tableau. Sisense uses multi-faceted in-memory innovation that is intended for Big Data. It additionally has an implanted man-made reasoning motor with prescient examination, permitting you to handily envision information drifts and find stowed away examples in your information.

#### 6. Plotly

Plotly is a data representation device that is utilized to make intelligent diagrams, graphs, and guides. You can likewise utilize Plotly to make a visualization of a dataset, then share the connection of that representation with your perusers via virtual entertainment or on your blog [17]. Diagrams made on Plotly are intelligent and have a remarkable URL, so they're simple for you to share. Pursuers can investigate how you made them by floating over data of interest and survey data about them.

pursuers can likewise investigate every one of the information intuitively as opposed to attempting to interpret your code, which makes it ideal for sharing both intelligent plots and datasets with your crowd. Plotly's connection point is not difficult to utilize, so you can make delightful diagrams quicker than any time in recent memory. Likewise, Plotly highlights a huge library of open-source representation types, permitting you to browse various plots and guides.

#### 7. ZOHO Analytics

ZOHO Analytics is a business data examination stage that uses a collection of devices, which consolidate KPI (Key Progress Indicator) contraptions, pivot tables, and inconceivable view parts that engage it to create reports that go with critical business encounters. The stage, in the past known as Zoho Reports, propels joint exertion, engaging clients and their partners to participate in report improvement and dynamic. To make things shockingly better, the game plan allows clients to embed essentially any report or dashboard in their applications, locales, and online diaries [18].

The dealer makes its guarantee to structure security by using simply top tier wellbeing endeavours, which consolidate an

encoded affiliation and best it among all Data Visualization apparatuses. Application creators and ISVs and specialists can, in this way benefit by involving the item as it licenses them to coordinate and manufacture uncovering and precise capacities into their establishment. ZOHO Analytics free fundamental licenses you to find a useful speed incorporates first hand at no cost and without obligation.

#### 8. Qlikview

QlikView is Qlik's exemplary examination answer for quickly growing exceptionally intelligent directed investigation applications and dashboards, conveying knowledge to settle business challenges. The cutting edge examination period genuinely started with the send off of QlikView and the game-changing Cooperative Motor it is based on. Reforming the manner in which associations use information with natural visual revelation and bragging a client base 36,000, QlikView put Business Knowledge (BI) under the control of additional individuals than any time in recent memory.

QlikView is a BI (Business Intelligence) information revelation item for making directed investigation applications and dashboards tailor-made for business challenges. A key part in the information perception market, Qlikview gives answers for north of 40,000 clients in 100 nations. Qlikview's information perception instrument, other than empowering sped up, redid representations, likewise integrates a scope of strong highlights, including investigation, undertaking revealing, and Business Intelligence capacities [19].

The Pros of QlikView:

- Easy to use interface
- Engaging, bright representations
- Inconvenience free upkeep
- A practical arrangement

The Cons of QlikView:

- Smash constraints
- Unfortunate client care
- Does exclude the 'simplified' include

## V. TYPES OF DATA VISUALIZATION

In the beginning of discernment, the most generally perceived portrayal methodology was using a Microsoft Excel accounting sheet to change the information into a table, visual chart or pie diagram [20]. While these discernment methodologies are still commonly used, more astounding techniques are by and by available, including the accompanying:

There are various procedures and devices you can use to imagine information, so you need to know which ones to utilize and when. Here are probably the main information perception procedures all experts ought to be aware.

The uses of data visualization as follows:

1. Strong method for investigating information with respectable outcomes.
2. Essential use is the pre-handling part of the information mining process.
3. Upholds the information cleaning process by tracking down wrong and missing qualities.
4. For variable deduction and determination means to figure out which variable to incorporate and disposed of in the examination.
5. Likewise assume a part in consolidating classifications as a feature of the information decrease process.

### 5.1 Data Visualization Techniques

Contingent upon these variables, you can pick various information perception methods and design their highlights. Here are the normal kinds of information representation strategies:

**Chart:** The most straightforward method for showing the improvement of one or a few informational indexes is a diagram. Graphs fluctuate from bar and line diagrams that show the connection between components over the long run to pie outlines that exhibit the parts or extents between the components of one entirety.

**Plots:** Plots permit to circulate at least two informational indexes over a 2D or even 3D space to show the connection between these sets and the boundaries on the plot. Plots likewise fluctuate. Dissipate and bubble plots are the absolute most broadly utilized representations. With regards to enormous information, investigators frequently utilize more complicated box plots to picture the connections between huge volumes of information.

**Maps:** Maps are well known methods utilized for information representation in various businesses. They permit finding components on important articles and regions — geological guides, building plans, site designs, and so on. Among the most well-known map representations are heat maps, spot dissemination maps, cartograms.

**Diagrams and Matrices:** Diagrams are normally used to exhibit complex information connections and connections and remember different sorts of information for one visual portrayal. They can be various leveled, multi-layered, tree-like. Matrices is one of the high-level information perception methods that assist with deciding the connection between's numerous



continually refreshing (steaming) informational collections.

## About Dataset

Gap Minder gathers information from a small bunch of sources, including the Institute for Health Metrics and Evaluation, the US Census Bureau's International Database, the United Nations Statistics Division, and the World Bank.

## Variable Name and Description of Indicator:

**country:** Unique Identifier

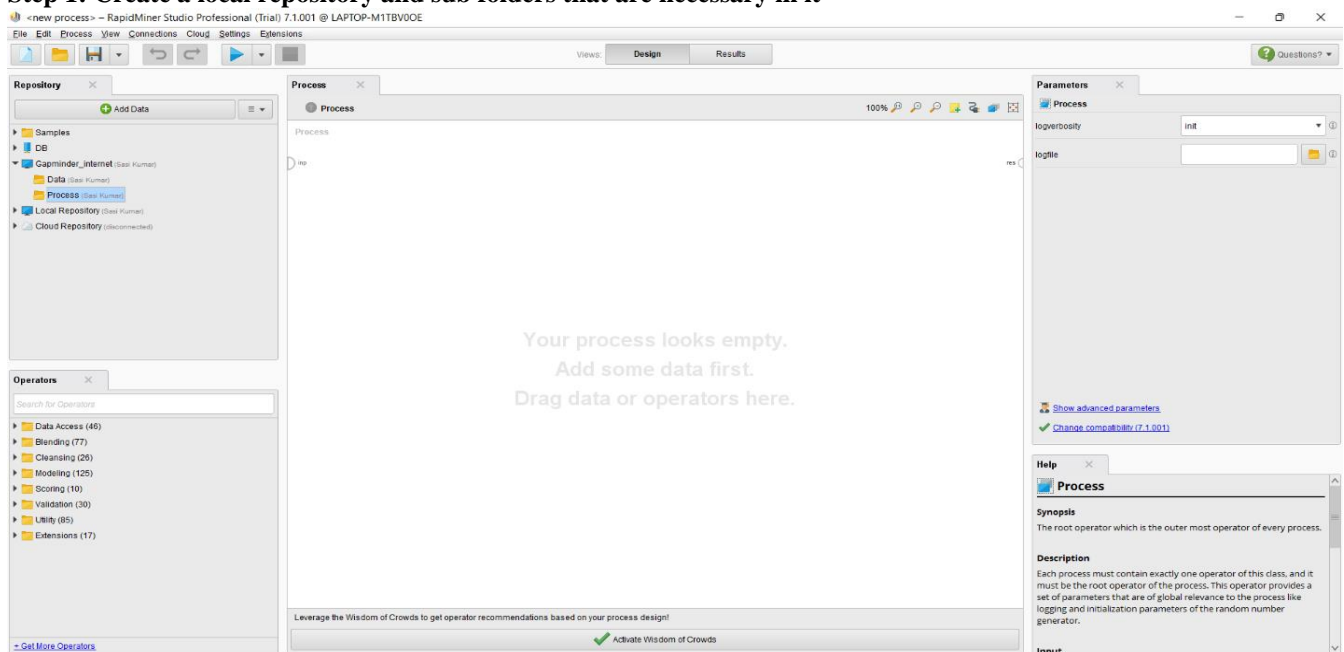
**incomeperperson:** Gross Domestic Product per capita in consistent 2000 US\$. The expansion yet not the distinctions in that frame of mind of living between nations has been considered.

**internetuserate:** Internet clients (per 100 individuals). Web clients are individuals with admittance to the overall organization.

**urbanrate:** Urban populace (% of aggregate) Urban populace alludes to individuals living in metropolitan regions as characterized by public measurable workplaces (determined utilizing World Bank populace gauges and metropolitan proportions from the United Nations World Urbanization Prospects).

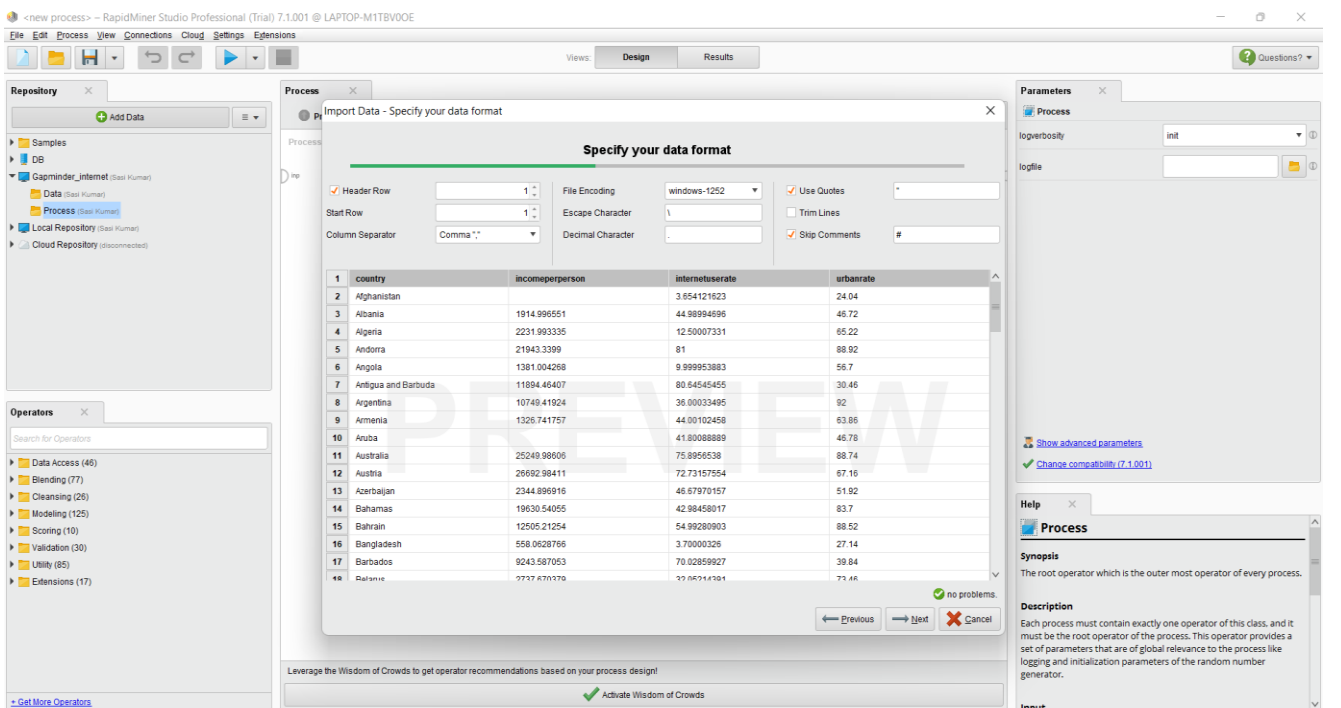
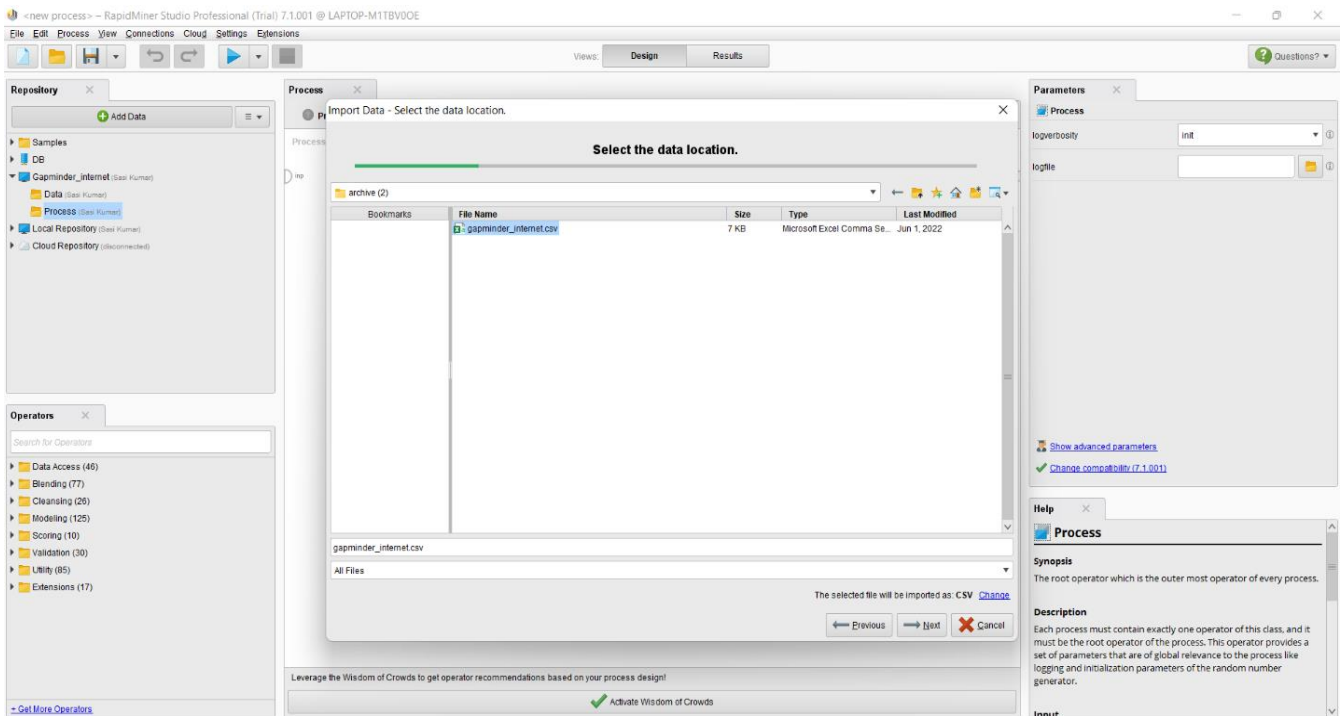
We have done the data visualization by using Rapid Miner

## Step 1: Create a local repository and sub folders that are necessary in it



**Step 2: Import the data set in the repository either in the csv or xlsx format. Change the data format of your data by column wise accordingly and select the sub folder that you want to import**





new process - RapidMiner Studio Professional (Trial) 7.1.001 @ LAPTOP-M1TBV0OE

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Questions?

Repository

- Samples
- DB
- Gapminder\_internet (Sasi Kumar)
  - Data (Sasi Kumar)
  - Process (Sasi Kumar)
- Local Repository (Sasi Kumar)
- Cloud Repository (disconnected)

Operators

Search for Operators

- Data Access (46)
- Blending (77)
- Cleansing (26)
- Modeling (125)
- Scoring (10)
- Validation (30)
- Utility (85)
- Extensions (17)

Process

Import Data - Format your columns.

Format your columns.

Date format: MMM d, yyyy h:mm:ss a z  Replace errors with missing values: ?

country	incomeperperson	internetuserate	urbanrate
polynomial	real	real	real
1 Afghanistan	?	3.654	24.040
2 Albania	1914.997	44.990	46.720
3 Algeria	2231.993	12.500	65.220
4 Andorra	21943.340	81.000	88.920
5 Angola	1381.004	10.000	56.700
6 Antigua and Barbuda	11894.464	80.645	30.460
7 Argentina	10749.419	36.000	92.000
8 Armenia	1326.742	44.001	63.860
9 Aruba	?	41.801	46.700
10 Australia	25249.986	75.895	88.740
11 Austria	26692.984	72.732	67.160
12 Azerbaijan	2344.897	46.680	51.920
13 Bahamas	19630.541	42.985	83.700
14 Bahrain	12505.213	54.993	88.520
15 Bangladesh	558.063	3.700	27.140
16 Barbados	9243.587	70.029	39.840
17 Belarus	2737.670	32.952	73.460
18 Belgium	24496.048	73.734	97.360
19 Belize	1645.657	12.646	51.700

no problems.

Previous Next Cancel

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Parameters

Process

logverbosity: init

logfile: [empty]

Show advanced parameters

Change compatibility (7.1.001)

Help

Process

Synopsis

The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random number generator.

Inout

new process - RapidMiner Studio Professional (Trial) 7.1.001 @ LAPTOP-M1TBV0OE

File Edit Process View Connections Cloud Settings Extensions

Views: Design Results

Questions?

Repository

- Samples
- DB
- Gapminder\_internet (Sasi Kumar)
  - Data (Sasi Kumar)
  - Process (Sasi Kumar)
- Local Repository (Sasi Kumar)
- Cloud Repository (disconnected)

Operators

Search for Operators

- Data Access (46)
- Blending (77)
- Cleansing (26)
- Modeling (125)
- Scoring (10)
- Validation (30)
- Utility (85)
- Extensions (17)

Process

Import Data - Where to store the data?

Where to store the data?

- Gapminder\_internet (Sasi Kumar)
  - Data (Sasi Kumar)
  - Process (Sasi Kumar)
- Local Repository (Sasi Kumar)
- Cloud Repository (disconnected)

Name: gapminder\_internet

Location: /Gapminder\_internetData/gapminder\_internet

Previous Finish Cancel

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Parameters

Process

logverbosity: init

logfile: [empty]

Show advanced parameters

Change compatibility (7.1.001)

Help

Process

Synopsis

The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random number generator.

Inout

ExampleSet (213 examples, 0 special attributes, 4 regular attributes)

Row No.	country	incomeperp...	internetuser...	urbanrate
1	Afghanistan	?	3.654	24.040
2	Albania	1914.997	44.990	46.720
3	Algeria	2231.993	12.500	65.220
4	Andorra	21943.240	81	88.920
5	Angola	1381.004	10.000	56.700
6	Antigua and ...	11894.464	80.645	30.460
7	Argentina	10749.419	36.000	92
8	Armenia	1326.742	44.001	63.860
9	Aruba	?	41.801	46.780
10	Australia	25249.986	75.895	88.740
11	Austria	26692.984	72.732	67.160
12	Azerbaijan	2344.897	46.680	51.920
13	Bahamas	19630.541	42.985	83.700
14	Bahrain	12505.213	54.993	88.520
15	Bangladesh	558.063	3.700	27.140
16	Barbados	9243.587	70.029	39.840
17	Belarus	2737.670	32.052	73.460
18	Belgium	24496.048	73.734	97.360
19	Belize	3545.652	12.640	51.700
20	Benin	377.040	3.130	41.200
21	Bermuda	62682.147	84.655	100
22	Bhutan	1324.195	13.599	34.480
23	Bolivia	1232.794	20.002	65.580
24	Bosnia and H...	2183.345	52.002	47.440
25	Botswana	4189.437	6.000	59.580
26	Brazil	4699.411	40.650	85.580

### Step 3: Replace the missing values in your data set and do the pre-processing activities that are necessary.

The screenshot shows the RapidMiner Studio Professional interface. The main workspace displays a workflow with three operators: 'Retrieve gapminder\_internet', 'Replace Missing Values', and 'Filter Examples'. The 'Replace Missing Values' operator is currently selected, and its parameters are visible in the 'Parameters' pane on the right. The 'Parameters' pane shows 'repository entry' set to 'me/Data/gapminder\_internet'. The 'Operators' pane on the left shows the 'Replace Missing Values' operator selected under the 'Missing' category. The 'Help' pane on the right provides a synopsis and description for the 'Retrieve' operator.

The screenshot shows the 'Filter Examples' operator output in the 'Results' view. The output is a table with 26 rows and 5 columns: 'Row No.', 'country', 'incomeperp...', 'internetuser...', and 'urbanrate'. The table contains data for various countries, including Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Aruba, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bermuda, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, and Brazil.

Row No.	country	incomeperp...	internetuser...	urbanrate
1	Afghanistan	8740.966	3.654	24.040
2	Albania	1914.997	44.990	48.720
3	Algeria	2231.993	12.500	65.220
4	Andorra	21943.340	81	88.920
5	Angola	1381.004	10.000	56.700
6	Antigua and ...	11894.464	80.645	30.460
7	Argentina	10749.419	36.000	92
8	Armenia	1326.742	44.001	63.860
9	Aruba	8740.966	41.801	46.780
10	Australia	25249.966	75.896	88.740
11	Austria	26692.984	72.732	67.160
12	Azerbaijan	2344.897	46.680	51.920
13	Bahamas	19630.541	42.985	83.700
14	Bahrain	12505.213	54.993	88.520
15	Bangladesh	558.063	3.700	27.140
16	Barbados	9243.587	70.029	39.840
17	Belarus	2737.670	32.052	73.460
18	Belgium	24496.048	73.734	97.360
19	Belize	3545.652	12.646	51.700
20	Benin	377.040	3.130	41.200
21	Bermuda	62682.147	84.655	100
22	Bhutan	1324.195	13.599	34.480
23	Bolivia	1232.794	20.002	65.580
24	Bosnia and H...	2183.345	52.002	47.440
25	Botswana	4189.437	6.000	59.580
26	Brazil	4699.411	40.650	85.580

### 1. PIE CHART

Pie charts are perhaps of the most well-known and fundamental datum perception procedures, utilized across a great many applications. Pie diagrams are great for outlining extents, or part-to-entire correlations.

Since pie chart are moderately basic and simple to peruse, they're the most ideal for crowds who may be new to the data or are just inspired by the key action items. For watchers who require a more careful clarification of the information, pie diagrams miss the mark in their capacity to show complex data.

Pie charts are among the most popular data visualisations, but they're not always the best for comparing data values. This is mostly caused by the size of each slice and the quantity of represented data categories. Doughnut charts, which are typically preferable for users to compare the size of each slice or arc, can also be used to illustrate pie charts.

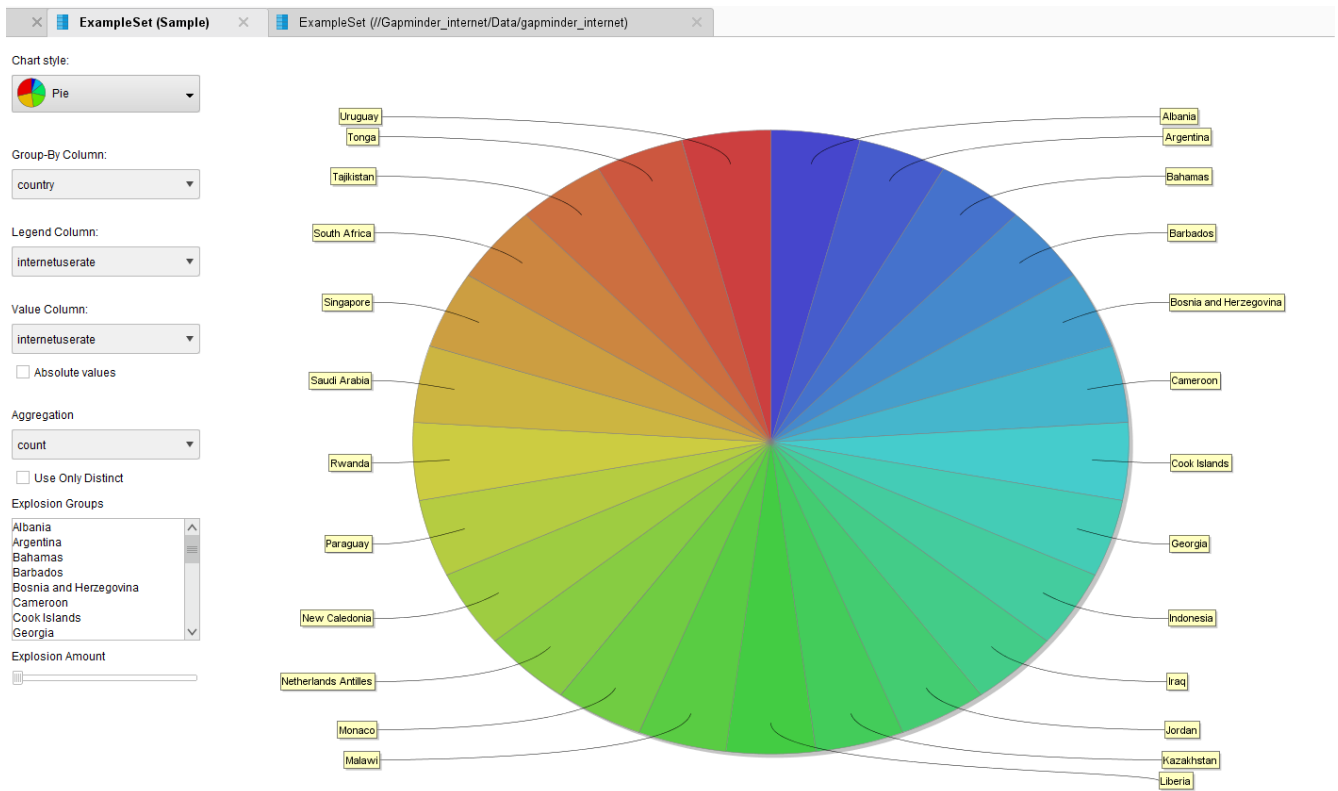


Figure 1. Pie Chart

## 2. BAR GRAPH

The exemplary bar diagram, or reference chart, is another normal and simple to-utilize technique for information representation. In this kind of representation, one hub of the outline shows the classes being looked at, and the other, a deliberate worth. The length of the bar demonstrates how each gathering measures as per the worth.

One disadvantage is that marking and lucidity can become tricky when there are an excessive number of classifications included. Like pie diagrams, they can likewise be excessively straightforward for additional intricate informational collections.

Subsequently, upsides of a class are tended to with the guide of bars and they can be planned with vertical or level bars, with the length or level of each bar tending to the worth.

To look at information over the long run or the information is gathered in various areas like various businesses, assortment of food, and so on, a Visual chart is the most ideal choice for certain qualities or a few kinds of careful thoughts.

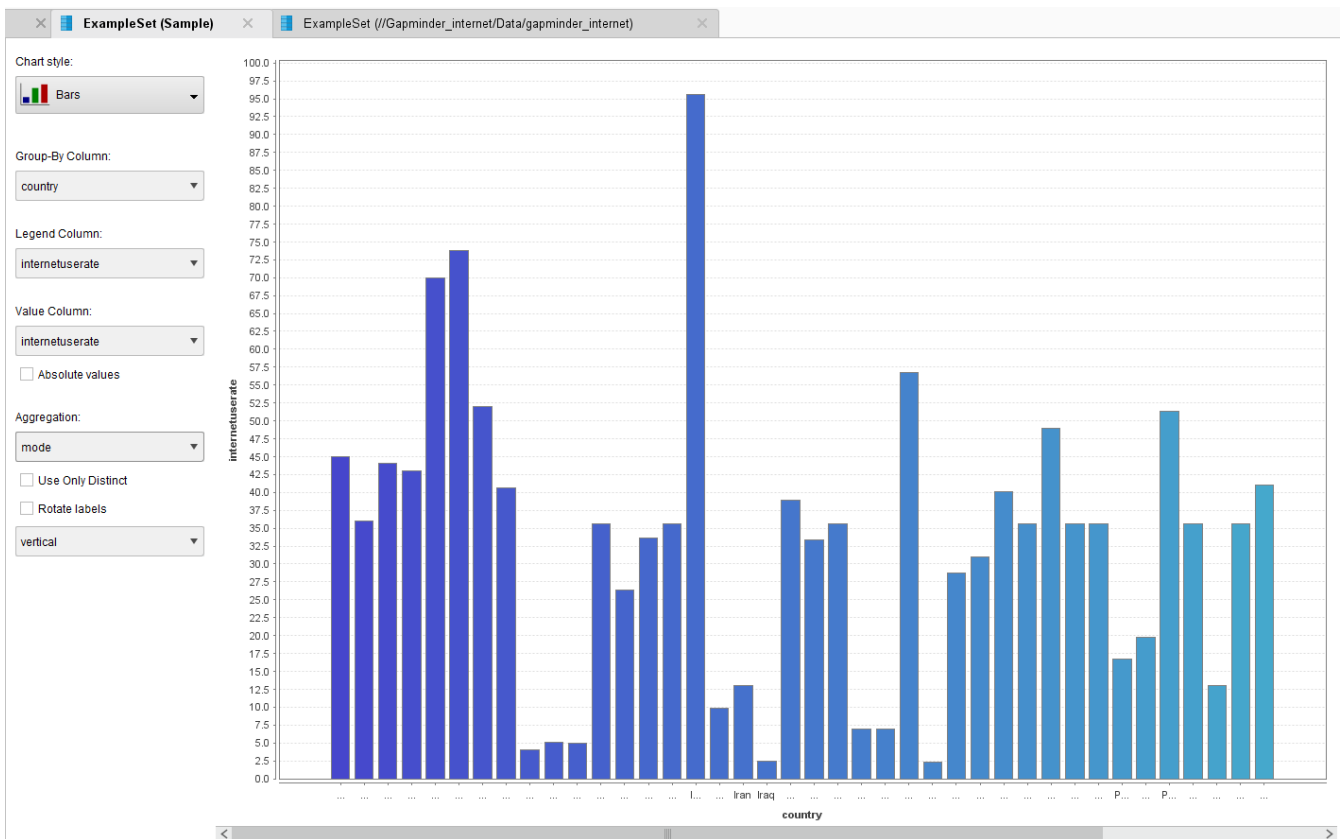


Figure 2. Bar Chart

### 3. HISTOGRAM

Not at all like bar diagrams, histograms show the circulation of information over a constant span or characterized period. These perceptions are useful in distinguishing where values are concentrated, as well as where there are holes or surprising qualities.

Histograms are particularly helpful for showing the recurrence of a specific event. For example, on the off chance that you might want to show the number of snaps your site that got every day over the course of the past week, you can utilize a histogram. From this perception, you can rapidly figure out which days your site saw the best and least number of snaps.

- Taller bars show that more information falls there.
- A histogram shows the shape and spread of nonstop example information.
- A plot permits you to find, and show, the fundamental recurrence dispersion (state) of a bunch of persistent information.
- This allows the evaluation of the information for its fundamental dispersion, skewness, anomalies, etc.
- It is a careful depiction of the dispersion of numerical information and it relates only a solitary variable.
- Consolidates receptacle or pail the scope of values that parcel the entire scope of values into a movement of stretches and a while later really look at the quantity of values that fall into each span.
- Containers are successive, non-covering time frames. As the adjoining containers leave no holes, the square shapes states of the histogram reach each other to show that the primary worth is persistent.



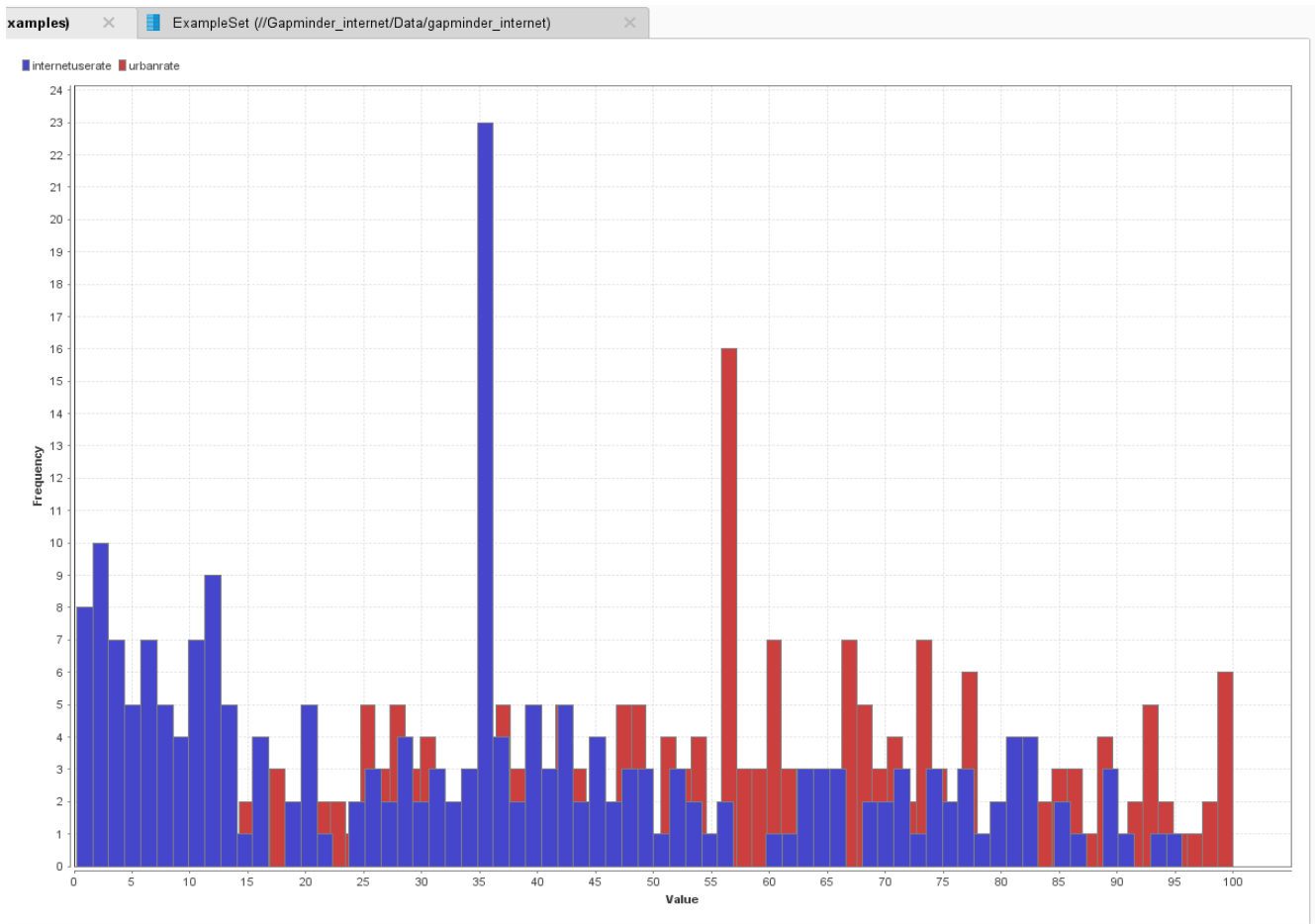


Figure 3. Histogram

#### 4. SCATTER PLOT

One more strategy regularly used to show information is a disperse plot. A disperse plot shows information for two factors as addressed by focuses plotted against the level and vertical hub. This kind of information representation is valuable in showing the connections that exist among factors and can be utilized to recognize patterns or relationships in information.

Dissipate plots are best for genuinely huge informational indexes, since it's frequently simpler to distinguish patterns when there are more information focuses present. Moreover, the nearer the information focuses are gathered, the more grounded the relationship or pattern will in general be.

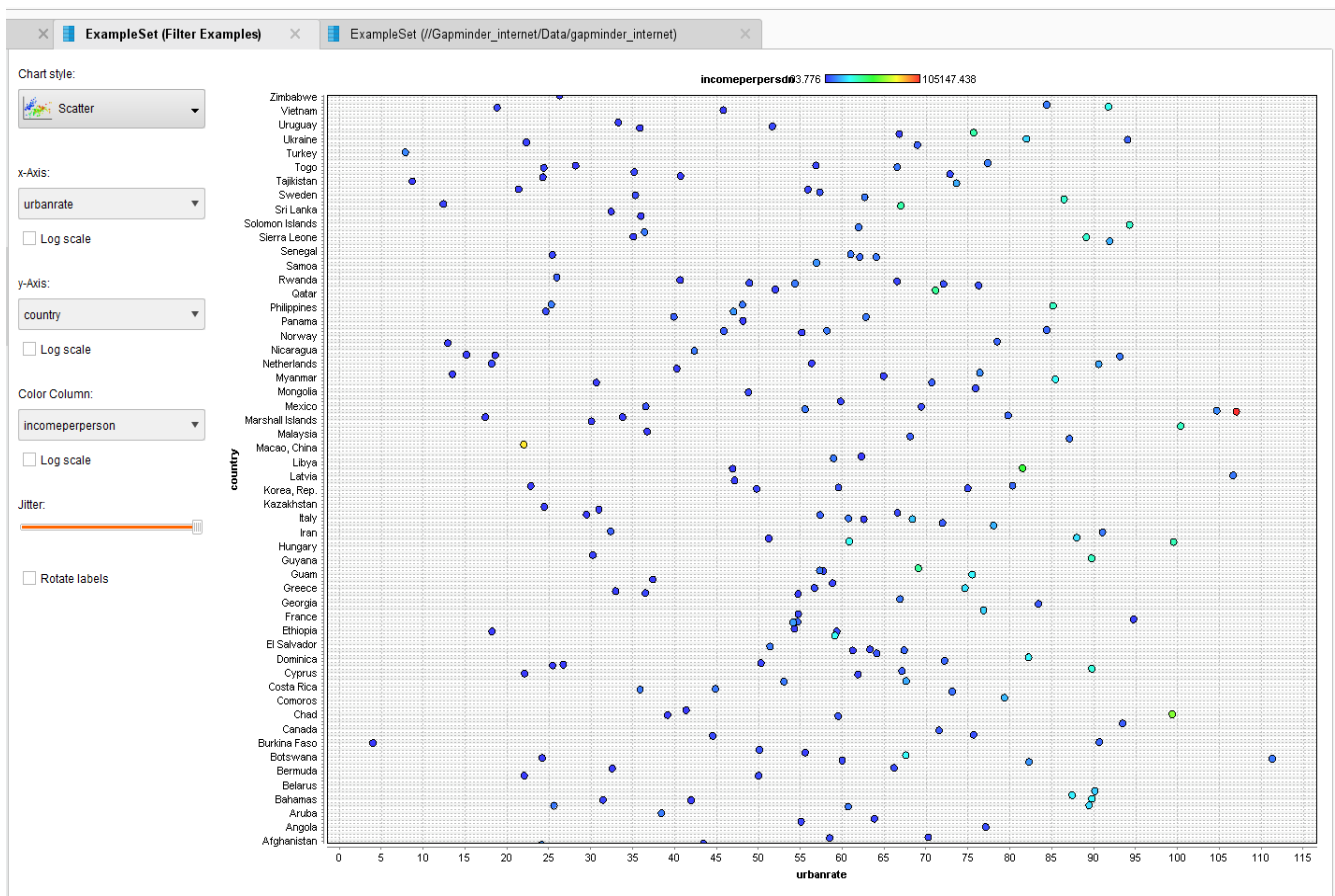


Figure 4. Scatter Plot

## 5. TIME LINES

Timelines are the best method for imagining a succession of occasions in sequential request. They're ordinarily direct, with key occasions illustrated along the hub. Courses of events are utilized to impart time-related data and show verifiable information.

