Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques

A dissertation-I submitted to the Mahatma Gandhi Central University
in partially fulfillment of the requirements
for the award of the degree of

MASTER OF TECHONOLGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

SUBHAM KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

March 12, 2023

Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques

A dissertation-I submitted to the Mahatma Gandhi Central University
in partially fulfillment of the requirements
for the award of the degree of

MASTER OF TECHONOLGY

IN

COMPUTER SCIENCE & ENGINEERING

BY

SUBHAM KUMAR

(MGCU2021CSIT4029)

Under the Supervision of

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

MAHATMA GANDHI CENTRAL UNIVERSITY,

MOTIHARI, BIHAR - 845401, INDIA

March 12, 2023

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department Of Computer Science And Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय,
बिहार-८४५४०१
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

DECLARATION

This is to certify that the dissertation entitled **" Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques"** is being submitted to the **Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India** in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science & Engineering**, is a record of bonafide work carried out by me under the supervision of **"Dr. Vipin Kumar, Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India."**

The matter embodied in the dissertation has not been submitted in part or full to any University or Institution for the award of any other degree or diploma.

Subham Kumar
(MGCU2021CSIT4029)
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University,
Motihari, Bihar - 845401, India
Email id: - subh700454@gmail.com

"I am enough of an artist to draw freely upon my imagination. Imagination is more important than knowledge. For knowledge is limited, whereas imagination encircles the world."

Albert Einstein

कंप्यूटर विज्ञान और सूचना प्रौद्योगिकी विभाग
Department Of Computer Science And Information Technology
महात्मा गाँधी केन्द्रीय विश्वविद्यालय,
बिहार-८४५४०१
MAHATMA GANDHI CENTRAL UNIVERSITY,
MOTIHARI, BIHAR - 845401, INDIA

CERTIFICATE

This is to certify that the dissertation entitled **"Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques"** submitted by **Subham Kumar** to the **Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India** for the award of the degree of **Master of Technology** in **Computer Science & Engineering**, is a research work carried out by him under the supervision of **"Dr. Vipin Kumar, Department Of Computer Science And Information Technology, Mahatma Gandhi Central University, Motihari, Bihar - 845401, India."**

Head of Department

Prof. (Dr.) Vikash Pareek
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University, Motihari, Bihar - 845401, India

Supervisor

Dr. Vipin Kumar
Department Of Computer Science And Information Technology
Mahatma Gandhi Central University, Motihari, Bihar - 845401, India

Abstract

Leaf is the essential part of plant which have significant feature for the identification of Plant. But without an expert, it is a very difficult task to find the type of plant. Fruits Plant have significant role in the fulfillment of proper nutrition. Therefore, automatic fruit plant detection is highly required and it can be accomplished using image processing techniques. In this dissertation, the authors themselves have collected 30 categories of fruit plant leaf (i.e., 11321 RGB images) leaves using a smartphone device and comparative studies have been performed for the classification of the leaf of the fruit plant using multiple machine learning (ML) models i.e., Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision tree (DT), Multi-layer Perceptron (MLP), and Naive Bayes (NB). By using MLP of Machine learning models we achieved the highest accuracy of 93.67%, while with SVM (89.30%) we obtained the second-highest, and for L.R. (83.90%) it is the third highest, on the original dataset, while on Augmentation dataset to MLP (93.18%), SVM (88.53%), LR (83.73%) are the first, second, and third highest, respectively. The performance analysis has concluded that MLP has performed better than other classifiers used such as LR, KNN, SVM, DT and NB. In this disseratation, also used D.L. models are ResNet, AlexNet, VggNet, SqueezeNet, GoogLeNet, ShuffleNet, Resnext50, and DenseNet. On the original data, the SuffleNet (99.96%) models have higher performance and the second-highest DenseNet (99.69%) model is closest to the SuffleNet model than other models, and it is the third highest, on GoogleNet (99.47%), while on the augmented dataset, SuffleNet (98.34%), DenseNet (97.77%). Based on the analysis of D.L. models, we can conclude that ShuffleNet has performed better than the others i.e., 99.96% accuracy. Therefore, DL models better performed than ML classifiers.

Acknowledgements

This M.Tech dissertation is the result of hard work, upon which many people have contributed and given their support. I have made this dissertation on the topic **"Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques."** I have also tried my best in this dissertation to explain all the related detail. I would like to express my sincere gratitude towards my Superviser **Dr. Vipin Kumar**, Department of CS & IT , for providing excellent guidance, encouragement, inspiration, and constant and timely support throughout this M.Techdissertation work. He taught me how to pursue the right aim towards the work, and showed me differnt ways to approach the research problem. His wide knowledge and logical ways of thinking have been great value for me, and his understanding and guidance have provided the successful completion of the Dissertation work.

First and foremost, I would like to express my gratitude to our beloved Dean of the Computational Sciences, Information and Communication Technology and Head of Department of Computer Science and Information Technology **Prof. (Dr.) Vikash Pareek**, for providing his kind support in various aspects. A special thanks to all the Respected Teachers **Dr. Atul Tripathi**, **Dr. Sunil Kumar Singh**, and **Mr. Shubham Kumar**, of the Department of Computer Science and Information Technology. Next, I am also very glad to express my sincere thank and gratitude to **Dr. Pawan Kumar Chaurasia**, the Ex-Head of Department, Department of Computer Science & Information Technology, Mahatma Gandhi Central University, Bihar, India, who taught me many things about Research when I dont have any knowledge about it. And, I will give my special thanks to **Dr. Kundan Kishor Rajak**, Assistant Professor, the Department of Zoology, Mahatma Gandhi Central University, Bihar for finding places where we have found these fruit plants.

I am always grateful to the university, our Honble Vice chancellor **Prof. (Dr.) Anand Praksh** for providing such a good research environment and thanks should

also go to the Administration, OSD Administration, OSD Finance, and Librarians for giving every possible resource.

Special thanks to all my friends, especially **Aditya Kumar**, **Gaurav Kumar**, **Chitranjan Kumar**, **Anchit Akshaansh**, **Ankit Prakash**, **Alka Rani**, **Aparna**, **Bittu Kumar Aman**, **Naushad Ahmad**, **Prashant Prabhakar**, **Rifat Naj**, **Ritika Singh**, **Sushant Raj**, **Umesh Kumar**, **Vinita Kumari**, **Watan Kishor Verma**, and all my lovely juniors for their invaluable feedbacks, care, and moral support during this endeavor.

Mother and **Father**, it is impossible to thanks adequately for everything you have done, from loving me unconditionally to rising me in a stable household, where your persistent efforts and traditional values taught your children to celebrate and embrace life. I could not have asked for better parents or role-models. You showed me that anything is possible with faith, hard work and determination. Word fails me to express my appreciation and thanks to **Aprajeeta Kumari**, her constant encouragements and persistent confidence in me, has given me passion to carry on my dissertation work.

Subham Kumar
(MGCU2021CSIT4029)
M.Tech(CSE)

List of Publications

1. Classification of Fruit Plants Leaf and Comparative Analysis of Machine Learning Algorithms
International Conference on Networking, Communication and Computing Technology (ICNCC-2022), indexed by Scopus and ESCI.
Authors - **Aaryan Kumar Hrithik** and Vipin Kumar

2. Classification of Fruit Plants Leaf and Comparative Analysis of Machine Learning and Deep Learning Algorithms
IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS-2022), IEEE Xplore.
Authors - **Aaryan Kumar Hrithik** and Vipin Kumar

Contents

List of Figures

List of Tables

List of Abbreviations

ML	**Machine Learning**
DL	**Deep Learning**
CNN	**Convolutional Neural Network**
EPA	**Environmental Protection Agency**
Lc	**Low spatial correlation prediction sites**
Hc	**High spatial correlation prediction sites**
GA	**Genetic Algorithm**
GRU	**Gated Recurrent Unit**
LSTM	**Long Short Term Memory**
BiLSTM	**Bi-directional long short term memory**
TL	**Transfer Learning**
GNN	**Graph Neural Network**
RNN	**Recurrent Neural Network**
MLR	**Multivariate linear regression**
ANN	**Artificial neural network**
CLSTM	**Classic Long Short Term Memory**
$PM_{2.5}$	**Particulate Matter 2.5**
CPCB	**Central Pollution Control Board**

List of Symbols

D	**Dataset**
x_i	**Individual image**
Y_i	**Corresponding label**
I_i	**Associated label**
k	**Number of sample**
M	**Data matrix**
ω	**Dimention**
$\varnothing$	**ML Classifier**
$P_i^{\varnothing}$	**Prediction**
m	**Number of test sample**
$\mathcal{L}$	**Zero loss function**
ψ	**DL model**
δ	**Error**
α	**Accuracy**

Dedicated
to
Maa, and Papajee

Chapter 1

Forecasting $PM_{2.5}$ of Some Indian Cites Using Deep Learning Techniques

1.1 Introduction

Pollution is defined as the presence or introduction of dangerous chemicals into the environment that have a negative impact on living creatures and the natural world. Pollution may take many forms, including air pollution, water pollution, soil contamination, and noise pollution, and it can be produced by a variety of human activities such as industrial operations, transportation, and agriculture. Air pollution, for example, can cause respiratory issues, heart illness, and cancer, whilst water pollution can cause disease spread and harm to aquatic ecosystems. Soil pollution can have an impact on agricultural and food safety, while noise pollution can cause hearing loss and other health issues. Pollution is a huge worldwide problem that demands quick action to alleviate its detrimental impact on the environment and human health.

The presence of dangerous compounds in the air, which can have a detrimental influence on human health and the environment, is referred to as air pollution. These air contaminants can originate from both human and natural sources. Particulate matter (PM), nitrogen oxides (NOx), sulphur oxides (SOx), carbon monoxide (CO), volatile organic compounds (VOCs), and ozone are some of the

most frequent air pollutants (O3). PM is composed of microscopic particles that can be inhaled and cause respiratory issues. NOx and SOx are emitted by the combustion of fossil fuels and can contribute to acid rain and smog.CO is a toxic gas that can induce headaches, nausea, and dizziness. VOCs are emitted by a variety of sources, including paint, cleaning chemicals, and gasoline, and can contribute to the creation of ozone, which can cause respiratory difficulties. Overall, air pollution is a severe problem that requires both individual and governmental effort to mitigate its detrimental consequences.

