

Quantile Regression and Deep Learning Models for Air Quality Analysis and Prediction in Delhi City

by

Gaurav Jha
202111059

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

INFORMATION AND COMMUNICATION TECHNOLOGY

to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY



June, 2023

Declaration

I hereby declare that

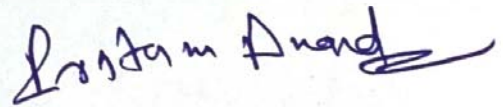
- i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,
- ii) due acknowledgment has been made in the text to all the reference material used.



Gaurav Jha

Certificate

This is to certify that the thesis work on Quantile Regression and Deep Learning Models for Air Quality Analysis and Prediction in Delhi City has been carried out by gaurav jha for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.



Dr. Pritam Anand
Thesis Supervisor

Acknowledgments

I am extremely grateful to my supervisor, Dr. Pritam Anand, for their invaluable advice, continuous support, and patience during my MTech Thesis. Their immense knowledge and ample experience have encouraged me during my academic research.

I want to thank my peers and friends Yash Joshi, Harsh Savaliya, Jaymin Vekariya and Rohit Mishra who were always there in my ups and downs. Their kind help and support have made my study and life at DA-IICT a wonderful and memorable experience.

Finally, I would like to express my gratitude to my parents and family for their love, care, and invaluable support throughout my life.

Contents

Abstract	v
List of Principal Symbols and Acronyms	vi
List of Tables	vii
List of Figures	viii
1 Introduction	1
1.1 Motivation	1
1.2 Overview of the (AQI) System	2
1.3 Contribution	3
1.4 Thesis Organization	5
2 Literature Survey	6
2.1 Pollution Analysis	6
2.2 Estimating PM2.5 concentration using temperature variables	7
2.3 Comparative analysis of pollutant gas levels in metro city	8
2.4 Seasonal Trends of air pollution in india: causes and consequences	8
2.5 Air pollution levels and compliance with international norms	9
2.6 A Comparative Study and Application	11
3 A statistically study of polluted gases in delhi city:	12
3.1 Analysis of polluted gases	12
3.2 Linear Regression Approach	15
3.3 Quadratic regression Approach	17
3.4 Gaussian Kernel Function Approach	18
3.5 Quantile regression model with Pin Loss Fuunction	20
4 Time Series Forecasting using Deep Learning Models	22
4.1 GRU Model	23

4.2	Vanilla LSTM Model	24
4.3	Simple LSTM Model	26
4.4	CNN-LSTM Model	27
4.5	SVR Model	29
5	Conclusions	36
	References	37

Abstract

Quantile regression models have gained popularity among researchers these days. The mean regression model estimates the mean of y_i given x . But in some applications, estimation of the quantiles of y_i given x is not very useful. This thesis presents a data-driven analysis and prediction of air quality in Delhi metro city using quantile regression and deep learning models.

The main objectives are to investigate the monthly trend and correlation of PM2.5, PM10, NO2 and SO2 concentration and temperature, to compare different regression models such as linear, quadratic, kernel, and quantile regression to estimate the PM2.5, PM10, NO2 and SO2 concentration using the temperature variables, and to compare different deep learning models such as gated recurrent units (GRUs), vanilla(LSTM), simple long short-term memory (LSTM) networks, convolutional neural network - long short-term memory (CNN-LSTM) networks, and support vector regression (SVR) for time series forecasting of pollution levels. The data used in this study is the Delhi air quality data from 2015 to 2020, which contains various pollutants and environmental factors.

The results show that quantile regression is more flexible, robust, and informative than other models, and can capture the variability and diversity of the PM2.5, PM10, NO2 and SO2 distribution over distinct quantiles or percentiles. The results also show that deep learning models are effective and powerful tools for time series forecasting on pollution data. Among them, the SVR model is superior to other models. The study aims to contribute to the scientific knowledge and practical solutions for air quality prediction and analysis.

Index Terms: *Quantile Regression, Quantile estimation, Regression Models, Deep learning models*

List of Principal Symbols and Acronyms

AQI	Air Quality Index
CP	Coverage probability
CPCB	Central Pollution Control Board
Exe-Time	Execution Time
IQ	Interquartile Value
LSTM	Long short-term memory
NO ₂	Nitrogen-Dioxide
PM ₁₀	Particulate Matter 10
PM _{2.5}	Particulate Matter 2.5
RMSE	Root Mean Square Error
SO ₂	Sulfur-Dioxide
SSE	Sum of Squared Error
SVR	Support Vector Regression
Var	Variance
WHO	World Health Organization

List of Tables

3.1	Monthly Var, Min, Max, IO and Median Values 2015-2020	13
3.2	Correlation between PM2.5 and Temperature	14
3.3	RMSE and SSE value on Regression models.	20
3.4	Coverage Probability Result	21
4.1	Result on Pollution Dataset (PM2.5)	35
4.2	Result on Pollution Dataset (PM10)	35
4.3	Result on Pollution Dataset (NO2)	35
4.4	Result on Pollution Dataset (SO2)	35

List of Figures

1.1	Environment and Health Risk	2
1.2	AQI Index	3
3.1	Monthly box plot of PM2.5 concentration	13
3.2	Shows the plot obtained by linear regression carried out by us	16
3.3	Linear Regression Model	16
3.4	Quadratic regression Model	18
3.5	Kernel Function	19
3.6	Training data with a 0.10 τ and 0.90 τ value	21
4.1	GRU Model Architecture	23
4.2	Vanilla LSTM Model Architecture	25
4.3	LSTM Model Architecture	26
4.4	LSTM Model Architecture	28
4.5	Time Series Forecasting using GRU Model	30
4.6	Time Series Forecasting using LSTM Model	31
4.7	Time Series Forecasting using Vanilla LSTM Model	32
4.8	Time Series Forecasting using CNN-LSTM Model	33
4.9	Time Series Forecasting using SVR Model	34