Timetables permit you to feature the main occasions that happened, or have to happen from here on out, and make it simple for the watcher to recognize any examples showing up inside the chose time span. While timetables are in many cases generally straightforward direct perceptions, they can be made all the more outwardly engaging by adding pictures, tones, textual styles, and brightening shapes.

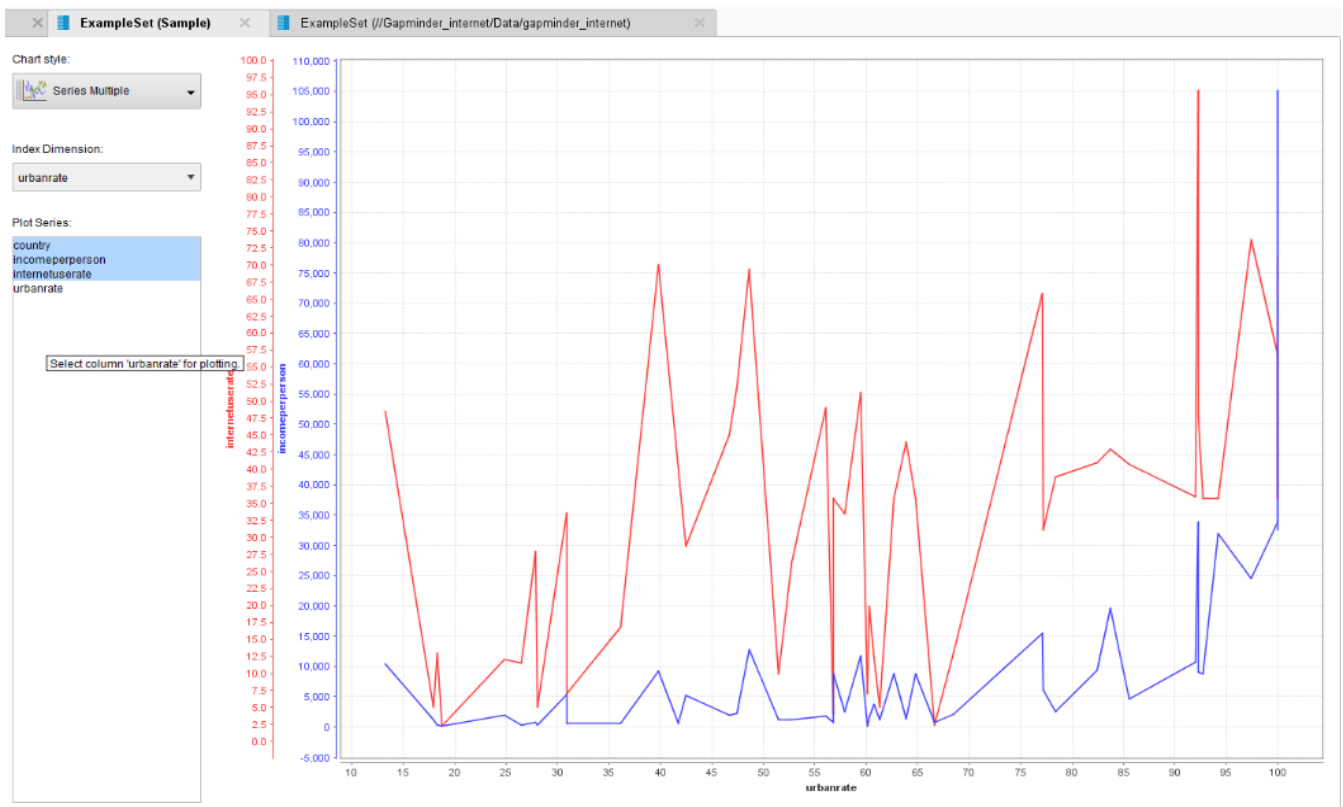


Figure 5. Timeline Chart

## 6. AREA CHART

In data visualization, an area diagram is an expansion of a line chart. It joins the line graph and the bar diagram to uncover how at least one gathering's numeric quality change north of a subsequent variable. An area chart is an extraordinary graph to imagine a volume change throughout some undefined time frame. It gives a feeling of summation of the quantitative information.

Data is plotted on the x-and y-hub. D qualities are plotted utilizing information focuses that are associated utilizing line sections. Dissimilar to inline outlines, the region between the line and x-hub is loaded up with variety or concealing in a space diagram.

An area chart combines the line chart and bar chart to show how one or more groups' numeric values change over the progression of a second variable, typically that of time. An area chart is distinguished from a line chart by the addition of shading between lines and a baseline, like in a bar chart. Basically, the X axis represents time or an ordered variable, and the Y axis gives the value of another variable. Data points are connected by straight line segments and the area between the x axis and the line is filled in with colour or shading.

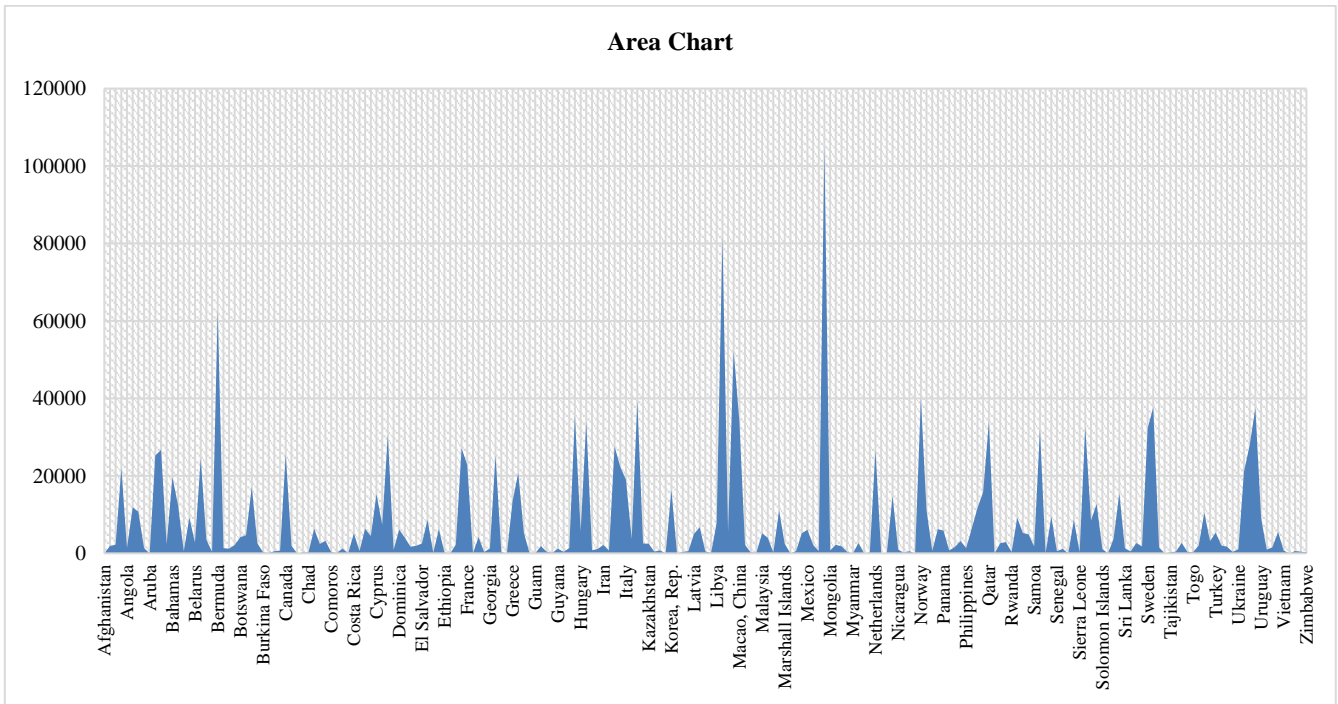


Figure 6. Area Chart

## 7. NETWORK DIAGRAMS

Network outlines are a sort of information perception that address connections between subjective data of interest. These perceptions are made out of hubs and connections, likewise called edges. Hubs are solitary information focuses that are associated with different hubs through edges, which show the connection between various hubs.

There are many use cases for network outlines, including portraying informal communities, featuring the connections between workers at an association, or envisioning item deals across geographic districts.

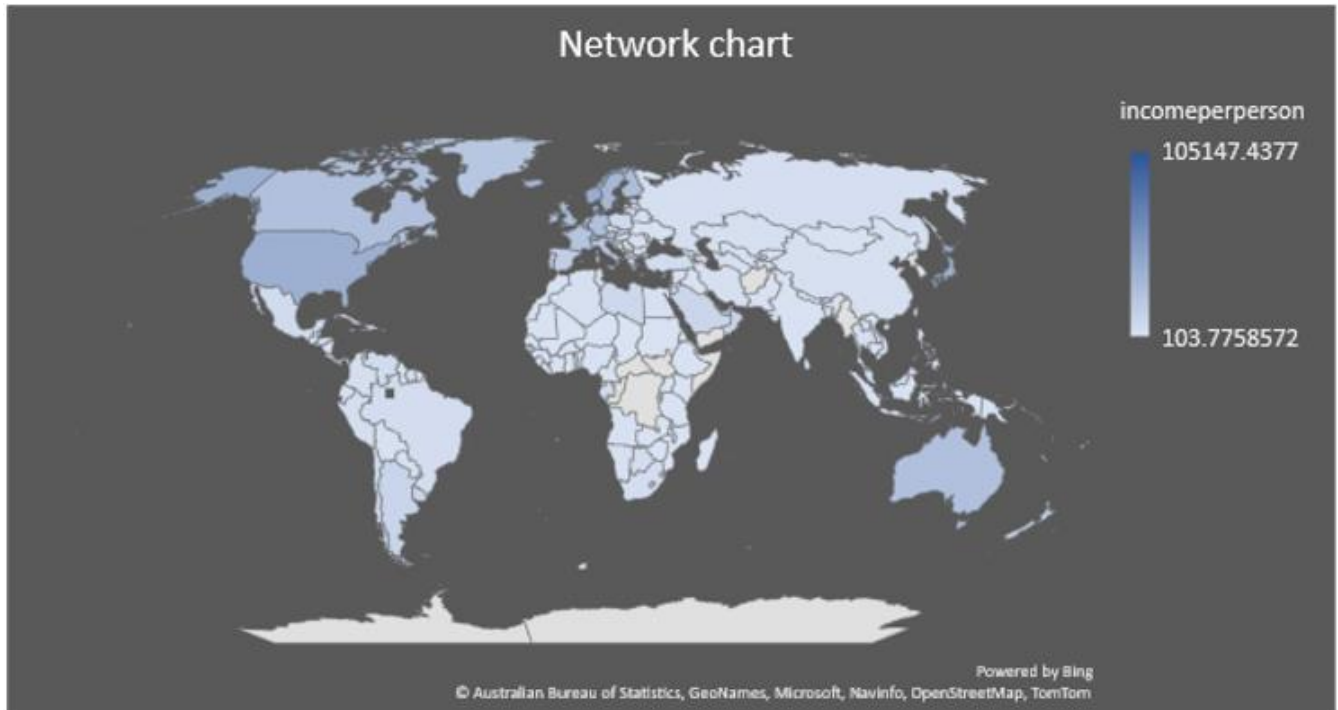


Figure 7. Network Diagrams

## 8. TREE MAPS

Tree maps are representations for various levelled information. They are made of a progression of settled square shapes of sizes relative to the comparing information esteem. An enormous square shape addresses a part of an information tree, and it is partitioned into more modest square shapes that address the size of every hub inside that branch.

Tree maps are generally tracked down on information dashboards. Originators frequently pick them to include visual assortment a thick dashboard. Notwithstanding, tree maps are a mind-boggling perception and present numerous impediments to fast understanding (which is the fundamental necessity for any data showed on a dashboard). Tree maps are frequently utilized for deals information, as they catch relative sizes of information classes, taking into account fast impression of the things that are enormous supporters of every classification.

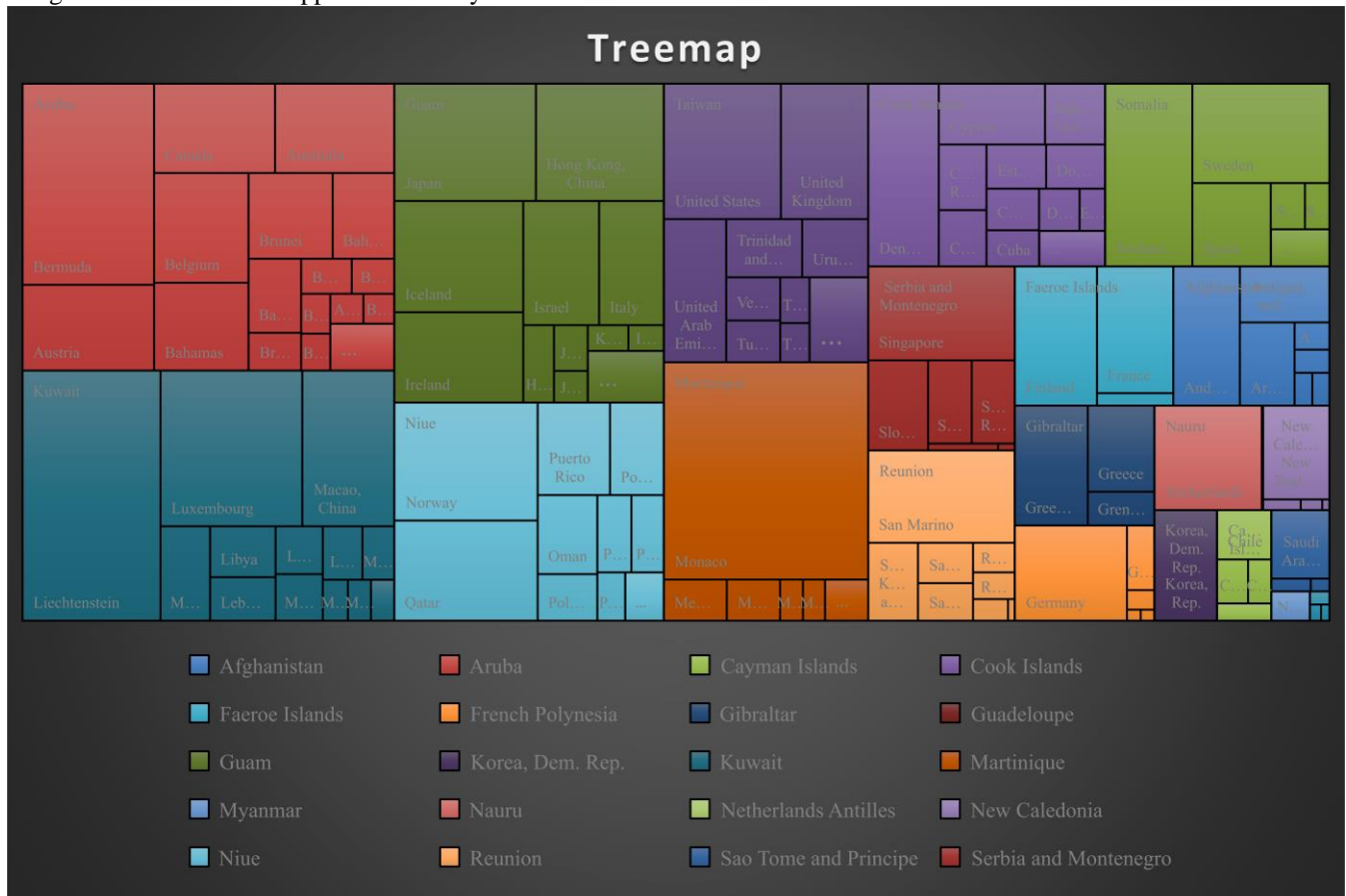


Figure 8. Tree Maps

## VI. CONCLUSION

In the modern world, information is all over and brands should have the option to unravel and impart their message in a successful way. Also, for information researchers, learning and staying aware of the relative multitude of most recent information representation instruments is fundamental, and solely after they ace this craftsmanship, they can stay aware of the speed of enormous information, and the quick domains of man-made intelligence and ML Data visualization is showing information in the illustrations structure so that it very well may be effectively reasonable. Various information perception methods are utilized to show information in visual structure. It has arisen as a strong and generally pertinent device for breaking down and deciphering enormous and complex information. It has turned into a fast, simple method for conveying ideas in a widespread configuration. It should discuss complex thoughts with lucidity, precision, and proficiency. These advantages have permitted information perception to be helpful in many fields of study. we have incorporated the dataset of Gap Minder accumulates data from a little pack of sources, including the Institute for Health Metrics and Evaluation, the US Census Bureau's International Database, the United Nations Statistics Division, and the World Bank. we utilized Rapid digger device and imported the dataset in the apparatus and diminished the missing qualities introduced in the dataset and imagined the measurements and drawn the various kinds of information visualization.

## REFERENCES

- [1] Chen CH, Härdle WK, Unwin A, editors. Handbook of data visualization. Springer Science & Business Media; 2007 Dec 18.
- [2] Shweta Srivastav, Simon Lannon, Donald k. Alexander, and phil jones, —A Review and Comparison of Data Visualization Techniques Used in Building Design and in Building Simulationl, Eleventh international Ibpsa conference Glasgow, Scotland, 2009
- [3] Ben Shneiderman, Catherine Plaisant, Strategies for Evaluating Information Visualization Tools: —Multidimensional In-depth Long-term Case Studiesl, Proceedings of the BELIV'06 workshop Advanced Visual Interfaces Conference 2006, Venice
- [4] S. Few. Benefitting infovis with visual difficulties? Provocation without a cause. —Visual Business Intelligence Newsletter, 2011
- [5] Meloncon L, Warner E. Data visualizations: A literature review and opportunities for technical and professional communication. In2017 IEEE International Professional Communication Conference (ProComm) 2017 Jul 23 (pp. 1-9). IEEE.
- [6] Michelle A. Borkin, Student Member, IEEE, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Student Member, IEEE, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister, Senior Member, IEEE, —What Makes a Visualization Memorable?l, Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
- [7] Daniel A. Keim, —Information Visualization And Visual Data Miningl, —Ieee Transactions On Visualization And Computer Graphicsl, Vol. 7, No. 1, January-March 2002
- [8] Liu J, Tang T, Wang W, Xu B, Kong X, Xia F. A survey of scholarly data visualization. IEEE access. 2018 Mar 12;6:19205-21.
- [9] Grainger S, Mao F, Buytaert W. Environmental data visualisation for non-scientific contexts: Literature review and design framework. Environmental Modelling & Software. 2016 Nov 1;85:299-318.
- [10] Muzammil Khan, Sarwar Shah Khan, Data and Information Visualization Methods, and Interactive Mechanisms: —A Survey, International Journal of Computer Applicationl (0975-8887), Volume 34– No.1, November 2011
- [11] R. R. Laher, “Thoth: software for data visualization and statistics,” Astronomy and Computing, vol. 17, 2016, pp. 177-185.
- [12] Melanie Tory And Torsten Moller, Human Factors In Visualization Research, —Ieee Transactions On Visualization And Computer Graphicsl, Vol. 10, No. 1, January/February 2004
- [13] Paul kent, Visualization D. Making Big Data Approachable and Valuable. Whitepaper, Source: IDG Research Services. 2012 Aug.
- [14] Ma S, Chowdhury SK. Application of LC–high-resolution MS with ‘intelligent’ data mining tools for screening reactive drug metabolites. Bioanalysis. 2012 Mar;4(5):501-10.
- [15] Murphy SA. Data visualization and rapid analytics: Applying tableau desktop to support library decision-making. Journal of Web Librarianship. 2013 Oct 1;7(4):465-76.
- [16] Ono JP, Freire J, Silva CT. Interactive data visualization in jupyter notebooks. Computing in Science & Engineering. 2021 Mar 31;23(2):99-106.
- [17] Sievert C. Interactive web-based data visualization with R, plotly, and shiny. CRC Press; 2020 Jan 30.
- [18] Nunes F, Correa C, Jandrey A, Barcelos A, Reyes D, Bernardes M, Sales A, Silveira MS. Data visualization on focus: exploring communicability of dashboards generated from BI tools. InProceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems 2020 Oct 26 (pp. 1-6).
- [19] Pover K. Mastering QlikView Data Visualization. Packt Publishing Ltd; 2016 Apr 25.
- [20] Sadiku M, Shadare AE, Musa SM, Akujuobi CM, Perry R. Data visualization. International Journal of Engineering Research And Advanced Technology (IJERAT). 2016 Dec;2(12):11-6.



# Chapter - 14

## Data Analytics for Disease Prediction

S. Menaga<sup>1</sup>, Dr.G.Kalaiarasi<sup>2</sup>, Dr. R.Vanithamani<sup>3</sup>, M.Nivetha<sup>4</sup>

<sup>1,4</sup> Assistant Professor, Department of Electronics and Communication Engineering,  
Jai Shriram Engineering College, Tiruppur, India

<sup>2</sup> Associate Professor, Department of Electronics and Communication Engineering, VSB Engineering College, Karur, India

<sup>3</sup> Professor, Department of Biomedical Instrumentation Engineering, School of Engineering,  
Avinashilingam Institute for Home science and Higher Education for Women, India

E-mail: [1sri.ece09@gmail.com](mailto:1sri.ece09@gmail.com), [2kalaiibe@gmail.com](mailto:2kalaiibe@gmail.com), [3vanithamani\\_bmie@avinuty.ac.in](mailto:3vanithamani_bmie@avinuty.ac.in), [4niveathaee@gmail.com](mailto:4niveathaee@gmail.com)

*Abstract— Human life in the modern era is influenced by a large number of diseases, which are the major causes of death. When patients exhibit symptoms clearly indicating abnormalities, healthcare systems can treat them. Diagnoses of intense diseases during the early stages allow patients to be treated, thus reducing their risk. In the absence of treatment, chronic conditions develop, sometimes resulting in death. Diagnosis of intensive diseases causes 59 percent of deaths annually. Medical services is a complex structure, containing a wide range of areas that are challenging to manage with excellent accuracy, while at the same time patients demand reasonable prices. In the medical services industry, fresh innovations are being incorporated. Predictive analytics and data science are changing industries because they can predict future outcomes and mitigate risks. Healthcare organizations can use these technologies to gain actionable insights into their patients' data as well as outcomes in order to lower total healthcare costs, recognize individuals at high risk more rapidly, produce real-time notifications, etc. The subjects of clinical decision assistance systems, diagnostic image interpretation, forecasting models, and universal healthcare are highlighted in this chapter.*

*Keywords— Data collection, Big data analytics, Biomedical image analysis, Disease Prediction models, Universal healthcare*

### I. INTRODUCTION

Based entirely on selected truths. In today's society, health forecasts are very significant. Every healthcare application must handle a large amount of data in various formats, and data type, data size and other features are critical to the data handling process. Detecting and predicting diseases, such as diabetes, lung cancer, brain cancer, heart diseases, and liver diseases, requires massive tests, which leads to an increase in patient medical data. Medical data are digitized thereby reforming their dimensions, increasing data size and enhancing the value of analytics. Healthcare data is both fascinating and challenging due to the variety of types of data including health surveys, patient illness information, insurance claims, electronic health records, and administrative information. There are many forms of data in healthcare, including structured, unstructured, and semi-structured data. It brings together all information from various sources, like claims, medical records, and laboratory records. With the help of statistical analysis or big data analytics, we can predict hidden information, and a healthcare analytics system can deliver multiple advantages to patients. The information will allow clinicians to make more accurate decisions and early diagnoses, enabling them to begin treatment more quickly and minimize any long-term damage caused by this disease.

As a result of data science, disease diagnostics and treatment methods can be changed and diseases can be prevented in the future. Using predictive analytics, we can tell when autoimmune patients will have flare-ups, whether their condition is improving or worsening, and how they respond to different treatments.

In order to understand such huge volumes of data, high-end computing resources and algorithms based on Artificial Intelligence (AI) are needed. In order to achieve automatic decision-making, Machine Learning (ML) approaches combine fuzzy logic and neural networks. Data management strategies that are innovative and efficient, cloud-based applications that are intelligent and effective, and user-friendly visualization are essential for gaining practical insight from big data. Data mining is the computing approach to detecting patterns in massive data sets by combining machine learning techniques, statistics and database gadgets. Sincerely, prediction is a forecast of an uncertain event and it depends on certain fact

### II. THE NATURE OF BIG DATA IN THE HEALTH CARE SYSTEM

The term "Big Data" in medical services refers to the large amounts of information generated from many assets, including Electronic Health Records (EHRs), scientific imaging, payer information, drug research, genomic sequencing, wearable sensor gadgets, and clinical gadgets.

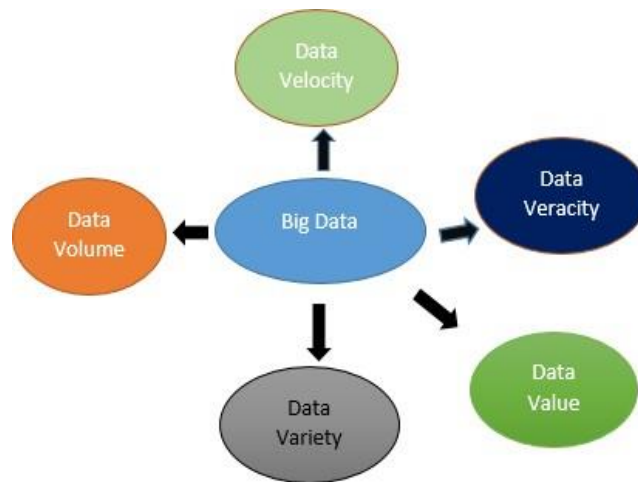
---

© 2022 Technoarete Publishing

S. Menaga – “Data Analytics for Disease Prediction” Pg no: 185 – 202.

<https://dx.doi.org/10.36647/MLAIDA/2022.12.B1.Ch014>

The healthcare industry has traditionally stored the huge amount of data it generates on hard drives. Compared to traditional electronic medical data, it has 3 distinct traits: volumes of information are extraordinarily high and circulate at high prices, it is a complete digital universe for the health industry, and its shape and nature are tremendously variable since it originates from numerous resources [1].



**Figure.1.** Components of Big Data Analytics

The medical industry benefited greatly from the use of big data analytics. Volume of data, data velocity, data diversity, data veracity, and data value are some of the traits of big data. The concept of 5V's in Big Data is illustrated in Figure 1.

In recent years, information has spread exponentially. Data in Big Data are unstructured, composite, noisy, mixed, and represent both size and vision. Below is a detailed discussion of these 5V's.

- **Data Volume:** The word "Big Data" implies an abundance of information. In the past, personal data was created by an individuals. Due to the digital generation of data on platforms like social media, there is a vast amount of data that needs to be processed today.
- **Data Velocity:** A measure of speed is Data Velocity, which refers to flows of data from the origins of data such as professional systems, machines, organizations, and human communication, such as social media and movable devices. As a result, there are a lot of data points and they are constant.
- **Data Veracity:** Bias, noise, and unstructured data are considered to be key elements of veracity in big data. As compared to volume and velocity, big data believes veracity is the key to data analysis.
- **Data Variety:** It refers to the numerous kinds of data that have been produced. Data can be structured or unstructured and this concept draws attention to diversity of sources and classifications. Databases, file systems, spreadsheets, etc., are used to supply data. E-mail, photos, video, PDF, audio, and more are some of the forms of data we receive today.
- **Data Value:** Data value refers to the importance or value of data contained within information. Big Data is essential to understanding value. With a massive amount of data and a variety of formats, it provides quality analytics that allow you to make informed decisions. It provides the actual technology.

### 2.1 APPLICATION, BENEFITS, AND PROSPECTS OF DATA ANALYSIS IN HEALTHCARE

Big data and its utility in medical services and clinical sciences have become more basic with the rise of web-based entertainment (stagesok and Twitter) and cell phone applications that can screen individual well-being boundaries utilizing sensors and analyzers. The purpose of information mining is to find unrivaled treatment and care for clients by their stored data. A comprehensive virtual platform that provides patients with assistance has been developed by data scientists using disease predictive modeling. This platform allows patients to enter their symptoms and receive insights about potential diseases based on their confidence rate. Besides, patients who experience the ill effects of mental issues like misery, uneasiness and neurodegenerative diseases like Alzheimer's can utilize virtual applications to help them with their day-to-day errands. Some notable examples of remote helpers are Ada - A startup situated in Berlin that predicts illness in light of the client's side effects. Furthermore, WoeBot, a chatbot created at Stanford University, gives depression treatment medicines to patients suffering from the condition [1],[8].

### III. DATA COLLECTION

The goal of data collection is to gather and measure information from as many sources as possible. Data collection allows us to analyze past events to find patterns that recur, as well as capture past events so that we can analyze them. In order to predict future changes, machine learning algorithms are used to build predictive models based on those patterns.

Since predictive models can only be as powerful as the data they are based on, effective data collection procedures are essential to the creation of strong predictive models. The data has to be error-free (garbage in, garbage out) and include relevant details for the task in hand in order to get the intended results.

Here are a couple of data sources we shall try:

- Dataset Search by Google – It allows you to search not only by keywords, but also by the type of dataset (e.g., tables, images, text) or by the availability of the dataset for free.
- Data Discovery - specializes in Computer Vision datasets, all of which can be easily categorized and filtered.
- Open ML – In addition to sharing data, this resource allows co-working with other data scientists and solving problems in collaboration
- UC Irvine Machine Learning Repository – In the top 100 most cited computer science resources, there is a collection of datasets and data generators
- Awesome Public Datasets on Github - It would be strange if Github didn't have its own list of datasets, separated by category.
- Kaggle - Sort data sets into categories that have usability scores (an indicator that the dataset is well- documented).
- Amazon Datasets – Using Amazon Web Services, there are many datasets available in S3

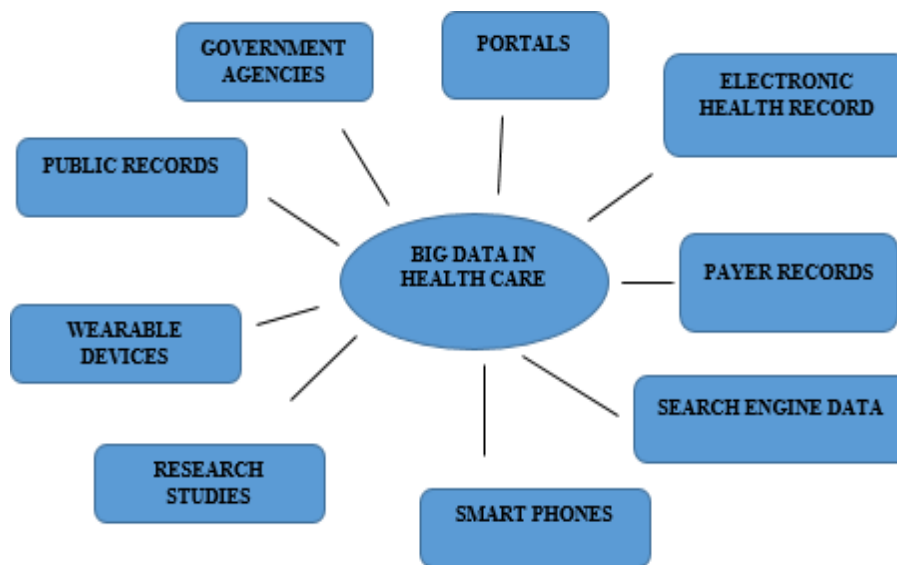


Figure.2 Sources of Big data in Health care

Figure.2 shows the various data sources in the healthcare. Data preparation, Data gathering, cleansing, analysis, graphing, and feature extraction, takes up a large portion of time when utilizing ML in its entirety.

Data collection can be time-consuming due to all of these steps, but has recently become more challenging due to the following reasons [2]:

- **In precise data** - The data collected may not be pertinent to the problem statement.
- **Missing data** - Sub-data might not be present. Some classes of predictions might have empty columns or missing images.
- **Data imbalance** - The number of related samples for some categories or classes with in the data may be extremely high or very low. They run the danger of being underrepresented as a result.
- **Data bias** - Based on the selection of data, subjects, and labels, models may propagate ingrained biases.

Several techniques can be applied to address those problems:

- **Pre-cleaned, freely available datasets** - Use existing, clean, properly formulated datasets if the problem statement (say, image classification, object recognition) aligns with existing, open-source expertise.
- **Web crawling and scraping** - Data can be scraped and crawled by bots, stateless browsers, and automated processes.
- **Private data** - ML engineers can create their own data. The model could be trained with a limited amount of data and the problem statement cannot be generalized to open-source datasets.
- **Custom data** - Data can be created or crowd sourced by agencies for a fee.

#### IV. TYPES OF DATA

The Data type is broadly classified as Quantitative and Qualitative. The figure 3 represents the different types of data.

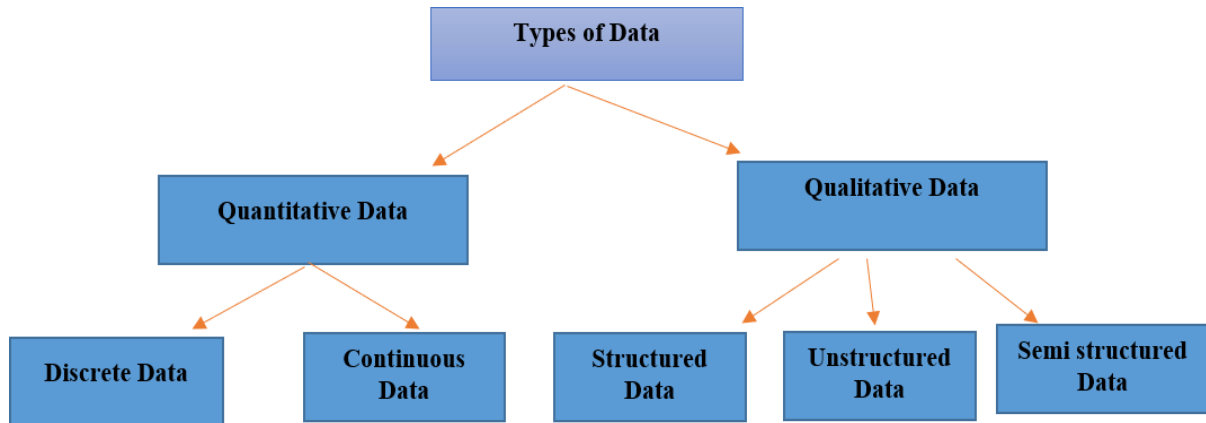


Figure.3 Types of Data

##### 4.1 Quantitative Data Type

Data sets are composed of numerical values associated with a definite numerical value, and their value is measured by numbers or counts [3].

Example: Data projection, Census, Annual income

##### Discrete Data Type:

This category includes numeric data that has discrete values or whole numbers. Explicitly expressing this type of variable value in decimal format would be meaningless. Their Values Can Be Counted.

Examples of discrete data types are shown in figure 4: Number of cars and laptops you have, Number of marbles in containers, Students in A Class, Etc.

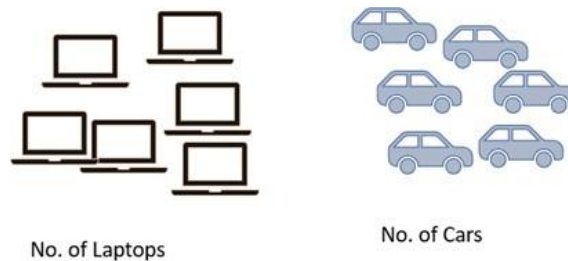


Figure.4. Examples of Discrete type data

##### Continuous Data Type:

The numerical measures which are capable of taking values within a certain range. The value of this type of variable has true meaning when expressed in decimal format. Their Values are measured but cannot be counted. The value is infinite one. Examples of continuous data types are shown in figure 5 : Height difference of people and time.

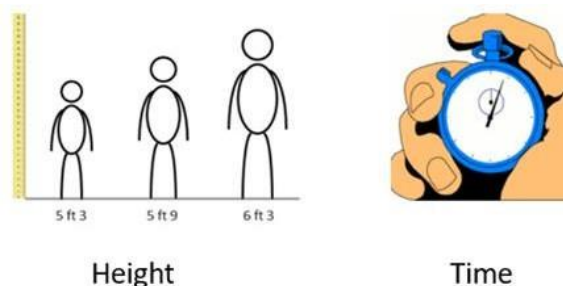


Figure.5. Examples of Continuous type data

##### 4.2 Qualitative Data Type:

Information that could not be expressed, measured and countered well by using numbers in qualitative data. The data is gathered from images, audios and texts and these are shared through visualization tools like timelines, graph databases, word

clouds, concept maps and info graphics.