$PM_{2.5}$ refers to microscopic particles in the air that are less than 2.5 microns in diameter, and can be particularly detrimental to human health as they can penetrate deep into the lungs. $PM_{2.5}$ pollution is a major problem in many metropolitan areas across the world, particularly in developing nations with a lack of legislation and infrastructure to control emissions. Transportation, industry, and the combustion of biomass for cooking and heating are all causes of $PM_{2.5}$ pollution.Governments may undertake measures to minimise emissions from these sources, such as boosting public transportation, shifting to cleaner energy sources, and enforcing tougher emissions limits for industry, to reduce $PM_{2.5}$ pollution. People can also take actions to decrease their exposure to $PM_{2.5}$, such as utilising air purifiers, wearing masks, and limiting outdoor activity during periods of heavy pollution. We can help lessen the adverse impacts of $PM_{2.5}$ pollution on human health and the environment by following these activities.

1.2 Literature Review

in the paper by Drewil Data Acquire from Kaggle. This data is a multivariate time series. serise hourly data. In this paper, hyperparameter settings are used as input time steps. 24; epoch 200; batch size 32; learning rate 0.0001; and optimizer Adam. and prose model compare with BiLSTM and CLSTM and get RMSE-9.582 **drewil2022air**.

in the paper By Sarkar Data Acquire from CPCB (India). this data is multivierot time-series Daliy data. This data contains approximately 2482 samples. and When compared to LR, KNN, SVM, and LSTM, the proposed model has a RMSE of 52.03 **sarkar2022air**.

In the paper by Nath Data Acquire from CPCB (India), the acquired data is indian city of Kolkata. This is multivariate time series data. approximately 4 years data held in the dataset and the proposed model compare with stacked LSTM, bi-LSTM, and Convolutional LSTM and get RMSE-18.8 **nath2021long**.

J. Ma uses a transfer learning approach in this paper. unvierot time series hourly data used in this paper. dataset holds three years of data. and the proposed model compare with ARIMA, RNN, LSTM, and CNN-LSTM, and get RMSE-8.65 **ma2019improving**.

In the paper of X. Li Data Acquare from the Ministry of Environmental Protection Beijing,China . This data is univieriot time serise Hourly data. approximately 2 years data held in the dataset and proposed model compare with LSTM-NN, ARMA, and SVR And get RMSE-12.6 **li2017long**.

D. Qin uses a hybrid model approach in this paper. unvierot time series hourly data used in this paper. and proposed model compare with BP, CNN, RNN, and LSTM. get RMSE-14.3 **qin2019novel**.

In the paper of Shengdong Du Data Acquare UCI repository. This data is univieriot time series Hourly data. Approximately 4 years of data are held in the dataset. and The proposed model compares with CNN, RNN, LSTM, and GRU and gets a RMSE of 77.38 **du2019deep**.

"Jing Li Data Acquired from the US EPA," according to the paper. this data is univieriot time serise Hourly data. Approximately nine months of data are held in

the dataset **li2023nested**.

In the paper of C. Erden Data acquired from the Kathane Observation Center (Istanbul). This data is univieriot time serise Hourly data. Approximately 4 years of data hold in dataset. and proposed model compare with LSTM, CNN-LSTM, BiLSTM, ANN and CNN, and get RMSE-4.53 **erden2023genetic**.

In the paper of Mingying Zhu, Data Acquired from Nine Air Quality Monitoring stations,Shanghai(China). This is a univariate time series of hourly data. aproxmatly Six years of data are held in the dataset. and propose a model to compare with BiLSTM, and get RMSE-4.24 **zhu2023investigation**.

In the paper of So-Young Park, Acquired Data from Gwangyang Container Port (Korea). This is a multivariate time series of hourly data. Approximately 6900 samples hold in dataset. and proposed model compare with RNN and LSTM **park2023predicting**.

In the paper by Beytullah Eren Data acquired from Kathmandu Air Quality Monitoring Station (Istanbul). This is a multivariate time series of hourly data. Approximately 4 years of data are held in the dataset. and proposes a model to compare with the LSTM, GRU, RNN, and LSTM+GRU, and get RMSE-4.785 **eren2023predicting**.

In the paper of Beytullah Eren Data Acquire from the UCI repository This data is Multivariate time series Hourly data. Approximately 4 years of data are held in the dataset. and The proposed model is compared with BiLSTM, BiLSTM-GRU, and CNN-BiLSTM. And get RMSE-17.2 **hu2023optimized**.

Table **??** Summarises of some Related works performed and evluated to pridict $PM_{2.5}$.

Paper	Proposed Model	Data Source	Forecasting Object	Benchmark Models	Results
G.I.Drewil et al (2022) **drewil2022air**	GA−LSTM	Kaggle (Hourly data) (INDIA) { 2017-2020}	$PM_{2.5}$, PM_{10}, CO, NO_x	Bi-LSTM, CLSTM, WLSTME	RMSE=9.582, MAE=19.164
N. Sarkar et al (2022)**sarkar2022air**	LSTM-GRU	CPCB (Daily Data) (Delhi India) {Mar.2015-dec.2021}	NO, NO_2, NO_x, SO_2, CO,O_3, $PM_{2.5}$	LR, KNN, SVM, LSTM, GRU	R^2=0.84, RMSE=52.03, MAE=36.11
P. Nath et al (2021)**nath2021long**	LSTM Auto-Encoder Based	CPCB (Daily Data) (Kolkata India) {Jan 2016-Feb 2020}	$PM_{2.5}$, PM_{10}	Stacked LSTM, Bi-LSTM, Convolutional LSTM	RMSE=18.8±0.19, MAE=15.88±0.19
J. Ma et al(2019)**ma2019improving**	TL-BiLSTM	Guangdong province (Hourly data) (Guangdong China) {3 year}	$PM_{2.5}$	ARIMA, RNN, LSTM, CNN-LSTM	RMSE=8.6529, MAE=6.184, MAPE=27.909
X. Li et al (2017)**li2017long**	LSTME (LSTM NN extended)	Ministry of Environmental protection (Hourly Data) (Beijing China) {Jan 2014-May 2016}	$PM_{2.5}$	LSTM-NN, ARMA, SVR	RMSE=12.6, MAE=5.46, MAPE=11.93
D. Qin et al (2019)**qin2019novel**	CNN +LSTM	(Daily Data) (Shanghi) {2015-2017}	$PM_{2.5}$	BP, CNN, RNN, LSTM	RMSE=14.3
Shengdong Du et al (2019)**du2019deep**	DAQFF	UCI repository (Hourly Data) (Beijing China) {2010-2014}	$PM_{2.5}$	CNN, RNN, LSTM,GRU	RMSE=77.38, MAE=54.58

Paper	Proposed Model	Data Source	Forecasting Object	Benchmark Models	Results
Jing Li et al (2023)**li2023nested**	GNN-LSTM	US EPA (Hourly Data) (North America) {jan 2021-sep 2021}	$PM_{2.5}$		MAE(Lc)=3, MAE(Hc)=2.81
C. Erden et al (2023)**erden2023geneticCNN**	LSTM-GRU-CNN	Kathane observation center (Hourly Data) (Istanbul) {2015-2019}	$PM_{2.5}$	LSTM, CNN-LSTM, BiLSTM, ANN, CNN	RMSE=4.53, R^2=0.91
Mingying Zhu et al (2022)**zhu2023investigation**	1D-CNN-biLSTM-v2	nine air quality monitoring stations (Hourly Data) (Shanghai China) {2014-2020}	$PM_{2.5}$	BiLSTM	RMSE=4.2489, MAE=2.7198, R^2=0.9297
So-Young Park et al (2023)**park2023predicting**	MLR	Gwangyang Container Port (Hourly Data) (Korea) {Sep 2020-Dec 2021}	$PM_{2.5}$, PM_{10}	RNN, LSTM	$PM_{2.5}$-R^2=0.434, PM_{10}-R^2=0.352
Beytullah Eren et al (2023)**eren2023predicting**	LSTM+LSTM	Kathane Air Quality Monitoring Station (Hourly Data) (Istanbul) {2015-2019}	$PM_{2.5}$, SO_2, NO, NO_2, NO_X, O_3	LSTM, GRU, RNN, LSTM+GRU	MAE=3.086, RMSE=4.785, R^2=0.97
Juntao Hu et al (2023)**hu2023optimized**	CNN-BiLSTM-GRU	UCI repository (Hourly Data) (Beijing China) {2013-2017}	$PM_{2.5}$, O_3	BiLSTM, BiLSTM-GRU, CNN-BiLSTM	MAE=9.8, RMSE=17.2, R^2=0.9589

TABLE 1.1: Summarizing of Related work to pridict $PM_{2.5}$

1.3 Basics Related Roncepts

1.3.1 Machine Learning

One of the Subfield of Artifical Intelligent(AI) is Machine Learning(ML) that focuses on explicitiy Programmed witch devloping algorithms and models that enable computers learn from Data.

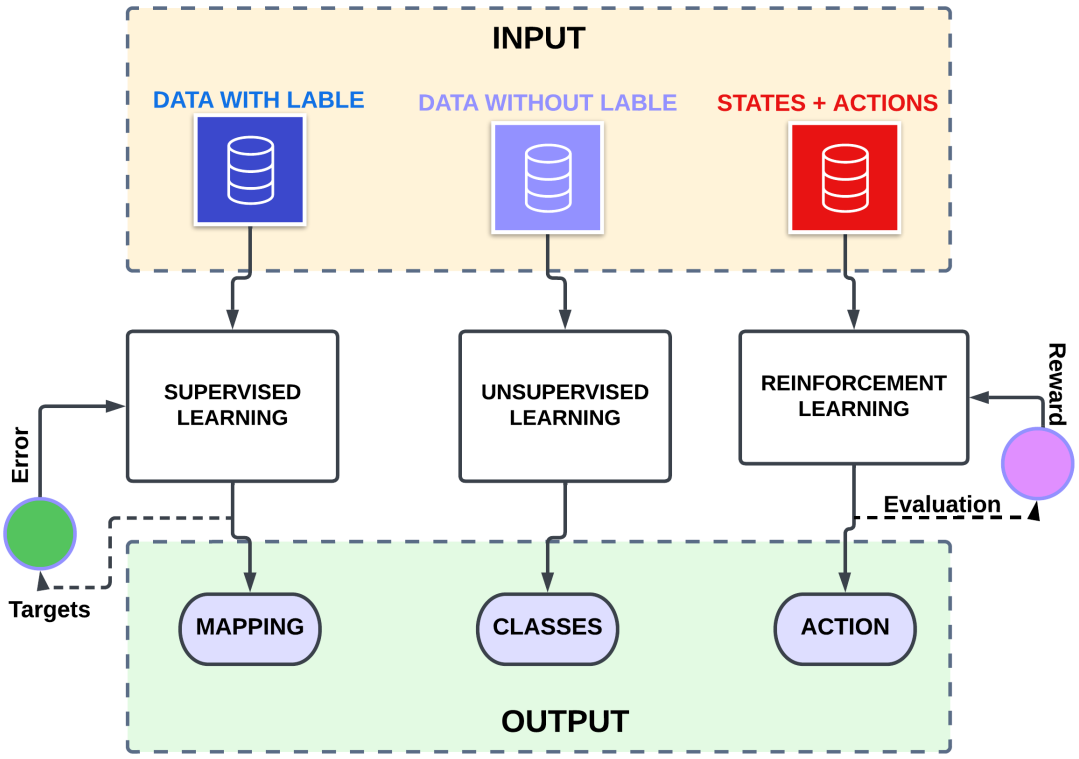


FIGURE 1.1: Architecture of Machine Learning

- **History:** Throughout the 1980s and 1990s, machine learning began to gain momentum as academics built increasingly powerful algorithms and applied them to real-world issues. The emergence of huge data and sophisticated computer resources in the twenty-first century has resulted in an explosion of interest in machine learning, with applications in industries like as healthcare, finance, and transportation.

- **Key Concepts:** Their are three brances of ML: "Supervised Learning, Unsupervised Learning, Reinforcement Learning".
  - **Supervised Learning:** A Supervised learning algorithm learns from labeled data. The objective is to minimise the difference between projected and actual output while learning a mapping function from input variables to output variables. Several domains use supervised learning, including picture and audio recognition, natural language processing, fraud detection, and recommendation systems.**geron2022hands**
  - **Unsupervised Learning:** The purpose is to find patterns and organisation in the data, such as clustering similar data points together or lowering data dimensionality. Several domains use unsupervised learning, including image and video analysis, anomaly detection, and market segmentation.**geron2022hands**
  - **Reinforcement Learning:** Reinforcement learning seeks to discover a strategy that maximises a cumulative reward signal over time. Reinforcement learning offers a wide range of applications, including robots, games, and autonomous vehicles.**geron2022hands**

1.3.2 Time Series

On the basices of regluar time intervals collected sequencel data point is Known as time series Data. in the time series analysis statistical technique using for examin the trends and patterns.

- **Key Concepts :**
  - Time series data collected over time.
  - Time series data has use time column use as index.
  - In time series analysis stationarity is an importent concept.
  - In time series data presence of regular patterns repeat over fixed periods of time in cycles is called Seasonality.

– On the Basics of time series data predicting future values as known as Forecasting.

1.3.3 Deep Learning for Time-Series

The term "deep learning" was initially used in the mid-2000s to describe neural networks with numerous layers of hidden units that were shown to be more successful at solving complicated tasks.

Deep learning advances began in 2006, when Geoffrey Hinton, Yoshua Bengio, and Yann LeCun demonstrated that deep neural networks outperformed standard machine learning approaches in key benchmark tasks, including image and speech recognition. Deep learning has gotten a lot of attention since then and has become a dominating method in several domains, including computer vision, natural language processing, and speech recognition.

The availability of massive datasets and strong computer resources was a crucial component in deep learning's success. Deep learning models may now be trained on millions or billions of instances because to the availability of huge quantities of data and advancements in technology, resulting in considerable increases in accuracy and performance.

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs) are three of the most common deep learning models.

Deep Learning Algoritham for Time-Series

- **CNN:** A convolutional neural network (CNN) is a form of neural network that is extensively employed in signal processing and time series analysis. Unlike standard fully connected neural networks, which process inputs as vectors, 1D CNNs use a sliding window to extract local features via convolution operations.

  A one-dimensional signal, such as audio or a time series of sensor measurements, can be used as the input of a 1D CNN. Several filters are used to extract features from the input signal, while pooling layers are used to downsample the output of the convolutional layers. Lastly, the network might include one

or more fully connected layers to perform classification or regression tasks. Simple Architecture of 1D CNN in Figure **?? rticle**.

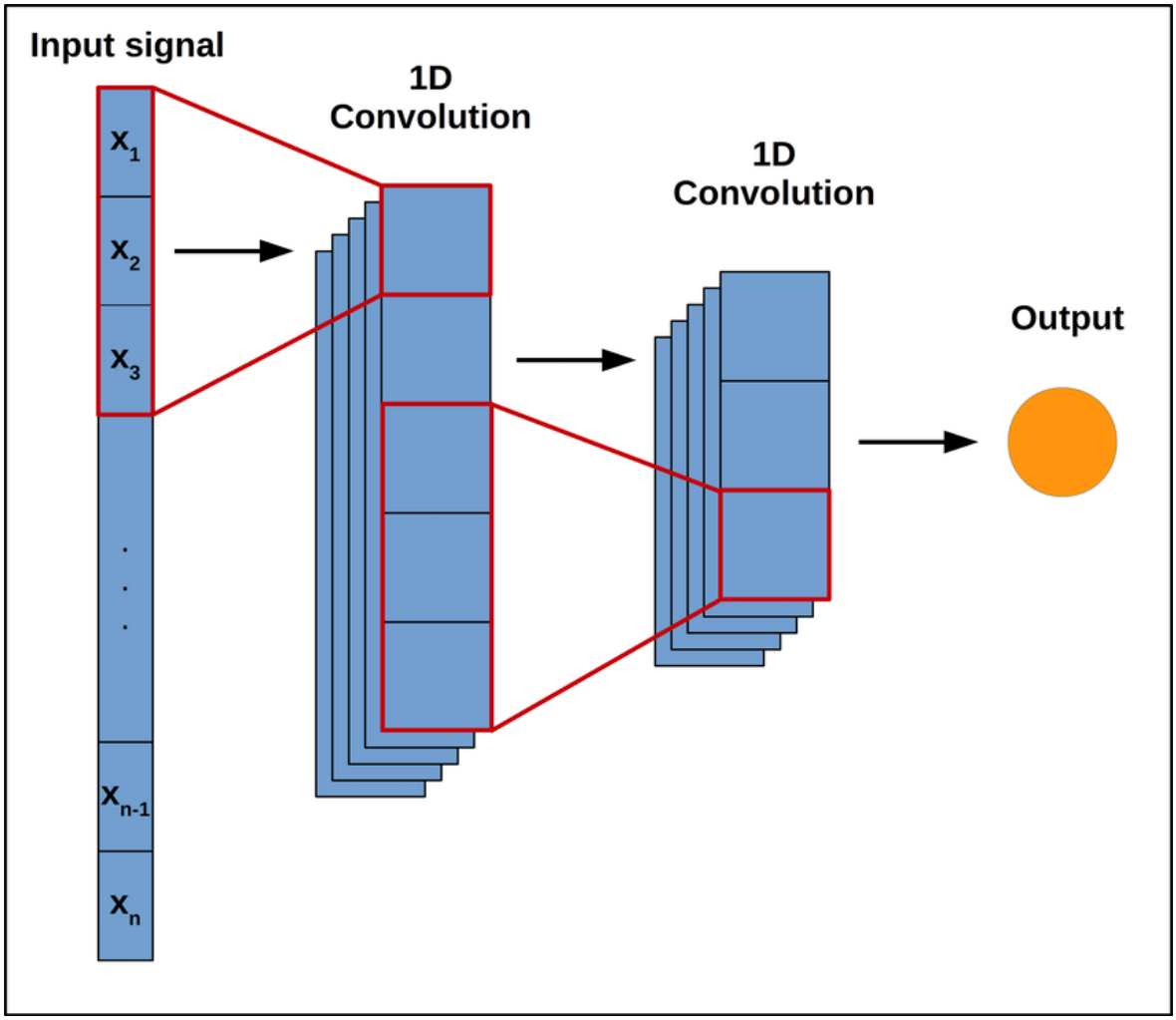


FIGURE 1.2: Simple 1D CNN architecture with two convolutional layers.

- **GRU:** GRU (Gated Recurrent Unit) is a recurrent neural network (RNN) architecture used for sequential data processing. Cho et al. introduced GRU, a variation of the more extensively utilised Long Short-Term Memory (LSTM) network, in 2014.

The use of gating methods to regulate the flow of information via the network is a major element of GRU. These gating mechanisms govern the amount of information that is preserved and discarded at each time step, allowing GRU to selectively update its hidden state based on the current input and the prior hidden state. Architecture of GRU in Figure **?? phdthesis**.

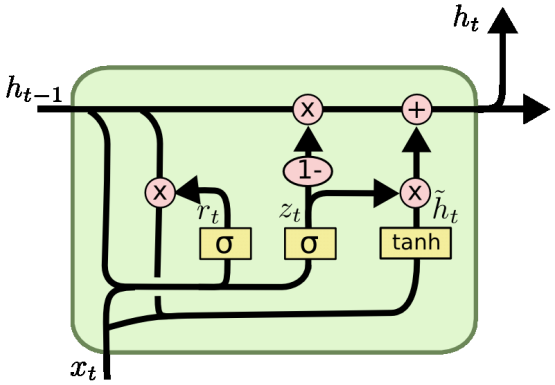


FIGURE 1.3: Architecture of Gated Recurrent Unit (GRU).