#### 4.2.1 Structured Data:

This type of data is either number or words. Numeric values can be input into this, but no mathematical operations can be performed on it. Data that is factual, concise, and well-organized is structured data. It is easy to search, analyze and has a predefined format. The structured data shown in figure 6 has been managed by the programming language SQL (Structured Query Language). To handle relational database and warehouse SQL is used and it was developed by IBM in 1970s. [4]



Figure.6. Structured Data

#### Pros:

- Algorithms that use Machine Learning (ML) can easily use it: As structured data is organized and specific, ML data can easily be manipulated and queried.
- Business users will find it easy to use: There is no need to understand how each type of data works or how structured data works in order to understand structured data. Users can easily access and interpret data if they have a basic understanding of the topic.
- Tools that are more accessible: For structured data, there are more tools available for analyzing and using them than for unstructured data because structured data is older.

#### Cons:

- Limited Usage: Its flexibility and usability are limited since it cannot be used for anything other than its intended purpose.
- Options for limited storage: The data storage system with rigid scheme is used to store the structured data. So all structured data must be updated due to these changes in data requirements, which consumes significant time and resources.

#### Structured data tools

- [PostgreSQL](#): supports high-level programming languages, JSON, SQL, and querying (Python, Java, C/C++ etc.).
- [OLAP](#): Centralized data stores, Multidimensional data analysis from a single sources, Performs high speed, Multidimensional data analysis
- [MySQL](#): Embeds data into widely used software, especially in high-volume production systems, mission-critical.
- [SQLite](#): Establishes a self-contained, Cloud - hosted, linear transaction database engine, Zero-configuration.

#### 4.2.2 Unstructured Data:

This type of data does not have the proper format. This comprises textual data, sounds, images, videos, etc. It is common for some organizations to have a lot of data, but they don't know how to extract value from raw data such as text, images, audio, and video files. In addition to requiring a lot of storage space, it is difficult to maintain security. For that reason analyzing, managing or searching for unstructured data is hard. The information is stored in a non-relational database, also known as NoSQL (Not only SQL) [5]. The model for unstructured has shown in figure 7.



Figure.7. Unstructured Data

#### Pros

- Native format: As long as unstructured data remains in its native format; it cannot be defined until it is needed. The database is able to handle a wide range of file formats, so it making easier for data scientists to prepare as well as analyze data according to their needs because of its adaptability.
- Quick accumulation rates: As the data does not need to be defined in advance, it is quick and easy to collect.
- Data lake storage: With massive storage and pay-per-use pricing, costs can be cut and scalability is made easier.

## Cons

- Requires expertise: Preparing and analyzing unstructured data requires data science expertise due to its non-formatted and undefined nature. Unspecialized business users may find it difficult to understand specialized topics or use the data in this scenario. Data analysts benefit from this situation, but unspecialized business users may not.
- Specialized tools: Data managers have limited product options due to the need for specialized tools to manipulate unstructured data.

## Unstructured data tools

- [MongoDB](#): Applications and services are based on flexible documents that can be used to process data across platforms.
- [DynamoDB](#): Ensures single-digit millisecond performance despite scalability thanks to in-memory caching, backup & restoration capabilities and security.
- [Hadoop](#): In order to analyse massive amounts of data distributed, Hadoop could be utilized for simple programming models.
- [Azure](#): Using Microsoft's data centers, the user can create and manage apps using agile cloud computing.

### 4.2.3 Semi-structured Data:

- Apart from structured and unstructured data, next introduce a third category as semi structured data, which is essentially a mix of both. There are some defining characteristics of a semi-structured database, but unlike a relational database, it does not conform to a rigid structure. Due to these factors, some organization features, such as semantic tags and metadata, make organizing data easier, but fluidity still exists [6].
- A few examples are semi-structured emails by Sender, Recipient, Subject, Date, and so on, or emails are automatically categorized into folders based on machine learning, such as Inbox, Spam, Promotions, and so on.

## Pros & Cons

It has the following advantages and drawbacks, such as:

- There is no single architecture for semi-structured data. NoSQL databases can also scale across a variety of data formats to store vast amounts of information. As a result, analyzing data becomes much more challenging [7].
- Semi-structured data can be stored more easily, but its storage costs are higher than structured data, since semi-structured data can be mobilized and stored more easily.
- In addition to being versatile, it allows you to change the schema as you want. It is still essential to know the data you are attempting to retrieve in advance when conducting queries due to the tight relationships between schema and data.

## V. SICKNESS PREDICTION USING ML

Sickness Prediction by ML is the framework that is utilized to anticipate the illnesses from the side effects which are given by the patients or any client. The framework processes the side effects gave by the client as info and provides the result as the likelihood of the illness. Naive Bayes classifier is utilized in the expectation of diagnosis, which is a managed AI calculation. The likelihood of sickness is determined by the Naive Bayes calculation. With an expansion in biomedical and medical care information, detailed examination of clinical information helps early illness location and patient consideration.

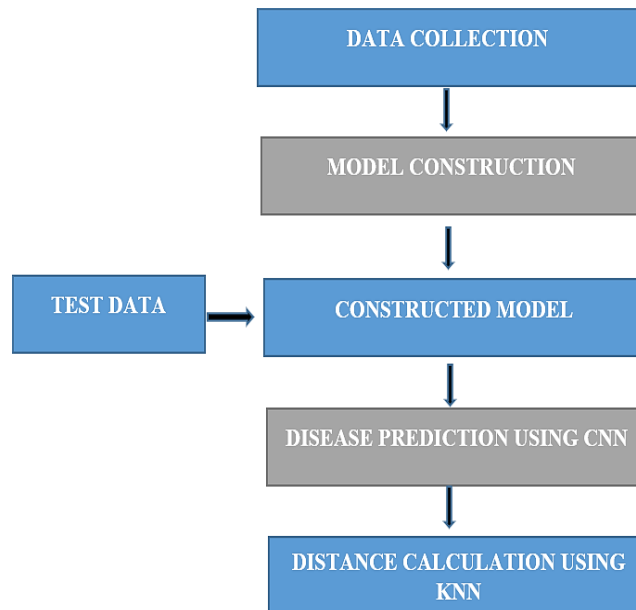
By utilizing the direct relapse and choice tree we are anticipating infections like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

To implement a strong machine learning technique that may expeditiously predict the illness of a person's, based on the symptoms that the person has exhibits. The ML calculation has two stages: 1) Training and 2) Testing. To foresee the sickness from a patient's side effects and from the historical backdrop of the patient, AI innovation has been fighting for many years. Medical care issues can be addressed more effectively by utilizing Machine Learning Technology. For S type information, the framework is utilizing Machine Learning methods such as K-Nearest Neighbors, Decision Tree, Naive Bayesian [9]. The figure 8 illustrates the process involved in prediction of sickness using ML.

### 5.1 Identification of Chronic diseases

Chronic illnesses are a common issue in the medical services space. As per the clinical assertion, because of ongoing illnesses, the death rate of people increases. Over 70% of a patient's pay is spent on medicine for this infection. As indicated by US National Center for Health Statistics, constant illnesses are infections that persist for a significant stretch of time, or at least, over 90 days. These infections are neither treated by drugs nor prevented by inoculations. The most important reason for constant sicknesses is the consumption of tobacco, eating undesirable food, and lack of actual work. Likewise, this illness can arise from maturing.





**Figure.8.** Steps in illness Prediction using ML

Ongoing illnesses incorporate cardiovascular sickness, malignant growth, joint pain, diabetes, stoutness, epilepsy and seizures, and issues in oral well-being. Thus, it is imperative to limit the patient's risk factor. The development of clinical investigation simplifies the gathering of health-related information. The medical services information includes the social economics, clinical examination reports, and the historical background of illness of the patient. The illnesses caused could vary according to the locales and the living environments around there. Therefore, the patient's environmental conditions and living environment should also be kept in the data set along with his or her illness information [10].

**Information Collection**

The actual information that includes organized information, for example, lab test results, living environment, patient essential data including social economics, and the unstructured information, for example, the side effects of the illness looked by the patient and their counsel with the specialist. To ensure the privacy of patients, the informational index rejects their identifying details such as their name, ID, and area.

**Preprocessing**

A large number of the gathered data are preprocessed to ensure accessibility of missing characteristics. Therefore, it is necessary to complete the missing information or eliminate or change them in order to improve the nature of the informational index. Additionally, commas, accents, and void spaces are removed during preprocessing. As soon as the data has been preprocessed, it is then exposed to highlights extraction, followed by an expectation of illness.

**Model Description**

As discussed previously, the informational index comprises of both organized and unstructured information. The organized information contains patient socioeconomic and the information connected with the reason for the illness like age, orientation level, weight, etc. It also contains patient's living environment, lab test results, and the sickness that illness they are experiencing in plain configuration. Data from the cross-examination with specialists in text design are included in the unstructured information about the patients' illness and side effects. As a result of the expectation undertaking, additional precise results can be obtained with the unstructured information. The informational index is spitted into 80% for preparation and 20% for testing.

**Infection Prediction Using Convolutional Neural Network (CNN)**

CNN calculation under constant illness is included in the proposed framework. Right away, the informational collection is restructured into a vector structure, using word installing to embrace no qualities for filling the information.

It is then given to the convolution layer. This layer takes contributions from the convolution layer and follows maximum pooling activity. Max pooling results are given to the completely associated layer, and the result layer provides output arrangements. Figure 9 shows the Convolutional brain organization flow chart.

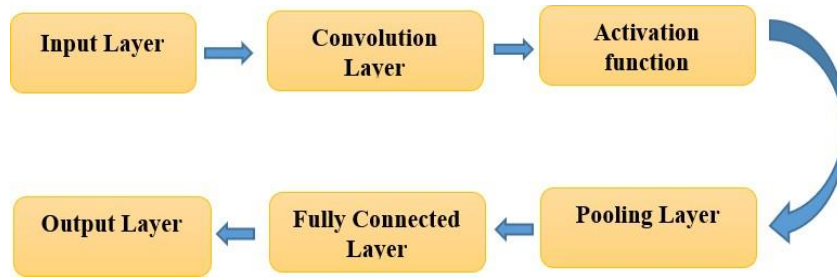


Figure.9. Process in illness prediction using CNN

### K-Nearest Neighbour (KNN) used for Distance Calculation

The most effective data mining strategy for classification issues is K-Nearest-Neighbor (KNN). Also because training data must be kept in memory during execution, the technique is known as memory-based classification. When working with continuous characteristics, the Euclidean distance is used to calculate the difference between the attributes. The value of K is known in K-Nearest Neighbor (KNN), and the closest neighbour are those whose highlights are similar to K's value. The closest distance between them is calculated by selecting the neighbor with the greatest K worth as its closest neighbor. In the specific match, the component with the lowest distance esteem is the most recent illness forecast yield. Although it is capable of handling discrete elements, KNN typically deals with continuous data. When working with discrete attributes, the difference between two instances of samples, A and B, is equal to one if their attribute values differ. Alternatively, it is equal to zero.

### 5.2 ML in cancer prediction and prognosis based on data analyst

In the past 20 years, disease detection and analysis have been supported by data science and Artificial Neural Networks (ANNs). Today ML techniques are being used for a wide array of applications, from finding a cancer through X-ray and CT scans to creating classifications for cancer cells based on proteomic and genomic analysis (micro array tests) [13]. More than 1500 studies have been published on the subject of ML and cancer, according to the most recent PubMed statistics. Nevertheless, the bulk of research projects focus on applying ML techniques to identify, classify, or recognize cancers and other malignancies. The use of machine learning has been used to guide disease analysis and discovery. Disease experts have only recently attempted to apply ML to predict and estimate the likelihood of malignant development. As a result, there isn't much literature on the topic of AI and predicting the proliferation of cancer cells. The main goals of predicting and predicting cancer progression can be distinguished from those of locating and discovering the disease. Three predictive foci are causes for concern in disease forecasting or guesswork:

1. The forecast of disease spread.
2. The expectation of disease growth and survivability.
3. The expectation of cancer disease growth and helplessness (for example risk appraisal)

In some circumstances, determining the likelihood of promoting a particular type of malignant growth before the illness manifests itself is a goal. In the second case, one aims to forecast the likelihood of recurrent malignant growth following the disease's distinct objectives. In the third scenario, after the diagnosis of the illness, one is trying to predict a result (future, survivorship, mobility, growing drug consciousness). The prognostic forecast's evolution in the last two cases is evidently somewhat influenced by the choice's result or nature. In any case, a prognostic expectation should take into account more than just a fundamental conclusion since a disease prognosis can follow a clinical discovery.

Unquestionably, a disease guess typically entails a variety of physicians from various fortes using a variety of subsets of biomarkers and various clinical variables include the cancer's grade and size, the patient's age and general health, the location and type of malignant development, and more. Commonly, the attending physician needs meticulously coordinate segment (population based), clinical (patient-based), and histological (cell-based) data to think of a reasonable forecast. In any case, doing this isn't difficult for the most skilled doctor. Addressing the matter of sickness anticipation and cancer cell growth weakening anticipation, there are also comparable difficulties for both doctors and patients. Age, food, weight (heaviness), high-risk behaviour (smoking, heavy drinking), and exposure to naturally occurring carcinogens (UV radiation, radon, asbestos) all have a role in predicting a person's chance of contracting disease. Tragically, these routine "macro scale" medical, environmental, and social constraints rarely provide enough information to create trustworthy predictions or anticipations. In an ideal situation, what is needed are a few definite subatomic clues about the tumor's development or perhaps the patient's own genetic make-up. Looking at the different forecasts or visualizations created, we can find that the most majority (86%) have to do with predicting disease mortality (44%) and cancer cell growth repetition (42%). Despite this, an increasing number of researchers are increasingly concentrating on anticipating the development of disease or the risk factors connected with causing malignant growth. When in doubt, AI techniques appear to work on the precision of expectations by and normal of 15–25% above alternative options or conventional methodology, regardless of the AI technique used, the type of forecast being made, or the type of malignant growth being assessed [11].

### 5.3 Prediction of liver diseases using data mining approach

Among the organs inside the human body, the liver occupies the second largest place. As well as metabolizing cholesterol, glucose, and iron, it produces protein and clots blood in the body. Liver is important to maintain survival because it also functions to remove poisons from the body. Many bodily activities cannot be carried out effectively when the liver is not functioning, which results in serious harm to the body. If the liver becomes infected with a virus, is targeted by its own immune system, or is exposed to chemicals, it may suffer damage. Hepatotropic viruses such the hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus can lead to chronic liver damage, which can be fatal. In order to contain the illness, a more thorough physical examination is needed. This can be done by automatically determining patient data records that are kept in hospital data systems or health care organizations. The data mining methods could be utilized to group the liver illness into intense or constant in light of the patients' side effects.

This permits the specialists or clinical suppliers to separate the right data to propose for viable clinical help. Analyses based on association rules, classification, or clustering can be categorized according to the type of data mining technique used. In association rule mining, symptoms are compared to the degree of co-occurrence in health records so that a better diagnosis can be made. By classifying the items, the target class is assigned, and the actual class is determined. The classification model is commonly used to classify data and to predict its behavior in the future. C4.5 Decision Trees, Support Vector Machines, and Nave Bayes are a few of the data mining algorithms used. Based on comparing the accuracy of each algorithm, the algorithms are used to predict liver diseases. A final method of clustering is to divide or cluster the symptoms of each patient according to how similar they are. Clustering analysis will influence the clustering outcomes directly as it groups comparable records into clusters. Data regression techniques as poisson regression, linear regression methods have been used to predict the liver disease. Finally clustered data records are explored by different classification algorithms as Naive Bayes, Neural Network, Kstar etc. to classify the liver diseases [12]. Figure 10 flow out the steps involved in prediction of liver disease using data mining.

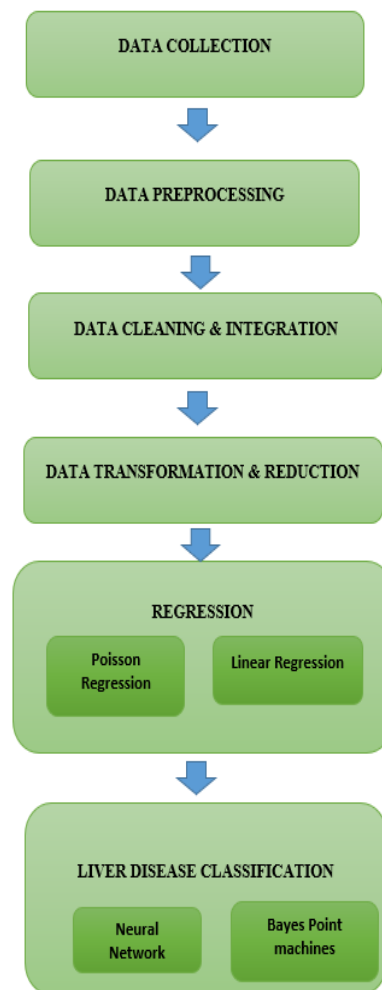
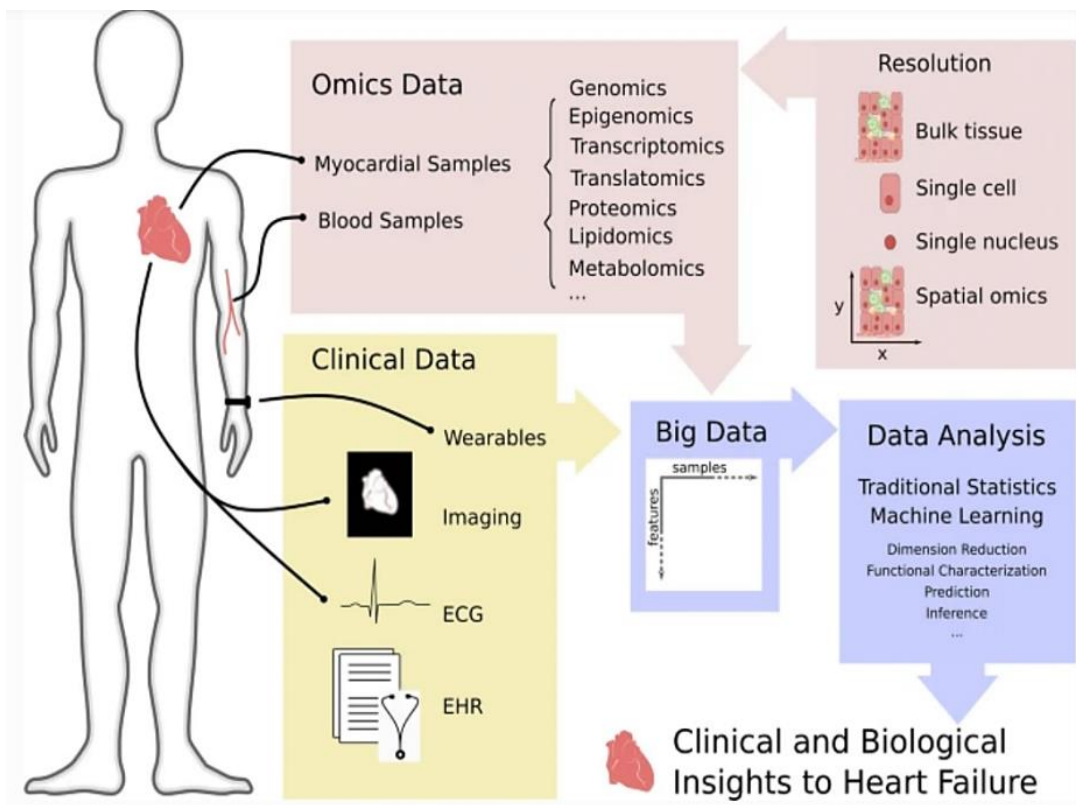


Figure.10. Steps in prediction of Liver diseases based on data mining approach

#### 5.4 Heart diseases analysis using data mining:

The heart is the second important organ after the brain, which is more significant in the human body. In the medical field the forecast of disease in the heart is one of the toughest problems. Heart disease is the extensive disease which leads to reduce the lifetime among the humans in these recent days. Due to that the medical centers, data analytics can help predict multiple diseases by evaluating more information. Also which includes Challenging aspects and more. For the information which has been mentioned in the data base can be utilize for prediction of disease in the future.

These include Naive Bayes, Support Vector Machine, K-Nearest Neighbors (KNN), Fuzzy Logic, Decision Tree, and Artificial Neural Network (ANN) (SVM). This report looks of a different algorithms and it presents an overall summary of the previous work [14]. The figure 11[16] shows the clear steps involved in heart failure analysis using big data.



**Figure.11.** Heart Diseases analysis using data mining

Due to the technologies in the digital side is fleetly growing, Healthcare facilities that store significant amounts of data in databases that are incredibly complicated and difficult to evaluate. In the medical industry, machine learning methodologies and data mining techniques are extremely important. The algorithms and the techniques used in these techniques can be used for managing this kind of challenges. So, data mining and the algorithms using machine learning is the ultra-major technique in the digital technology. The data mining techniques and the machine learning have a major role in reducing of the burdens among the dataset. The purpose of the data mining technique is to provide the classification accuracy of the heart disease and which avoids the upcoming behavior of the disease. In addition to increasing the size of data, it also expands its dimensions by digitalized system in the medical field.

The disease co-occurrence is diagnosed by patient discharge description details. For reducing the expenditure of the health care system care monitoring of the PoA (Patient Data at the time of Admission) can be analyzed. By following this strategy the system of health care treatment could improvised. In addition to chest discomfort, women are more likely to suffer from nausea, extreme fatigue and shortness of breath. So it is the most severe case in the heart case, so by using this machine learning technology and data mining technology it is more likely preferable.

#### 5.5 Lung disease in data mining technique

An inexpensive, yet highly effective medical imaging procedure, thoracic radiography (chest X-ray) is widely used. Due to the lack of radiology technique the applicability for this disease is reduced. The figure 11 illustrate the data mining techniques in X-ray images [15].

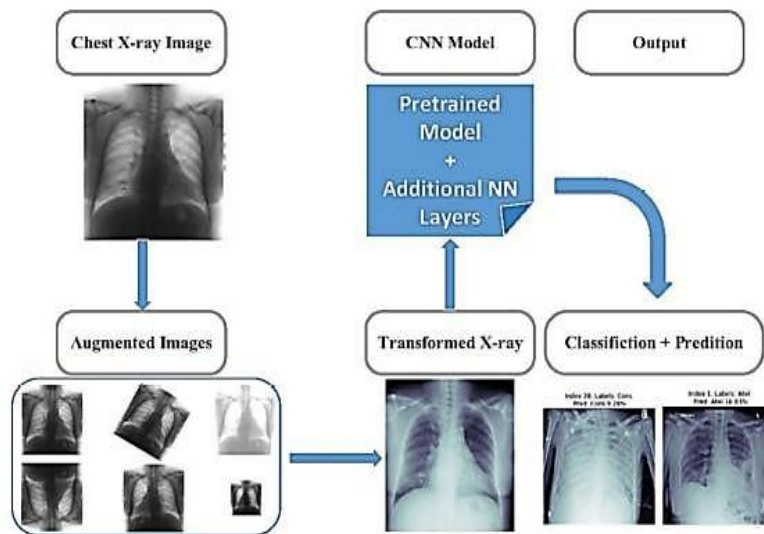


Figure.11. MobileNet V2 Architecture

For training systems other than those that can be harvested in a large scale, annotations are used to label bounding boxes. A modified model of Mobile Net V2 is used in this study to predict lung pathology in X-rays of the frontal thoracic region of the lung. Based on their area under the Receiver Operating Characteristic Curve (ROCC) statistics, classifiers were compared. Once the dataset is resampled, the model's performance considerably increases. Since they can be integrated by using smaller IoT devices, we aimed to develop a model in this study that can be taught and changed on devices with minimal processing power [15]. Various lung diseases have an influence on people around the world, which increase the lungs' susceptibility to various physical issues and air pollution. Lung function is hampered as a result. Some lung conditions, including emphysema, asthma, pleural effusion, tuberculosis, aspiration fibrosis, pneumonia, and lung cancers, cause the lungs to lose their adaptability, which results in a variety of symptoms. [15].

The MobileNet V2 is the design used to associate the mobile and on-board applications. In our recent days deep learning is not only used in the computer techniques it also used in the electronics and data analytical techniques. Also, it is used in the IoTs (Internet of Things), NLP (Natural Language Processing) and medical image processing application kind of fields. The figure 3 shown above is the architecture of the MobileNet V2 and as some hidden layers based on the blocks in residual and also the depth wise separation of convolutions and which could minimize the total various parameters and which leads as the lightweight of the neural networks. Those convolutions networks are different from the normal convolution. A popular deep learning architecture for mobile devices is MobileNet, which is a mobile-centric network for doing computationally and efficiently high performance.

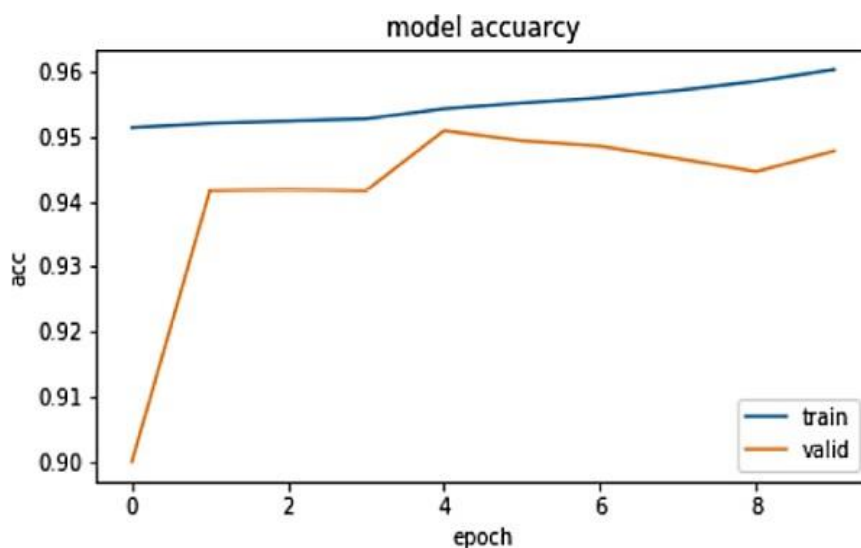


Figure.12. Model accuracy chart between the validation and over a 10 epoch



The experiment was severely constrained by a lack of computer capacity, thus in this technique we edited and experimented with the model we mentioned using the "Kaggle" open free cloud [15]. Although this could offer a free GPU (Graphics Processing Unit) and the RAM (Random Access Memory) is of 15GB, it will be not suitable for the techniques over ten epochs. For getting the good results it is managed by validating the different epochs that has been shown in the figure 12.

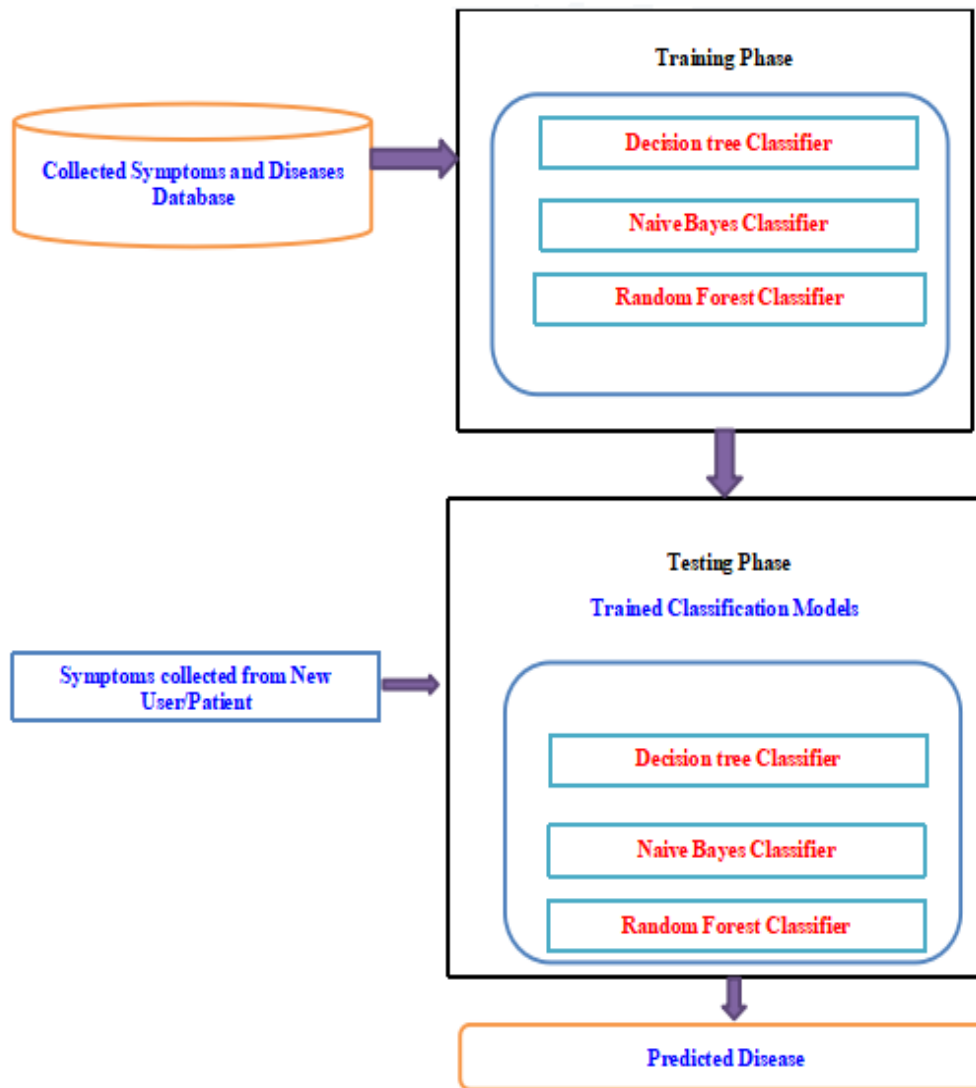
### 5.6 Medical image analysis from detection to diagnosis based on ML

Clinical imaging is the use of imaging modalities and cycles to take pictures of the human body that can aid in patient identification and therapy. It can also be used to monitor any ongoing problems, which will help with treatment approaches. There are quite a large number various sorts of clinical imaging methods, which utilize various innovations to create pictures for various purposes. Here the most widely recognized imaging procedures involves AI in radiology shows how these methods, blended in with AI, will coordinate the way for more precise imaging. Different clinical imaging modalities and sophisticated clinical images, such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), single photon emission computed tomography (SPECT), and others, can provide the patient being imaged with precise information. Research in clinical picture handling predominantly focuses to remove significant elements that may be challenging to survey with the unaided eye. A histology slide is a picture record of a couple of megabytes while a solitary MRI might be two or three hundred megabytes. This has specialized ramifications on how the information is pre-handled, and on the plan of a calculation's design, with regards to processor and memory constraints. ML approaches are progressively fruitful in picture based analysis, illness visualization, and risk identification. Supervised and unsupervised learning are the two basic types of machine learning. Supervised learning uses training examples comprised of inputs and outputs, with each example having its own input and output value. Machine learning algorithms are trained with supervised machine learning systems to support their judgment in the future. Training models is done with input data and matched labels. A mathematical model links input data to matched labels, and a predictive model is verified with unobserved data. A type of machine learning techniques called unsupervised learning is used to identify patterns in data. A non-supervised algorithm is given unlabeled data, meaning that only inputs (X) are provided, and no outputs (Y) are provided. Supervised learning which gains a planning from input data set to yield (names) from a bunch of preparing models, have shown extraordinary commitment in clinical image examination. Classification of pattern has previously been utilized for quite a long time to identify, and later portray, irregularities, for example, masses in mammograms and knobs in chest radiographs in light of elements depicting neighborhood picture appearance. With upgrades in PC equipment, it has become possible to prepare an ever increasing number of mind boggling models on additional information, and over the most recent couple of years, the utilization of supervised learning in image division, recognition, and enlistment has sped up. The use of trained appearance models in segmentation systems has replaced simple intensity and gradient models, and the application of statistical models to characterize the typical shape and variation has replaced free form deformable models. By using multivariate classification or regression on imaging data, several unique techniques are created to learn to recognise disease in a data-driven manner. Unlike conventional quantitative analysis based on straightforward volume or density measurements, these methods are not constrained by familiarity with radiological patterns associated with disease [13].

### 5.7 Big Data Analytics for Disease Prediction Based on Symptoms

Around the globe, there are numerous techniques to treat different illnesses. One method for predicting and diagnosing diseases is machine learning. By using the patients symptoms as inputs for machine learning to forecast diseases. Additionally, it accurately forecasts the user's or the patient's disease based on the data or symptoms entered into the system and returns findings accordingly. If the customer simply wants to know the type of ailment the patient has experienced and the condition is not particularly significant. If the patient basically wants to recognize the type of sickness the user has experienced and the condition is not particularly significant. In predicting the disease based on the user symptoms, the first basic process is the collection of data. After gathering the data, which is in default as raw data that could be used to train the ML model. Then prepared that data according for machine learning models by using certain tools like Python tools, NumPy and pandas. The data have been trained by choosing the various unsupervised machine learning technique including Decision Tree Classifier, Random Forest Classifier, and Naive Bayes algorithms. After implementing different techniques, then must decide the one which best fits for given dataset and the system would provides with greater accuracy. So, to map out the correctness of each model, researchers employed a confusion matrix. The classifiers is applied to perform both Classification and regression. A classifier technique uses many tree structure on different sets of the supplied dataset and uses the typical to increase the dataset's predicted accuracy. The confusion matrix would help to differentiate the actual and predicted values to show the accuracy of disease prediction based on the symptoms data. [17-19].





**Figure.13.** Disease prediction based on User Symptoms

**Table-1.** Symptoms collected from patients

Patient No	SYMPTOMS			
	S1	S2	S3	S4
Patient 1	Orthopnea	Fatigue	Dyspnea	Shortness of Breathe
Patient 2	Drowsiness	Chest Pain	Pressure chest	Angina pectoris
Patient 3	Hematuria	Tumor cell invasion	Anosmia	Pain
Patient 4	Wheezing	Cough	Chest Congestion	Distress respiratory

**Table-2.** Disease forecasting based on Symptoms

Patient No	PREDICTION BASED ON SYMPTOMS		
	Decision Tree	Naive Bayes	Random Forest
Patient 1	Carcinoma of lung	Exanthema	Adenocarcin oma
Patient 2	Encephalopath y	Encephalopath y	Encephalopath y
Patient 3	Malignant tumor of colon	Encephalopath y	Pancreatitis
Patient 4	Exanthema	Sickle cell anemia	Asthma

The table.1 Shows the example of collected patients symptoms. The big data analytics helps us to handle the data content. It then applied as input for training and test data which could be given to the machine learning approaches like Decision-Tree, Random-Forest, and Naive-Bayes classifiers for forecasting the victim disease. The figure 13 illustrates the steps in forecasting of disease based on user symptoms. In addition to using training and test data, the classifier method provides three disease forecasts based on the supplied symptoms, as demonstrated in the accompanying table.2. The anticipated disease and the symptoms are record in the database, that would given suggestion to the victim for healing of diseases.It minimizes the need for victims to repeated visit to the hospitals.

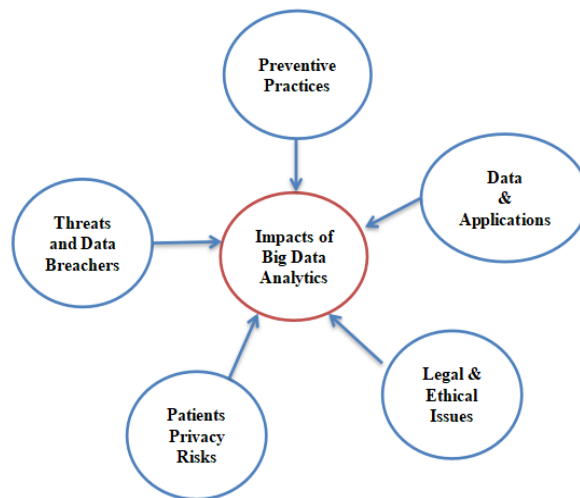
## VI. DIAGNOSIS VARIOUS DISEASES USING ML AND DEEP LEARNING BASED ON BIG DATA

Machine learning and deep learning-based techniques can be used to diagnose a variety of different types of diseases

- Cancer prognosis and detection: Advanced signal and image processing, classification techniques, and deep learning-based approaches can be applied to predict the existence of cancerous tissue early on and the prognosis of the disease. Deep neural networks which are trained with x-ray and mammography pictures are another example.
- Cardiovascular diseases: Data scientists can apply computer vision algorithms to identify heart illness. For instance, Coronary Artery Disease (CAD) develops when plaque gathers in the blood channels feeding the heart and brain with oxygen-rich blood. In order to display CAD, researchers can train a convolutional neural network to recognize illnesses using pictures from CT (computed tomography) or MRI (magnetic resonance imaging) scans.
- Liver illnesses: Using sickness characteristics like liver texture or echogenicity, data scientists have trained deep cognitive networks to evaluate ultrasound pictures in order to detect fatty liver disease or hepatic steatosis.
- Lung diseases: Infection detection methods may be applied to scans of the lungs to identify conditions like tuberculosis, pulmonary nodules, and lung masses.
- Autistic disorder: Diagnostic techniques such as neuroimaging and machine learning can be used to diagnose autism spectrum disorder. By the analysis of the eye movements of children and applying computer vision techniques like face recognition, face image assessment, or eye tracking, it is possible to identify autism condition.
- Diabetic retinopathy disease diagnosis: It is brought on by damage to the retina's tiny blood vessels pushed on by high blood sugar levels. In order to diagnose and prognosis this disease, AI can be used deep learning-based machine learning algorithms to segment images, classify diseases, and use deep learning-based machine learning algorithms (e.g.,CNN).
- Psoriasis disease diagnosis: Image analysis can be utilized to create disease diagnostic algorithms for the identification of psoriasis illness.
- Alzheimer’s disease: By extracting information from genuine speech, machine learning methods like deep neural networks or Support Vector Machines (SVMs) can be used to diagnose Alzheimer's disease.
- Parkinson’s disease: Movement-related features of the human body have been taken into account when using deep learning algorithms to diagnose Parkinson's disease. Convolutional Neural Networks can monitor stress and depression, which are typical signs of Parkinson's disease, and predict the course of the disease using video data.