- **LSTM:** Long Short-Term Memory is an example of a recurrent neural network (RNN) architecture, and its name is LSTM. LSTMs are very helpful in natural language processing (NLP) jobs because they are built to handle long-term dependencies in sequential input. The ability of LSTMs to keep a long-term memory of prior inputs and selectively update this memory using a gating mechanism is its distinguishing characteristic. Input, forget, and output gates make up this gating system, which regulates the information flow into and out of the LSTM. Overall, LSTMs have shown to be an effective method for handling sequential data and have been used for a variety of NLP applications, such as sentiment analysis, speech recognition, and machine translation. Architecture of LSTM in Figure **?? drewil2022air article**.

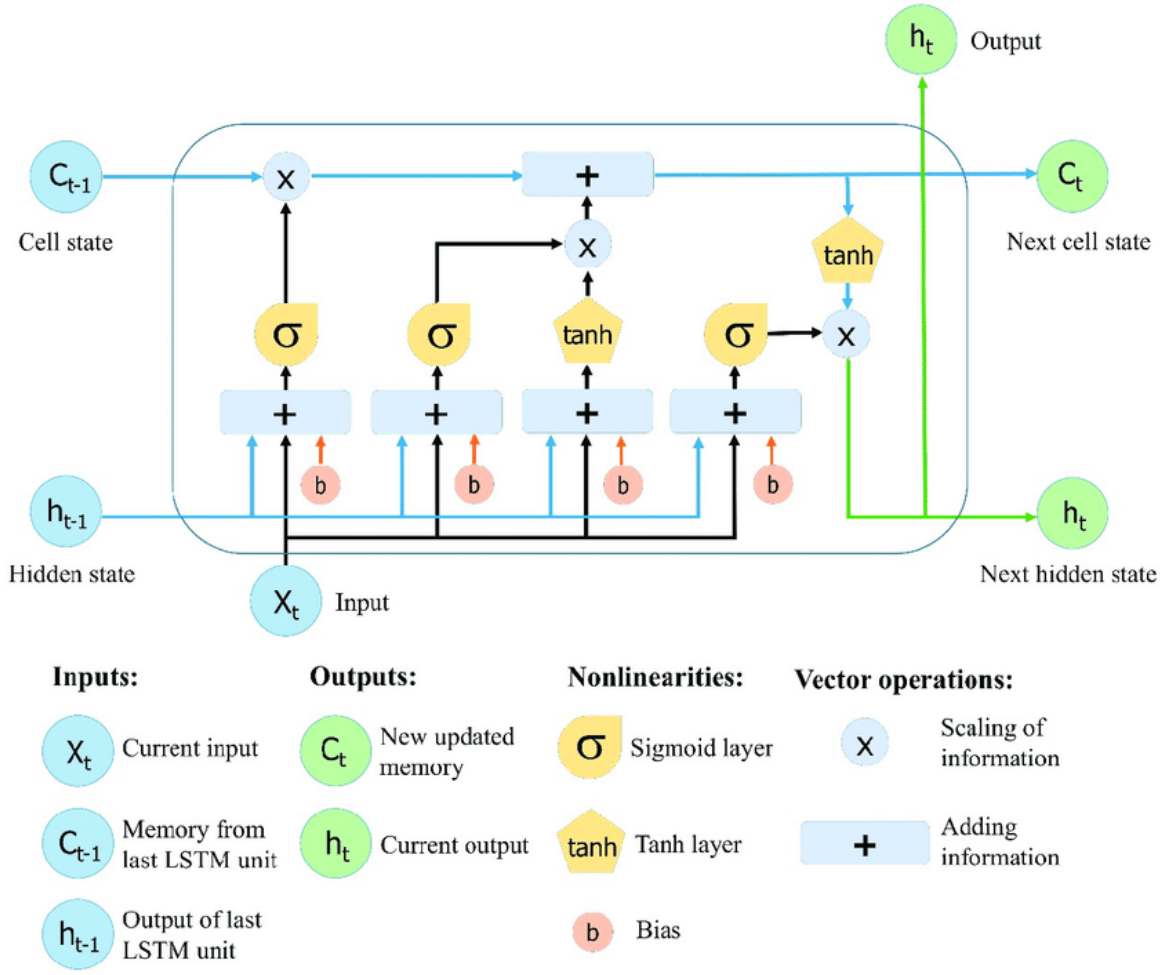


FIGURE 1.4: Architecture of LSTM

Accuracy Measures for Time-Series We are use some accuracy measures in this report. e.t -RMSE,MAE and MAPE.

- **RMSE :** The RMSE (Root Mean Square Error) statistic is often used to assess the accuracy of a regression model. It computes the root mean square of the discrepancies between the expected and actual numbers to calculate the deviation between the two. RMSE calculates the average distance between expected and actual values in the same units as the data. A lower RMSE number suggests greater model performance, whereas a larger RMSE value indicates poorer model performance. It is a widely used statistic in many

industries, including engineering, finance, and data science.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} equ : RMSE \tag{1.1}$$

where N is the number of data points, y(i) is the i-th measurement, and $\hat{y}(i)$ is its corresponding prediction.

The root mean square error (RMSE) is a helpful statistic for evaluating the effectiveness of regression models since it measures how distant the predictions are from the actual values in the same units as the data. A lower RMSE number suggests greater model performance, whereas a larger RMSE value indicates poorer model performance.

- **MAE :** MAE (Mean Absolute Error) is a popular statistic for assessing the performance of a regression model. It computes the average of the absolute discrepancies between the projected values and the actual values. MAE can be calculated using the following formula:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1.2}$$

where y_i=Prediction,x_i=True value and n=Total number of data points.

MAE measures the average magnitude of the errors in anticipated values in the same units as the data. A lower MAE number suggests greater model performance, whereas a higher MAE value indicates poorer model performance. It is widely utilised in many different disciplines, including banking, engineering, and machine learning. MAE, unlike RMSE, is not affected by outliers in the data.

- **MAPE :** MAPE (Mean Absolute Percentage Error) is a popular indicator for assessing a regression model's performance. It computes the percentage difference between expected and actual values. MAPE is computed by averaging the absolute percentage errors between anticipated and actual

values. MAPE can be calculated using the following formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{1.3}$$

where n=number of times the summation iteration happens , A_t=Actual Value and F_t=Forecast Vaiue.

MAPE calculates the average percentage difference between anticipated and actual values. It is frequently utilised in business and financial applications where the accuracy of projections or predictions must be measured. A lower MAPE number suggests better model performance, whereas a greater MAPE value indicates poorer model performance. Nevertheless, when the real numbers are near to zero, MAPE might cause division by zero mistakes or very big percentage errors. Other measures, such as Symmetric Mean Absolute Percentage Error (SMAPE), may be utilised in such instances.

1.4 Methodology

In this part of the work, we discuss the method of forecasting $PM_{2.5}$ with the help of various deep learning models. CPCB hour univariate data was collected for this purpose. In this data, there is only one future $PM_{2.5}$. then possessed the collected data. Then various deep learning models were implemented, and their forecasts were obtained. Models are used to improve forecasting accuracy. The whole method has been illustrated with the help of a flowchart, shown in Figure **??**. In the process, the steps involved are described below.

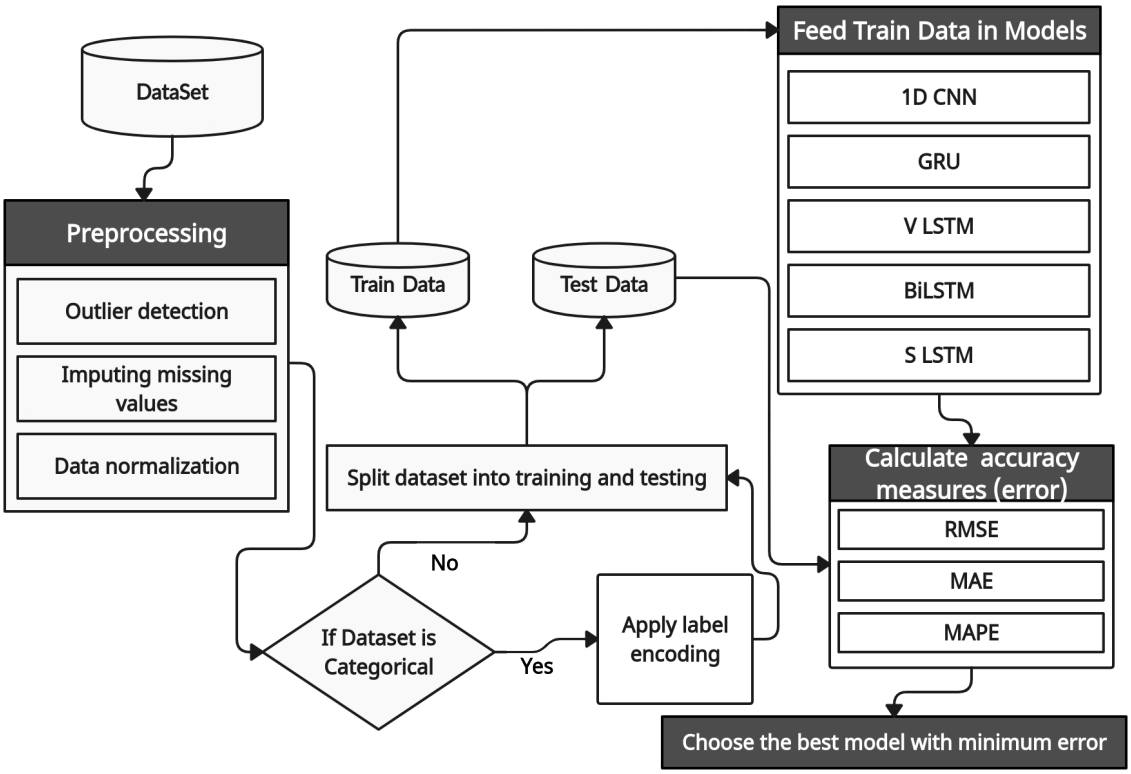


FIGURE 1.5: Process flow of developing on $PM_{2.5}$ prediction model

1.4.1 DataSets

Data collection is the process of collecting and evaluating the information in this study. In 17 datasets used in this study, data was collected from the Indian government portal CPCB (Central Pollution Control Board). The collected data is univariate hourly data. Data is available for 17 Indian cities, with most cities having around 4 years of data available in each dataset. Shown in Table **??**.

DataSets	**No. Of Samples**	**Days**	**Years**
AMBALA	34174	1423.9	3.9
ANKLESHWAR	33535	1397.3	3.8
BHIWADI	43394	1808.1	5.0
BULANDSHAHR	39869	1661.2	4.6
CHARKHI_DADRI	24099	1004.1	2.8
DHARUHERA	34265	1427.7	3.9
DURGAPUR	17434	726.4	2.0
FATEHABAD	34160	1423.3	3.9
HISAR	34143	1422.6	3.9
JIND	34145	1422.7	3.9
JODHPUR	61409	2558.7	7.0
KURUKSHETRA	34208	1425.3	3.9
LUDHIANA	49010	2042.1	5.6
MUZAFFARNAGAR	38786	1616.1	4.4
SINGRAULI	43695	1820.6	5.0
SONIPAT	34362	1431.8	3.9
YAMUNA_NAGAR	34299	1429.1	3.9

TABLE 1.3: 1C-totel Sempls available in DataSets,2C-totel no. of Days available in DataSets, 3C-totel no. of years available in DataSets

1.4.2 Framework

Step 1 :

"Data Input" Data download from CPCB

Step 2 :

Step 3 :

Step 4 :

Step 5 :

1.4.3 Exprement Design :

1.4.4 Exprement Setup :

Hardware

Software

	Cities	Count	Mean	Std	Min	25%	50%	75%	Max			
BHIWADI	43394	108.0337483	79.7575935	0.02	55.22	97.32	135.355	999.99				
JODHPUR	61409	84.30884555	56.17507578	0.18	53.25	84.30884555	93.42	999.99				
SINGRAULI	43695	84.07888484	78.32602035	0.25	32.25	66	111.25	985				
ANKLESHWAR	33535	58.47126979	35.83229258	0.51	32.75	58.47126979	72.24	977.39				
LUDHIANA	49010	54.18070537	41.72936492	0.07	29.7	47.66	64.88	999.99				
DURGAPUR	17434	71.66516634	46.19819069	0.33	37.4725	62.045	98.0275	565.41				
YAMUNA_NAGAR	34299	77.8633741	52.30959419	0.1	43.8	69.91	94.28	930				
CHARKHI_DADRI	24099	80.19422853	62.81330763	0.01	39.535	77.92	94.485	995.1				
JIND	34145	81.20790101	71.19703393	0.2	38.99	61.45	98.25	845.6				
KURUKSHETRA	34208	68.75483661	53.80132654	0.46	33.33	56.38	87.56	962.7				
SONIPAT	34362	54.88284717	43.2094932	0.02	27.87	49.4	62.72	543.1				
DHARUHERA	34265	78.86179491	59.20936284	0.02	40.9	70.32	92.85	838.9				
AMBALA	34174	61.58184731	45.39340564	0.02	32.94	51.27	76.1775	754.89				
HISAR	34143	86.22030417	71.02050965	0.63	42.62	69.33	102.885	999.99				
FATEHABAD	34160	63.01257933	60.46487543	0.07	32.63	49.01	72.5	999.99				
BULANDSHAHR	39869	90.53344553	85.08059871	0.25	34	63.75	120.25	985				
MUZAFFARNAGAR	38786	89.28704635	72.8395854	1	42.75	81.25	102.25	986				

TABLE 1.4: All Datasets Discription.

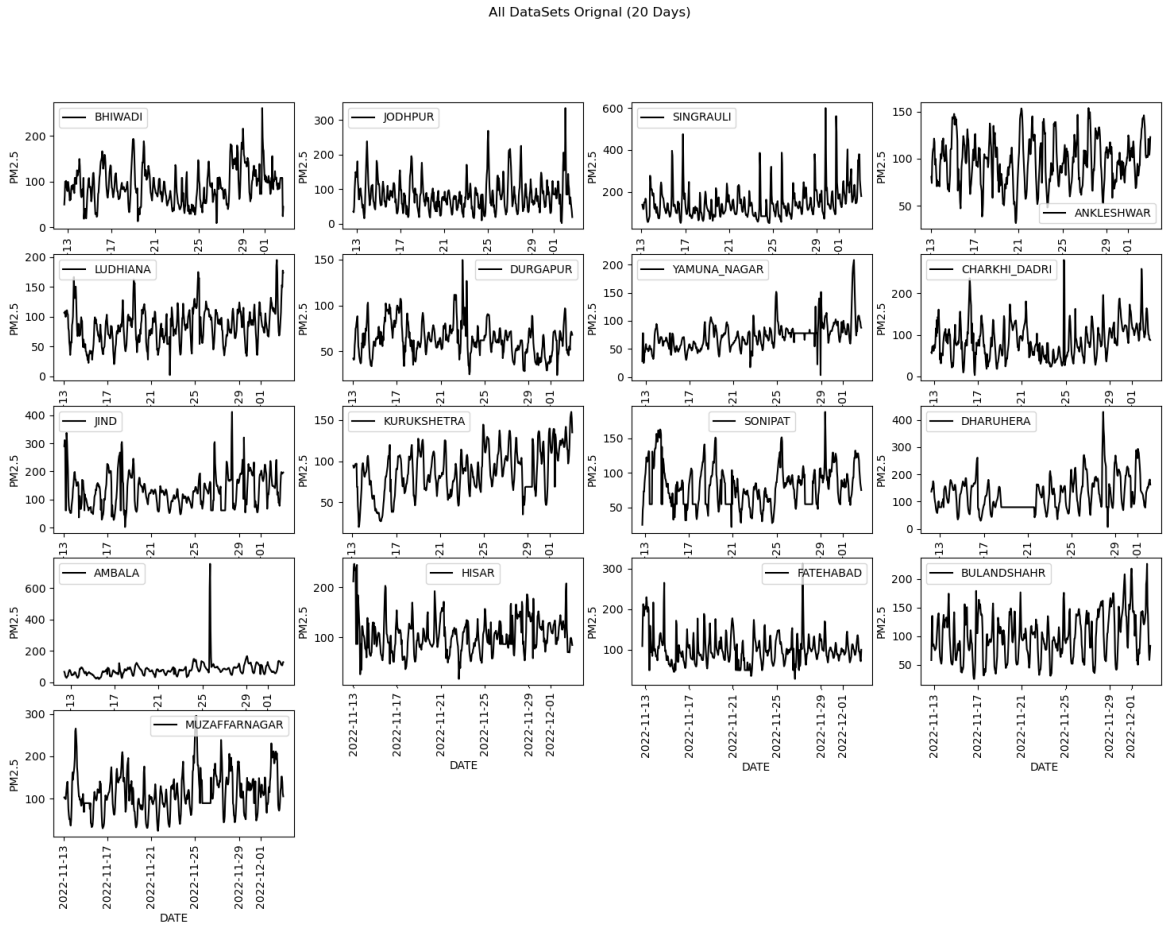


FIGURE 1.6: 20 Days Origanal Data Graph.

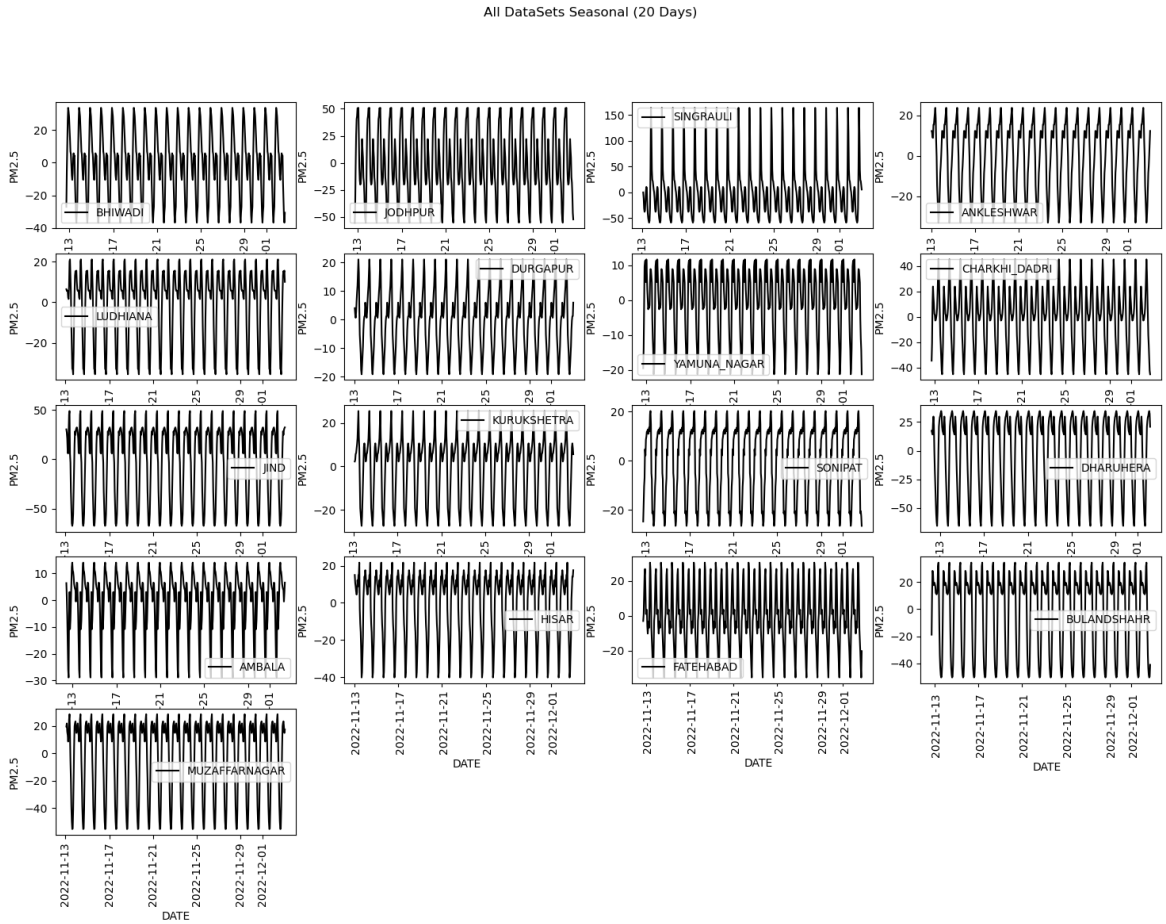


FIGURE 1.7: 20 Days Seasonal Graph.