## VII. BIG DATA ANALYTICS ASSOCIATED ETHICAL IMPACTS

Both positive and negative implications of big data could have a big impact on quality of life.The three most ethical principles are respecting individuals autonomy, ensuring equity, and preserving privacy that may be most frequently threatened by utilization of big data



**Figure.14.** Big Data in Portfolio Management- Ethical Impacts

.From the time big data is extracted until it is put to use, ethical decisions and legal implications must be taken into account. The figure.14 shows the five factors can influence big data analytics to produce significant positive effects for society[20]. The web, cloud services, and data pooling in big data increases the prospect of unwanted access to the data by third parties. Medical firms' assessments of security issues are mostly focused with malware attacks, lost or stolen devices, and hacking. Privacy of the patients is significantly impacted by security breaches since they expose personal details, violate the law. Illegal attacks, lost or stolen devices, negligent employee behaviour, blunders by third parties, system design flaws, malicious cyber activity, and intentional non-malicious employee behaviour are the main sources of data breaches in medical businesses. For big data analysis to study healthcare quality for disease prevention and predictive modeling, the inclusion of significant amounts of essential clinical, budgetary, genetic, social, and atmospheric data is essential. Big data ethical considerations have included potential for misrepresenting the quality of data or its constraints, as well as the use of the data for damaging or unneeded objectives. Big data might potentially forecast useless information, for instance, and service providers might utilize that information to offer pointless services. Organizations must devote enough tools to prevent or find unlawful access to patient information.

### VIII . BIG DATA ANALYTICS FOR PREVENTIVE MEDICINE

Data from medical records is one of the most rewarding and yet most complex sources of information. Data analytics promises to find useful trends in the healthcare industry by examining heterogeneous , unstructured, non-standardized and incomplete data. This tool helps with decision-making as well as forecasting, and it has emerged as a key component of ongoing developments aiming to lower healthcare costs and boost patient care quality. Diseases can be prevented more sensitively, effectively and economically with the growth of healthcare data. Due to its primary focus on generating healthcare benefits, a standard preventative measure generates enormous amounts of data that are challenging to analyze. Researchers in data analytics are set to make significant improvements in medical care. Applications of data analytics in the healthcare field have enormous promise. These days, early disease detection and treatment are made possible by data mining, data analytics and machine learning. In a number of nations, such as the USA's Bio Sense, Canada's CDPAC, Australia's AIHW, France's Senti Web, and others, diseases are monitored and early detection is achieved [21].

#### 8.1. Different types of Datasets:

**Cardiovascular dataset :** It is an analysis of risk factors associated with cardiovascular diseases and stroke in individuals over 65 years of age is known as the Cardiovascular Heart Study (CHS). In addition to cardiovascular disease risk factors, it includes stroke risk factors.[24]. A significant fraction of missing values is found in 5201 instances, and there are a lot of

**Physio Bank database:** A sizable and expanding repository of physiological data is the Physio Bank database [22]. In more than 80 databases, there are more than 90,000 recordings, or more than 4 gigabytes, of digitalized physiologic signals. Waveform archives, Clinical records, ECG archives, multiparameter archives, interbeat interval archives, other cardiac databases, datasets from computation-based cardiology challenges, synthetic data, gait and balance databases, and neuroelectric and myoelectric databases are all included in a Physio Bank archive.

**Thyroid illness dataset :** It is an additional dataset from UCI's machine learning library in the medical field. Because it has three classes (subnormal functioning , hyper functioning and normal), 3584 training cases, 3328 testing instances, and 14 category and 7 real attributes, this version of the model is appropriate for ANN training. Raw patient measurements include

missing values as well as categorical features.[25-26]

**Diabetes dataset :** In order to examine factors associated with readmission as well as other outcomes pertaining to patients with diabetes, a diabetes dataset was taken from the Health Facts national database of 128 US hospitals [27]. There are 75 (integer) attributes from 110K instances that are missing their values.

**Breast cancer wisconsin diagnostic (WDBC):** The breast cancer dataset includes features that were calculated from digital breast mass pictures obtained through Fine Needle Aspiration (FNA) and that define the properties of the cell nuclei [23]. The dataset consists of 548 instances (348 benign and 200 malignant) with 29 attributes and 31 real-valued input features). Patients with cancer and healthy patients are divided in the UCI dataset. The Wisconsin Prognosis Breast Cancer (WPBC) dataset has 187 cases and 33 attributes (29 real-valued input features).

**FertilityDataset:** There are 102 instances of genuine values from 19 to 35 years old were gathered to aid with classification and regression [28]. The 165 samples in the hepatitis UCI dataset each have 20 characteristics (14 binary and 6 are discrete) [29]. The goal of the dataset is to forecast the existence of the hepatitis virus based on the findings of several patient-submitted medical tests.

**Heparin induced thrombocytopenia (HIT) dataset** collected from 4273 records of post-surgical cardiac patients treated [31].

**Liver ILPD (Indian Liver Patient Dataset) :** It consists of 12 (integer and real) features of 573 (425 liver patients and 148 non-liver patients) for classification purposes. Examples in [30].

## IX. ANALYSIS OF DATA MINING TOOLS FOR DISEASE PREDICTION

### 9.1.Fundamentals of Data Mining

Data mining is the process of extracting unknown knowledge using computation from large datasets. To detect and treat diseases at the earliest possible stage, it is imperative to be able to extract usable knowledge from the vast data sets and deliver decision-making outcomes. Various diseases can be predicted and analyzed through data mining. The medical domain contains a wealth of data sets that can be mined for patterns that can be discovered. Health care data can be mined using various data mining techniques depending on their suitability. There is great potential and effectiveness for data mining applications in health care. It automatically locates predictive information using massive datasets. To identify a disease, a patient must undergo a number of tests. Data mining techniques, however, can be used to decrease the number of tests. Doctors can identify which characteristics, such as age, weight, and symptoms, are more essential to diagnosis. Doctors will be able to identify the illness more accurately in this manner [ 32]. Figure 15 highlights the many steps of the knowledge discovery process in databases.

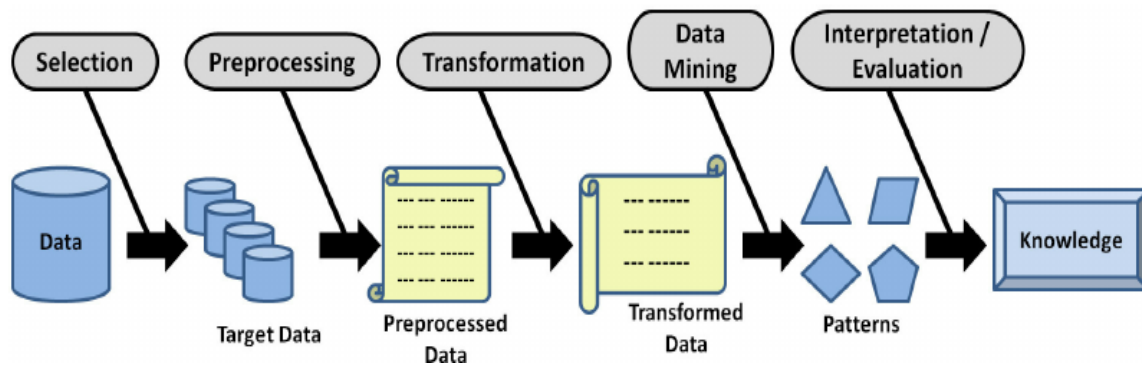


Figure 15. Healthcare Knowledge Discovery process

### 9.2. Data Mining Techniques

Data mining techniques including clustering , classification and association algorithms are frequently used by illness data analysts to examine their data.

**Classification:** In data mining, machine learning is used to categorise data. In order to classify a set of data, information must be divided into predetermined categories or groups. Utilizing methods like decision trees, linear programming, neural networks, and statistics, data is divided into many groups. With the use of contemporary classification systems, diseases can now be predicted more accurately [17]. . Discriminates ,naive methods decision trees , support vector machines, linear regressions, and non linear regressions can all be used to categorise data.

**Clustering:** In clustering, classes are established, and items are inserted into them when no preset class exists. K-means, Gaussian mixture , hierarchical , fuzzy C-means (FCM), , Rough-Fuzzy C-means (RFCM), Rough C-means (RCM)and Robust

RFCM (RRFCM) are a few examples of different clustering approaches.

**Association Rule Mining:** Association rule learning is a powerful and well-researched method for identifying intriguing links between various pieces of data in massive databases. The input data set is used to identify important procedures that are used to find well-built rules in databases. In association rule mining, rules are discovered by spotting common correlations, relationships, patterns, or causal structures among groups of elements that are linked to associations and causal objects. Determine the products that are related to and correlate with the customers' "shopping baskets". Three of association rule mining's key uses are the cross-marketing, analysis of basket data and catalogue design. The aforementioned data mining [30].

### 9.3. Data Mining Tools

Various data mining approaches are carried out using data mining tools like Rapidminer, Orange, and Knime, Weka

**RAPIDMINER: (RM)** [33] The programme is open source and offers a favourable environment for data mining procedures. Here the drag and drop operations are used to build the data flow. It can open a variety of file formats such as classification regression and clustering tasks can be performed using a variety of learning technique. Rapid Miner provides a wide range of tools, including classification and regression methods, decision trees, association rules, and clustering algorithms, for preprocessing, normalising, filtering, and analysing data. It may import information from various conventional and standardized databases

**ORANGE:** It was created by the Bioinformatics Laboratory at the University of Ljubljana as an open source data mining tool [36]. Applications can be created using visual programming and scripting. . The Python library allows for the change of widgets and data. Programming is done by connecting the inputs and outputs of widgets to the canvas. This tool can be used to create algorithms for data mining and machine learning. Both professional data mining researchers and novice users interested in creating and testing their own algorithms can use it.

**KNIME (Konstanz Information Miner):** It's a multipurpose open source data mining tool that was created and is currently maintained by a Swiss firm. This platform can integrate, process, examine, and analyse data because it is based on the Eclipse platform. KNIME is compatible with R and WEKA, two other data mining programmes [35].

**WEKA:** The Waikato Environment for Knowledge Analysis (WEKA) is open source software and a machine learning toolkit created by the Waikato University in New Zealand [29]. WEKA supports a number of common data mining operations, including feature selection, data preprocessing, clustering, classification, and regression. The New methods can also be built utilising WEKA and currently used data mining and machine learning approaches. Data can be loaded from a variety of sources with WEKA, including files, databases, and URLs. It supports file formats as CSV, Lib SVMs, WEKA's proprietary ARFF format, and C4.5. Additional evaluation criteria offered by WEKA include confusion matrix, precision, recall, true positive, and false negative. This tool has many benefits, such as being open source, portable, and platform-independent, having a graphical user interface, and providing a huge variety of different data mining techniques

**RAPIDMINER: (RM)** [33] is open source software which provides a good environment for data mining processes. The dataflow is constructed by dragging-and-dropping. It support different file formats. A variety of learning algorithms can be used for regression, classification, and clustering tasks. For preprocessing, normalizing, filtering, and analyzing data, Rapid Miner offers many tools, such as classification and regression algorithms, decision trees, association rules, and clustering algorithms. It can import data from different traditional and standard databases.

**ORANGE** [34] is an open source data mining tool developed at the Bioinformatics Laboratory at the University of Ljubljana.

Scripting and visual programming can be used to implement applications. . Data manipulation and widget alteration are possible with the Python library. Connecting widgets' inputs and outputs to the canvas is how programming is done. Data mining and machine learning algorithms can be implemented using this tool. It can be used by both advanced data mining researchers and inexperienced users who are interested in developing and testing algorithms of their own.

**KNIME (Konstanz Information Miner)** [35] is a general purpose open source data mining tool developed and maintained by the Swiss company. This platform is built on the Eclipse platform and has the ability to integrate, process, explore, and analyze data. KNIME can be integrated with other data mining tools such as R and WEKA.

## VIII. CONCLUSION

Taking everything into account, as distinguished through the survey, accept just a modest success is achieved in the creation of predictive models for illness patients and consequently the accuracy of predicting the early stages of sickness, combinational and more complex models are therefore required. The clinical information is organized and controlled so that healthcare practitioners can use it to enhance outcomes, lower costs, and make strategic business decisions. Due to the cross-industry infiltration of big data, many sectors are now required to acquire technology that supports more advanced analytics. This pattern is not unusual in the healthcare sector. According to marketing strategy, the global big data in healthcare infrastructure is anticipated to grow at a CAGR of 22.07% from 2017 to 2022, reaching \$34.27 billion. Big data enables healthcare firms to develop full, tailored, all-encompassing informing future and identify diseases at far earlier stages for more effective treatment. Big data examination in medical care is developing into a promising field for giving understanding from extremely enormous



informational collections and further developing results while lessening costs. Its true capacity is perfect; but there remain difficulties to survive.

## REFERENCES

- [1] Saranya, P. and Asha, P., 2019, November. Survey on Big Data Analytics in health care. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 46-51). IEEE.
- [2] Yuji Roh, Geon Heo, Steven Euijong Whang, Senior Member, IEEE, A Survey on Data Collection for Machine Learning and Big Data - AI Integration Perspective, <https://arxiv.org/pdf/1811.03402,2019>
- [3] Yuji Roa, Geon Heo, A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective, IEEE Transactions on Knowledge and Data Engineering PP(99):1-1, October 2021, DOI:10.1109/TKDE.2019.2946162
- [4] <https://mldoodles.com/statistical-data-types-used-in-machine-learning>
- [5] Thashmee Karunaratne; Henrik Bostrom; Ulf Norinder, Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization - A Case Study with Medicinal Chemistry Datasets, IEEE Explore, Dec 2010, DOI: 10.1109/ICMLA.2010.128.
- [6] <https://www.ibm.com/cloud/blog/structured-vs-unstructured-data>
- [7] <https://www.geeksforgeeks.org/what-is-semi-structured-data>
- [8] Houda Ahmad, Shokoh Kermanshahani, Ana Simonet & Michel Simonet, Data Warehouse Based Approach to the Integration of Semi-structured Data, Springer nature, Lecture Notes in Computer Science book series (LNISA, volume 5731), 2020
- [9] Dennis M. Dimiduk, Elizabeth A. Holm & Stephen R. Niezgoda Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial Intelligence on Materials, Processes, and Structures Engineering, Springer Nature, Integrating Materials and Manufacturing Innovation volume 7, pages157–172 (2018)
- [11] Alanazi, R., 2022. Identification and prediction of chronic diseases using machine learning approach. Journal of Healthcare Engineering, 2022.
- [12] Cruz, J.A. and Wishart, D.S., 2006. Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, p.117693510600200030.
- [13] Razali, N., Mustapha, A., Abd Wahab, M.H., Mostafa, S.A. and Rostam, S.K., 2020, April. A data mining approach to prediction of liver diseases. In Journal of Physics: Conference Series (Vol. 1529, No. 3, p. 032002). IOP Publishing.
- [14] Tchito Tchapgaa, C., Mih, T.A., Tchagna Kouanou, A., Fozin Fonzin, T., Kuetche Fogang, P., Mezatio, B.A. and Tchiotop, D., 2021. Biomedical image classification in a big data architecture using machine learning algorithms. Journal of Healthcare Engineering, 2021.
- [15] Saranya, P., and P. Asha. "Survey on Big Data Analytics in health care." 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2019.
- [16] Souid, Abdelbaki, Nizar Sakli, and Hedi Sakli. "Classification and predictions of lung diseases from chest x-rays using mobilenet v2." Applied Sciences 11.6 (2021): 2751.
- [17] Lanzer, J.D., Leuschner, F., Kramann, R., Levinson, R.T. and Saez-Rodriguez, J., 2020. Big data approaches in heart failure research. Current Heart Failure Reports, 17(5), pp.213-224.
- [18] Kanchanamala, P., Das, S. and Neelima, G., 2022. Symptoms-Based Disease Prediction Using Big data Analytics. In Innovations in Computer Science and Engineering (pp. 339-346). Springer, Singapore.
- [19] Hamsagayathri, P. and Vigneshwaran, S., 2021, February. Symptoms Based Disease Prediction Using Machine Learning Techniques. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 747-752). IEEE.
- [20] Menaga, S. and Paruvathavardhini, J., 2022. AI in Healthcare. Smart Systems for Industrial Applications, pp.115-140.
- [21] Fox, M. and Vaidyanathan, G., 2016. Impacts of healthcare Big Data: a Framework with Legal and Ethical insights. Issues in Information Systems, 17(3).
- [22] .Muhammad Imran Razzak,1 Muhammad Imran and Guandong Xu1"Big data analytics for preventive medicine", Neural Comput Appl. 2020; 32(9): 4417–4451. Doi: 10.1007/s00521-019-04095-y
- [23] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ch IP, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220.
- [24] Frank A, Asuncion A (2010) UCI machine learning repository. <http://archive.ics.uci.edu/ml>. University of California, Irvine, CA. School of Information and Computer Science, 213.
- [25] Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, OLeary DH, Psaty B, Rautaharju P,
- [26] Tracy RP, Fried LP, Borhani NO, Weiler PG (1991) The cardiovascular health study: design and rationale. Ann Epidemiol 1(3):263–276
- [27] Ghamdi HA, Alshammari R, Razzak MI (2016) An ontologybased system to predict hospital readmission within 30 days. Int J Healthc Manag 9(4):236–244
- [28] Al-Qarny ZA, Alshammari R, Razzak MI (2015) Impact of sharing health information related to diabetes through the social media network: ontology. Int J Behav Healthc Res 5(3–4):162–171
- [29] Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN, Strack B, DeShazo JP (2014) Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int. <https://doi.org/10.1155/2014/781670>
- [30] Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M (2012) Predicting seminal quality with artificial intelligence methods. Expert Syst Appl 39(16):12564–12573.
- [31] Asuncion A, Newman D (2007) UCI machine learning repository
- [32] Babu MSP, Ramana BV, Venkateswarlu NB (2012) A critical comparative study of liver patients from USA and India: an exploratory analysis. Int J Comput Sci 9:506.
- [33] Valko M, Hauskrecht M (2008) Distance metric learning for conditional anomaly detection. In: FLAIRS conference, pp 684–689.
- [34] Policy brief on ageing no. 3, older persons as consumers (2009)
- [35] Slawson DL, Fitzgerald N, Morgan KT (2013) Position of the academy of nutrition and dietetics: the role of nutrition in health promotion and chronic disease prevention. J Acad Nutr Diet 113(7):972–979.
- [36] World Health Organization (1990) Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group. Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group, p 797
- [37] Willett WC, Koplan JP, Nugent R, Dusenbury C, Puska P, Gaziano TA (2006) Prevention of chronic disease by means of diet and lifestyle changes. Disease Control Priorities in Developing Countries, pp 833–850



# Chapter - 15

## Integrating Smart Wearables and Exploratory Data Analysis for Disease Prediction

Dr.D.Satheesh Kumar<sup>1</sup>, Sai Raam V<sup>2</sup>

<sup>1</sup> Associate Professor, Dept. of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India

<sup>2</sup> Department of Computer Science, Hindusthan College of Engineering and Technology, Tamil Nadu, India

E-mail: <sup>1</sup> [dsatheeshme@gmail.com](mailto:dsatheeshme@gmail.com), <sup>2</sup> [srinrealyf@gmail.com](mailto:srinrealyf@gmail.com)

*Abstract— In today's modern world with a population of more than 7.5 billion people, the advancement in predicting and diagnosing patients has become preeminent. The most common problem doctors face across the globe in treating the patient is identifying the disease at the eleventh hour or even later. This is not the sole problem of the patients themselves. At the time when the patient consults with the doctor for the disease, mostly it will be untreatable or will be treated just to get side effects post treatment. In the 21<sup>st</sup> century, while treatment has been modernized and a cure for thousands of diseases has been found compared to the past 4-5 decades, still Coronary Artery Disease (CAD) tolls nearly 15.5% of annual deaths which is nearly 8.8 Million in 2015. This rate is said to be increased quarterly by now. However untreated CAD or treating CAD in the annihilate stage ends up in chest pain, and arrhythmia and even becomes fatal with heart failure [1]. However, this ratio can be changed to a nominal level if the symptoms or the disease are identified at an earlier stage. In this chapter, we will discuss one of the effective methods for identifying Cardiovascular Diseases at an earlier stage using smart wearables and other non-invasive products. Smart wearables from smartwatches to smart suits and other medical products like non-invasive medical patches will be discussed on how the data collected from these devices helps in predicting CAD after performing exploratory data analysis on the collected dataset. This area of research is currently limited to predicting cardiovascular diseases as only the dataset for CAD is been found with the most accurate data which is easier to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.*

*Keywords— Smart Wearables, Disease Prediction, Exploratory Data Analysis, Cardiovascular Disease, Coronary Artery Disease*

### I. INTRODUCTION

The popularity of fitness bands and smartwatches has made it easier to collect physiological indicators like heart rate, activity, sleep, etc. from these wearable biosensors [2-3]. Smart wearables were owned by more than 100 million people as of 2019 and the figure is still rising. Wearable device data can offer people more unbiased information about their health status than huge data from search engines on the web. For instance, once users contract a disease like influenza, their physiological indications might change.

They utilized the heart rate and sleep data from the wearable devices to improve upon the standard models. The prediction results have strong correlation with the official data. Li et al. also investigated the role of physiological changes measured with wearable devices on the diagnosis and analysis of disease [4]. The researchers established a personalized disease detection framework, which identifies abnormal physical signs, e.g., from Lyme disease and other inflammatory responses, from the longitudinal data of the individuals.

Before diving into Exploratory Data Analysis (EDA), it's necessary to grasp the ideas and different phenomena of Data Mining. Data processing is additionally familiarised with applications than the essential nature of the underlying phenomena. In straightforward terms, data processing may be a process (or associate degree activity) to get new patterns and prices during a dataset. Varying stages of data discovery in the databases method are described as follows. In the selection stage, it obtains information from totally different resources. In the pre-processing stage, it removes the unwanted missing and screaming information and volume the clean information which may format to a standard format in the transformation stage. Then data processing techniques are applied to induce desired output. Finally, within the interpretation stage, it'll render the result to the end-user in a meaningful manner.

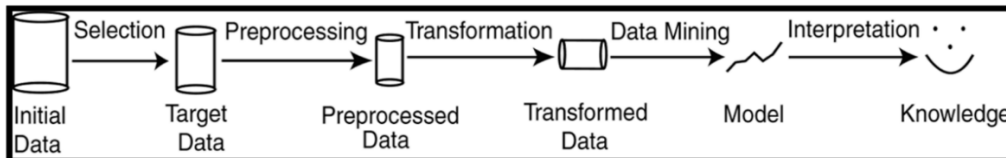


Figure 1.1 Data Mining

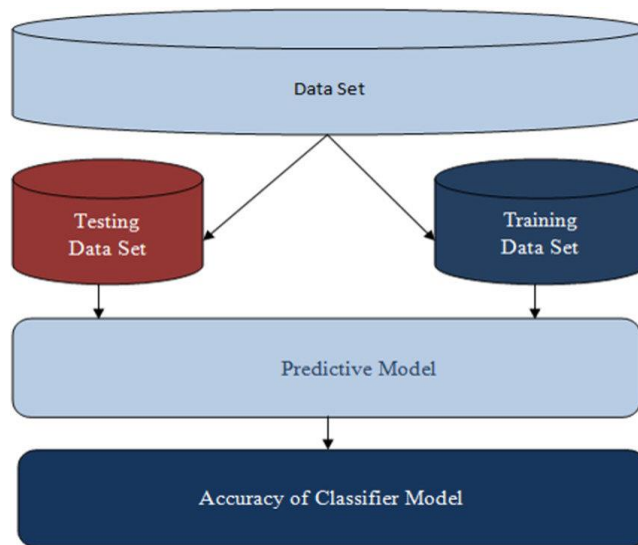
Data mining refers to the method of computationally extracting unknown data from immense sets of information. Extraction of helpful data from the big information sets and providing decision-making results for the designation and treatment of diseases is extremely vital. Data processing is often used to extract data by analyzing and predicting varied diseases. Healthcare data processing has huge potential to get hidden patterns within the information sets of the medical domain. Varied data processing techniques are available on the market with their quality obsessed with the health care information. Data processing applications in health care will have a beautiful potential and effectiveness. It automates the method of finding prophetic data in immense databases.

## II. ROLE OF DATA MINING

Data Mining plays a vital role in disease prediction. Back in the era when computational data was not used in predicting diseases, the patient has to undergo dozens of painful tests just to know if he/she has the chance to become vulnerable or already affected by the disease. This traditional way of diagnosing the patient changed in the early 90s in medically developed countries whereas Data Analysis for Disease Prediction comes into the play for rest of the countries only after the rise of the 21<sup>st</sup> century.

The finding of an illness needs the performance of a variety of tests on the patient. However, the use of data mining techniques will scale back the number of tests. This reduced test plays an important role in performance and time. Health care data processing is a necessary task as a result it permits doctors to visualize that attributes are a lot of important for a designation like age, weight, symptoms, etc. this can facilitate the doctors diagnose the illness a lot of with efficiency. Data knowledge in information bases is the method of finding helpful data and patterns in data. The data discovery in databases may be done through data mining. It uses algorithms to extract the information and patterns derived from the knowledge discovery in databases method. The algorithms for extracting the data and analyzing the patterns tend to vary supporting the kind of dataset and therefore the quantity of classification needed for the disease prediction.

The idea of data mining has originated from 3 completely different techniques viz. Statistics, Artificial Intelligence and Machine Learning. many heuristics are projected to perk up the competency of the data mining method. Classification may be a dominant data processing technique. In general, categoryfication is classified as single or multi-class. In a single category, there's just one category label that must be recognized. The weather that belongs to the category is called normal and the remainder of the elements is classified as anomalies. The regulation procedures are predicated upon training and testing knowledge sets. Within the system is trained by exploitation of existing tagged knowledge. The coaching knowledge set is then accustomed to segregate {the knowledge|the info|the information} into various classes primarily based upon the parameters and results of the coaching data set. In alternative words, coaching knowledge is one of the foremost phases of the classification method and is meant to impart some kind of intelligence primarily based upon that the information is deep-mined. Technically, training may be a machine learning method. The best data mining is predicated upon the character and level of the trained data set. Training data set is employed to develop a classification or prophetic model. To shun fitting quandary and to distill the classification model, sometimes, the coaching knowledge set is rotten into the coaching set and validation set. The responsibility of the validation set is to perk up the performance of the beneath construction framework. Finally, the accuracy of the developed system is checked with test knowledge. The accuracy is often measured by the exploitation confusion matrix that represents the accuracy of recognizing knowledge set as true negative, true positive, false positive, and false negative. There are many many techniques. The unremarkably used classification approaches are given below. call tree induction assists in learning ideas and dealing with call trees. A call tree may be a hierarchic constitution consisting of leaf, non-leaf nodes, and edges. The testing condition is described by the non-leaf node. Edge is employed to represent the result. The call tree classification approach is painless to implement. However, for a fancy, it needs immeasurable training data. The number of information is accrued because the dimension of classification is accrued. category labels are recognized by leaf nodes. ID3 (*Iterative Dichotomiser 3*), C4.5 (extension of ID3 algorithm), and CART (Classification and Regression Tree) are some milestones of DTI (Decision Tree Techniques). A number of the foremost characteristics of DTI are not any would-like domain information.



**Figure 1.2** Classification of Dataset

### 1.1.1 Suitable for investigative knowledge

Naive Bayes classification scrounges its plan from statistics and probability. It's primarily based on posterior probability and previous probability. One of the foremost anomalies during this approach is once one gets zero likelihood. To handle the zero probability case, the construct of 'Laplace' estimation was introduced. the simplest part of this approach is that normally it's the minimum rate or error. Rule primarily-based approach is predicated upon a set of rules. The foundations are portrayed within the type of 'IF-Then', extracted from the call tree or generated by employing an ordered covering approach. The working rule of rule primarily based on technique is predicated upon the antecedent (Left facet of the rule) and resultant (the right facet of the rule). One of the key rules primarily based on approach is sharp brought to a halt. To beat sharp bring to a halt drawback, the construct of formal logic is employed. SVM (Support Vector Machine) is employed to classify each linear and non-linear kind of data. In SVM, the size of the coaching knowledge square measure is enlarged by victimization non-linear mapping. The credit for SVM goes to Vladimir Vapnik. One of the simplest aspects of SVM is that it ends up with a high rate of accuracy. It is effectively applicable for each prediction, and classification method, conspicuously employed in digit recognition, voice identification, and seeing. SVM is powerfully primarily based upon hyperplane. It is accustomed completely differentiating components from different categories. For linear knowledge, the largest marginal hyperplane is portrayed as

$$d(X^T) = \sum_{i=1}^k C_i \alpha_i S_i X^T + b_0 \quad \text{-----1}$$

Here,  $C_i$  is the category label.  $\alpha_i$  is the Lagrangian multiplier factor.  $S_i$  is a support vector.  $X^T$  is test data. A neural network could be a soft computing technique that scrounges its plan from the human mind. These are units used for each single and multi-class issue. it appeared in 1943. However, came into action in the Nineteen Eighties. Besides, resolving the matter, the neural network also can learn from the previous system or applications. NN (Neural Network) area unit self-organizing and reconciling in learning. Genetic Algorithm (GA) borrows its essential options from natural biology and lets a population serene of various individual chromosomes grow below demarcated choice rules to breed a state that is optimizing the target performance. It typically employs some heuristics like 'Selection', 'Crossover', and 'Mutation' to develop higher solutions. GA is capable of being applied to ANN (Artificial Neural Network) tremendously big selection of issues like Task planning, Image process, Machine Learning, data processing, Medical Sciences, etc. It starts engaging with a collection of resolutions instead of one solution. The initial population is generated willy-nilly.

### III. CARDIOVASCULAR DISEASES

Cardiovascular disease (heart disease) refers to a gaggle of diseases that affect the guts and blood vessels of your body. These diseases will affect one or several components of your heart and /or blood vessels. an individual could also be symptomatic (physically expertise the disease) or be well (not feel something at all). heart condition includes heart or vas issues of those types: Abnormal heart rhythms, Heart valve sickness, narrowing of the blood vessels in your heart, different organs or throughout your body with plaque, Heart compression, and relaxation difficulties, Heart and vas issues that an individual born with and issues with the heart's outer lining [5].

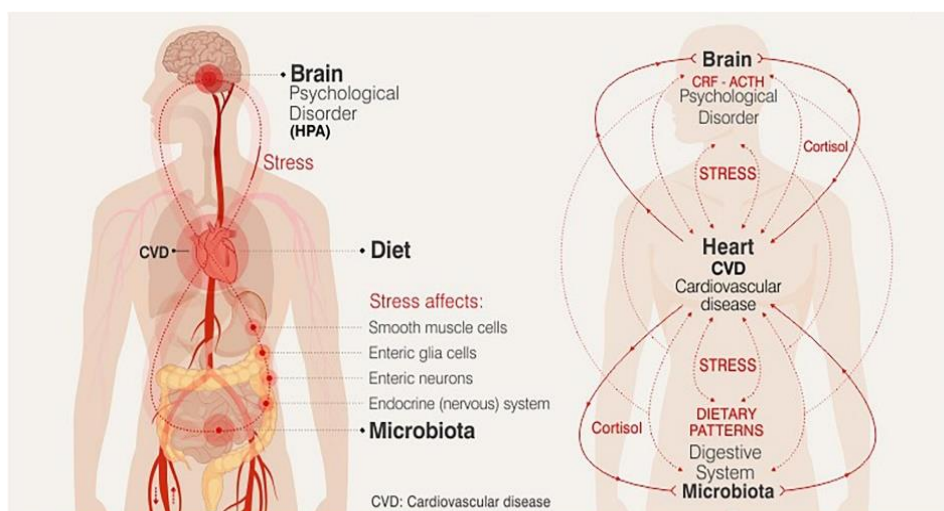


Figure 1.3 Cardiovascular Disease

### 1.2.1 Afflictions in CVD

Cardiovascular diseases (CVD) square measure as the leading reason for death where it representing the world's death rate of thirty percent with calculable human deaths of nearly 17.9 million in 2019. Of those deaths, eighty-five percent were due to coronary failure and stroke [6]. There are many forms of CVD diseases as however not restricted to;

- **Arrhythmia:** the downside with the conduction system of your heart which may cause abnormal heart rhythms or heart rates.
- **Valve disease:** the downside together with your heart valves (structures that enable blood to be due one chamber to a different chamber or blood vessel), like valve modification or leak.
- **Coronary artery disease:** the downside with the blood vessels of your heart, like blockages.
- **Heart failure:** the downside with heart pumping/relaxing functions, that cause fluid build-up and shortness of breath.
- **Peripheral artery disease:** the downside with the blood vessels of your arms, legs, or abdominal organs, like narrowing or blockages.
- **Aortic disease:** the downside with the massive vas that directs blood from your heart to your brain and therefore the remainder of your body, like dilatation or cardiovascular disease.
- **Congenital heart disease:** Heart downside that you're born with, which may affect completely different components of the guts.
- **Pericardial disease:** the downside with the liner of your heart, as well as carditis and pericardiac effusion.
- **Cerebrovascular disease:** the downside with the blood vessels that deliver blood to your brain, like narrowing or blockages.
- **Deep vein thrombosis:** Blockage within the veins, vessels that bring blood back from your brain/body to your heart.

### 1.2.2 Causes of Heart Disease

The causes of CVD will vary counting on the particular variety of disorders. As an example, arteriosclerosis (plaque build-up within the arteries) causes artery disease and peripheral artery disease. Artery disease, scarring of the guts muscle, genetic issues, or medications will cause arrhythmias. Aging, infections, and rheumatic unwellness will cause valve disease. the chance of developing CAD will increase with age and includes ages >45 years in men and >55 years in girls. A case history of early cardiovascular disease is additionally a risk issue, like a cardiovascular disease within the father or a brother diagnosed before age fifty-five years and within the mother or a sister diagnosed before age sixty-five years. Acute coronary and cerebrovascular will occur suddenly and can become typically fatal before medical aid is given. It is important to understand these risks to scale back incapacity and premature deaths from CVD, who have not yet experienced a cardiovascular event. People with established CVD are at very high risk of recurrent events.

A current study recommends screening distinctive symptomless people in danger of developing CVD. The objectives of those pointers area units are to scale back the incidence of initial or continual clinical events because of CHD (Coronary Heartery Disease), cerebrovascular accidents, and peripheral artery disease. the main focus is on the interference of incapacity and early death. the rules emphasize the importance of mode changes and the use of various prophylactic drug therapies in the management of risks. The understanding of such risk factors is crucial to the interference of cardiovascular morbidities and mortality[7].

### 1.2.3 Clinical tests for confirmation of heart disease

Some common tests to diagnose cardiovascular disease include:

- **Blood work** measures substances in the blood that indicate cardiovascular health, such as cholesterol and specific proteins.
- **Electrocardiogram (EKG)** records the electrical activity in your heart.
- **Ambulatory monitoring** uses wearable devices that track your heart rhythm and rates.
- **Echocardiogram** uses sound waves to create an image of your heartbeat and blood flow.
- **Cardiac CT** uses X-rays to create images of your heart and blood vessels.
- **Cardiac MRI** uses magnets and radio waves to create images of your heart.
- **Stress tests** use different ways to stress the heart in a controlled way (exercise or medications) to determine how your heart responds through EKGs and/or images.
- **Cardiac catheterization** uses a catheter (thin, hollow tube) to measure pressure and blood flow in your heart.

## IV. WEARABLE DEVICES FOR REAL-TIME DISEASE MONITORING

Wearable devices are getting widespread in an exceedingly wide selection of applications, from aid to medical specialty observation systems, that alter the continuous measure of crucial biomarkers for medical specialty, physiological health observation, and analysis. particularly because the older population grows globally, varied chronic and acute diseases become more and more necessary, and therefore the medical business is ever-changing dramatically thanks to the requirement for point-of-care (POC) diagnosing and period observation of semipermanent health conditions. wearable devices have evolved step by step within the variety of accessories, integrated articles of clothing, body attachments, and body inserts. Over the past few decades, the tremendous development of physical science, biocompatible materials, and nanomaterials has resulted in the development of implantable devices that alter the diagnosing and prognosis through little sensors and medical specialty devices, and greatly improve the standard and efficaciousness of medical services.

Beyond infection control, intelligent risk prediction models provide options for public health that include targeted health promotion and treatment adherence. In the future, therapies for chronic illnesses including heart disease, diabetes, and obesity will be informed by patient and wearable data. For instance, self-reported diet and weight information can be combined with daily step count and heart rate data to identify at-risk individuals and track the effectiveness of lifestyle changes in real-time. Deidentified population-level data can also be employed to more precisely detect care gaps, identify at-risk populations, and monitor risk indicators [8].

## V. MANUAL EDA FOR DISEASE PREDICTION

To begin with the Exploratory Data Analysis for disease prediction, The work of Aishah Ismail from the UCI Heart Disease dataset is used to find the anomalies from the given dataset and identify the disease in an efficient method. Initially, the dataset contains 76 features or attributes from 303 patients; however, the published study chose only 14 features that are relevant in predicting heart disease. Hence, here we will be using the dataset consisting of 303 patients with 14 feature sets.