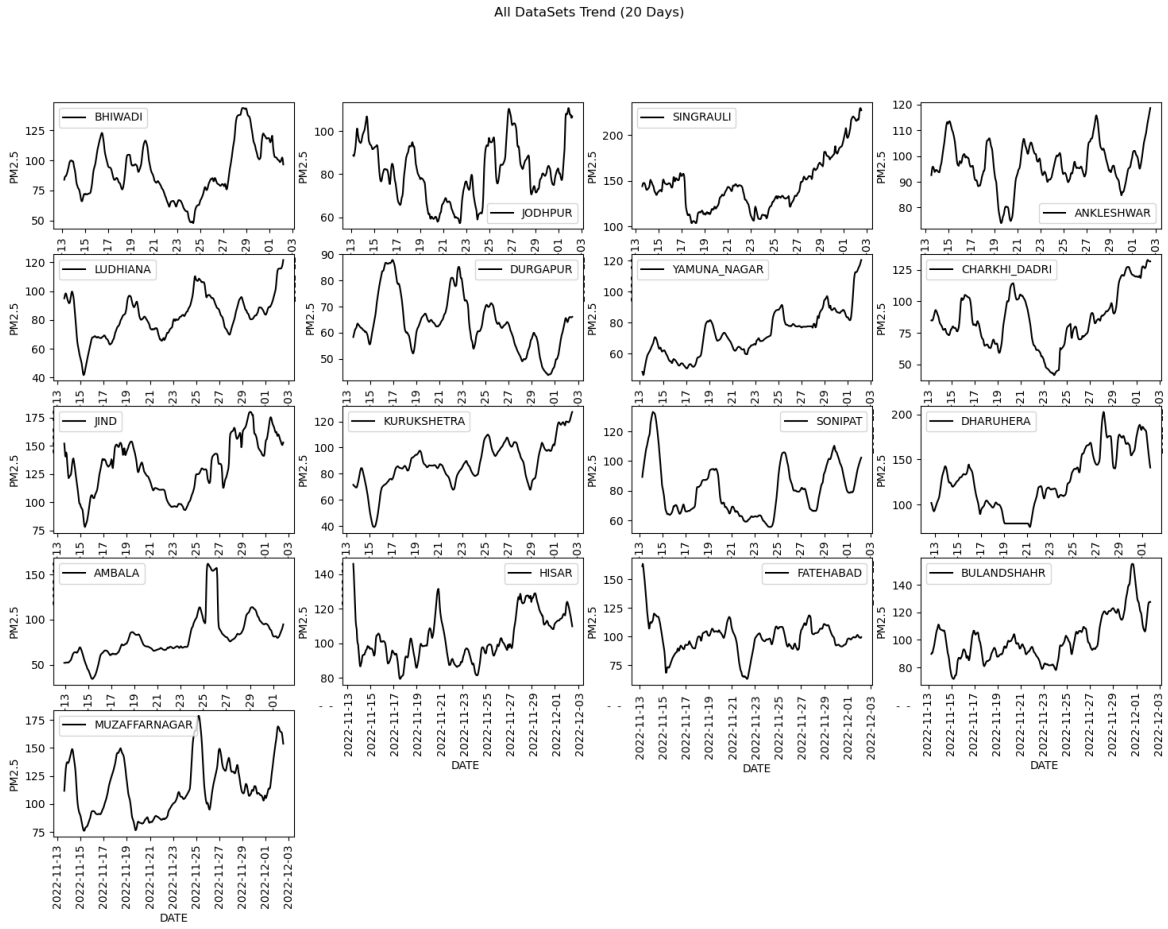


FIGURE 1.8: 20 Days Trend Graph.

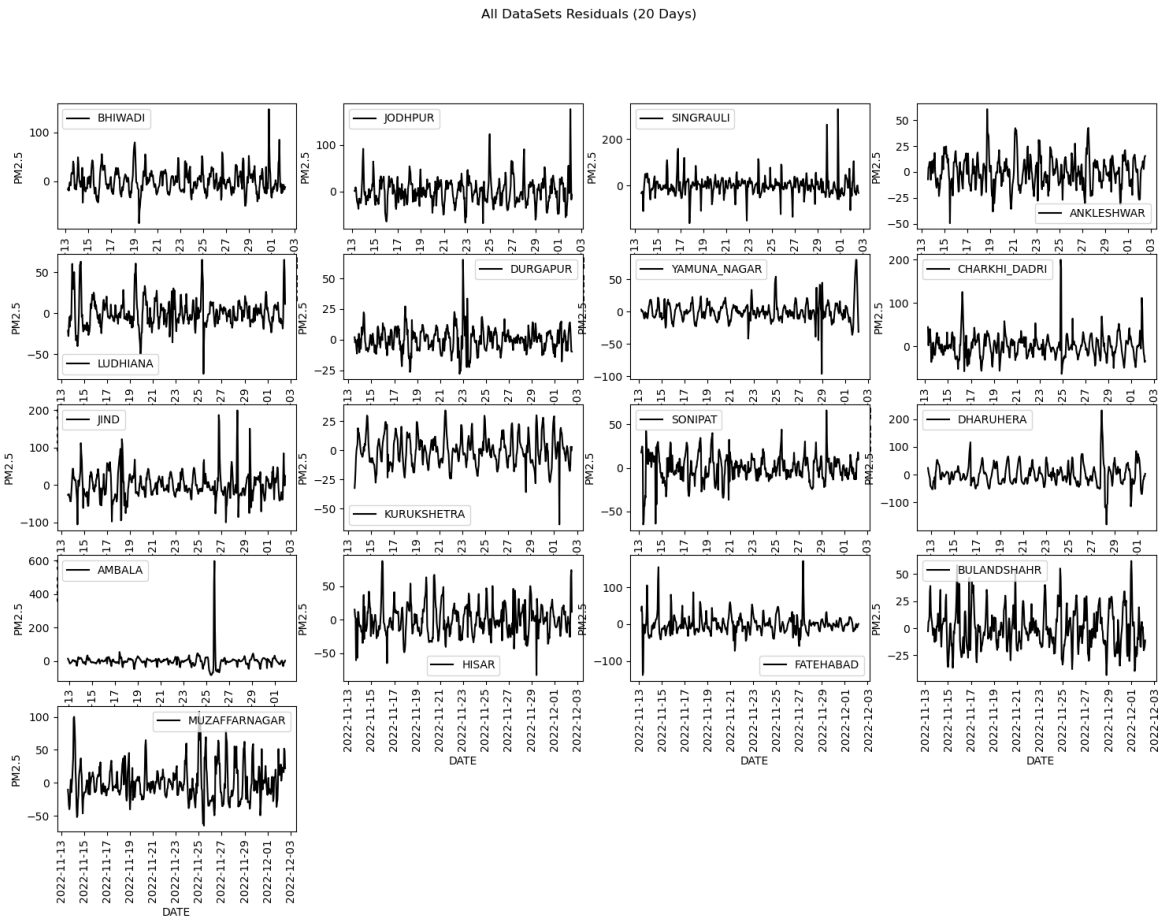


FIGURE 1.9: 20 Days Residuals Graph.

in

Step 1: The Dataset:

Collected fruit plant leaves have been represented as dataset D, shown in equation (1).

$$D = \{x_i, y_i\}_{i=1}^{k} \tag{1.4}$$

Where D is denoted as Dataset, X_i is an individual image of a fruit plant leaf, and Y_i its corresponding label. k denotes the number of samples in the Dataset. For all Y_i belongs to Y; where Y is several classes.

Step 2: Resizing of the Dataset:

dataset D has been resized into a new dimension. The function representing the dimension is omega. shown in equation (2).

$$D_{new} = \omega\left(D, p \times q\right) \tag{1.5}$$

Where D_{new} is the newly resized image of the Dataset of the fruit plant leaf. Here, p x q are the new dimension of the Dataset.

Step 3: Splitting Dataset:

the preprocessed Dataset has been divided into three parts: training (T_{train}), testing (T_{test}), and validation (V_{valid}). The Dataset is mutually exclusive in equation(4) and exhaustive in equation(3).

$$T_{train} \cup T_{test} \cup V_{valid} = D_{new} \tag{1.6}$$

$$\{T_{train} \cup V_{valid}\} \cap \{T_{train} \cup T_{test}\} \cap \{T_{test} \cup V_{valid}\} = \varnothing \tag{1.7}$$

Step 4: Feature Extraction using PCA:

Principal component analysis (PCA) is the technique of computing the principal components and utilizing them to modify the foundation of the data, often simply using the first few and disregarding the rest **jolliffe2002springer**. To obtain a linear transformation that determines the directions of the highest variance data. The axes with higher explained variance produce a distribution that assists the separation of classes when sample patterns are projected onto a high-dimensional feature space. A larger explained variance is expressed by the axis with the highest eigenvalues.

$$D_{PCA} = \omega\left(D, p \times q\right) \tag{1.8}$$

where n is the number of features after extraction.

Step 5: Applying Machine Learning Classifiers:

the classifier $\emptyset\,(T_{train}\,,\,V_{valid})$ has been applied. In the classifier's training and validation part, the prediction has been obtained for the test sample. Prediction is obtained:$\left\{P_i^{\emptyset}\right\}_i^m$, where m is the number of test samples **omerintroduction**.

Step 6: Accuracy:

Accuracy can be calculated with zero-one loss function equation(6).

$$acc \;=\; \mathcal{L}\left(\{y_{test_i}\}_{i=1}^{k}\,,\,\left\{P_i^{\emptyset}\right\}_{i=1}^{k}\right) \tag{1.9}$$

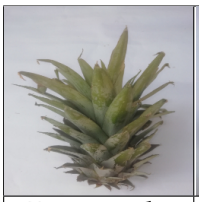
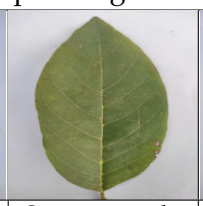
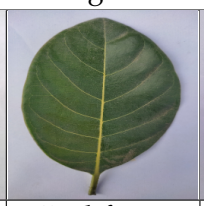
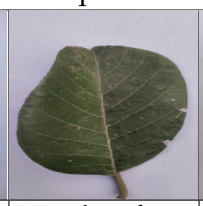
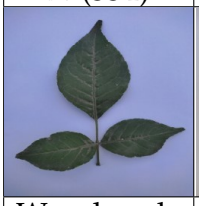
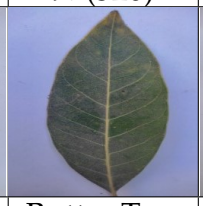
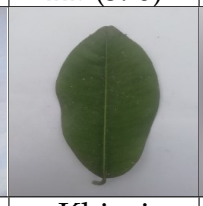
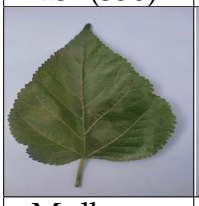
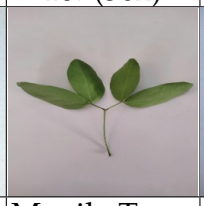
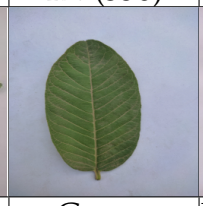
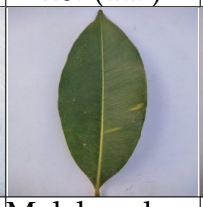
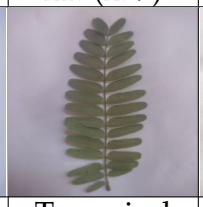
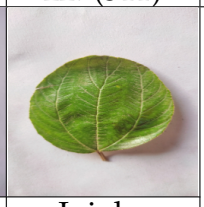
Where *y_test* is test Sample, $P_i^{\emptyset}$ is a prediction of classifier $\emptyset$, and k is the number of test samples.

1.5 Experiments

1.5.1 Dataset Description

In our experiment, we utilized 300 samples randomly picked from each species as average precision to evaluate the effectiveness of leaf image retrieval. The leaves were chosen to include a variety of sizes, from the tiniest to the largest. For all the selected leaves, Numerical images were taken with a smartphone camera. The micro lens was situated in the center of its path to reduce image distortion as much as possible. High-quality JPEG images were shot at a distance of 10 - 25 cm from the leaves. For better analysis, we utilized a consistent white background. Depending on the leaf size, one to numerous leaves per photo were recorded. Afterward when the size of each leaf was calculated using image analysis of the entire pixel region. Leaf length was measured using Optimal software to assess the numerical photograph method's robustness compared to the measured leaf length **leroy2007**. To test the accuracy of our models, we utilized 300 pieces of leaves from testing sets for each category of plant.

TABLE 1.5: Sample images of 28 categories of fruit plants leaf

					
Pineapple *(Ananas Comosus)* 1. (356)	Custardapple *(Ananas Reticulata)* 2. (361)	Sugarapple *(Annona Squmosa)* 3. (335)	Jackfruit *(Artocarpus Heterophyllus)* 4. (302)	Monkeyfruit *(Artocarpus Lacucha)* 5. (356)	Key Lime *(Citrus Aurantiifolia)* 6. (306)
					
Lemon *(Citrus Limon)* 7. (334)	Pummelo *(Citrus Maxima)* 8. (422)	Ponderosa *(Citrus Pyriformis)* 9. (320)	Assyranplum *(Cordia Myxa)* 10. (388)	Cluster Fig *(Ficus Racemosa)* 11. (390)	Phalsa *(Grewia Asiatica)* 12. (312)
					
Woodapple *(Limonia Acidissima)* 13. (356)	Lychee *(Litchi Chinesis)* 14. (361)	Butter Tree *(Madhuca Longfolia)* 15. (335)	Mango *(Mangifera Indica)* 16. (302)	Khirni *(Manilkara Hexandra)* 17. (356)	Sapodilla *(Manilkara Zapota)* 18. (306)
					
Mulberry *(Morus Alba)* 19. (312)	Kadam *(Neolamarckio Cadamba)* 20. (447)	Gooseberry *(Phyllanthus Emblica)* 21. (279)	Manila Tama *(Pithecellobu Dulce)* 22. (311)	Guava *(Psidium Guajava)* 23. (418)	Pomegranate *(Punica Granatum)* 24. (333)
					
Tomato *(Solanum Iycopersicum)* 25. (297)	Malabarplum *(Syzygium Jambos)* 26. (357)	Tamarind *(Tamarindus Indica)* 27. (492)	Jujube *(Zizphus Mauritiana)* 28. (345)		

We compared our algorithm's accuracy to that of existing leaf-shape-only different classifiers. Our algorithm's accuracy is comparable to other approaches.

1.5.2 Experiment Design

Figure 1.1 is the experimental flowchart, which shows the whole process from data collection to the classification of a fruit plant leaf. The first phase is the data set. In this, we collected more than 300 images of each category of fruit plant.

After the data collection, we went for the data preprocessing phase, in which the collected images were cropped, resized, and organized based on their class. First, we collected more than 10000 images of fruit plant leaves. Then the images were cropped to remove the data images' abnormalities as much as possible. After this, the images were resized into 250X250 size, and each category of images was stored in a separate individual folder with their names.

Different classifiers have been applied to the comparative analysis and the best performing classifier has been found. Once the data is preprocessed, it will be ready for the ML program to do the execution, and the features of these leaf images have been extracted in the feature extraction part. After feature extraction, each class is divided into train, validation, and test set by the algorithm. Different classifiers had applied one after another to get the performance in accuracy, precision, recall, etc. And lastly, in the fruit plant classification stage, we were able to do the classification based on the result obtained from the previous stage.

1.6 Results and Analysis

1.6.1 Results

The Bar Plot showed in Figure 1.2. It has shown the classification performance in the form of the accuracy of six classifiers, names, LR, KNN, SVM, DT, MLP, and NB. In this Figure, the x-axis represents the name of the classifiers, and the y-axis represents their respected accuracy. For a clear view, the accuracy of each classifier has been shown just below their names. While Figure 1.3. It shows the Box Plot

of all the same six classifiers. Here the x-axis represents the name of the classifiers, and the y-axis represents their respected accuracy. The classification performance (Precision, Recall, f1-score, support) of the best classifier (MLP) has been shown in table-1.2

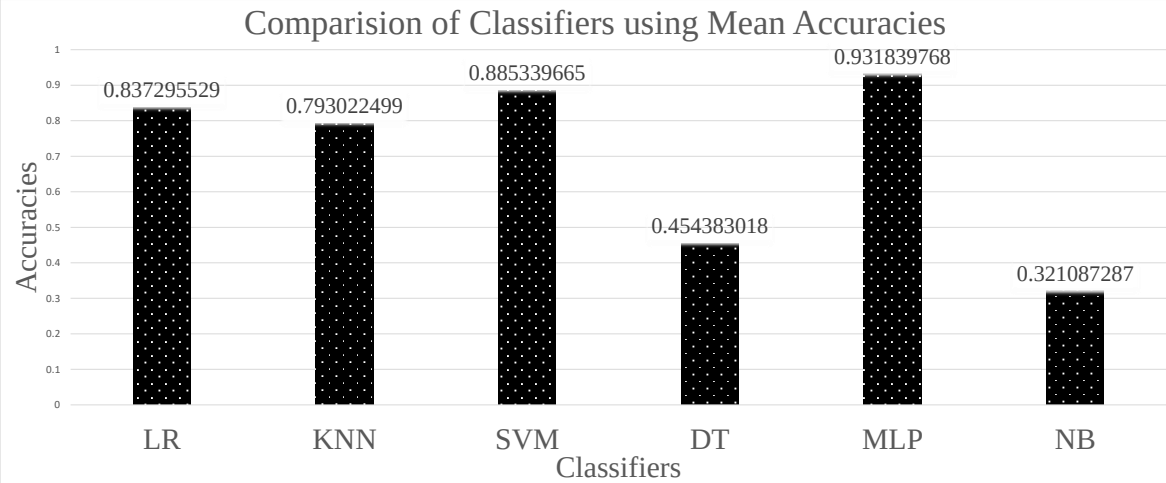


FIGURE 1.10: Barplot of fruit leaf image classification

Figure-1.4 is the confusion matrix using a heat map for our herb's Dataset. It shows the classification and misclassification amongst different categories of Fruits image plant [19], [20]. The x-axis and y-axis bot contain the name of the fruit plants of which classification has been done in this paper. The color range of this heat map varies from 0 to 4. For the diagonal elements, which starts from left-top to right bottom, the brighter (closer to 4) color of a cell shows that the category has that much better correlation with itself. The classification has been done by the classifier that much successfully, but if these diagonal cells' color is darker (closer to 0), much misclassification has occurred. But for the cells except for this diagonal cell, the story will be just the opposite. That is darker (closer to 0). The color shows that lest the correlation of that class with corresponding another type of fruit.