**The outline for EDA is as follows:**

- **Import and get to know the data**
- **Data Cleaning**
- **Distributions and Relationship**
- **Automated EDA using pandas profiling report**

### 1.4.1 Variables or Features in the Dataset Explanations:

1. **age** (Age in years)
2. **sex** (1 = male, 0 = female)
3. **cp** (Chest Pain Type): [ 0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina]
4. **trestbps** (Resting Blood Pressure in mm/hg )
5. **chol** (Serum Cholesterol in mg/dl)
6. **fbs** (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes]
7. **restecg** (Resting ECG): [0: showing probable or definite left ventricular hypertrophy by Estes' criteria, 1: normal, 2: having ST-T wave abnormality]
8. **thalach** (maximum heart rate achieved)
9. **exang** (Exercise Induced Angina): [1 = yes, 0 = no]
10. **oldpeak** (ST depression induced by exercise relative to rest)



11. **slope** (the slope of the peak exercise ST segment): [0: downsloping; 1: flat; 2: upsloping]
12. **ca** [number of major vessels (0–3)]
13. **thal** [1 = normal, 2 = fixed defect, 3 = reversible defect]
14. **target** [0 = disease, 1 = no disease]

#### 1.4.2 Working with the dataset

Import the dataset to python and import all the necessary functions and libraries such as panda, numpy, os, matplotlib.pyplot, seaborn, missingno, and others [9]. After importing the dataset, now we have 303 rows with 14 variables. [Table 1.1]

#### Data Cleaning

First, check the data type of the variables.

The variable types are;

- Binary data type: sex, fbs, exang, target
- Categorical data type: cp, restecg, slope, ca, thal
- Continuous data type: age, trestbps, chol, thalach, oldpeak

```
[3] df = pd.read_csv('/content/drive/My Drive/csv/github/heart.csv')
df.head(3)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1

```
[ ] df.shape
(303, 14)

[ ] df.columns
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
      'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   age         303 non-null    int64
1   sex         303 non-null    int64
2   cp          303 non-null    int64
3   trestbps    303 non-null    int64
4   chol        303 non-null    int64
5   fbs         303 non-null    int64
6   restecg     303 non-null    int64
7   thalach     303 non-null    int64
8   exang       303 non-null    int64
9   oldpeak     303 non-null    float64
10  slope       303 non-null    int64
11  ca          303 non-null    int64
12  thal        303 non-null    int64
13  target      303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

**Table 1.1** After importing the dataset to python



## Check for the Data Characters Mistakes

The feature 'ca' ranges from 0–3, however, `df.nunique()` listed 0–4. So let us find the '4' and change them to NaN. [Table 2]

```
[ ] df['ca'].unique()
Out: array([0, 2, 1, 3, 4], dtype=object)

# to count the number in of each category descending order
df.ca.value_counts()
Out: 0    175
     1     65
     2     38
     3     20
     4         5
     Name: ca, dtype: int64

# To find the row for '4'
df[df['ca']==4]
Out:   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  target
     92   52   1   2     138   223   0         1    169     0     0.0   2   4   2     1
    158   58   1   1    125   220   0         1    144     0     0.4   1   4   3     1
    163   38   1   2     138   175   0         1    173     0     0.0   2   4   2     1
    164   38   1   2     138   175   0         1    173     0     0.0   2   4   2     1
    251   43   1   0     132   247   1         0    143     1     0.1   1   4   3     0

df.loc[df['ca']==4, 'ca'] = np.NaN

df['ca'].unique()
Out: array([0, 2, 1, 3, nan], dtype=object)
```

**Table 1.2** The range of ca - `df.nunique()` changed to NaN

Feature 'thal' ranges from 1–3, however, `df.nunique()` listed 0–3. There are two values of '0'. So let us change them to NaN. [Table 3]

```
[5] df.thal.value_counts()
Out: 2    166
     3    117
     1     18
     0         2
     Name: thal, dtype: int64

[10] df.loc[df['thal']==0, 'thal'] = np.NaN

[11] df[df['thal']==0]
Out:   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  target
     92   52   1   2     138   223   0         1    169     0     0.0   2   4   2     1
    158   58   1   1    125   220   0         1    144     0     0.4   1   4   3     1
    163   38   1   2     138   175   0         1    173     0     0.0   2   4   2     1
    164   38   1   2     138   175   0         1    173     0     0.0   2   4   2     1
    251   43   1   0     132   247   1         0    143     1     0.1   1   4   3     0

[12] df['thal'].unique()
Out: array([1.0, 2.0, 3.0, nan], dtype=object)
```

**Table 1.3** The value of thal - `df.nunique` changed to NaN

## Check for missing values and replace them

```
# to check missing values
df.isnull().sum()
Out: age          0
     sex          0
     cp           0
     trestbps     0
     chol         0
     fbs          0
     restecg      0
     thalach      0
     exang        0
     oldpeak      0
     slope        0
     ca           5
     thal         2
     target       0
     dtype: int64
```

**Table 1.4** Replacing Missing Values

**Visualize** the missing values using the Missingno library. The missing values are represented by horizontal lines. This library provides an informative way of visualizing the missing values located in each column and seeing whether there is any correlation between missing values of different columns.

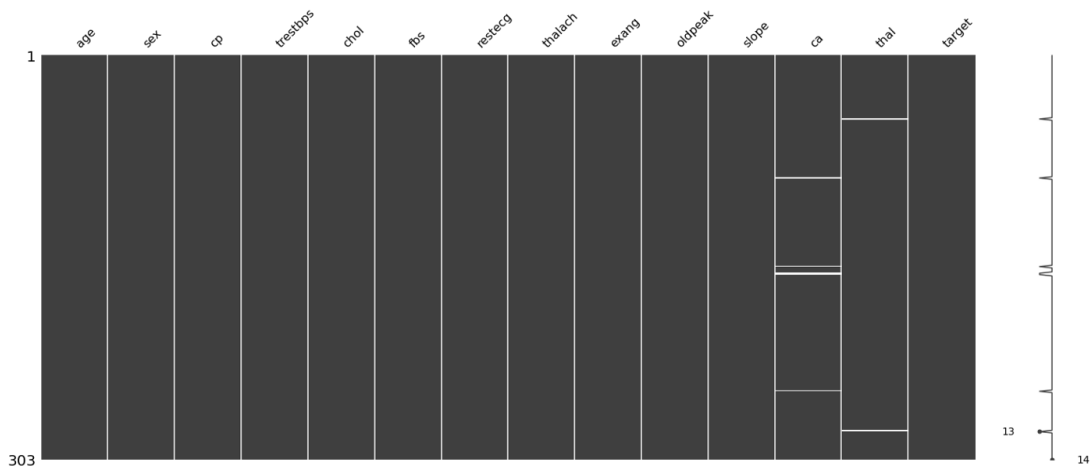


Figure 1.4 Visualization using Missingno library

### Replace the NaN with the median and check for duplicates

```
[ ] duplicated = df.duplicated().sum()
    if duplicated:
        print('Duplicates Rows in Dataset are : {}'.format(duplicated))
    else:
        print('Dataset contains no Duplicate Values')
```

Duplicates Rows in Dataset are : 1

```
[ ] duplicated = df[df.duplicated(keep=False)]
    duplicated.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
163	38	1	2	138	175	0	1	173	0	0.0	2	0.0	2.0	1
164	38	1	2	138	175	0	1	173	0	0.0	2	0.0	2.0	1

Table 1.5 Post duplicate check

### Statistics Summary

```
[13] # to know the basic stats
    df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.663366	2.326733	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	0.934375	0.583020	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	3.000000	3.000000	1.000000

Table 1.6 Statistics of the current dataset are displayed

Basically, with `df.describe()`, we should check on the min and max values for the categorical variables (min-max). Sex (0–1), cp (0–3), fbs (0–1), restecg (0–2), exang (0–1), slope (0–2), ca (0–3), thal (0–3). We should also observe the mean, std, 25%, and 75% on the continuous variables.

There are also other several ways of plotting **Boxplot** or using **Seaborn**.

Pass the following code to plot using Boxplot

```
fig = px.box(df, x="target", y="chol")
fig.show()
```

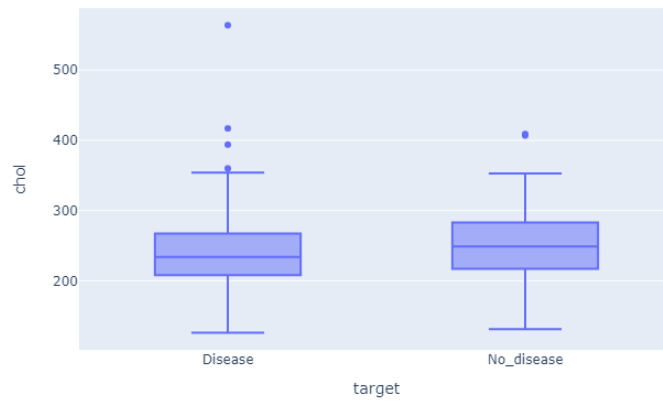


Figure 1.5 Boxplot

Pass the following code to plot using Seaborn  
`sns.boxplot(x='target', y='oldpeak', data=df)`

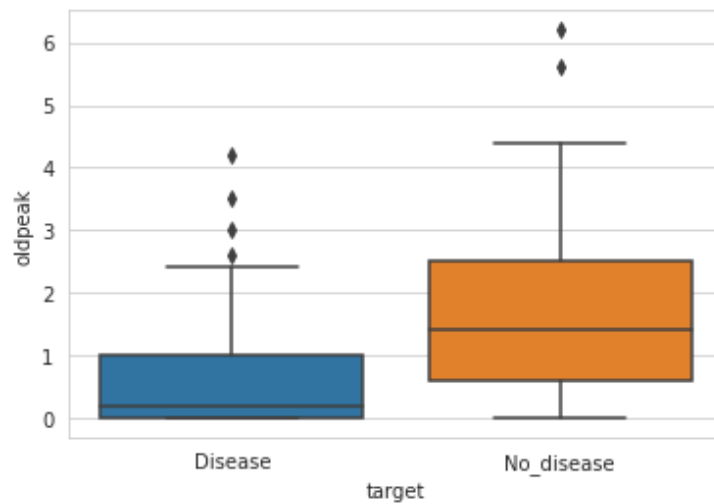


Figure 1.6 Seaborn

### 1.4.3 Distributions and Relationships

#### A] Target variable distribution

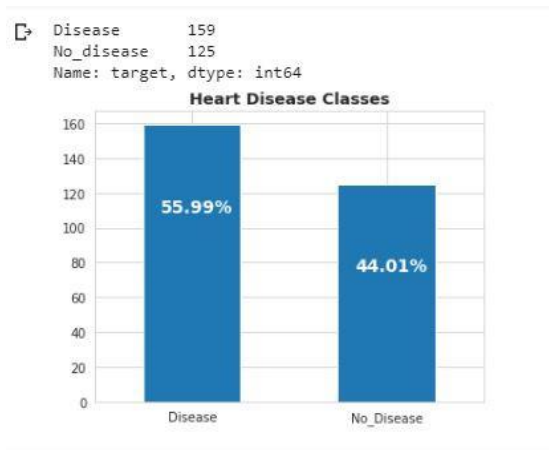
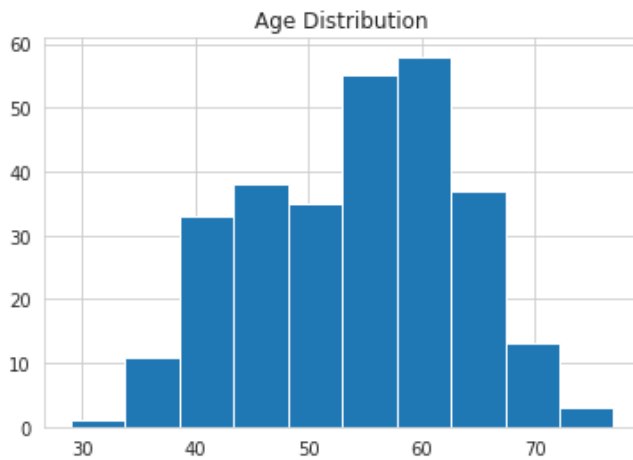


Figure 1.7 Target Variable Distribution

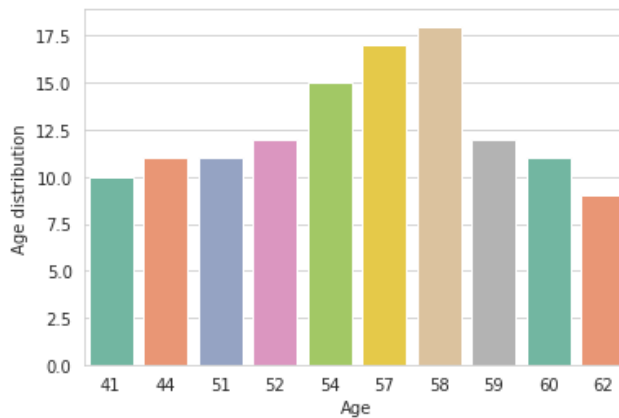
From the dataset, it is observed that there are more diseased than healthier ones.

**B] Age Variable Distribution**



**Figure 1.8** Age Variable Distribution

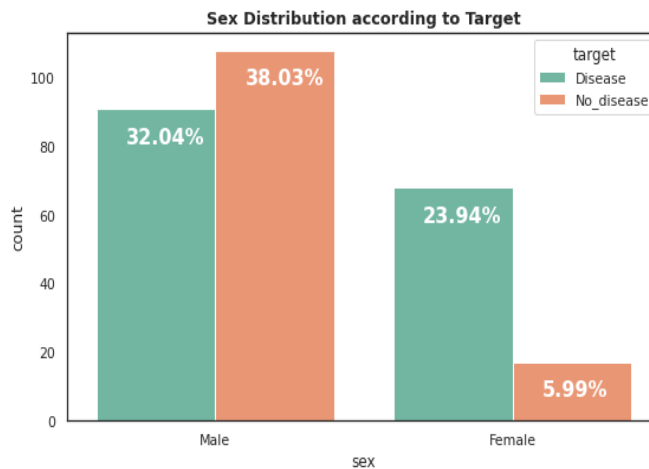
The age is normally distributed across the dataset.



**Figure 1.9** Age Variable Distribution

The majority of the patients are in the range of 50s to 60s. Let's take a quick look at basic stats. The mean age is about 54 years with  $\pm 9.08$  std, the youngest is at 29 and the oldest is at 77.

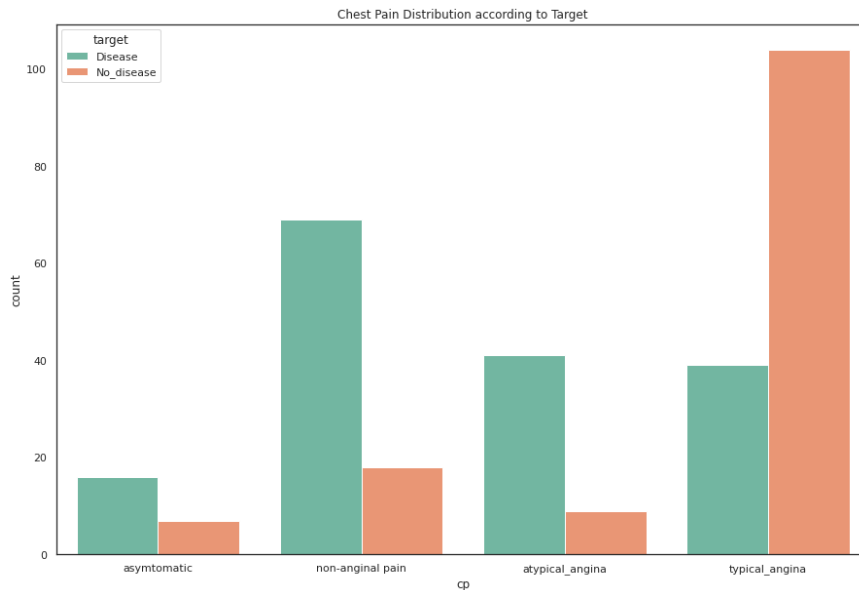
**C] Gender Distribution according to the target variable**



**Figure 1.10** Gender Distribution

We can point out that, males are most affected than females by heart disease.

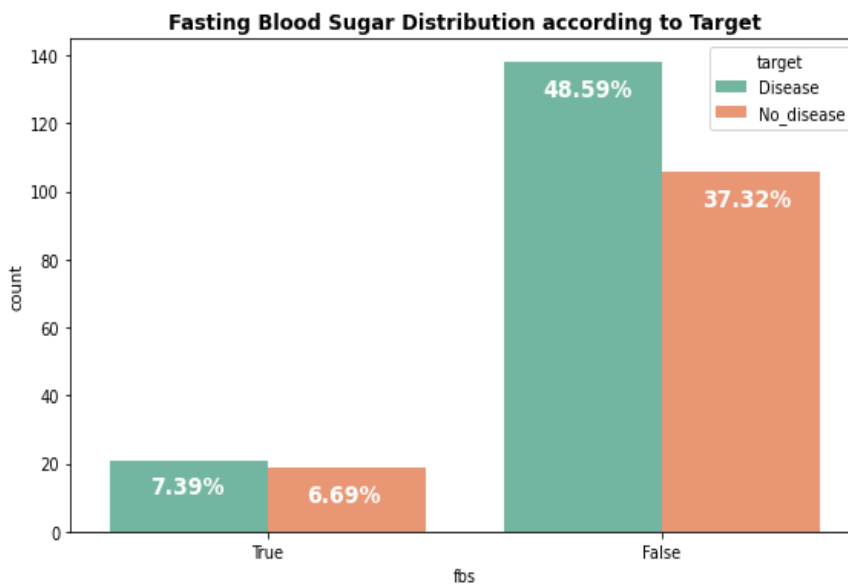
**D] Chest Pain Distribution according to the target variable**



**Figure 1.11** Chest Pain Distribution

Chest pain (cp) or angina may be a kind of discomfort caused by muscular tissue when it doesn't receive enough O<sub>2</sub> in blood, which triggered discomfort in the arms, shoulders, neck, etc. However, staring at the bar chart on top, raised a matter of the upper variety of healthy subjects having typical\_angina. Pain will be subjective because of stress, physical activities, and lots of a lot and varies between gender. women and elderly patients sometimes have atypical symptoms with a history of illness [10].

**E] Fasting Blood Sugar Distribution according to the target variable**



**Figure 1.12** Fasting Blood Sugar

Fasting glucose or fbs could be a polygenic disease indicator with fbs >120 mg/d taken into account of diabetes (True class). Here, we tend to observe that the amount for sophistication true, is lower compared to the category false. However, if we glance closely, there are a higher variety of cardiopathy patients while not have a polygenic disease. This gives a sign that fbs won't be a powerful feature in differentiating between cardiopathy and non-cardiopathy patients.

### F] Distribution Plot on continuous variables

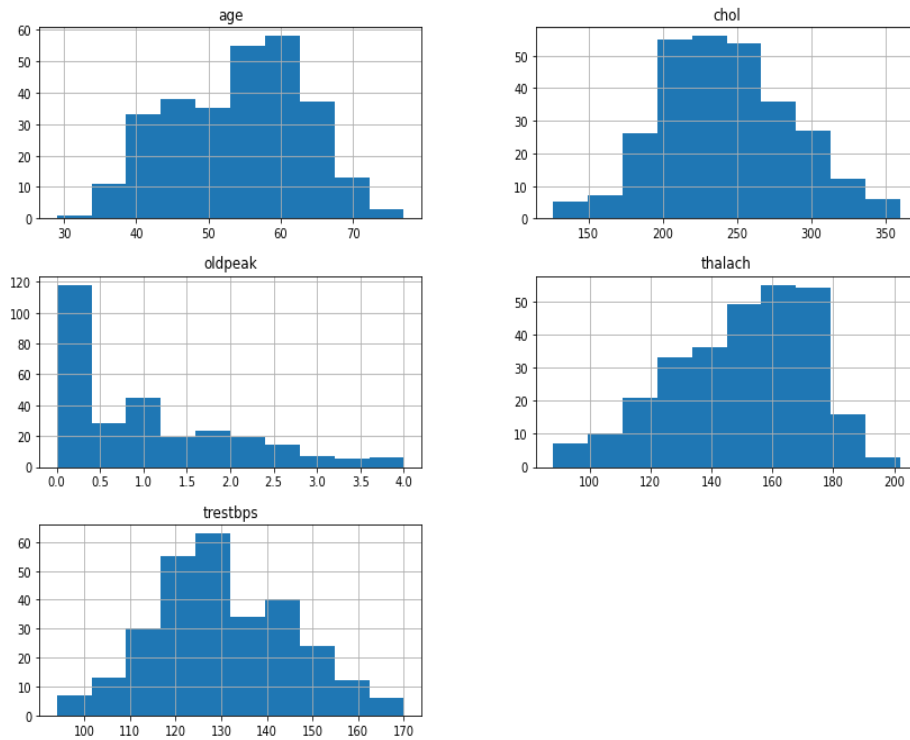


Figure 1.13 Distribution Plot

- normal distribution for age, trestbps, and almost for chol
- oldpeak is left-skewed
- thalach is right-skewed

### G] SnS pair plot to visualize the distribution

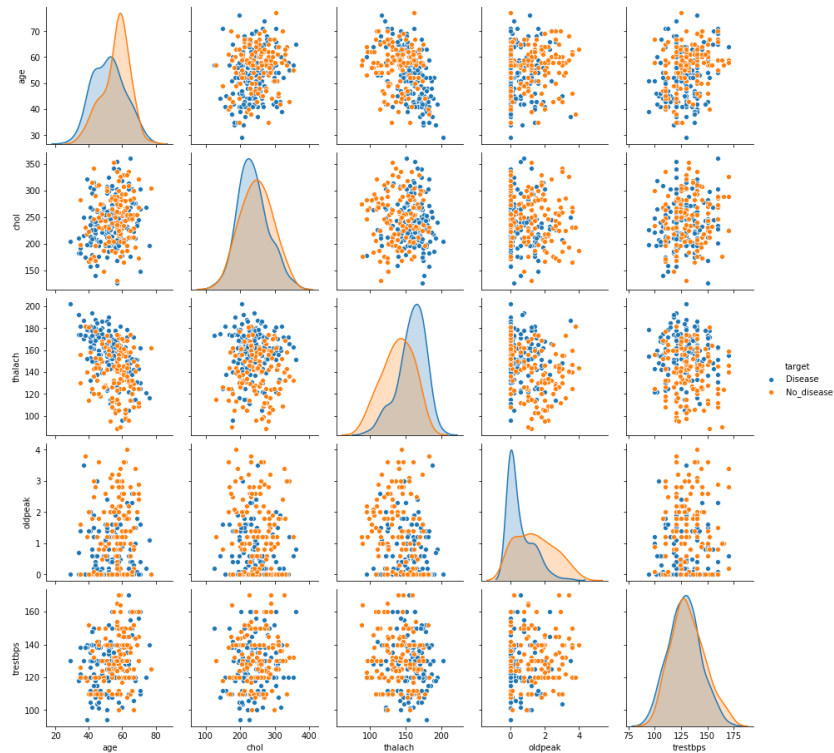


Figure 1.14 SnS Pairplot



- oldpeak having a linear separation relation between disease and non-disease.
- thalach having a mild separation relation between disease and non-disease.
- Other features don't form any clear separation

### H] Slope Distribution according to the target variable

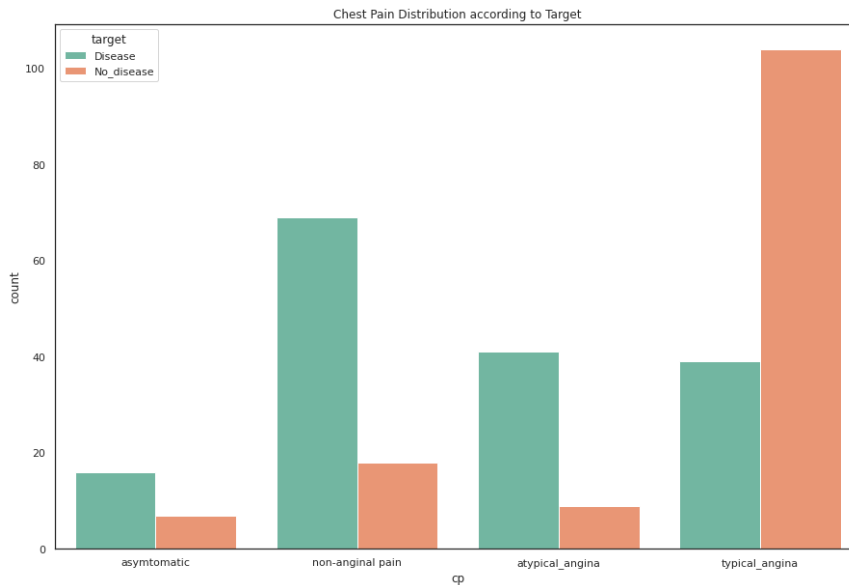


Figure 1.15 Slope Distribution

- normal distribution for asymptomatic
- non-anginal pain and atypical\_angina is left-skewed
- typical\_angina is right-skewed

### I] Correlation

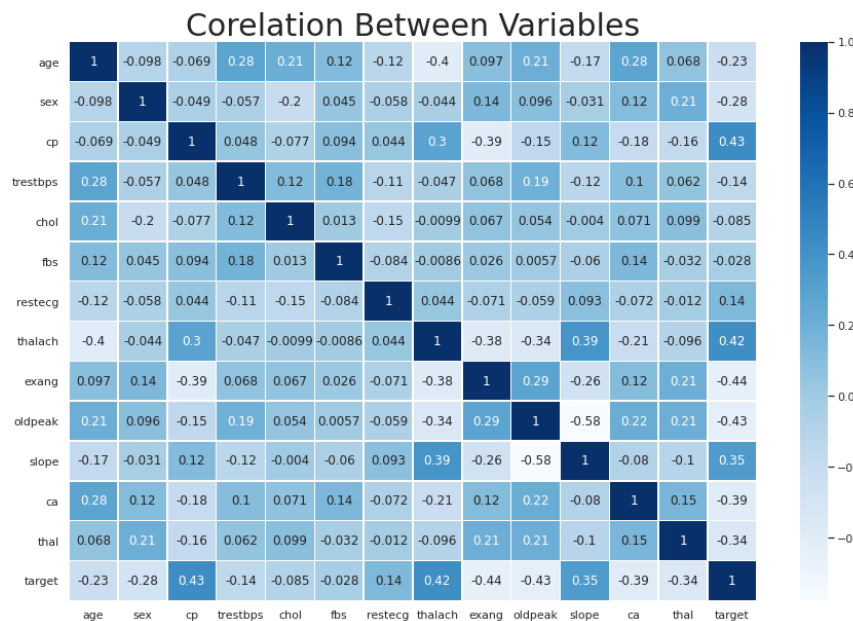


Figure 1.16 Correlation

- 'cp', 'thalach', and 'slope' shows a good positive correlation with the target
- 'oldpeak', 'exang', 'ca', 'thal', 'sex', and 'age' shows a good negative correlation with the target
- 'fbs', 'chol', 'trestbps', and 'restecg' has a low correlation with our target

## VI. INTEGRATING SMART WEARABLES AND EXPLORATORY DATA ANALYSIS

Any object that can be worn on the body and use various sorts of sensors to collect noninvasive signals from the human body is considered a wearable device. There are a variety of well-known signals and indicators that are drawn from the human body in literature in order to determine the vital signs and other details regarding the subject's health or mental state. The skin temperature sensor used in and the electrodermal activity (EDA) sensor, also referred to as the galvanic skin response (GSR) sensor, are two examples of these sensors. These sensors are used on the skin to record the skin conductance that varies with the subject's sympathetic state. Another example is the electrocardiogram (ECG) sensor used in to record electrical changes in the skin that correlate to heartbeats.

The upper hand of Exploratory Data Analysis (EDA) in the medical field has been proven with many real-time scenarios and the working of EDA through Aishah Ismail's work on EDA for Heart Disease is presented in this paper to ideal the effective working result of EDA. Integrating Exploratory Data Analysis with Smart Wearables has more advantages than can not be anticipated. By, smart-wearables it is not limited to wrist-mounted devices but also other non-invasive objects. Some of them include smart watches, smart glasses, wearable skin patches, and much more.

To get started with EDA for heart disease, the following data are very vital to perform the task. The required features are age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG, maximum heart rate in a period, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of peak exercise, and the number of major vessels. For diagnosing a patient with Cardiovascular Disease, a patient has to take several tests which give the details of the above-mentioned data. By using smart wearables, the features that are required to diagnose CVD-based disease can be done at ease for a finite level[11].

Wrist-mount devices for physiological observation are developed commercially with improvements in battery longevity and shrinking of hardware for changing raw signals to period explainable knowledge. Wrist-mounted devices, like fitness bands and smartwatches, are moving from basic accelerometer-based "smart pedometers" to incorporate biometric sensing. Typical noninvasive observation devices do 2 functions: (1) Communication with electronic devices and (2) observation of human physiological signals and human action signals. Blood pressure measure is one of the foremost necessary physiological indicators of an Associate in Nursing individual's health standing. standard pulse wave sensors used cuffs to non-invasively monitor pressure and enclosed optical, pressure, and ECG (ECG) sensors. However, these sensors are giant, tough to handle, and can't be accurately measured once the topic moves throughout the pressure measurement.

To unravel this downside, Lee's cluster developed a wearable device with a Hall device that may discover the minute changes in the force field of the magnet and acquire the heartbeat wave knowledge. This device will be worn on the radiocarpal joint, and maybe a pulsimeter while not a cuff. Hsu et al. conferred an example skin-surface-coupled personal wearable health observation system that captures hi-fi pressure waveforms in real-time and communicates with wireless devices like sensible phones and laptops. Recently, varied applications employing a negatron imaging (PPG - Photoplethysmography) based pulse sensing element mounted on the radiocarpal joint are projected. The bracelet-type PPG pulse sensing element developed by Ishikawa et al. detects changes in pulse and shows the likelihood of overcoming motion artifacts in daily activities. standardization of noise-free pulse detection was measured victimization noise reduction pulse signals supported peak detection and autocorrelation strategies [12].

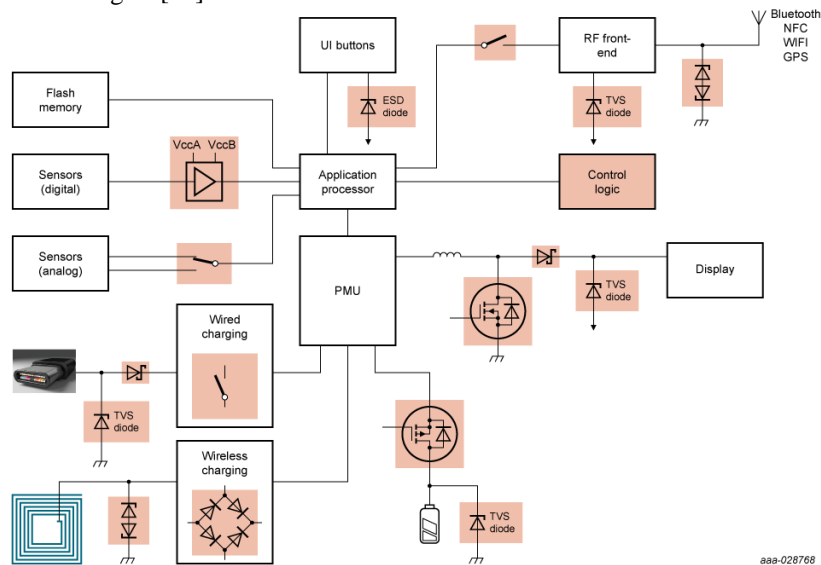


Figure 1.17 Working of Smart Wearable

Smartwatches are one of the foremost fashionable wearable device varieties, and GlucoWatch® writer (Cygnus opposition., Redwood town, CA, USA) is that the initial to possess a commercially approved non-invasive aldohexose monitor by the Food and Drug Administration (FDA). It electrochemically acquires info regarding aldohexose concentration extracted by reverse electromotive drug administration from skin extracellular fluid. Glennon et al. introduced a watch as well as fluid systems and storage systems, which might monitor atomic number 11 content within the body from sweat in real-time. additionally, the wrist-mounted device is applied within the measure of daily activity as well as motion, gesture, rotation, acceleration, and patient observation.

For observation of Parkinson’s malady (PD) patients, the smartwatches will be used to analyze tremor and balance dysfunction with a gyro or measuring instrument. Roberto’s cluster assessed sensible watches for quantification of tremor in metal patients, analysis of clinical correlation, and its acceptance and reliability as an observation instrument. As a result, the sensible watch has the likelihood as a clinical tool and sensible acceptance by patients. additionally, Tison’s cluster used the sensible devices for developing the Associate in Nursinging rule to discover fibrillation (AF- Atrial Fibrillation) from the information of pulse measured with PPG sensing element and step count with the measuring instrument. the most explanation for stroke is AF, and patients in danger of stroke will brace themselves for the malady by endlessly observing AF [13].

Wearable skin patches have become more and more pervasive at intervals in the wearables market. Soft, versatile, and stretchy electronic devices square measure connected to soft tissue to produce a replacement platform for robotic feedback and management, regenerative drugs, and continuous tending. Skin patches square measure ideal wearables, as a result, they will be obscured by wear and may record additional correct knowledge while not being disturbed by movement. wearable patches worn on the human skin are utilized as vas, sweat, strain, and temperature sensors.[14]

Monitoring CVD signals like pressure and also the pulse of patients receiving medical aid is extremely necessary. A thin, versatile, and patch-type continuous pressure (BP) watching detector is built with a bedded structure of a ferroelectric film, specially designed electrodes, and versatile electronic circuits, that along with change synchronous EKG (ECG - Electrocardiogram) and graph (BCG - Ballistocardiograph) measurements on the human chest while not discomfort. during a practicability study mistreatment of the developed detector, the estimates of heartbeat {blood pressure|vital sign|pressure|pressure level|force per unit square measurea} are in sensible agreement with the reference worth, and also the coefficient of correlation of the category was 0.95 ( $p < 0.01$ ).

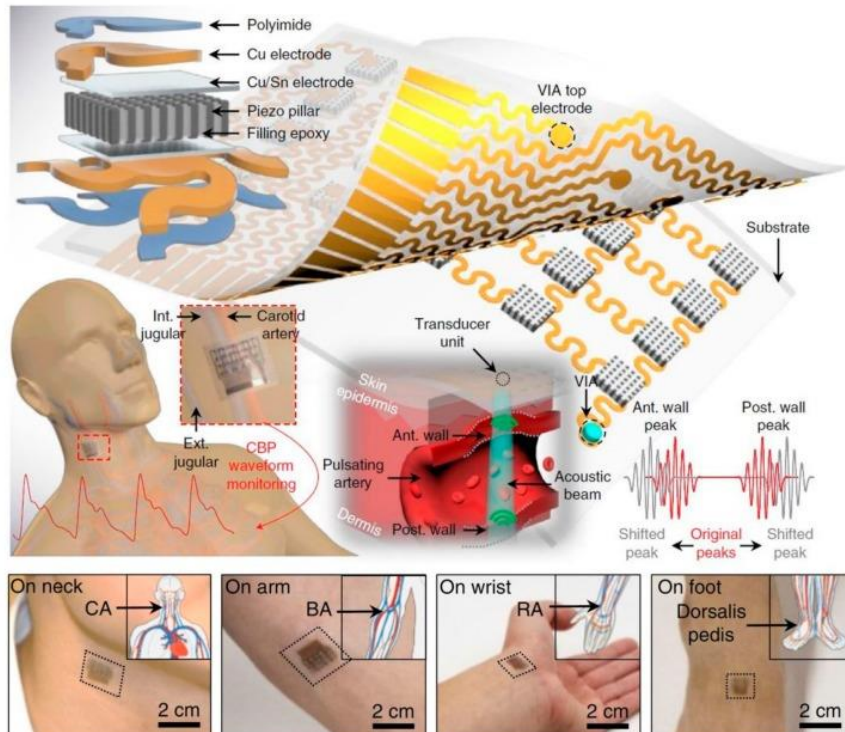


Figure 1.18 Smart Skin Patches

As reported in Advanced useful Materials, a wearable patch detector incorporating a versatile flexible piezoresistive sensor (FPS) detector and dermal graph sensors for cuffless pressure watching has been developed. The system at the same time measures dermal pulse signals and also the graph and obtains bit-to-bit BP knowledge in real-time through the heartbeat transit

time (PTT – Partial Thromboplastin Time) methodology. to get a stable surface pulse signal, a constant quantity model of the FPS detection mechanism was developed, and also the operating conditions were optimized. above all, this sensing patch will operate at ultra-low power (3 nW) and detects refined physiological changes like before and once exercise to produce promising solutions for the period and home-based BP (Blood Pressure) watching[15].

Electrooculogram (EOG), which is produced by eye movements and may be monitored with electrodes put around the eye [16], and electrogastrogram (EGG), which captures the electrical activity of the stomach, are other signals that have been utilised sparingly in literature. Olfactory inputs, in particular, are difficult to model, although it has been found that the human body reacts differently to different odours in terms of its autonomic reactions, which can be examined by GSR and ECG signals [17]. These inputs can be used in applications such as customised treatments for eating and neuropsychiatric problems based on flavours and foods. Other inputs could require visual monitoring.

Some literature studies use statistical values such as mean, minimum, maximum, mode, variance, standard deviation, entropy, and kurtosis. However, it is often hard to interpret how some of these statistical features affect the classification or the outcome variable. Additionally, model's accuracy is usually negatively affected by adding more ir-relevant features as more is not always better, and domain-specific features that are expressive achieve better performance [43]. Estimating heart rate and breathing rate as features from the PPG signal, change in acceleration magnitude, jerk of motion, and transient changes in skin resistance for seizure detection are examples of domain specific features. Some applications are concerned with changes happening over a long time period, and some are concerned with transient changes due to certain events such as fall detection and emotion recognition.