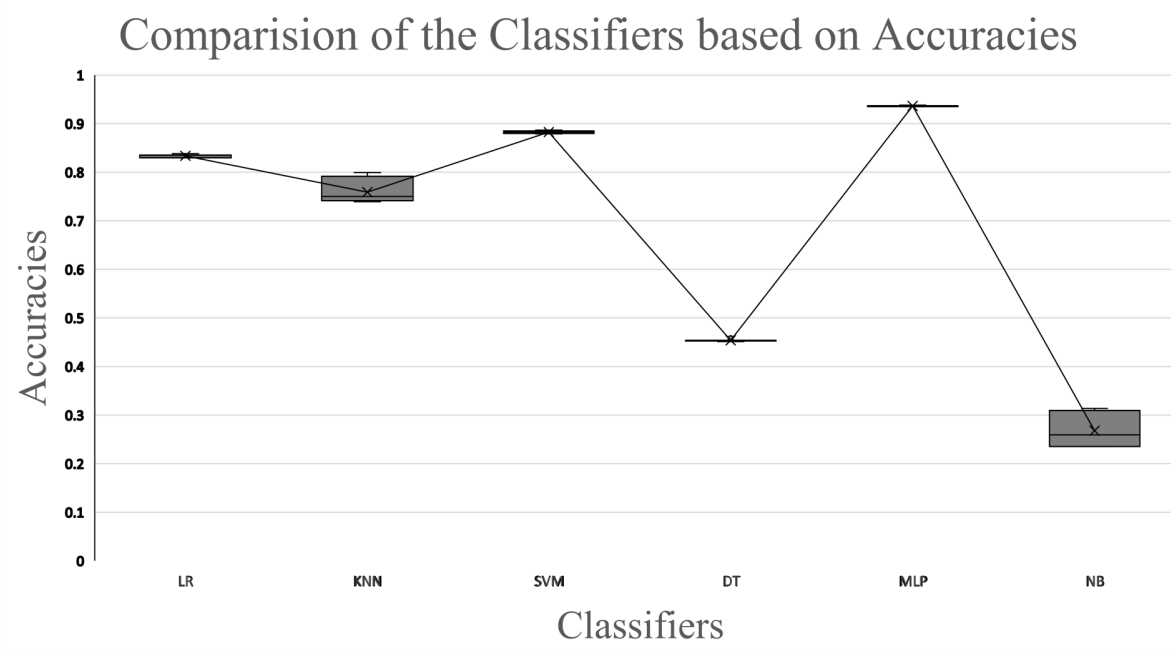


FIGURE 1.11: Boxplot of fruit leaf image classification

1.6.2 Analysis of Results

Classification comparison using bar plot:

As we can see from the bar plot in Figure-2, Linear Regression (LR) has 83.40%, k-Nearest Neighbours (KNN) has 74.52%, Support vector Machine (SVM) has 88.18%, Decision Tree (DT) has 45.92%, Neural Network (MLP-Multi-layer perceptron) has 93.18%, and the Naïve Bayes (NB) has 26.54% classification accuracy. So, on basis of classification accuracy if we will choose MLP, which has the maximum accuracy, then this classifier for the classification of herbal plants will give the best result as compared to other classifiers. And the NB will give the worst classification report because it has the least accuracy.

Classification comparison using box plot:

The evaluation criteria for the analysis of a box plot are Max Maximization, Min Maximization, Size of a box plot, and the median of the box plots. From Figure-3

TABLE 1.6: Performance based on Precision, Recall, F1-score, and Support

Category	Precision	Recall	F1-score	Support
1	0.95	0.97	0.96	86
2	0.93	0.9	0.91	98
3	0.87	0.96	0.91	78
4	1	0.94	0.97	66
5	0.94	0.94	0.94	93
6	0.88	0.91	0.89	87
7	0.97	0.99	0.98	90
8	0.98	0.97	0.98	111
9	0.94	0.84	0.89	76
10	0.97	0.98	0.98	100
11	0.93	0.97	0.95	93
12	0.96	0.96	0.96	77
13	0.97	0.99	0.98	75
14	0.98	0.94	0.96	124
15	0.92	0.99	0.95	85
16	0.92	0.9	0.91	73
17	0.95	0.91	0.93	104
18	0.92	0.99	0.96	**141**
19	0.96	0.94	0.95	93
20	0.95	0.97	0.96	115
21	0.97	0.94	0.96	69
22	**0.89**	**0.89**	**0.89**	72
23	0.98	0.91	0.95	116
24	0.93	0.89	0.91	85
25	0.98	0.9	0.94	**63**
26	0.92	0.92	0.92	87
27	**0.97**	**1**	**0.98**	127
28	0.95	0.97	0.96	73
Accuracy			**0.95**	2557
Macro Avg.	**0.95**	**0.94**	**0.94**	2557
Weighted Avg.	**0.95**	**0.95**	**0.95**	2557

we can see that MLP has the maximum max, the maximum min, the median has a maximum value, and it has the smallest box size as compared to other classifiers. So again, the MLP classifier is performing best based on all four evaluation criteria of a

box plot.

Classification comparison based on performance:

The performance of the MLP classifier has been shown in Table-1.2. Category-28 of fruit plant leaf has the best performance because it has a maximum value of precision, recall, and f1-score. While the category-22 leaf has the least performance due to the lowest value of the precision, recall, and f1-score. Based on support, category-18 has the maximum performance and category-25 leaf has minimum performance evaluation. These results are validated clearly from the heatmap of Figure-1.4.

Conclusions and Future Work

A novel fruit leaf dataset has been collected for classification purposes. This research is beneficial in identifying fruit plants based on leaf automatically. For this purpose, ML models are implemented over the collected images of fruit plant leaves. A rigorous analysis has been done of ML models, and their performance has been recorded. The performance analysis has concluded that MLP has performed better than other classifiers such as LR, KNN, SVM, DT and NB, i.e., 93.18.00% accuracy. The experiment has been conducted only for fruit plant leaves; other classes such as herbals **grvDL2022**, flowers

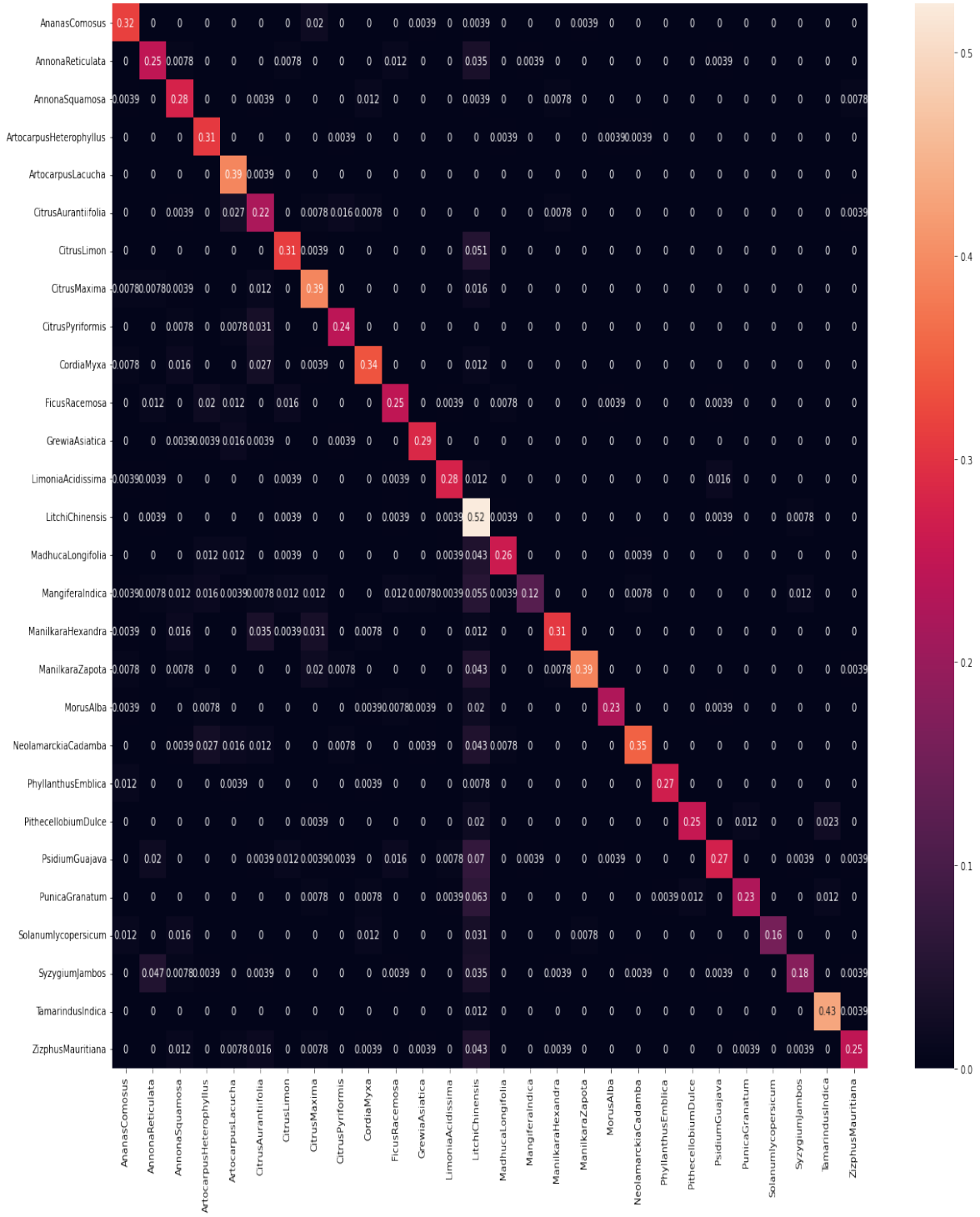


FIGURE 1.12: Confusion matrix using heatmap of ML classifiers.