Statistical values including mean, minimum, maximum, mode, variance, standard deviation, entropy, and kurtosis are used in some literary studies. However, it is frequently difficult to determine how some of these statistical traits affect the result variable or the categorization. As more is not necessarily better and expressive domain-specific features yield greater performance, adding more ir-relevant features typically has a detrimental impact on a model's accuracy [19]. Examples of domain-specific features include estimation of heart rate and breathing rate from the PPG signal, change in acceleration magnitude, jerk of motion, and transient changes in skin resistance for seizure detection. While some applications are focused on long-term changes, others are focused on momentary changes brought on by specific events, such as fall detection and emotion recognition.

## VII. CONCLUSION

With the advancement of modern sensors in a tiny form factor, real-time EDA technology, and processing terabytes of data in a nanosecond, the world has revolutionized alongside the growth of the technology. By integrating smart wearables and the data collected with a cloud-based application that runs exploratory data analysis in the backend, a patient's health condition especially CVD and CAD patients can be keenly monitored 24/7 by the application as well as the doctors[16]. The application can be programmed to send alerts to doctors themselves or nearby hospitals in case of any irregularities (medical conditions) found in the patient's body. Since many smartwatches now have ECG feature built in, it is now time to develop smartwatches with more features that helps to monitor a person's health. Appending Ballistocardiogram (BCG) to smartwatches helps to observe more features accurately. While smart patches are adhesive based as of now, they can be changed to patch type where the user can clip and un-clip based on the requirement. Various Multinational Companies are spending millions of dollars in R&D to find a smartwatch that can accurately predict CVD. But everything comes at a cost. We need more data on users and the health record of a patient varies from one to another it is highly impossible to predict accurately but making a quicker diagnosis is highly possible[17]. This is more than enough for the user to consult with the doctor and get treatment at an earlier stage which is more effective than other methods. While the data for the diagnosis is collected from the smart wearables, Exploratory Data Analysis (EDA) plays a vital role in processing data to give the output in a most accurate form than any other data mining technique. This feature will be a showstopper in the upcoming years.

## REFERENCES

- [1] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning, and data mining methods in diabetes research. *Computational and structural biotechnology journal*. 2017; 104-116.
- [2] Y. Liu, H. Wang, W. Zhao, M. Zhang, H. Qin, and Y. Xie, "Flexible, stretchable sensors for wearable health monitoring: sensing mechanisms, materials, fabrication strategies and features," *Sensors*, vol. 18, no. 2, p. 645,
- [3] T. Arakawa, "Recent research and developing trends of wearable sensors for detecting blood pressure," *Sensors*, vol. 18, no. 9, p. 2772, 2018.
- [4] X. Li, J. Dunn, D. Salins et al., "Digital Health: tracking physiomes and activity using wearable biosensors reveals useful health-related information," *PLoS Biology*, vol. 15, no. 1, Article ID e2001402, 2017.
- [5] Patil PH, Thube S, Ratnaparkhi B, Rajeswari K. Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining. *International Journal of Computer Applications*. 2014; 93(8):35-39.
- [6] American Heart Association. Cardiovascular diseases affect nearly half of American adults, statistics show. (<https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show>) Accessed 9/28/2021.

- [7] Centers for Disease Control and Prevention. Heart Disease. (<https://www.cdc.gov/heartdisease/about.htm>) Accessed 9/28/2021.
- [8] Centers for Disease Control and Prevention. Heart Disease Facts. (<https://www.cdc.gov/heartdisease/facts.htm>) Accessed 9/28/2021.
- [9] World Health Organization. Cardiovascular diseases (CVDs). ([https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds%29\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds%29))) Accessed 9/28/2021.
- [10] ACC/AHA Clinical Practice Guideline. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. (<https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000000678>) Accessed 9/28/2021.
- [11] Tsao CW, et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. *Circulation* 2020. (<https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000000757>) Accessed 9/28/2021.
- [12] Cho L, Davis M, Elgendy I, et al. Summary of Updated Recommendations for Primary Prevention of Cardiovascular Disease in Women JACC State-of-the-Art Review. (<https://www.jacc.org/doi/abs/10.1016/j.jacc.2020.03.060>) *J Am Coll Cardiol* 2020 May, 75(20):2602-2618. Accessed 9/28/2021.
- [13] Exploratory Data Analysis on Heart Disease UCI data set - A complete step-by-step exploratory data analysis with a simple explanation by Aishah Ismail
- [14] Chan M., Esteve D., Fourniols J.Y., Escriba C., Campo E. Smart wearable systems: Current status and future challenges. *Artif. Intell. Med.* 2012;56:137–156
- [15] Patel S., Park H., Bonato P., Chan L., Rodgers M. A review of wearable sensors and systems with application in rehabilitation. *J. Neuroeng. Rehabil.* 2012;9:1–17.
- [16] Khan Y., Ostfeld A.E., Lochner C.M., Pierre A., Arias A.C. Monitoring of vital signs with flexible and wearable medical devices. *Adv. Mater.* 2016;28:4373–4395. doi: 10.1002/adma.201504366.
- [17] Hwang I., Kim H.N., Seon G.M., Lee S.H., Kang M., Yi H., Bae W.G., Kwak M.K., Jeong H.E. Multifunctional smart skin adhesive patches for advanced health care. *Adv. Healthc. Mater.* 2018;7:1800275. doi: 10.1002/adhm.201800275.
- [18] Kamisalic A., Fister I., Jr., Turkanovic M., Karakatic S. Sensors and functionalities of non-invasive wrist-wearable devices: A review. *Sensors.* 2018;18:1714. doi: 10.3390/s18061714.

## Compilation of References

- [1] [Aman and R. S. Chhillar, "Disease predictive models for healthcare by using data mining techniques: State of the art," *Int. J. Eng. Trends Technol.*, vol. 68, no. 10, pp. 52–57, 2020, doi: 10.14445/22315381/IJETT-V68I10P209.
- [2] A literature review of 2019 novel coronavirus (SARS-CoV2) infection in neonates and children Matteo Di Nardo, Grace van Leeuwen, Alessandra Loreti, Maria Antonietta Barbieri, Yit Guner, Franco Locatelli and Vito Marco Ranieri. *Pediatric research*. 2020
- a. Al-Ajlan, A., "The comparison between forward and backward chaining. *international journal of machine learning and computing*, "5, 2nd ser. 2015
- [3] Al-Ajlan, A., "The comparison between forward and backward chaining. *international journal of machine learning and computing*, "5, 2nd ser. 2015
- [4] Ali et al., "Network Intrusion Detection Leveraging Machine Learning and Feature Selection," 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), 2020, pp. 49-53.
- [5] Bayata. "Review on nutritional value of cassava for use as a staple food". *Sci J Anal Chem*, Vol.7, No. 4, pp.83-91, Sep. 2019. doi: : 10.11648/j.sjac.20190704.12.
- [6] Bayata. "Review on nutritional value of cassava for use as a staple food". *Sci J Anal Chem*, Vol.7, No. 4, pp.83-91, Sep. 2019. doi: : 10.11648/j.sjac.20190704.12.
- [7] Camargo, and J.S. Smith, "An image-processing based algorithm to automatically identify plant disease visual symptoms," *Biosyst. Eng*, Vol. 102, No.1, pp.9-21. Jan.2009, doi : 10.1016/j.biosystemseng.2008.09.030.
- [8] Camargo, and J.S. Smith, "An image-processing based algorithm to automatically identify plant disease visual symptoms," *Biosyst. Eng*, Vol. 102, No.1, pp.9-21. Jan.2009, doi : 10.1016/j.biosystemseng.2008.09.030.
- [9] Darolia and R. S. Chhillar, "Analyzing Three Predictive Algorithms for Diabetes Mellitus Against the Pima Indians Dataset," *ECS Trans.*, vol. 107, no. 1, pp. 2697–2704, 2022, doi: 10.1149/10701.2697ecst.
- [10] Fuentes, D.H. Im, S. Yoon and D.S. Park "Spectral analysis of CNN for tomato disease identification," In *International Conference on Artificial Intelligence and Soft Computing*, pp. 40 – 51, Springer, Cham, 2017.
- [11] Fuentes, D.H. Im, S. Yoon and D.S. Park "Spectral analysis of CNN for tomato disease identification," In *International Conference on Artificial Intelligence and Soft Computing*, pp. 40 – 51, Springer, Cham, 2017.
- [12] Kliot, et al., "Fluorescence in situ hybridizations (FISH) for the localization of viruses and endosymbiotic bacteria in plant and insect tissues," *J. Vis. Exp*, Vol. 84, p. e51030, Feb. 2014, doi : 10.3791/51030.
- [13] Kliot, et al., "Fluorescence in situ hybridizations (FISH) for the localization of viruses and endosymbiotic bacteria in plant and insect tissues," *J. Vis. Exp*, Vol. 84, p. e51030, Feb. 2014, doi : 10.3791/51030.
- [14] Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, arXiv preprint arXiv:1404.5997. [Online]. Available <https://arxiv.org/abs/1404.5997>.
- [15] Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, arXiv preprint arXiv:1404.5997. [Online]. Available <https://arxiv.org/abs/1404.5997>.
- [16] Muchtar, D. Nur, E. Tungadi, and M.N.Y Utomo, "Perancangan Back-End Server Menggunakan Arsitektur Rest dan Platform Node. JS (Studi Kasus : Sistem Pendaftaran Ujian Masuk Politeknik Negeri Ujung Pandang), " *Seminar Nasional Teknik Elektro dan Informatika (SNTEI)*, pp. 72-77, Oct. 2020.
- [17] Muchtar, D. Nur, E. Tungadi, and M.N.Y Utomo, "Perancangan Back-End Server Menggunakan Arsitektur Rest dan Platform Node. JS (Studi Kasus : Sistem Pendaftaran Ujian Masuk Politeknik Negeri Ujung Pandang), " *Seminar Nasional Teknik Elektro dan Informatika (SNTEI)*, pp. 72-77, Oct. 2020.
- [18] R. Jamasb, B. Day, ~ Ta ~ Lina Cangea, P. Liò, and T. L. Blundell, "Chapter 16 Deep Learning for Protein-Protein Interaction Site Prediction," doi: 10.1007/978-1-0716-1641-3\_16.
- [19] Wang, "Internet of Things Computer Network Security and Remote Control Technology Application," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 1814-1817.
- [20] Zhang, E. Jakku, R. Llewellyn, and E.A. Bake, "Surveying the needs and drivers for digital agriculture in Australia," *Farm Policy J*, Vol.15, No,1, pp. 25-39, 2018.
- [21] Zhang, E. Jakku, R. Llewellyn, and E.A. Bake, "Surveying the needs and drivers for digital agriculture in Australia," *Farm Policy J*, Vol.15, No,1, pp. 25-39, 2018.



- [22] A.J. Rozaqi, A. Sunyoto, and M.R. Arief, "Deteksi Penyakit Pada Daun Kentang Menggunakan Pengolahan Citra dengan Metode Convolutional Neural Network," *Creat. Inf. Technol. J.*, Volume 8, No.1, pp. 22-31, Mar. 2021, doi : 10.24076/citec.2021v8il.263.
- [23] A.J. Rozaqi, A. Sunyoto, and M.R. Arief, "Deteksi Penyakit Pada Daun Kentang Menggunakan Pengolahan Citra dengan Metode Convolutional Neural Network," *Creat. Inf. Technol. J.*, Volume 8, No.1, pp. 22-31, Mar. 2021, doi : 10.24076/citec.2021v8il.263.
- [24] A.K.Rumpf, et al., "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance," *Comput Electron Agric*, Vol.74, No.1, pp. 91–99. Oct..2010, doi : 10.1016/j.compag.2010.06.009
- [25] A.K.Rumpf, et al., "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance," *Comput Electron Agric*, Vol.74, No.1, pp. 91–99. Oct..2010, doi : 10.1016/j.compag.2010.06.009
- [26] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and Privacy Challenges in Industrial Internet of Things," in 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC). IEEE, 2015, pp. 1–6.
- [27] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, Rashedur M Rahman (2015) "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh" 978-1-4799-86767, IEEE SNPD .
- [28] Abdolvahab, E.R., & Kumar, Y.H.S. 2010. Leaf recognition for plant classification using GLCM and PCA methods. *International Journal of Computer Science & Technology*, 3(1): 31-36.
- [29] Abrar, Z. Ayub, F. Masoodi and A. M. Bamhdi, "A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 919-924.
- [30] ACC/AHA Clinical Practice Guideline. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. (<https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000000678>) Accessed 9/28/2021.
- [31] Akash Raj N, Balaji Srinivasan, Deepit Abhishek D, Sarath Jeyavanth J, Vinith Kannan A, "IoT based Agro Automation System using Machine Learning Algorithms", *International Journal of Innovative Research in Science, Engineering and Technology* November 2016, pp. 19938-19342
- [32] Alanazi, R., 2022. Identification and prediction of chronic diseases using machine learning approach. *Journal of Healthcare Engineering*, 2022.
- [33] Al-Qarny ZA, Alshammari R, Razzak MI (2015) Impact of sharing health information related to diabetes through the social media network: ontology. *Int J Behav Healthc Res* 5(3–4):162–171
- [34] Altalak, Maha, Amal Alajmi, and Alwaseemah Rizg. "Smart Agriculture Applications Using Deep Learning Technologies: A Survey." *Applied Sciences* 12, no. 12 (2022): 5919.
- [35] Aman and R. S. Chhillar, "Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease using WEKA Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, p. 2021, Oct. 2021.
- [36] American Heart Association. Cardiovascular diseases affect nearly half of American adults, statistics show. (<https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show>) Accessed 9/28/2021.
- [37] Anish Halimaa A, K. Sundarakantham: Machine Learning Based Intrusion Detection System. In: *Proceedings of the Third International Conference on Trends in Electronics and Informatics*, pp. 916–920. IEEE Xplore, Tirunelveli, India (2019).
- [38] Anita, Priscilla, Mary, M., & Josephine, M.S. (2018), Analysis and Forecasting Of Electrical Energy a Literature Review. *International Journal of Pure and Applied Mathematics*, 119(15), 289-293.
- [39] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb (2018) — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, *Computers and Electronics in Agriculture* 151 61–69, Elsevier.
- [40] Aranganayagi, S., & Thangavel, K. (2007). Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)* (pp. 13-17).
- [41] Asuncion A, Newman D (2007) UCI machine learning repository
- [42] Atonu Ghosh, Koushik Majumder, Debashis De. "Chapter 2 A Systematic Review of Digital, Cloud and IoT Forensics"

- [43] Audun, Josang. & Jochen, Haller. (2007, April). Dirichlet Reputation Systems, Paper presented at the Second International Conference on Availability, Reliability and Security (ARES'07), Vienna, Austria
- [44] Ge and J. Xu, "Analysis of Computer Network Security Technology and Preventive Measures under the Information Environment," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 1978-1981.
- [45] Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," 2015 International Conference on Signal Processing and Communication Engineering Systems, 2015, pp. 92-96.
- [46] Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Front. Genet.*, vol. 10, no. MAR, pp. 1–10, 2019, doi: 10.3389/fgene.2019.00214.
- [47] Wen et al., "Deep Learning in Proteomics," *Proteomics*, vol. 20, no. 21–22, Nov. 2020, doi: 10.1002/PMIC.201900335.
- [48] Xu, S. Chen, H. Zhang and T. Wu, "Incremental k-NN SVM method in intrusion detection," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 712-717.
- [49] B.R. Hastilestari, C.F. Pantouw, S. Nugroho and A. Estiati. "Uji ketahanan padi transgenik mengandung gen Cry 1B dibawah kontrol promoter terinduksi pelukaan Mpi terhadap hama penggerek batang kuning (*Scirpophaga Incertula* WK.) pada fase vegetatif". Prosiding Seminar Nasional 2013 : Inovasi Teknologi Padi Adaptif Perubahan Iklim Global Mendukung Surplus 10 Juta Ton Beras 2014. Balai Penelitian dan Pengembangan Pertanian Kementerian Pertanian, pp. 215 – 223, Jul. 2014.
- [50] B.R. Hastilestari, C.F. Pantouw, S. Nugroho and A. Estiati. "Uji ketahanan padi transgenik mengandung gen Cry 1B dibawah kontrol promoter terinduksi pelukaan Mpi terhadap hama penggerek batang kuning (*Scirpophaga Incertula* WK.) pada fase vegetatif". Prosiding Seminar Nasional 2013 : Inovasi Teknologi Padi Adaptif Perubahan Iklim Global Mendukung Surplus 10 Juta Ton Beras 2014. Balai Penelitian dan Pengembangan Pertanian Kementerian Pertanian, pp. 215 – 223, Jul. 2014.
- [51] B.R.Hastilestari, D. Astuti, A. Estiati and S. Nugroho, "Sequence analysis of ORF IV RTBV isolated from tungro infected *Oryza sativa* L. cv Ciherang". AIP Conference Proceedings, Vol. 1677, No. 1, p, 090013, Sep, 2015. <https://doi.org/10.1063/1.4930758>.
- [52] B.R.Hastilestari, D. Astuti, A. Estiati and S. Nugroho, "Sequence analysis of ORF IV RTBV isolated from tungro infected *Oryza sativa* L. cv Ciherang". AIP Conference Proceedings, Vol. 1677, No. 1, p, 090013, Sep, 2015. <https://doi.org/10.1063/1.4930758>.
- [53] Babu MSP, Ramana BV, Venkateswarlu NB (2012) A critical comparative study of liver patients from USA and India: an exploratory analysis. *Int J Comput Sci* 9:506.
- [54] Barbon G, Margolis M, Palumbo F, et al. 2016. "Taking Arduino to the Internet of Things: The ASIP programming model[J]". *Computer Communications*, 2016, s 89–90:128-140.
- [55] Basavaraj, S. A., Suvarna, S. N., & Govardhan, A. 2010. A combined color, texture and edge features-based approach for identification and classification of Indian medical plants. *International Journal of Computer Applications*, 6(12): 45-51.
- [56] Basumatary, Jwngsar., Pratap, Singh, Brijendra., Gore, M. M. (2018, January). Demand Side Management of a University Load in Smart Grid Environment, Paper presented at the Workshops ICDCN '18, Varanasi, India.
- [57] Ben, J. M. Jason, M. D., Naomi, S. B., Mitzi, L. D., & Dudley, R. A. 2014. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, 21(5): 871-875.
- [58] Benyue, Su, Wang Guangjun, and Zhang Jian. "Smart home system based on internet of things and Kinect sensor." *Journal of Central South University (Science and Technology)* 44, no. Suppl 1 (2013): 182-184.
- [59] Bhardwaj, A., Kaur, M., & Kumar, A. 2013. Recognition of plants by leaf image using moment invariant and texture analysis. *International Journal of Innovation and Application Studies*, 3(1): 237-248.
- [60] Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*. 3. 601-608.
- [61] Bodhwani, V., Acharjya, D. P., & Bodhwani, U. 2019. Deep residual networks for plant identification. *Procedia Computer Science*, 152: 186–194.
- [62] Borah, J. W. G. S., Hines, E. L., Leeson, M. S., Iliescu, D. D., & Bhuyan, M. 2008. Neural network based electronic nose for classification of tea aroma. *Univ. Warwick Institutional Repos*, 2(1): 7-14.
- [63] Perera, C. H. Liu, and S. Jayawardena, "The Emerging Internet of Things Marketplace from an Industrial Perspective: A Survey," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 585–598, 2015.

- [64] Puttamadappa and B.D. Parameshachari, "Demand side management of small scale loads in a smart grid using glow-worm swarm optimization technique," *Microprocessors Microsystems*, vol 71, pp. 102886, 2019.
- [65] Puttamadappa and B.D. Parameshachari, "Demand side management of small scale loads in a smart grid using glow-worm swarm optimization technique," *Microprocessors Microsystems*, vol 71, pp. 102886, 2019.
- [66] Simon, K. Davidsen, C. Hansen, E. Seymour, M. B. Barnkob, and L. R. Olsen, "BioReader : a text mining tool for performing classification of biomedical literature," vol. 19, no. Suppl 13, 2019.
- [67] Zhang, Y. Lu, and T. Zang, "CNN - DDI : a learning - based method for predicting drug – drug interactions using convolution neural networks," pp. 1–11, 2022.
- [68] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). *New Avenues in Opinion Mining and Sentiment Analysis*. *IEEE Intelligent Systems*, 2, 15–21. <https://doi.org/10.1109/mis.2013.30>.
- [69] Centers for Disease Control and Prevention. Heart Disease Facts. (<https://www.cdc.gov/heartdisease/facts.htm>) Accessed 9/28/2021.
- [70] Centers for Disease Control and Prevention. Heart Disease. (<https://www.cdc.gov/heartdisease/about.htm>) Accessed 9/28/2021.
- [71] Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. *IEEE Access*, 90225–90265. <https://doi.org/10.1109/access.2020.2992341>.
- [72] Chan M., Esteve D., Fourniols J.Y., Escriba C., Campo E. Smart wearable systems: Current status and future challenges. *Artif. Intell. Med.* 2012;56:137–156
- [73] Chen, B.;Wan, J. Emerging trends of ml-based intelligent services for industrial internet of things (iiot). In *Proceedings of the 2019 Computing, Communications and IoT Applications (ComComAp)*, Shenzhen, China, 26–28 October 2019; pp. 135–139.
- [74] Chengdu University of Technology. Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia, "An ensemble forecasting method for the aggregated load with subprofiles," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3906–3908, Feb. 2018.
- [75] Chiu, S. W., & Tang, K. T. 2013. Towards a chemiresistive sensor-integrated electronic nose: a review. *Sensors*, 13(10): 14214-14247.
- [76] Cho L, Davis M, Elgendy I, et al. Summary of Updated Recommendations for Primary Prevention of Cardiovascular Disease in Women *JACC State-of-the-Art Review*. (<https://www.jacc.org/doi/abs/10.1016/j.jacc.2020.03.060>) *J Am Coll Cardiol* 2020 May, 75(20):2602-2618. Accessed 9/28/2021.
- [77] Clark A, Jit M, Warren-Gash C, Guthrie B, Wang HHX, Mercer SW, Sanderson C, McKee M, Troeger C, Ong KL, Checchi F, Perel P, Joseph S, Gibbs HP, Banerjee A, Eggo RM., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob Health*. 2020 Aug;8(8):e1003-e1017
- [78] Clinical and immunological features of severe and moderate coronavirus disease 2019. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, Wang T, Zhang X, Chen H, Yu H, Zhang X, Zhang M, Wu S, Song J, Chen T, Han M, Li S, Luo X, Zhao J, Ning Q *J Clin Invest*. 2020 May 1; 130(5):2620-2629.
- [79] Connelly, L. 2020. Logistic regression. *Medsurg Nursing*; Pitman, 29(5): 353-354.
- [80] Coronavirus Occurrence and Transmission Over 8 Years in the HIVE Cohort of Households in Michigan.Monto AS, DeJonge PM, Callear AP, Bazzi LA, Capriola SB, Malosh RE, Martin ET, Petrie JG , *J Infect Dis*. 2020;222(1):9
- [81] Cotfas, L.-A., Delcea, C., Roxin, I., Ioanas, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month Following the First Vaccine Announcement. *IEEE Access*, 33203–33223. <https://doi.org/10.1109/access.2021.3059821>.
- [82] Cremer, F., Sheehan, B., Fortmann, M. et al. Cyber risk and cybersecurity: a systematic review of data availability. *Geneva Pap Risk Insur Issues Pract* 47, 698–736 (2022).
- [83] Cruz, J.A. and Wishart, D.S., 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, p.117693510600200030.
- [84] Cui, S., Inocente, E. A. A., Acosta, N., Keener, H. M., Zhu, H., & Ling, P.P. 2019. Development of fast e-nose system for early-stage diagnosis of aphid-stressed tomato plants. *Sensors*, 19(3480): 1-14.

- [85] D S. Jambekar, S. Nema and Z. Saquib, "Prediction of Crop Production in India Using Data Mining Techniques," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5. doi: 10.1109/ICCUBEA.2018.8697446
- [86] Griffith and A. S. Holehouse, "PARROT is a flexible recurrent neural network framework for analysis of large protein datasets," pp. 1–17, 2021.
- [87] Gupta, A. Julka, S. Jain, T. Aggarwal, A. Khanna, N. Arunkumar, and N.V.C. De Albuquerque,"Optimized cuttlefish algorithm for diagnosis of Parkinson's disease,"Cognition System Research, 52: 36e48, 2018, doi : 10.11648/j.sjac.20190704.12.
- [88] Gupta, A. Julka, S. Jain, T. Aggarwal, A. Khanna, N. Arunkumar, and N.V.C. De Albuquerque,"Optimized cuttlefish algorithm for diagnosis of Parkinson's disease,"Cognition System Research, 52: 36e48, 2018, doi : 10.11648/j.sjac.20190704.12.
- [89] Li, Z. Fu, and Y. & Duan," Fish-Expert: a web-based expert system for fish disease diagnosis," Expert System Application, vol. 23, no. 3, pp. 311–320. 2022.
- [90] Li, Z. Fu, and Y. & Duan," Fish-Expert: a web-based expert system for fish disease diagnosis," Expert System Application, vol. 23, no. 3, pp. 311–320. 2022.
- [91] Novaliendry, and C.H.Y. Yang," The expert system application for diagnosing human vitamin deficiency through forward chaining method," Inf. and Comm. Tech. Conv. (ICTC), pp. 53-58. 2015. DOI: 10.1109/ICTC.2015.7354493.
- [92] D. Novaliendry, and C.H.Y. Yang," The expert system application for diagnosing human vitamin deficiency through forward chaining method," Inf. and Comm. Tech. Conv. (ICTC), pp. 53-58. 2015. DOI: 10.1109/ICTC.2015.7354493.
- [93] D. Sharma and A. Sai Sabitha, "Identification of Influential Factors for Productivity and Sustainability of Crops Using Data Mining Techniques," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 322-328.doi: 10.1109/SPIN.2019.8711630
- [94] D.Klauser "Challenges in monitoring and managing plant diseases in developing countries," J Plant Dis Prot, Vol. 125, No.3, pp. 235-237, Jan. 2018, .doi :/10.1007/s41348-018-0145-9.
- [95] D.Klauser "Challenges in monitoring and managing plant diseases in developing countries," J Plant Dis Prot, Vol. 125, No.3, pp. 235-237, Jan. 2018, .doi :/10.1007/s41348-018-0145-9.
- [96] D.L. Vu, T.K. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, and P.H. Phung, "HIT4Mal: hybrid image transformation for malware classification," Transportation Emerging Telecommunication Technology, vol. 31, no. 11, pp. e3789, 2020.
- [97] D.L. Vu, T.K. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, and P.H. Phung, "HIT4Mal: hybrid image transformation for malware classification," Transportation Emerging Telecommunication Technology, vol. 31, no. 11, pp. e3789, 2020.
- [98] Dai XiGuo. 2017. "Research on human attitude recognition based on Convolutional neuralNetwork"
- [99] Dakshayini Patil et al (2017),"Rice Crop Yield Prediction using Data Mining Techniques: An Overview", International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 7, Issue 5.
- [100] Daniel, S. P., Ferri, C., & Ramirez, M. J. 2017. Improving performance of multiclass classification by inducing class hierarchies. *Procedia Computer Science*, 108: 1692-1701.
- [101] dataset are: <https://www.kaggle.com/srinivas1/agriculture-crops-production-in-india> for crop yield data.
- [102] dataset are: <https://en.tutiempo.net/> for weather data
- [103] David," Sistem pakar diagnosa penyakit ikan lele dumbo. konferensi nasional sistem & informatika," STMIK STIKOM Bali, 9 – 10 Oktober. 2015, pp. 107-112.
- [104] David," Sistem pakar diagnosa penyakit ikan lele dumbo. konferensi nasional sistem & informatika," STMIK STIKOM Bali, 9 – 10 Oktober. 2015, pp. 107-112.
- [105] Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. 2016. Efficient kNN classification algorithm for big data. *Neurocomputing*, 195(C):143-148.
- [106] Dennis M. Dimiduk, Elizabeth A. Holm & Stephen R. Niezgod Perspectives on the Impact of Machine Learning, Deep Learning, and Artificial
- [107] Disha Garg , Samiya Khan, and Mansaf Alam, "Integrative Use of IoT and Deep Learning for Agricultural Applications", Springer, pp. 521–531,2020.
- [108] Dogra, Ajay Kumar, and Jagdeep Kaur. "Moving towards smart transportation with machine learning and Internet of Things (IoT): A review." *Journal of Smart Environments and Green Computing* 2, no. 1 (2022): 3-18.

- [109] Dr.V.Sivakumar, Bakkachenna Ranadeep, Swathi, "IOT enabled Agriculture in Smart Drip Irrigation System" in Grenze International Journal of Engineering and Technology (GIJET), ISSN: 2395-5295, 2022 January, Volume no: 8, Issue No: 1, Page No: 581-586 URL: <http://thegrenze.com/index.php?display=page&view=journalabstract&absid=1082&id=8> OR <http://thegrenze.com/pages/servej.php?fn=70.pdf&name=IOT%20Enabled%20Agriculture%20in%20Smart%20Drip%20IrrigationSystem&id=1082&association=GRENZE&journal=GIJET&year=2022&volume=8&issue=1>
- [110] D. Alalade, "Intrusion Detection System in Smart Home Network Using Artificial Immune System and Extreme Learning Machine Hybrid Approach," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-2.
- [111] Elbasani, S. N. Njimbouom, T. J. Oh, E. H. Kim, H. Lee, and J. D. Kim, "GCRNN : graph convolutional recurrent neural network for compound – protein interaction prediction," pp. 1–13, 2021.
- [112] Gavin," Discusses about machine learning: an introduction," 2018. [Online]. Available: <https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0>.
- [113] Gavin," Discusses about machine learning: an introduction," 2018. [Online]. Available: <https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0>.
- [114] E.F. DeLong, G.S. Wickham, and N.R. Pace. "Phylogenetic stains: Ribosomal RNA-based probes for the identification of single cells," *Science*, Vol. 243, No. 4896, pp. 1360–1363, Mar. 1989, doi: 10.1126/science.2466341.
- [115] E.F. DeLong, G.S. Wickham, and N.R. Pace. "Phylogenetic stains: Ribosomal RNA-based probes for the identification of single cells," *Science*, Vol. 243, No. 4896, pp. 1360–1363, Mar. 1989, doi: 10.1126/science.2466341.
- [116] Elfani and A. Pujiyanta," Sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website. sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website," vol. 1, no. 1, pp. 42–50. 2013.
- [117] Elfani and A. Pujiyanta," Sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website. sistem pakar mendiagnosa penyakit pada ikan konsumsi air tawar berbasis website," vol. 1, no. 1, pp. 42–50. 2013.
- [118] en.wikipedia.org, Internet Source
- [119] enam.uac.bj, Internet Source
- [120] Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, OLeary DH, Psaty B, Rautaharju P,
- [121] Epidemiology of Seasonal Coronaviruses: Establishing the Context for the Emergence of Coronavirus Disease 2019.Nickbakhsh S, Ho A, Marques DFP, McMenamin J, Gunson RN, Murcia PR , *J Infect Dis.* 2020;222(1):17.
- [122] Erik Wiener, Jan O. Pedersen, & Andreas S. Weigend. (1995). *A Neural Network Approach to Topic Spotting.*
- [123] Exploratory Data Analysis on Heart Disease UCI data set - A complete step-by-step exploratory data analysis with a simple explanation by Aishah Ismail
- [124] Martinelli, et al., "Advanced methods of plant disease detection. A review," *Agron Sustain Dev*, Vol. 35, pp. 1-25, Sep.2014, doi: 10.1007/s13593-014-0246-1.
- [125] Martinelli, et al., "Advanced methods of plant disease detection. A review," *Agron Sustain Dev*, Vol. 35, pp. 1-25, Sep.2014, doi: 10.1007/s13593-014-0246-1.
- [126] Pazos, Y. Santos, A.R. Macías, S. Núñez, and A.E. Toranzo, "Evaluation of media for the successful culture of *Flexibacter maritimus*," *J. of Fish Dis.*, vol.19, pp. 193-197. 1996.
- [127] Pazos, Y. Santos, A.R. Macías, S. Núñez, and A.E. Toranzo, "Evaluation of media for the successful culture of *Flexibacter maritimus*," *J. of Fish Dis.*, vol.19, pp. 193-197. 1996.
- [128] F. Z. Belgrana, N. Benamrane, M. A. Hamaida, A. Mohamed Chaabani and A. Taleb-Ahmed, "Network Intrusion Detection System Using Neural Network and Condensed Nearest Neighbors with Selection of NSL-KDD Influencing Features," 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), 2021, pp. 23-29.
- [129] Fan, G., Yang, Z., Lin, Q., Zhao, S., Yang, L., & He, D. (2020). Decreased Case Fatality Rate of COVID-19 in the Second Wave: A study in 53 countries or regions. *Transboundary and Emerging Diseases*, 2, 213–215. <https://doi.org/10.1111/tbed.13819>.
- [130] FarhanaParvin, SkAjim Ali, S. Najmul Islam Hashmi, Ateeque Ahmad, Spatial prediction and mapping of the COVID-19 hotspot in India using geo-statistical technique, *Korean Spatial Information Society* 2021.

- [131] Felix Nikolaus Wirth, Marco Johns, Thierry Meurers, Fabian Prasser, “Citizen-Centered Mobile Health Apps Collecting Individual-Level Spatial Data for Infectious Disease Management: Scoping Review”, *JMIR MHEALTH AND UHEALTH*, 2020.
- [132] Fox, M. and Vaidyanathan, G., 2016. Impacts of healthcare Big Data: a Framework with Legal and Ethical insights. *Issues in Information Systems*, 17(3).
- [133] Frank A, Asuncion A (2010) UCI machine learning repository. <http://archive.ics.uci.edu/ml>. University of California, Irvine, CA. School of Information and Computer Science, 213.
- [134] Fumo, Nelson., & Biswas, Rafe, M.A.(2015). Regression analysis for prediction of residential energy consumption. *Elsevier Renewable and Sustainable Energy Reviews*, 7(47), 332-343.
- [135] Èerne, D. Dovžan, and I. Škrjanc, “Short-term load forecasting by separating daily profiles and using a single fuzzy model across the entire domain,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 9, pp. 7406–7415, Sep. 2018.
- [136] Engin, B. Aksoyer, M. Avdagic, D. Bozanli, U. Hanay, d. Maden, and G. Ertek, “Rule-based expert systems for supporting university students,” *Proc. Com. Sci.*, vol. 31, pp. 22-31. 2014. DOI: 10.1016/j.procs.2014.05.241.
- [137] Engin, B. Aksoyer, M. Avdagic, D. Bozanli, U. Hanay, d. Maden, and G. Ertek, “Rule-based expert systems for supporting university students,” *Proc. Com. Sci.*, vol. 31, pp. 22-31. 2014. DOI: 10.1016/j.procs.2014.05.241.
- [138] Market, “Genomics Market by Product & Service (System & Software, Consumables, Services), Technology (Sequencing, PCR), Application (Drug Discovery & Development, Diagnostic, Agriculture), End User (Hospital & Clinics, Research Centers) – Global Forecast to 2025.” <https://www.marketsandmarkets.com/Market-Reports/genomics-market-613.html>.
- [139] G. Wang, Y. Sun, and J. Wang, “Automatic image-based plant disease severity estimation using deep learning”. *Comput. Intell. Neurosci.*, Vol.2017, pp. 1–8, Jul. 2017, doi : 10.1155/2017/2917536.
- [140] G. Wang, Y. Sun, and J. Wang, “Automatic image-based plant disease severity estimation using deep learning”. *Comput. Intell. Neurosci.*, Vol.2017, pp. 1–8, Jul. 2017, doi : 10.1155/2017/2917536.
- [141] Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN, Strack B, DeShazo JP (2014) Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Res Int*. <https://doi.org/10.1155/2014/781670>
- [142] Ghamdi HA, Alshammari R, Razzak MI (2016) An ontologybased system to predict hospital readmission within 30 days. *Int J Healthc Manag* 9(4):236–244
- [143] Ghosh, S., Singh, A., K., Jhanjhi, N. Z., Masud, M., & Aljahdali, S. 2022. SVM and KNN based CNN architectures for plant classification. *Computers, Materials & Continua*, 71(3): 4257–4274.
- [144] Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M (2012) Predicting seminal quality with artificial intelligence methods. *Expert Syst Appl* 39(16):12564–12573.
- [145] Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) (pp. 257-261).
- [146] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ch IP, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220.
- [147] Guo Zhe, Chen Peitou, Hu Mengkai, et al. 2016. “Kinect-based Smart Home System”, *Modern Electronic Technology*, 2016, 39 (18): 149-152.
- [148] Chakravorty, P. Rituraj P., and P. Das, “Image Processing Technique to Detect Fish Disease,” *Intl. J. of Com. Sci. and Sec. (IJCSS)*, vol. 9, no. 2, pp. 121-131. 2015.
- [149] Chakravorty, P. Rituraj P., and P. Das, “Image Processing Technique to Detect Fish Disease,” *Intl. J. of Com. Sci. and Sec. (IJCSS)*, vol. 9, no. 2, pp. 121-131. 2015.
- [150] S. Basavegowda and G. Dagnev, “Deep learning approach for microarray cancer data classification,” vol. 5, pp. 22–33, 2020, doi: 10.1049/trit.2019.0028.
- [151] H.T. Sihotang, “Sistem pakar untuk mendiagnosa penyakit pada tanaman jagung dengan metode bayes,” *Journal of Informatic Pelita Nusantara*, Vol. 3, No. 1, pp. 17-22, 2018.
- [152] H.T. Sihotang, “Sistem pakar untuk mendiagnosa penyakit pada tanaman jagung dengan metode bayes,” *Journal of Informatic Pelita Nusantara*, Vol. 3, No. 1, pp. 17-22, 2018.
- [153] H.Tyagi, S. Rajasubramaniam, M.V. Rajam, and I. Dasgupta, “RNA-interference in rice against Rice tungro bacilliform virus results in its decreased accumulation in inoculated rice plants”. *Transgenic Res.*, Vol. 17, No.5, pp.897-904, Feb. 2008, doi: 10.1007/s11248-008-9174-7.



- [154] H.Tyagi, S. Rajasubramaniam, M.V. Rajam, and I. Dasgupta, "RNA-interference in rice against Rice tungro bacilliform virus results in its decreased accumulation in inoculated rice plants". *Transgenic Res.*, Vol. 17, No.5, pp.897-904, Feb. 2008, doi: 10.1007/s11248-008-9174-7.
- [155] Hamsagayathri, P. and Vigneshwaran, S., 2021, February. Symptoms Based Disease Prediction Using Machine Learning Techniques. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 747-752). IEEE.
- [156] Hao, Hu., Rongxing, Lu., Zonghua, Zhang. (2015, December). Vtrust: A robust trust framework for relay selection in hybrid vehicular communications, IEEE Global Communications Conference, GLOBECOM 2015, San Diego, CA, USA.
- [157] Hari Shankar Gangwar, P.K. Champati Ray, Geographic information system-based analysis of COVID-19 cases in India during pre-lockdown, lockdown, and unlock phases, *International Journal of Infectious Diseases*, 2021.
- [158] Harvey, B. S., & Flores-Sarnat, L. 2019. Development of the human olfactory system. *Handbook of Clinical Neurology*, 164: 29-45, 2019.
- [159] Haryono, Anam, K., & Saleh, A. 2020. Autentikasi daun herbal menggunakan convolutional neural network dan raspberry pi. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 9(3): 278 – 286.
- [160] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. In 2014 47th Hawaii International Conference on System Sciences (pp. 1833-1842).
- [161] Hlaing, Win., Thepphaeng, Somchai., Nontaboot, Varunyoo., Tangsun, Natthan., Sangsuwan, Tanayoot., Chaiyod, Pira. (2017, March). Implementation of WiFi-based single phase smart meter for Internet of Things (IoT), International Electrical Engineering Congress (iEECON), Pattaya, Thailand
- [162] Hou Yuyi, Yang Dongtao, Liu Yan, et al. 2016. "Smart Home Life and Security System Based on Wireless Bluetooth Technology". *Journal of Jiaying University*, 2016, 34 (5): 36-40.
- [163] Houda Ahmad, Shokoh Kermanshahani, Ana Simonet & Michel Simonet, Data Warehouse Based Approach to the Integration of Semi-structured Data, Springer nature, Lecture Notes in Computer Science book series (LNISA, volume 5731), 2020
- [164] <https://mldoodles.com/statistical-data-types-used-in-machine-learning>
- [165] <https://www.businesswire.com/news/home/20190516005700/en/Strategy-Analytics-Internet-of-Things-Now-Numbers-22-Billion-Devices-But-Where-Is-The-Revenue>.
- [166] <https://www.geeksforgeeks.org/what-is-semi-structured-data>
- [167] <https://www.ibm.com/cloud/blog/structured-vs-unstructured-data>
- [168] Huang, S., Cai, N., Pedro, P.P, Narrandes, S., Wang, Y., & Xu, W. 2018. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics & Proteomics*, 15(1): 41-51.
- [169] Human aminopeptidase N is a receptor for human coronavirus 229E. Yeager CL, Ashmun RA, Williams RK, Cardellicchio CB, Shapiro LH, Look AT, Holmes KV ,*Nature*. 1992;357(6377):420
- [170] Human and bovine coronaviruses recognize sialic acid-containing receptors similar to those of influenza C viruses. Vlasak R, Luytjes W, Spaan W, Palese P , *ProcNatAcadSci U S A*. 1988;85(12):4526.
- [171] Human Coronavirus in Hospitalized Children with Respiratory Tract Infections: A 9-Year Population-Based Study from Norway. Heimdal I, Moe N, Krokstad S, Christensen A, Skanke LH, Nordbø SA, Døllner H , *J Infect Dis*. 2019;219(8):1198.
- [172] Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. Hofmann H, Pyrc K, van der Hoek L, Geier M, Berkhout B, Pöhlmann S, *ProcNatAcadSci U S A*. 2005;102(22):7988. Epub 2005 May 16.
- [173] Hwang I., Kim H.N., Seon G.M., Lee S.H., Kang M., Yi H., Bae W.G., Kwak M.K., Jeong H.E. Multifunctional smart skin adhesive patches for advanced health care. *Adv. Healthc. Mater*. 2018;7:1800275. doi: 10.1002/adhm.201800275.
- [174] Akil,"Analisa efektifitas metode forward chaining dan backward chaining pada sistem pakar," *J. Pilar Nusa Man.*, p. 13. 2017.
- [175] Akil,"Analisa efektifitas metode forward chaining dan backward chaining pada sistem pakar," *J. Pilar Nusa Man.*, p. 13. 2017.
- [176] Altinok and I. Kurt," Molecular Diagnosis of Fish Diseases: a Review," *Turkish J. of Fish. and Aquat. Sci.*, vol. 3, pp. 131-138. 2003.
- [177] Altinok and I. Kurt," Molecular Diagnosis of Fish Diseases: a Review," *Turkish J. of Fish. and Aquat. Sci.*, vol. 3, pp. 131-138. 2003.

- [178] M. Shofi, L.K. Wardhani, and G. Anisa, "Android Application for Diagnosing General Symptoms of Disease Using Forward Chaining Method," *Cyber and IT Service Management*, Bandung, Indonesia, 25-27 April. 2016. DOI: 10.1109/CITSM.2016.7577588.
- [179] M. Shofi, L.K. Wardhani, and G. Anisa, "Android Application for Diagnosing General Symptoms of Disease Using Forward Chaining Method," *Cyber and IT Service Management*, Bandung, Indonesia, 25-27 April. 2016. DOI: 10.1109/CITSM.2016.7577588.
- [180] S. Fotiou, P.G. Pappi, K.E. Efthimiou, N.I. Katis, and V.I. Maliogka, "Development of one-tube real-time RT-qPCR for the universal detection and quantification of Plum pox virus (PPV)," *J Virol. Methods*, Vol.263, pp. 10–13, Oct.2018, doi : 10.1016/j.viromet.2018.10.006.
- [181] S. Fotiou, P.G. Pappi, K.E. Efthimiou, N.I. Katis, and V.I. Maliogka, "Development of one-tube real-time RT-qPCR for the universal detection and quantification of Plum pox virus (PPV)," *J Virol. Methods*, Vol.263, pp. 10–13, Oct.2018, doi : 10.1016/j.viromet.2018.10.006.
- [182] I.S. Dewi, I.H. Somantri, D. Damayanti, A. Apriana and T.J. Santoso. Evaluasi tanaman padi transgenik Balitbio terhadap hama penggerek batang. Laporan Hasil Penelitian Balitbio, Bogor, pp. 141 – 150, Nov. 2002, <http://repository.pertanian.go.id/handle/123456789/12199>.
- [183] I.S. Dewi, I.H. Somantri, D. Damayanti, A. Apriana and T.J. Santoso. Evaluasi tanaman padi transgenik Balitbio terhadap hama penggerek batang. Laporan Hasil Penelitian Balitbio, Bogor, pp. 141 – 150, Nov. 2002, <http://repository.pertanian.go.id/handle/123456789/12199>.
- [184] *Intelligence on Materials, Processes, and Structures Engineering*, Springer Nature, Integrating Materials and Manufacturing Innovation volume 7, pages157–172 (2018)
- [185] Ivan Franch-Pardo, Brian M. Napoletano, Fernando Rosete-Verges, "Spatial analysis and GIS in the study of COVID-19. A review", *Science of the Total Environment*, 2020.
- [186] Chen, Q. Liu, and L. Gao, L, "Visual tea leaf disease recognition using a convolutional neural network model," *Symmetry*, Vol. 11, No.3, p. 343, Mar.2019, doi : 10.3390/sym11030343.
- [187] Chen, Q. Liu, and L. Gao, L, "Visual tea leaf disease recognition using a convolutional neural network model," *Symmetry*, Vol. 11, No.3, p. 343, Mar.2019, doi : 10.3390/sym11030343.
- [188] R, H. D and P. B, "A Machine Learning-based Approach for Crop Yield Prediction and Fertilizer Recommendation," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1330-1334.doi: 10.1109/ICOEI53556.2022.9777230
- [189] Schuster. "Big data ethics and the digital age of agriculture," *Resource Magazine*, Vol.24, No.1, pp. 20-21, 2017.
- [190] Schuster. "Big data ethics and the digital age of agriculture," *Resource Magazine*, Vol.24, No.1, pp. 20-21, 2017.
- [191] Yang, N. Li, S. Fang, K. Yu, and Y. Chen, "Semantic Features Prediction for Pulmonary Nodule Diagnosis Based on Online Streaming Feature Selection," *IEEE Access*, vol. 7, pp. 61121–61135, 2019, doi: 10.1109/ACCESS.2019.2903682.
- [192] J.A. Plumb," Health maintenance and principle microbial diseases of cultured fishes," Iowa State University Press. Ames, Iowa. 344 pp. 1999.
- [193] J.A. Plumb," Health maintenance and principle microbial diseases of cultured fishes," Iowa State University Press. Ames, Iowa. 344 pp. 1999.
- [194] J.A.Tomlinson, et al. "On-site DNA extraction and real-time PCR for detection of *Phytophthora ramorum* in the field," *Appl. Environ. Microbiol.*, Vol. 71, pp. 6702–6710, Nov. 2005, doi : 10.1128/AEM.71.11.6702-6710.2005.
- [195] J.A.Tomlinson, et al. "On-site DNA extraction and real-time PCR for detection of *Phytophthora ramorum* in the field," *Appl. Environ. Microbiol.*, Vol. 71, pp. 6702–6710, Nov. 2005, doi : 10.1128/AEM.71.11.6702-6710.2005.
- [196] J.F. Bernardetn, A.C. Campbell, J.A. Buswell,"*Flexibacter maritimus* is the agent of 'black patch necrosis' in Dover sole in Scotland," *Dis. in Aquat. Org.*, vol. 8, pp. 233-237. 1990.
- [197] J.F. Bernardetn, A.C. Campbell, J.A. Buswell,"*Flexibacter maritimus* is the agent of 'black patch necrosis' in Dover sole in Scotland," *Dis. in Aquat. Org.*, vol. 8, pp. 233-237. 1990.
- [198] J.M. Bertolini and J.S. Rohovec,"Electrophoretic detection of proteases from different *Flavobacterium columnare* strains and assessment of their variability,"*Dis. in Aquat. Org.*, vol. 12, pp. 121-128. 1992.
- [199] J.M. Bertolini and J.S. Rohovec,"Electrophoretic detection of proteases from different *Flavobacterium columnare* strains and assessment of their variability,"*Dis. in Aquat. Org.*, vol. 12, pp. 121-128. 1992.
- [200] J.M. Shewan and T.A. McMeekin,"Taxonomy and ecology of the *Flavobacterium* and related genera," *Ann. Rev. in Micr.* vol. 37, pp. 233-252. 1983

- [201] J.M. Shewan and T.A. McMeekin, "Taxonomy and ecology of the Flavobacterium and related genera," *Ann. Rev. in Micr.* vol. 37, pp. 233-252. 1983
- [202] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 8, 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- [203] Jain, B.; Brar, G.; Malhotra, J.; Rani, S.; Ahmed, S.H. A cross layer protocol for traffic management in Social Internet of Vehicles. *Future Gen. Comput. Syst.* 2018, 82, 707–714.
- [204] Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733-2742.
- [205] Jharna Majumdar, Sneha Naraseyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data", Springer journal, 2017.
- [206] Jia, W., Liang, G., Jiang, Z., & Jihua, W. 2019. Advances in electronic nose development for application to agricultural products. *Food Analytical Methods*, 12: 2226–2240.
- [207] Jig Han Jeong et al., "Random Forests for Global and Regional Crop Yield Predictions", *PLOS-ONE*, June 2016.
- [208] Jongeling, R., Datta, S., & Serebrenik, A. (2015). Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 531-535).
- [209] Muhammad, S. Khan, V. Palade, I. Mehmood, and V. H. C. De Albuquerque, "Edge intelligence-assisted smoke detection in foggy surveillance environments," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1067–1075, Feb. 2020.
- [210] Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [211] Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.
- [212] Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Trans. Intelligence & Transportation Systems*, vol. 22, no. 7, pp. 4337–4347, 2020.
- [213] Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Trans. Intelligence & Transportation Systems*, vol. 22, no. 7, pp. 4337–4347, 2020.
- [214] K.E. Eswari. L.Vinitha. (2018) "Crop Yield Prediction in Tamil Nadu Using Bayesian Network ", *International Journal of Intellectual Advancements and Research in Engineering Computations*, Vol-6, Issue-2, ISSN: 23482079.
- [215] K.P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput Electron Agric*, Vol. 145, pp. 311-318, Feb. 2018, doi: 10.1016/j.compag. 2018.01.009.
- [216] K.P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput Electron Agric*, Vol. 145, pp. 311-318, Feb. 2018, doi: 10.1016/j.compag. 2018.01.009.
- [217] K.Thenmozhi, And U.S. Reddy, "Crop pest classification based on deep convolutional neural network and transfer learning," *Comput. Electron. Agric*, Vol.164, p.104906, Aug. 2019, doi : 10.1016/j.compag.2019.104906.
- [218] K.Thenmozhi, And U.S. Reddy, "Crop pest classification based on deep convolutional neural network and transfer learning," *Comput. Electron. Agric*, Vol.164, p.104906, Aug. 2019, doi : 10.1016/j.compag.2019.104906.
- [219] K.Thongboonnak, and S. Sarapirome, "Integration of Artificial Neural Network And Geographic Information System For Agricultural Yield Prediction," *Suranaree J. Sci.Technol*, Vol.18, No.1, pp. 71-80, Jan, 2011.
- [220] K.Thongboonnak, and S. Sarapirome, "Integration of Artificial Neural Network And Geographic Information System For Agricultural Yield Prediction," *Suranaree J. Sci.Technol*, Vol.18, No.1, pp. 71-80, Jan, 2011.
- [221] Kamisalic A., Fister I., Jr., Turkanovic M., Karakatic S. Sensors and functionalities of non-invasive wrist-wearable devices: A review. *Sensors*. 2018;18:1714. doi: 10.3390/s18061714.
- [222] Kan, H. X., Jin, L., & Zhou, F. L. 2017. Classification of medical plant leaf image based on multi-feature extraction. *Pattern recognition and analysis*, 27(3): 581-587.
- [223] Kanchanamala, P., Das, S. and Neelima, G., 2022. Symptoms-Based Disease Prediction Using Big data Analytics. In *Innovations in Computer Science and Engineering* (pp. 339-346). Springer, Singapore.

- [224] Kaur, P., Robin, Mehta, R.G., Balbir, S., & Arora, S. 2019. Development of aqueous-based multi-herbal combination using principal component analysis and its functional significance in HepG2 cells. *BMC Complement Alternative Medicine*, 19(18):1-17.
- [225] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning, and data mining methods in diabetes research. *Computational and structural biotechnology journal*. 2017; 104-116.
- [226] Kepski M, Kwolek B. 2013. "Human Fall Detection Using Kinect Sensor[J]". 2013, 226:743-752.
- [227] Khalil, Ruhul Amin, Nasir Saeed, Mudassir Masood, Yasaman Moradi Fard, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri. "Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications." *IEEE Internet of Things Journal* 8, no. 14 (2021): 11016-11040.
- [228] Khan Y., Ostfeld A.E., Lochner C.M., Pierre A., Arias A.C. Monitoring of vital signs with flexible and wearable medical devices. *Adv. Mater.* 2016;28:4373–4395. doi: 10.1002/adma.201504366.
- [229] Kirtan Jha, Aalap Doshi and Poojan Patel, "Intelligent Irrigation System Using Artificial Intelligence And Machine Learning: A Comprehensive Review", Vol. 6, Issue 10, pp. 1493-1502 , 2018.
- [230] Klemenjak C, Egarter D, Elmenreich W. 2016. "YoMo: the Arduino-based smart meteringboard[J]". *Computer Science - Research and Development*, 2016, 31(1-2):97-103.
- [231] Kodinariya, Trupti & Makwana, Prashant. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*. 1. 90-95.
- [232] Kotenko, I. Saenko, O. Lauta and M. Karpov, "Situational Control of a Computer Network Security System in Conditions of Cyber Attacks," 2021 14th International Conference on Security of Information and Networks (SIN), 2021, pp. 1-8.
- [233] Krauss R. 2016. "Combining Raspberry Pi and Arduino to form a low-cost, real-time autonomous vehicle platform[C]", *American Control Conference*. IEEE, 2016:6628-6633.
- [234] Nie et al., "Intrusion Detection for Secure Social Internet of Things Based on Collaborative Edge Computing: A Generative Adversarial Network-Based Approach," in *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 134-145, Feb. 2022.
- [235] Owens," Diseases," in *Aquaculture. Farming Aquatic Animals and Plants*. J.S. Lucas, J.S. and P.C. Southgate, Eds., Blackwell Publishing. 2015, pp. 199-214.
- [236] Owens," Diseases," in *Aquaculture. Farming Aquatic Animals and Plants*. J.S. Lucas, J.S. and P.C. Southgate, Eds., Blackwell Publishing. 2015, pp. 199-214.
- [237] Lanzer, J.D., Leuschner, F., Kramann, R., Levinson, R.T. and Saez-Rodriguez, J., 2020. Big data approaches in heart failure research. *Current Heart Failure Reports*, 17(5), pp.213-224.
- [238] Latif, Shahid, Maha Driss, Wadii Boulila, Zil E. Huma, Sajjad Shaukat Jamal, Zeba Idrees, and Jawad Ahmad. "Deep learning for the industrial internet of things (IIoT): A comprehensive survey of techniques, implementation frameworks, potential applications, and future directions." *Sensors* 21, no. 22 (2021): 7518.
- [239] Lestari," Penerapan Metode Certainty Factor Pada Sistem Pakar Diagnosa Penyakit Ikan Gourami Berbasis Website (Studi kasus UPTD Balai Benih Kota Binjai)," Thesis. Universitas Pembangunan Panca Budi Medan, pp 77. 2019.
- [240] Lestari," Penerapan Metode Certainty Factor Pada Sistem Pakar Diagnosa Penyakit Ikan Gourami Berbasis Website (Studi kasus UPTD Balai Benih Kota Binjai)," Thesis. Universitas Pembangunan Panca Budi Medan, pp 77. 2019.
- [241] Leung, F., Lam, H., Ling, S., & Tam, P. (2003). Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Transactions on Neural Networks*, 14(1), 79-88.
- [242] Li Tao. 2014. "Design and implementation of Android based smart home APP". Suzhou:Soochow University.
- [243] Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., Duan, W., Tsoi, K. K., & Wang, F.-Y. (2020). Characterizing the Propagation of Situational Information in social media During COVID-19 Epidemic: A Case Study on Weibo. *IEEE Transactions on Computational Social Systems*, 2, 556–562. <https://doi.org/10.1109/tcss.2020.2980007>
- [244] Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 2, 451–461. [https://doi.org/10.1016/s0031-3203\(02\)00060-2](https://doi.org/10.1016/s0031-3203(02)00060-2).
- [245] Liu B. (2011) *Opinion Mining and Sentiment Analysis*. In: *Web Data Mining. Data-Centric Systems and Applications*. Springer, Berlin, Heidelberg.
- [246] Liu B., Zhang L. (2012) *A Survey of Opinion Mining and Sentiment Analysis*. In: Aggarwal C., Zhai C. (eds) *Mining Text Data*. Springer, Boston, MA.

- [247] Long, S. W., Olsen, R. J., Christensen, P. A., Bernard, D. W., Davis, et al. (2020). Molecular Architecture of Early Dissemination and Massive Second Wave of the SARS-CoV-2 Virus in a Major Metropolitan Area. *MBio*, 6. <https://doi.org/10.1128/mbio.02707-20>.
- [248] Alagappan and M. Kumaran, "Application of expert systems in fisheries sector – a review," *Res. J. Anim. Vet. Fish. Sci.*, vol. 1, no. 8, pp. 19–30. 2013.
- [249] Alagappan and M. Kumaran, "Application of expert systems in fisheries sector – a review," *Res. J. Anim. Vet. Fish. Sci.*, vol. 1, no. 8, pp. 19–30. 2013.
- [250] Arhami, "Konsep dasar sistem pakar," Penerbit Andi. Yogyakarta. 205 p. 2004.
- [251] Arhami, "Konsep dasar sistem pakar," Penerbit Andi. Yogyakarta. 205 p. 2004.
- [252] Føre, K. Frank, T. Norton, E. Svendsen, J.A. Alfredsen, T. Dempster, H. Eguiraun, W. Watson, A. Stahl, L.M. Sunde, C. Schellewald, K.R. Skøien, M.O. Alver, and D. Berckmans, "Precision fish farming: a new framework to improve production in aquaculture," *Bios. Eng.*, vol. 173, pp. 176–193. 2018.
- [253] M. Føre, K. Frank, T. Norton, E. Svendsen, J.A. Alfredsen, T. Dempster, H. Eguiraun, W. Watson, A. Stahl, L.M. Sunde, C. Schellewald, K.R. Skøien, M.O. Alver, and D. Berckmans, "Precision fish farming: a new framework to improve production in aquaculture," *Bios. Eng.*, vol. 173, pp. 176–193. 2018.
- [254] M. Hijri. "The use of Fluorescent in situ hybridisation in plant fungal identification and genotyping," In *Plant Pathology*, pp. 131-145. Humana Press, Totowa, NJ.
- [255] M. Hijri. "The use of Fluorescent in situ hybridisation in plant fungal identification and genotyping," In *Plant Pathology*, pp. 131-145. Humana Press, Totowa, NJ.
- [256] M. Manjunatha and A. Parkavi, "Estimation of Arecanut Yield in Various Climatic Zones of Karnataka using Data Mining Technique: A Survey," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-4. doi: 10.1109/ICCTCT.2018.8551083
- [257] M. Sharma, A.B. Shrivastav, Y.P. Sahni, Y.P., and G. Pandey, "Overviews of the treatment and control of common fish diseases. International Research," *J. of Pharmacy*, vol. 3, no. 7, pp. 123-127. 2012 [Online] Available: [www.irjponline.com](http://www.irjponline.com).
- [258] M. Sharma, A.B. Shrivastav, Y.P. Sahni, Y.P., and G. Pandey, "Overviews of the treatment and control of common fish diseases. International Research," *J. of Pharmacy*, vol. 3, no. 7, pp. 123-127. 2012 [Online] Available: [www.irjponline.com](http://www.irjponline.com).
- [259] M.F. Chen, D. Henry-Ford, and J.M. Groff, "Isolation and characterization of *Flexibacter maritimus* from marine fishes of California," *J. of Aquat. Anim. Health*, vol. 7, pp. 318- 326. 1995.
- [260] M.F. Chen, D. Henry-Ford, and J.M. Groff, "Isolation and characterization of *Flexibacter maritimus* from marine fishes of California," *J. of Aquat. Anim. Health*, vol. 7, pp. 318- 326. 1995.
- [261] M.F. Clark, and A.N. Adams. "Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses," *J Gen Virol*, Vol. 34, pp. 475–483, Mar.1977, doi : 10.1099/0022-1317-34-3-475.
- [262] M.F. Clark, and A.N. Adams. "Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses," *J Gen Virol*, Vol. 34, pp. 475–483, Mar.1977, doi : 10.1099/0022-1317-34-3-475.
- [263] M.M. López, et al., "Strategies for improving serological and molecular detection of plant pathogenic bacteria," In : De Boer, S.H. (eds) *Plant Pathogenic Bacteria*, Springer, Dordrecht, pp. 83–86, 2001, doi : 10.1007/978-94-010-0003-1\_15.
- [264] M.M. López, et al., "Strategies for improving serological and molecular detection of plant pathogenic bacteria," In : De Boer, S.H. (eds) *Plant Pathogenic Bacteria*, Springer, Dordrecht, pp. 83–86, 2001, doi : 10.1007/978-94-010-0003-1\_15.
- [265] M.M.Faizal Azizi and H.Y. Lau, "Advanced diagnostic approaches developed for the global menace of rice diseases: a review." *Canadian Journal of Plant Pathology*, Vol. 44, No.5, pp 627 – 651, Mar 2022, doi: 10.1080/07060661.2022.2053588.
- [266] M.M.Faizal Azizi and H.Y. Lau, "Advanced diagnostic approaches developed for the global menace of rice diseases: a review." *Canadian Journal of Plant Pathology*, Vol. 44, No.5, pp 627 – 651, Mar 2022, doi: 10.1080/07060661.2022.2053588.
- [267] M.N. Rachmatullah and I. Supriana, "Low Resolution Image Fish Classification Using Convolutional Neural Network 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA) pp 78-83. 2018.

- [268] M.N. Rachmatullah and I. Supriana, "Low Resolution Image Fish Classification Using Convolutional Neural Network 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA) pp 78-83. 2018.
- [269] M.S. Ahmed, T.T. Aurpa, and M.A.K. Azad, "Fish Disease Detection Using Image Based Machine Learning Technique in Aquaculture," *J. of King Saud Univ. – Com. and Inf. Sci.* 2021. doi: <https://doi.org/10.1016/j.jksuci.2021.05.003>.
- [270] M.S. Ahmed, T.T. Aurpa, and M.A.K. Azad, "Fish Disease Detection Using Image Based Machine Learning Technique in Aquaculture," *J. of King Saud Univ. – Com. and Inf. Sci.* 2021. doi: <https://doi.org/10.1016/j.jksuci.2021.05.003>.
- [271] M.Sharif, et al., "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *J Exp. Theor. Artif. Intell*, Vol. 33, No.4, pp.577-599, Feb, 2019, doi : 10.1080/0952813X.2019.1572657.
- [272] M.Sharif, et al., "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *J Exp. Theor. Artif. Intell*, Vol. 33, No.4, pp.577-599, Feb, 2019, doi : 10.1080/0952813X.2019.1572657.
- [273] Manojkumar, P., Surya, C. M., & Varun, P. G. 2017, Identification of ayurvedic medicinal plant by image processing of leaf samples, *Proceeding of International Conference on research in computational intelligence and communication network*, 3-5 November 2017, Kolkata, India: 351-355. USA: IEEE.
- [274] Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 228–233. <https://doi.org/10.1109/34.908974>.
- [275] Medical reviews. Coronaviruses. Monto AS, *Yale J Biol Med.* 1974;47(4):234
- [276] Menaga, S. and Paruvathavardhini, J., 2022. AI in Healthcare. *Smart Systems for Industrial Applications*, pp.115-140.
- [277] Mhadhbi, Zeineb., Zairi, Sajeh., Gueguen, Cedric., Zouari, Belhassen. (2018). Validation of a Distributed Energy Management Approach for Smart Grid Based on a Generic Colored Petri Nets Model, *Journal of Clean Energy Technologies*, 6(1), 20-25.
- [278] Mohamad Yusof, U. K. 2015. Development of electronic nose for herbs recognition based on artificial intelligent techniques. Unpublished Master Thesis, Faculty of Engineering, Universiti Putra Malaysia, Serdang, Selangor, Malaysia.
- [279] Mubarak Albarka Umar, Chen Zhanfang Effects of Feature Selection and Normalization on Network Intrusion Detection, *Communication, Networking and Broadcast Technologies*, 2020, 10.36227/techrxiv.12480425.v2.
- [280] Muhammad Imran Razzak, Muhammad Imran and Guandong Xu "Big data analytics for preventive medicine", *Neural Comput Appl.* 2020; 32(9): 4417–4451. DoI: 10.1007/s00521-019-04095-y
- [281] Muneer, A., & Fati, S.M. 2020. Efficient and automated herbs classification approach based on shape and texture features using deep learning. *IEEE Access*, 8 :196747-196764.
- [282] Muralitharan, K., Sakthivel, R., Shi, Y. (2015). Multiobjective Optimization Technique for Demand Side Management with Load Balancing Approach in Smart Grid, *Elsevier Neurocomputing*, 177, 110-119.
- [283] Mustafa, M. S., Husin, Z., Tan, W.K., Mavi, M. F., & Farook, R. S. M. 2020. Development of automated hybrid intelligent system for herbs plant classification and early herbs plant disease detection. *Neural Computing and Applications*, 32:11419-11441.
- [284] Gandhi and L. J. Armstrong, "Rice crop yield forecasting of tropical wet and dry climatic zone of India using data mining techniques," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 357-363. doi: 10.1109/ICACA.2016.7887981
- [285] Sapoval et al., "deep learning across the biosciences," pp. 1–12, 2022, doi: 10.1038/s41467-022-29268-7.
- [286] Naglic M, Souvent A. 2013. "Concept of Smart Home and Smart Grids integration". *Energy ,International Youth Conference on IEEE*, 2013:1-5.
- [287] Naresh, Y. G., & Nagendraswamy, H. S. 2016. Classification of medicinal plants: An approach using modified LBP with symbolic representation. *Neurocomputing*, 173: 1789–1797.
- [288] Nelli F. (2018) Textual Data Analysis with NLTK. In: *Python Data Analytics*. Apress, Berkeley, CA.
- [289] Niketa Gandhi et al. (2016), "Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques", *IEEE International Conference on Advances in Computer Applications (ICACA)* .
- [290] O.Azzam, and T.C. Chancellor, "The biology, epidemiology, and management of rice tungro disease in Asia". *Plant Dis.*, Vol. 86, No.2, pp. 88-100, Feb, 2007, doi : 10.1094/PDIS.2002.86.2.88.



- [291] O.Azzam, and T.C. Chancellor, "The biology, epidemiology, and management of rice tungro disease in Asia". *Plant Dis.*, Vol. 86, No.2, pp. 88-100, Feb,2007, doi : 10.1094/PDIS.2002.86.2.88.
- [292] Okafor, K.C., Ononiwu, G.C.,&Precious, U. (2017).Development of Arduino Based IoT Metering System for On-DemandEnergy Monitoring. *International Journal of Mechatronics, Electrical and Computer Technology*, 7(23), 3208-3224.
- [293] Ozbayoglu M, Kucukayan G, Dogdu E. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. 2016 IEEE International Conference on Big Data (Big Data); 2016 Dec 5-8; Washington, DC, USA. 2016.p. 1807-13.
- [294] Baldi, and N. La Porta, "Molecular approaches for low-cost point-of-care pathogen detection in agriculture and forestry," *Front Plant Sci*, Vol. 11, p.570862, Oct. 2020, doi : 10.3389/fpls.2020.570862.
- [295] Baldi, and N. La Porta, "Molecular approaches for low-cost point-of-care pathogen detection in agriculture and forestry," *Front Plant Sci*, Vol. 11, p.570862, Oct. 2020, doi : 10.3389/fpls.2020.570862.
- [296] Jayanthi," Machine learning and deep learning algorithms in disease prediction: future trends for the healthcare system,"In *Deep Learning for Medical Application with Unique Data*, pp. 123-152. 2022.
- [297] Jayanthi," Machine learning and deep learning algorithms in disease prediction: future trends for the healthcare system,"In *Deep Learning for Medical Application with Unique Data*, pp. 123-152. 2022.
- [298] Supriya, B. Marudamuthu, S. K. Soam, and C. S. Rao, "Trends and Application of Data Science in Bioinformatics BT - Trends of Data Science and Applications: Theory and Practices," S. S. Rautaray, P. Pemmaraju, and H. Mohanty, Eds. Singapore: Springer Singapore, 2021, pp. 227–244.
- [299] Thareja and R. S. Chhillar, "A Detailed Survey on Data Mining based Optimization Schemes for Bioinformatics Applications," 2021.
- [300] Thareja and R. S. Chhillar, "A review of data mining optimization techniques for bioinformatics applications," *Int. J. Eng. Trends Technol.*, vol. 68, no. 10, pp. 58–62, 2020, doi: 10.14445/22315381/IJETT-V68I10P210.
- [301] Thareja and R. S. Chhillar, "Comparative Analysis of Data Mining Algorithms for Cancer Gene Expression Data," vol. 12, no. 10, pp. 322–328, 2021, doi: <http://dx.doi.org/10.14569/IJACSA.2021.0121035>.
- [302] Wang, G. Zhang, Z. G. Yu, and G. Huang, "A Deep Learning and XGBoost-Based Method for Predicting Protein-Protein Interaction Sites," *Front. Genet.*, vol. 12, no. October, pp. 1–11, 2021, doi: 10.3389/fgene.2021.752732.
- [303] Zheng, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarak,S. Yu, X. Xu et al., "Smart Manufacturing Systems for Industry 4.0:Conceptual Framework, Scenarios, and Future Perspectives," *Front.Mech. Eng.*, vol. 13, no. 2, pp. 137–150, 2018.
- [304] P. Zhuang and H. Liang, "Hierarchical and decentralized stochastic energy management for smart distribution systems with high BESS penetration," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6516–6527, Nov. 2019.
- [305] P.I. Hidayati," Penerapan metode cf (certainty factor) pada diagnosa penyakit ikan nila," *Teknologi Informasi*, vol 8, no.2, pp. 127–134. 2017.
- [306] P.I. Hidayati," Penerapan metode cf (certainty factor) pada diagnosa penyakit ikan nila," *Teknologi Informasi*, vol 8, no.2, pp. 127–134. 2017.
- [307] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of LREC*. 10.
- [308] Patel S., Park H., Bonato P., Chan L., Rodgers M. A review of wearable sensors and systems with application in rehabilitation. *J. Neuroeng. Rehabil.* 2012;9:1–17.
- [309] Patil PH, Thube S, Ratnaparkhi B, Rajeswari K. Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining. *International Journal of Computer Applications*. 2014; 93(8):35-39.
- [310] Peng, Yanfei, Jianjun Peng, Jiping Li, and Ling Yu. "Smart home system based on deep learning algorithm." In *Journal of Physics: Conference Series*, vol. 1187, no. 3, p. 032086. IOP Publishing, 2019.
- [311] Policy brief on ageing no. 3, older persons as consumers (2009)
- [312] Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.-L.; Chen, S.C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. (CSUR)* 2018, 51, 1–36. [CrossRef]
- [313] Prabhakar, P., Shyamdew, K., Philip, V. S., Kishore, P., & Roopashree, S. 2016. Robust recognition and classification of herbal leaves. *International Journal of Research in Engineering and Technology*, 6(4):146-149.

- [314] Prasad, S., Kumar, P. S., & Ghosh, D. 2017. An efficient low vision plant leaf shape identification system for smart phones. *Multimedia Tools & Applications*, 76(5): 6915–6939.
- [315] Purnima Bholowalia, & Arvind Kumar (2014). Article: EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9), 17-24.
- [316] Qu, C., J. Sun, and J. Z. Wang. "Automatic detection of the fall of old people based on Kinect sensor." *Journal of sensor technology* 29, no. 3 (2016): 378-383.
- [317] R, Olivia, et al. "Broad-spectrum resistance to bacterial blight in rice using genome editing". *Nat biotechnol*, Vol. 37, No. 11, pp.1344-1350, Oct, 2019.
- [318] R, Olivia, et al. "Broad-spectrum resistance to bacterial blight in rice using genome editing". *Nat biotechnol*, Vol. 37, No. 11, pp.1344-1350, Oct, 2019.
- [319] Anyoha," The history of artificial intelligence (AI). Blog, special edition of artificial intelligence," 2017. [Online]. Available: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence>.
- [320] Anyoha," The history of artificial intelligence (AI). Blog, special edition of artificial intelligence," 2017. [Online]. Available: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence>.
- [321] Doshi, N. Aphorpe and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 29-35.
- [322] Loh, 2013. *Fish Pathology*, 4th edn Edited by Ronald J Roberts. Wiley Blackwell, Oxford. 597pp.
- [323] Loh, 2013. *Fish Pathology*, 4th edn Edited by Ronald J Roberts. Wiley Blackwell, Oxford. 597pp.
- [324] Moordiani, A. Wildani and S. Widayani, S. "Analisis Kebutuhan Penyuluh Pertanian Mendukung Jawa Tengah Menjadi Lumbung Pangan Nasional". In *Prosiding Seminar Nasional Fakultas Pertanian UNS Vol. 2, No. 1*, pp. C53 – C60, 2018.
- [325] Moordiani, A. Wildani and S. Widayani, S. "Analisis Kebutuhan Penyuluh Pertanian Mendukung Jawa Tengah Menjadi Lumbung Pangan Nasional". In *Prosiding Seminar Nasional Fakultas Pertanian UNS Vol. 2, No. 1*, pp. C53 – C60, 2018.
- [326] Syrlybaeva and E.-M. Strauch, "Deep learning of Protein Sequence Design of Protein-protein Interactions," doi: 10.1101/2022.01.28.478262.
- [327] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep learning in the biomedical applications: Recent and future status," *Appl. Sci.*, vol. 9, no. 8, Apr. 2019, doi: 10.3390/APP9081526.
- [328] R.P., Shaikh, and S.A. Dhole, "Citrus Leaf Unhealthy Region Detection by using Image Processing Technique, in: *IEEE International Conference on Electronics, Communication and Aerospace Technology*, pp. 420–423, 2017.
- [329] R.P., Shaikh, and S.A. Dhole, "Citrus Leaf Unhealthy Region Detection by using Image Processing Technique, in: *IEEE International Conference on Electronics, Communication and Aerospace Technology*, pp. 420–423, 2017.
- [330] R.R. Al Hakim, "Pencegahan Penularan Covid-19 Berbasis Aplikasi Android Sebagai Implementasi Kegiatan KKN Tematik Covid-19 di Sokanegara Purwokerto Banyumas," *Commun. Eng. and Emerg. J. (CEEJ)*, vol. 2, no.1, pp. 7–13. 2020.
- [331] R.R. Al Hakim, "Pencegahan Penularan Covid-19 Berbasis Aplikasi Android Sebagai Implementasi Kegiatan KKN Tematik Covid-19 di Sokanegara Purwokerto Banyumas," *Commun. Eng. and Emerg. J. (CEEJ)*, vol. 2, no.1, pp. 7–13. 2020.
- [332] R.R. Al Hakim, A. Pangestu, and A. Jaenul., A. 2021. Penerapan metode certainty factor dengan tingkat kepercayaan pada sistem pakar dalam mendiagnosis parasit pada ikan," *J. of Inf. Tech. Res.*, vol.2 no.1, pp. 27-37. 2021.
- [333] R.R. Al Hakim, A. Pangestu, and A. Jaenul., A. 2021. Penerapan metode certainty factor dengan tingkat kepercayaan pada sistem pakar dalam mendiagnosis parasit pada ikan," *J. of Inf. Tech. Res.*, vol.2 no.1, pp. 27-37. 2021.
- [334] R.R. Al Hakim, E. Rusdi, E., and M.A. Setiawan,"Android based expert system application for diagnose covid-19 disease: cases study of banyumas regency," *J. of Intell. Com. & Health Inf.*, vol. 1. No.2, pp.1–13. 2020.
- [335] R.R. Al Hakim, E. Rusdi, E., and M.A. Setiawan,"Android based expert system application for diagnose covid-19 disease: cases study of banyumas regency," *J. of Intell. Com. & Health Inf.*, vol. 1. No.2, pp.1–13. 2020.
- [336] Radhika, Narendiran, "Kind of Crops and Small Plants Prediction using IoT with Machine Learning," *International Journal of Computer & Mathematical Sciences* April 2018, pp. 93-97

- [337] Rajeshwari, sundar.,Santhoshs, Hebbar., Varaprasad, Golla. (2015). Implementing Intelligent Traffic Control System for Congestion Control, Ambulance Clearance, and Stolen Vehicle Detection, *IEEE Sensors Journal*, 15(2), 1109 – 1113.
- [338] Rajput,Rashika.,& Gupta, Amit. (2018). Power Grid System Management through Smart Grid inIndia.*International Journal on Recent Technologies in Mechanical and Electrical Engineering*, 5(1), 17-26.
- [339] Ramanan,Rajasekaran, G.,Manikandaraj, S., Kamaleshwar, R. (2017, February).Implementation of Machine Learning Algorithm for Predicting User Behavior and Smart Energy Management, *International Conference on Data Management, Analytics and Innovation*, Pune, India
- [340] Ramesh A. Medar (2014) "A survey on data mining techniques for crop yield prediction", *International Journal of advance in computer science and management studies*, ISSN:2231-7782, volume 2, Issue 9.
- [341] Rana, P., Liaw, S. Y., Lee, M. S., & Sheu, S. C. 2021. Discrimination of four Cinnamomum species with physico-functional properties and chemometric techniques: application of PCA and MDA models. *Foods*, 10(11): 2871, 2021.
- [342] Rashmi,Hegde.,Rohith,Sali, R., Indira, M. S.(2013). RFID and GPS based automatic lane clearance system for ambulance, *International Journal of Advanced Electrical and Electronics Engineering*, (IJAE), 2(3), 102–107.
- [343] Razali, N., Mustapha, A., Abd Wahab, M.H., Mostafa, S.A. and Rostam, S.K., 2020, April. A data mining approach to prediction of liver diseases. In *Journal of Physics: Conference Series* (Vol. 1529, No. 3, p. 032002). IOP Publishing.
- [344] repository.iitr.ac.in, Internet Source
- [345] Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R L, "Prediction of Crop Yield using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)* Feb 2018, pp. 2237-2239
- [346] Bhanumathi, M. Vineeth and N. Rohit, "Crop Yield Prediction and Efficient use of Fertilizers," 2019 *International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 0769-0773, doi: 10.1109/ICCSP.2019.8698087.
- [347] Budi," Kombinasi metode forward chaining dan certainty factor untuk mendiagnosa penyakit pada ikan cupang," *Tek-Sis. Inf. Unus. PGRI Kediri*, vol. 1, no.1, pp. 1–6. 2017.
- [348] Budi," Kombinasi metode forward chaining dan certainty factor untuk mendiagnosa penyakit pada ikan cupang," *Tek-Sis. Inf. Unus. PGRI Kediri*, vol. 1, no.1, pp. 1–6. 2017.
- [349] Dwibedi, M. Pujari and W. Sun, "A Comparative Study on Contemporary Intrusion Detection Datasets for Machine Learning Research," 2020 *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2020, pp. 1-6.
- [350] S. G L, N. V and S. U, "A Review on Prediction of Crop Yield using Machine Learning Techniques," 2022 *IEEE Region 10 Symposium (TENSYP)*, Mumbai, India, 2022, pp. 1-5. doi: 10.1109/TENSYP54529.2022.9864482
- [351] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein-protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018, doi: 10.1093/bioinformatics/bty573.
- [352] S. Hossain, et al. "Recognition and detection of tea leaf's diseases using support vector machine," In 2018 *IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 150-154, IEEE, Mar. 2018.
- [353] S. Hossain, et al. "Recognition and detection of tea leaf's diseases using support vector machine," In 2018 *IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 150-154, IEEE, Mar. 2018.
- [354] S. Iqbal, M.U. Ghani, T.Saba, and A. Rehman, "Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN)," *Microsc. Res. Tech*, Vol.81, No.4, pp. 419-427., Jan. 2018, doi : 10.1002/jemt.22994.
- [355] S. Iqbal, M.U. Ghani, T.Saba, and A. Rehman, "Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN)," *Microsc. Res. Tech*, Vol.81, No.4, pp. 419-427., Jan. 2018, doi : 10.1002/jemt.22994.
- [356] S. Kusrini,"Sistem pakar teori dan aplikasi," Andi offset, Yogyakarta. 2006.
- [357] S. Kusrini,"Sistem pakar teori dan aplikasi," Andi offset, Yogyakarta. 2006.
- [358] S. Kusumadewi,"Artificial Intelegence (Teknik dan Aplikasinya)," Yogyakarta: Graha Ilmu. 2003.
- [359] S. Kusumadewi,"Artificial Intelegence (Teknik dan Aplikasinya)," Yogyakarta: Graha Ilmu. 2003.

- [360] S. Mishra, P. Paygude, S. Chaudhary and S. Idate, "Use of data mining in crop yield prediction," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2018, pp. 796-802. doi: 10.1109/ICISC.2018.8398908
- [361] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion : a review," vol. 23, pp. 1–15, 2022.
- [362] S. Ramesh et al. "Plant disease detection using machine learning," 2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C), pp. 41-45, IEEE, Apr. 2018.
- [363] S. Ramesh et al. "Plant disease detection using machine learning," 2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C), pp. 41-45, IEEE, Apr. 2018.
- [364] S. Sahu, M. Chawla and N. Khare, "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 53-57. doi: 10.1109/CCAA.2017.8229770
- [365] S. Subbiah, K. S. M. Anbananthen, S. Thangaraj, S. Kannan and D. Chelliah, "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," in *Journal of Communications and Networks*, vol. 24, no. 2, pp. 264-273, April 2022.
- [366] S. Zahrah, R. Saptono, and E. Suryani, "Identifikasi Gejala Penyakit Padi Menggunakan Operasi Morfologi Citra". In *Seminar Nasional Ilmu Komputer (SNIK 2016)-Semarang* , Vol. 10, Oct, 2016 .
- [367] S. Zahrah, R. Saptono, and E. Suryani, "Identifikasi Gejala Penyakit Padi Menggunakan Operasi Morfologi Citra". In *Seminar Nasional Ilmu Komputer (SNIK 2016)-Semarang* , Vol. 10, Oct, 2016 .
- [368] S. Bartels, et al., "MAP Kinase phosphatase1 and protein tyrosine phosphatase1 are repressors of salicylic acid synthesis and SNC1-mediated responses in Arabidopsis," *The Plant Cell*, Vol. 21, No.9, pp. 2884-2897, Sep. 2009. doi : 10.1105/tpc.109.067678.
- [369] S. Bartels, et al., "MAP Kinase phosphatase1 and protein tyrosine phosphatase1 are repressors of salicylic acid synthesis and SNC1-mediated responses in Arabidopsis," *The Plant Cell*, Vol. 21, No.9, pp. 2884-2897, Sep. 2009. doi : 10.1105/tpc.109.067678.
- [370] S.D.Khirade and A.B. Patil "Plant Disease Detection Using Image Processing," 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 768 -771, doi: 10.1109/ICCUBEA.2015.153.
- [371] S.D.Khirade and A.B. Patil "Plant Disease Detection Using Image Processing," 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 768 -771, doi: 10.1109/ICCUBEA.2015.153.
- [372] S.H. Susilowati. "Fenomena penuaan petani dan berkurangnya tenaga kerja muda serta implikasinya bagi kebijakan pembangunan pertanian". *Forum Penelitian Agro Ekonomi*, Vol.34, No.1, pp.35-55, Jul. 2016.
- [373] S.H. Susilowati. "Fenomena penuaan petani dan berkurangnya tenaga kerja muda serta implikasinya bagi kebijakan pembangunan pertanian". *Forum Penelitian Agro Ekonomi*, Vol.34, No.1, pp.35-55, Jul. 2016.
- [374] S.J. Divinely, K Sivakami, and V. Jayaraj, "Fish diseases identification and classification using machine learning," *Intl. J. Adv. Res. Bas. Eng. Sci.Tech. (IJARBEST)*, vol. 5, pp. 46–51. 2019.
- [375] S.J. Divinely, K Sivakami, and V. Jayaraj, "Fish diseases identification and classification using machine learning," *Intl. J. Adv. Res. Bas. Eng. Sci.Tech. (IJARBEST)*, vol. 5, pp. 46–51. 2019.
- [376] S.kanaga Subba Raju et al.(2017), Demand based crop recommender system for farmers, International Conference on Technological Innovations in ICT For Agriculture and Rural Development.
- [377] S.W. Pyle and E.SupB. Shotts, "A new approach for differentiating flexibacteria isolated from cold water and warm water fish," *Can. Jour. of Fish and Aquat. Sci.*, vol. 37, pp. 1040-1042.1980.
- [378] S.W. Pyle and E.SupB. Shotts, "A new approach for differentiating flexibacteria isolated from cold water and warm water fish," *Can. Jour. of Fish and Aquat. Sci.*, vol. 37, pp. 1040-1042.1980.
- [379] Saarika, P.; Sandhya, K.; Sudha, T. Smart transportation system using IoT. In *Proceedings of the 2017 IEEE International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, Bangalore, KA, India, 17–19 August 2017; pp. 1104–1107.
- [380] Saif H., He Y., Alani H. (2012). Semantic Sentiment Analysis of Twitter. In: Cudré-Mauroux P. et al. (eds) *The Semantic Web – ISWC 2012*. ISWC 2012. Lecture Notes in Computer Science, vol 7649. Springer, Berlin, Heidelberg.
- [381] Salzberger, B., Glück, T., & Ehrenstein, B. (2020). Successful containment of COVID-19: the WHO-Report on the COVID-19 outbreak in China. *Infection*, 2, 151–153. <https://doi.org/10.1007/s15010-020-01409-4>.
- [382] Saranya, P. and Asha, P., 2019, November. Survey on Big Data Analytics in health care. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 46-51). IEEE.

- [383] Saranya, P., and P. Asha. "Survey on Big Data Analytics in health care." 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2019.
- [384] Satti, V., Satya, A., & Sharma, S. 2013. An automatic leaf recognition system for plant identification using machine vision technology. *International Journal of Engineering, Science and Technology*, 5(4): 874-879.
- [385] scholarworks.uni.edu, Internet Source
- [386] Sehgal, S., Singh, H., Agarwal, M., Bhasker, V., & Shantanu (2014). Data analysis using principal component analysis. In 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom) (pp. 45-48).
- [387] Seroepidemiologic studies of coronavirus infection in adults and children. McIntosh K, Kapikian AZ, Turner HC, Hartley JW, Parrott RH, Chanock RM, *Am J Epidemiol*. 1970;91(6):585
- [388] Shabanzade, M., Zahedi, M., & Aghami, S. A. 2011. Combination of local descriptors and global features for leaf recognition, signal and image processing. *Signal & Image Processing: An International Journal (SIPIJ)*, 2(3): 23-31.
- [389] Shen, K.-L., Yang, Y.-H., Jiang, R.-M., Wang, T.-Y., Zhao, D.-C., et al. (2020). Updated diagnosis, treatment and prevention of COVID-19 in children: experts' consensus statement (condensed version of the second edition). *World Journal of Pediatrics*, 3, 232–239. <https://doi.org/10.1007/s12519-020-00362-4>.
- [390] Shridhar Mhaiskar, Chinmay Patil, Piyush Wadhai, Aniket Patil, Vaishali Deshmukh, "A Survey on Predicting Suitable Crops for Cultivation Using IoT," *International Journal of Innovative Research in Computer and Communication Engineering* January 2017, pp. 318- 323
- [391] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, Sonali Idade. (2018) "Use of Data Mining in Crop Yield Prediction" IEEE Xplore ISBN:978-1-5386-0807-4; Part Number: CFP18J06.
- [392] Sivakumar Venu, Zubair Rahman, "Energy and cluster based efficient routing for broadcasting in mobile ad hoc networks", *Springer Cluster Computing*, 2018, Vol. 22, pp. 661-671. <https://doi.org/10.1007/s10586-018-2255-3>
- [393] Sivakumar Venu; A. M. J. Md. Zubair Rahman, "Effective Routine Analysis in MANET's Over FAODV" 2017 IEEE International Conference on "Power, Control, Signals and Instrumentation Engineering (ICPCSI)", ISBN: 978-1-5386-0813-5 on 21st & 22nd Sep 2017 Published in IEEE Conference publications, page no.2016-2020 <https://ieeexplore.ieee.org/document/8392068/>
- [394] Slawson DL, Fitzgerald N, Morgan KT (2013) Position of the academy of nutrition and dietetics: the role of nutrition in health promotion and chronic disease prevention. *J Acad Nutr Diet* 113(7):972–979.
- [395] Smisek J, Jancosek M, Pajdla T. 2013. 3D with Kinect[J]. "Advances in Computer Vision & Pattern Recognition", 2013, 21(5):1154-1160.
- [396] Snehal S. Dahikar, Dr. Sandeep V. Rode (2014), "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach", *International journal of innovative and research in electrical, instrumentation and control engineering*, volume 2, Issue 2.
- [397] Sontakke, S., Lohokare, J., Dani, R., & Shivagaje, P. 2018. Classification of cardiocography signals using machine learning, *Proceedings of the 2018 Intelligent Systems Conference*, 6-7 September 2018, London, UK: 1-6. USA: IEEE.
- [398] Souid, Abdelbaki, Nizar Sakli, and Hedi Sakli. "Classification and predictions of lung diseases from chest x-rays using mobilenet v2." *Applied Sciences* 11.6 (2021): 2751.
- [399] Springer Science and Business Media LLC, 2021.
- [400] Srivastava, D.K., & L. Bhambhu, L. 2010. Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, 12(1): 1-7.
- [401] Submitted to Ashoka University, Student Paper
- [402] Submitted to Asia Pacific University College of Technology and Innovation (UCTI), Student Paper
- [403] Submitted to Coventry University, Student Paper
- [404] Submitted to Durban University of Technology, Student Paper
- [405] Submitted to Federation University, Student Paper
- [406] Submitted to Fiji National University, Student Paper
- [407] Submitted to Florida Virtual School, Student Paper
- [408] Submitted to Liverpool John Moores University, Student Paper
- [409] Submitted to Middlesex University, Student Paper
- [410] Submitted to NCC Education, Student Paper
- [411] Submitted to Sim University, Student Paper
- [412] Submitted to Solihull College, West Midlands, Student Paper

- [413] Submitted to The Scientific & Technological, Research Council of Turkey (TUBITAK), Student Paper
- [414] Submitted to University of Bolton, Student Paper
- [415] Submitted to University of Greenwich, Student Paper
- [416] Submitted to University of Kentucky, Student Paper
- [417] Submitted to University of Lincoln, Student Paper
- [418] Submitted to University of Ulster, Student Paper
- [419] Submitted to University of Wollongong, Student Paper
- [420] Submitted to UOW Malaysia KDU University College Sdn. Bhd, Student Paper
- [421] Submitted to Victorian Institute of Technology, Student Paper
- [422] T Raghav Kumar, Bhagavatula Aiswarya, Aashish Suresh, Drishti Jain, Natesh Balaji, Varshini Sankaran, "Smart Management of Crop Cultivation using IOT and Machine Learning," *International Research Journal of Engineering and Technology (IRJET)* Nov 2018, pp. 845- 850
- [423] Arakawa, "Recent research and developing trends of wearable sensors for detecting blood pressure," *Sensors*, vol. 18, no. 9, p. 2772, 2018.
- [424] Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-assisted multi-view video summarization using CNN and bi-directional LSTM," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 77–86, Jan. 2020.
- [425] Hussain, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Intelligent embedded vision for summarization of multi-view videos in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2592–2602, Apr. 2020.
- [426] Hussain, K. Muhammad, S. Khan, A. Ullah, M. Y. Lee, and S. W. Baik, "Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers," *J. Artif. Intell. Syst.*, vol. 1, no. 15, p. 2019, 2019.
- [427] T.Boller and G. Felix, "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors," *Ann Rev Plant Biol*, Vol.60, pp.379 – 406. 2009, doi : 10.1146/annurev.arplant.57.032905.105346.
- [428] T.Boller and G. Felix, "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors," *Ann Rev Plant Biol*, Vol.60, pp.379 – 406. 2009, doi : 10.1146/annurev.arplant.57.032905.105346.
- [429] T.H. Yuniyanto, "Sistem pakar diagnosa penyakit pada ikan hias," pp. 17. 2013.
- [430] T.H. Yuniyanto, "Sistem pakar diagnosa penyakit pada ikan hias," pp. 17. 2013.
- [431] T.K. Malik, Shaveta, and A.K. Sahoo, "A novel approach to fish disease diagnostic system based on machine learning," *Adv. in Im. and Vid. Proc.*, vol. 5, no. 1, pp. 49–49. 2017.
- [432] T.K. Malik, Shaveta, and A.K. Sahoo, "A novel approach to fish disease diagnostic system based on machine learning," *Adv. in Im. and Vid. Proc.*, vol. 5, no. 1, pp. 49–49. 2017.
- [433] T.M. Voegel, and L.M. Nelson, "Quantification of *Agrobacterium vitis* from grapevine nursery stock and vineyard soil using droplet digital PCR," *Plant Dis.*, Vol. 102, No.11, pp. 2136-2141, Sep. 2018, doi : 10.1094/PDIS-02-18-0342-RE.
- [434] T.M. Voegel, and L.M. Nelson, "Quantification of *Agrobacterium vitis* from grapevine nursery stock and vineyard soil using droplet digital PCR," *Plant Dis.*, Vol. 102, No.11, pp. 2136-2141, Sep. 2018, doi : 10.1094/PDIS-02-18-0342-RE.
- [435] T.S. Dewi and R. Arnie, "Sistem Pakar Diagnosa Penyakit Ikan Patin Dengan Metode Certainty Factor Berbasis Web," *J. TIMES*, vol. 6, no. 1, pp. 1311–1448. 2017.
- [436] T.S. Dewi and R. Arnie, "Sistem Pakar Diagnosa Penyakit Ikan Patin Dengan Metode Certainty Factor Berbasis Web," *J. TIMES*, vol. 6, no. 1, pp. 1311–1448. 2017.
- [437] T.S. Saptadi and V.S. Sebukita, "Pengambilan keputusan dalam penerimaan karyawan bank dengan pendekatan terstruktur berbasis sistem pakar," *J. Tek. Kom. dan Inf.*, p. 81. 2012.
- [438] T.S. Saptadi and V.S. Sebukita, "Pengambilan keputusan dalam penerimaan karyawan bank dengan pendekatan terstruktur berbasis sistem pakar," *J. Tek. Kom. dan Inf.*, p. 81. 2012.
- [439] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, Sep. 2019.
- [440] Tan, T., & Xu, J. 2020. Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: A review. *Artificial Intelligence in Agriculture*, 4: 104-115.
- [441] Tchito Tchappa, C., Mih, T.A., Tchagna Kouanou, A., Fozin Fozin, T., Kuetche Fogang, P., Mezatio, B.A. and Tchiotop, D., 2021. Biomedical image classification in a big data architecture using machine learning algorithms. *Journal of Healthcare Engineering*, 2021.



- [442] Thashmee Karunaratne; Henrik Bostrom; Ulf Norinder, Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization - A Case Study with Medicinal Chemistry Datasets, IEEE Explore, Dec 2010, DOI: 10.1109/ICMLA.2010.128.
- [443] Thinsungnoen, Tippaya & Kaoungku, Nuntawut & Durongdumronchai, Pongsakorn & Kerdprasop, Kittisak & Kerdprasop, Nittaya. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. 44-51. 10.12792/iciae2015.012.
- [444] Tracy RP, Fried LP, Borhani NO, Weiler PG (1991) The cardiovascular health study: design and rationale. *Ann Epidemiol* 1(3):263–276
- [445] Tsai, J.T., Chou, J.H., Liu, T.K (2006). Tuning the structure and parameters of a neural network by using hybrid Taguchi-genetic algorithm. *IEEE Transactions on Neural Networks*, 17(1), 69-80.
- [446] Tsao CW, et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. *Circulation* 2020. (<https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000000757>) Accessed 9/28/2021.
- [447] Inyaem, "Construction Model Using Machine Learning Techniques for the Prediction of Rice Produce for Farmers," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 2018, pp. 870-874. doi: 10.1109/ICIVC.2018.8492883
- [448] S. Musa, M. Chhabra, A. Ali and M. Kaur, "Intrusion Detection System using Machine Learning Techniques: A Review," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 149-155.
- [449] Lyubchenko, R. Matarneh, O. Kobylin, and V. Lyashenko, "Digital image processing techniques for detection and diagnosis of fish diseases," *Intl. J. Of Adv. Res. in Com. Sci. and Soft. Eng.*, vol. 6, pp. 79–83. 2016.
- [450] Lyubchenko, R. Matarneh, O. Kobylin, and V. Lyashenko, "Digital image processing techniques for detection and diagnosis of fish diseases," *Intl. J. Of Adv. Res. in Com. Sci. and Soft. Eng.*, vol. 6, pp. 79–83. 2016.
- [451] Sivakumar, Anburajan. M. N, Aravind. R, ArunPrasath. R, Muniyasamy. K, "Packet Loss Detection in MANETs Using Modified Fine Grained Approach" in *International Journal of Management, Technology And Engineering*, Volume 9, Issue 4, April 2019, ISSN NO : 2249-7455 DOI:16.10089.IJMTE.2019.V9I4.19.27093 OR <https://app.box.com/s/y4g0pt4yf07canj8xk07q0k88g2efjsd>
- [452] V.A.J. Kempf, K. Trebesius and I.B. Autenrieth 2000. "Fluorescent in situ hybridization allows rapid identification of microorganisms in blood cultures," *Am Soc Microbiol*, Vol. 38, No. 2, pp. 830–838, Feb. 2000, doi : 10.1128/JCM.38.2.830-838.2000.
- [453] V.A.J. Kempf, K. Trebesius and I.B. Autenrieth 2000. "Fluorescent in situ hybridization allows rapid identification of microorganisms in blood cultures," *Am Soc Microbiol*, Vol. 38, No. 2, pp. 830–838, Feb. 2000, doi : 10.1128/JCM.38.2.830-838.2000.
- [454] V.Sivakumar, J.Kanimozhi, B.Keerthana, R.Muthu lakshmi "Capacity Enhancement using Delay-Sensitive Protocol in MANETs", *Springer Lecture Notes in Networks and Systems*, "Inventive Communication and Computational Technologies" Proceedings of ICICCT 2019, entitled Volume 89, Pages 901-910, Publisher: Springer, Singapore, ISSN 2367-3370 <https://www.springer.com/series/15179>
- [455] V.Sivakumar, R Swathi, Yuvaraj., "An IoT-Based Energy Meter for Energy Level Monitoring, Predicting" "Challenges and Opportunities for the Convergence of IoT, Big Data, and Cloud Computing", IGI Publisher, Chapter No. 4, Pages 48-65, 2021. DOI: 10.4018/978-1-7998-3111-2.ch004 or <https://www.igi-global.com/chapter/an-iot-based-energy-meter-for-energy-level-monitoring-predicting-and-optimization/269556>
- [456] Vahidy, F. S., Drews, A. L., Masud, F. N., Schwartz, R. L., Askary, B. "Billy," Boom, M. L., & Phillips, R. A. (2020). Characteristics and Outcomes of COVID-19 Patients During Initial Peak and Resurgence in the Houston Metropolitan Area. *JAMA*, 10, 998. <https://doi.org/10.1001/jama.2020.15301>.
- [457] Valko M, Hauskrecht M (2008) Distance metric learning for conditional anomaly detection. In: FLAIRS conference, pp 684–689.
- [458] Vignesh, G., Vishal, Narayanan., Prakash, S., Sivakumar, V.(2016, May). Automated Traffic Light Control System and Stolen Vehicle Detection, 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India. URL: <http://ieeexplore.ieee.org/document/7808101/>
- [459] Vo, A.H., Dang, H.T., Nguyen, B.T., & Pham, V.-H. 2019. Vietnamese herbal plant recognition using deep convolutional features. *International Journal Machine Learning Computing*, 9(3): 363–367.

- [460] Wang, M. Yang, M., and P.H. Seong, "Development of a rule-based diagnostic platform on an object-oriented expert system shell," *Annals of Nuc. En.* vol. 88, pp. 252-264. 2016.
- [461] Wang, M. Yang, M., and P.H. Seong, "Development of a rule-based diagnostic platform on an object-oriented expert system shell," *Annals of Nuc. En.* vol. 88, pp. 252-264. 2016.
- [462] Wang, X. Du, D. Shan, R. Qin and N. Wang, "Cloud Intrusion Detection Method Based on Stacked Contractive Auto-Encoder and Support Vector Machine," in *IEEE Transactions on Cloud Computing*, 2020.
- [463] Wagstaff, Kiri & Cardie, Claire & Rogers, Seth & Schrödl, Stefan. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of 18th International Conference on Machine Learning*. 577-584.
- [464] Wang, X., Ma, X., & Grimson, E. (2007). Unsupervised Activity Perception by Hierarchical Bayesian Models. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8).
- [465] Wang, Y., Wang, Y., Chen, Y. & Qin, Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *J. Med. Virol.* 92, 568–576 (2020).
- [466] Willett WC, Koplan JP, Nugent R, Dusenbury C, Puska P, Gaziano TA (2006) Prevention of chronic disease by means of diet and lifestyle changes. *Disease Control Priorities in Developing Countries*, pp 833–850
- [467] World Health Organization (1990) Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group. Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group, p 797
- [468] World Health Organization (WHO). 2004. *Waterborne Zoonosis: Identification, Causes and Control*. World Health Organization, Geneva.
- [469] World Health Organization (WHO). 2004. *Waterborne Zoonosis: Identification, Causes and Control*. World Health Organization, Geneva.
- [470] World Health Organization. Cardiovascular diseases (CVDs). ([https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds%29\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds%29))) Accessed 9/28/2021.
- [471] [www.entrepreneur.com](http://www.entrepreneur.com), Internet Source
- [472] [www.sensorsuae.com](http://www.sensorsuae.com), Internet Source
- [473] [www.sensortips.com](http://www.sensortips.com), Internet Source
- [474] Hu, C. Feng, Y. Zhou, A. Harrison, and M. Chen, "DeepTrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks," *Bioinformatics*, vol. 38, no. 3, pp. 694–702, 2022, doi: 10.1093/bioinformatics/btab737.
- [475] Li, J. Dunn, D. Salins et al., "Digital Health: tracking physiomes and activity using wearable biosensors reveals useful health-related information," *PLoS Biology*, vol. 15, no. 1, Article ID e2001402, 2017.
- [476] Wang, et al., "Current advances on genetic resistance to rice blast disease". In *Rice-Germplasm, genetics and improvement*, 2014, pp.195-217. InTech, Rijeka, Croatia.
- [477] Wang, et al., "Current advances on genetic resistance to rice blast disease". In *Rice-Germplasm, genetics and improvement*, 2014, pp.195-217. InTech, Rijeka, Croatia.
- [478] X. Zhang, J. Ran and J. Mi, "An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 2019, pp. 456-460.
- [479] X. Zhong, L. Xue-lu, L. Bing-hai, Z. Chang-yong, and W. Xue-feng, "Development of a sensitive and reliable droplet digital PCR assay for the detection of *Candidatus Liberibacter asiaticus*," *J. Integr. Agric.*, Vol.17, No.2, pp. 483–487, 2018, doi : 10.1016/S2095-3119(17)61815-X.
- [480] X. Zhong, L. Xue-lu, L. Bing-hai, Z. Chang-yong, and W. Xue-feng, "Development of a sensitive and reliable droplet digital PCR assay for the detection of *Candidatus Liberibacter asiaticus*," *J. Integr. Agric.*, Vol.17, No.2, pp. 483–487, 2018, doi : 10.1016/S2095-3119(17)61815-X.
- [481] X.Meng and S. Zhang, "MAPK cascades in plant disease resistance signaling," *Annu Rev Phytopathol*, Vol.51, No.1, pp. 245-266, May. 2013, doi : 10.1146/annurev-phyto-082712-102314.
- [482] X.Meng and S. Zhang, "MAPK cascades in plant disease resistance signaling," *Annu Rev Phytopathol*, Vol.51, No.1, pp. 245-266, May. 2013, doi : 10.1146/annurev-phyto-082712-102314.
- [483] X.Sun, S. Mu, Y. Xu, Z. Cao, and T.Su. "Image recognition of tea leaf diseases based on convolutional neural network," *2018 International Conference on Security, Pattern, Analysis, and Cybernetics (SPAC)*, 2018, pp. 304-309, doi : 10.1109/SPAC46244.2018.8965555.

- [484] X.Sun, S. Mu, Y. Xu, Z. Cao, and T.Su. "Image recognition of tea leaf diseases based on convolutional neural network," 2018 International Conference on Security, Pattern, Analysis, and Cybernetics (SPAC), 2018, pp. 304-309, doi : 10.1109/SPAC46244.2018.8965555.
- [485] Xu, M., Wang, J., & Zhu, L. 2021. Tea quality evaluation by applying E-nose combined with chemometrics methods. *Journal of Food Science and Technology*, 58(4): 1549–1561.
- [486] Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised K-means DDoS detection method using hybrid feature selection algorithm," *IEEE Access*, vol. 7, pp. 64351–64365, 2019.
- [487] He, G. Kang, X. Dong, Y. Fu, and Y. Yang, "Soft filter pruning for accelerating deep convolutional neural networks," 2018, arXiv:1808.06866. [Online]. Available: <http://arxiv.org/abs/1808.06866> .
- [488] He, G. Kang, X. Dong, Y. Fu, and Y. Yang, "Soft filter pruning for accelerating deep convolutional neural networks," 2018, arXiv:1808.06866. [Online]. Available: <http://arxiv.org/abs/1808.06866> .
- [489] Y. Huang et al., "LoadCNN: A efficient green deep learning model for day-ahead individual resident load forecasting," 2019. [Online]. Available: arXiv:1908.00298.
- [490] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, vol. 166, no. April 2019, pp. 4–21, 2019, doi: 10.1016/j.ymeth.2019.04.008.
- [491] Y. Liu, H. Wang, W. Zhao, M. Zhang, H. Qin, and Y. Xie, "Flexible, stretchable sensors for wearable health monitoring: sensing mechanisms, materials, fabrication strategies and features," *Sensors*, vol. 18, no. 2, p. 645,
- [492] Y.-F. Zhang and H.-D. Chiang, "Enhanced ELITE-load: A novel CMPSOATT methodology constructing short-term load forecasting model for industrial applications," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2325–2334, Apr. 2020.
- [493] Y.Fang, and R.P. Ramasamy, "Current and prospective methods for plant disease detection", *Biosensors*, Vol. 5, No.3, pp. 537-561, Aug. 2015, doi : 103390/bios5030537.
- [494] Y.Fang, and R.P. Ramasamy, "Current and prospective methods for plant disease detection", *Biosensors*, Vol. 5, No.3, pp. 537-561, Aug. 2015, doi : 103390/bios5030537.
- [495] Yadav, T. & Reddy, Dr & Prasad, Ram & Gopal, Pradeep. (2020). CROP YIELD AND FERTILIZERS PREDICTION USING DECISION TREE ALGORITHM. *International Journal of Engineering Applied Sciences and Technology*. 5. 187-193.
- [496] Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., & Lin, S. (2007). Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 40–51, <https://doi.org/10.1109/tpami.2007.250598>.
- [497] Yan, X., & Jia, M. 2018. A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. *Neurocomputing*, 313: 47-64.
- [498] Yang, J., Chen, X., Deng, X., Chen, Z., Gong, H., Yan, H., Wu, Q., Shi, H., Lai, S., Ajelli, M., Viboud, C., & Yu, P. H. (2020). Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China. *Nature Communications*, 1. <https://doi.org/10.1038/s41467-020-19238-2>.
- [499] Yogish D., Manjunath T.N., Hegadi R.S. (2019) Review on Natural Language Processing Trends and Techniques Using NLTK. In: Santosh K., Hegadi R. (eds) *Recent Trends in Image Processing and Pattern Recognition*. RTIP2R (2018). *Communications in Computer and Information Science*, vol 1037. Springer, Singapore.
- [500] Yu Zesheng. "Research and Design of Smart Home System Based on Kinect attitudeRecognition". Liaoning University of Science and Technology, 2017.
- [501] Yuan, Chunhui & Yang, Haitao. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J. 2*. 226-235. 10.3390/j2020016.
- [502] Yuji Roa, Geon Heo, A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective, *IEEE Transactions on Knowledge and Data Engineering PP(99):1-1*, October 2021, DOI:10.1109/TKDE.2019.2946162
- [503] Yuji Roh, Geon Heo, Steven Euijong Whang, Senior Member, IEEE, A Survey on Data Collection for Machine Learning and Big Data - AI Integration Perspective, <https://arxiv.org/pdf/1811.03402,2019>
- [504] Hakim and R. Rizky, "Sistem pakar diagnosis penyakit ikan mas menggunakan metode certainty factor di upt balai budidaya ikan air tawar dan hias kabupaten pandeglang banten," *J. Tek. Inf. Unis*, vol. 7, no.2, pp.164–169. 2020.
- [505] Hakim and R. Rizky, "Sistem pakar diagnosis penyakit ikan mas menggunakan metode certainty factor di upt balai budidaya ikan air tawar dan hias kabupaten pandeglang banten," *J. Tek. Inf. Unis*, vol. 7, no.2, pp.164–169. 2020.

- [506] Liao, G. Pan, C. Sun, and J. Tang, "Predicting subcellular location of protein with evolution information and sequence - based deep learning," pp. 1–22, 2021.
- [507] Zantalis, Fotios, Grigorios Koulouras, Sotiris Karabetsos, and Dionisis Kandris. "A review of machine learning and IoT in smart transportation." *Future Internet* 11, no. 4 (2019): 94.
- [508] Zaremba, Wojciech & Sutskever, Ilya & Vinyals, Oriol. (2014). Recurrent Neural Network Regularization.
- [509] Zhang, Monica, Xiaoou., Grolinger, Katarina., Capretz, Miriam, A.M. (2018, December). Forecasting Residential Energy Consumption: Single Household Perspective, 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA.
- [510] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. 2017. Learning k for kNN classification. *ACM Transactions on Intelligent Systems and Technology*, 8: 1-19.
- [511] Zhang, W., & Wen, J. 2021. Research on leaf image identification based on improved AlexNet neural network. *Journal of Physics*, 2031:1-13.
- [512] Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF\*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 3, 2758–2765. <https://doi.org/10.1016/j.eswa.2010.08.066>.
- [513] Zhang, Z. 2016. Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, 4(11):1-7.
- [514] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 8, 727–733. <https://doi.org/10.1056/nejmoa2001017>.
- [515] Zhu, N.; Liu, X.; Liu, Z.; Hu, K.; Wang, Y.; Tan, J.; Guo, Y. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *Int. J. Agric. Biol. Eng.* 2018, 11, 32–44. [CrossRef]

©Copyright Material

# Machine Learning Algorithms for Intelligent Data Analytics

## DESCRIPTION

Recent decade has seen a huge rise in the growth of massive amount of data. In the industry 40 digital era, there is large volumes of data in the form of mobile data, social media data, Internet of Things (IOT) data, cybersecurity data, financial data, medical data etc. This book aims to cover machine learning techniques for performing intelligent analytics on the data generated. Different types of machine learning algorithms such as supervised learning, semi – supervised learning, unsupervised learning and reinforcement learning could be applied to perform data analytics. Advanced learning technique of Artificial Neural Networks, the Deep learning technique could also be applied for performing intelligent analytics, which is also covered under the scope of this book.

## WHO THIS BOOK IS FOR

Researchers and graduate students in computer science, Artificial Intelligence, Machine Learning and Data Science, as well as industry scientific researchers, IT managers, and systems managers.