CHAPTER 1

Introduction

1.1 Motivation

Air pollution is an immense environmental and health problem that affects millions of people worldwide. Three Indian cities, Delhi, Mumbai, and Kolkata, are among the top ten most polluted in the world, according to the World Health Organisation (WHO) [10]. Air pollution can have a variety of negative impacts, including diminished lung function, asthma attacks, and premature death. As a result, it is essential to track and regulate air pollution levels in various locations and scenarios [13]. In this thesis, we propose to conduct detailed data-driven research on pollution trends in Delhi metropolitan city and their relationship to various weather conditions.

From 2015 to 2020, the Central Pollution Control Board, also known as the CPCB, gathered air quality data in India in order to establish future pollution control policies for Indian metro cities. Daily measurements of four primary pollutants are included in the data: PM_{2.5}, PM₁₀, SO₂, and NO₂. PM_{2.5} and PM₁₀ are particle sizes of less than 2.5 and 10 micrometres, respectively [17]. They can enter the lungs and cause issues with the heart and breathing. SO₂ and NO₂ are the chemical formulas for sulphur dioxide and nitrogen dioxide, respectively [21]. They are gases that can generate acid rain and irritate the eyes, nose, throat, and lungs.

We analyse data through different statistical approaches and visualisation tools to investigate patterns and trends in air pollution in major India metropolises [27]. We also look into the relationship between air pollution levels and environmental conditions like temperature. We hypothesise that weather conditions have a substantial impact on air pollution levels and vary across seasons and regions [22].

Pollutant	Environmental risks	Human health risks
Particulate matter(PM2.5,PM10)	Contributes to the formation of haze and acid rain, which affects the pH balance of rivers and harms plants, buildings, and historical places.	Discomfort of the lung organs, increased asthma, and irregular heartbeat
Nitrogen oxide (NO₂)	Damage to trees and plants; contributes to smog formation	Infections and irritation of breathing passages
Sulfur dioxide (SO₂)	A major contributor to the formation of acid rain, which affects plants, buildings, and monuments while also reacting to produce particulate matter.	Breathing issues, especially for those who have heart problems and asthma

Figure 1.1: Environment and Health Risk

1.2 Overview of the (AQI) System

The Air Quality Index (AQI) [15], a numerical measurement of the level of pollutants in the air, is one method for measuring air pollution. The air quality index (AQI) is calculated using levels of pollutants such as particulate matter (PM), ozone (O₃), nitrogen dioxide (NO₂), sulphur dioxide (SO₂), and carbon monoxide (CO) [26]. The Air Quality Index (AQI) provides a simple and standardised way for the public to understand the health effects of air pollution.,

The Air Quality Index (AQI) values and their respective categories, colours, and meanings are shown in Fig. 1.2. The AQI scale runs from 0 to 500, with higher numbers indicating greater pollution and more health concerns [12]. Green (good), light green (moderate), yellow (unhealthy for sensitive populations), orange (unhealthy), red (extremely unhealthy), and maroon (dangerous) are the AQI categories [19]. Each category has a different impact on the general population's and vulnerable groups' health, such as children, elderly people, or persons with lung or heart disease. When the AQI is in the red category, for example,

everyone may suffer from major health impacts, however when the AQI is in the yellow category, only the sensitive groups may suffer from health effects.

AQI Category (Range)	PM ₁₀ (24hr)	PM _{2.5} (24hr)	NO ₂ (24hr)	SO ₂ (24hr)
Good (0–50)	0–50	0–30	0–40	0–40
Satisfactory (51–100)	51–100	31–60	41–80	41–80
Moderate (101–200)	101–250	61–90	81–180	81–380
Poor (201–300)	251–350	91–120	181–280	381–800
Severe (301–400)	351–430	121–250	281–400	801–1600
Hazardous (401–500)	430+	250+	400+	1600+

Figure 1.2: AQI Index

1.3 Contribution

We used different methods to study the monthly trend, the effect of temperature, and the variability and diversity of PM_{2.5} concentration over different quantiles or percentiles [18]. We found that PM_{2.5} concentration was higher in winter months than summer and monsoon months, and that temperature had a positive correlation with PM_{2.5} concentration.

We used various statistical techniques to analyze the data on PM_{2.5} concentration and temperature in Delhi city during 2015-2020 [14]. We used boxplots [20] to show the median, quartiles, and outliers of PM_{2.5} concentration for each month, and scatter plots to show the parabolic pattern of PM_{2.5} and temperature.

We used correlation analysis [7] to measure the strength and direction of the linear relationship between PM2.5 and temperature, and regression analysis to model the relationship between PM2.5 and temperature using linear, quadratic [29], and kernel functions [6]. We used hypothesis tests, RMSE, SSE to evaluate the performance of these models. We also found that traditional regression models failed to capture the complexity of the relationship between PM2.5 and temperature.

We used quantile regression [4] to estimate the relationship between PM2.5 and temperature on different levels of the pollution distribution, such as the 10th or 90th percentile. We used different methods to construct confidence intervals for the slope coefficients of quantile regression [3] [2] [1]. We calculated coverage probability to assess the accuracy and reliability of these confidence intervals. We found that quantile regression had high coverage probability using different methods, and that it could capture the variability and diversity of PM2.5 over distinct quantiles or percentiles better than traditional regression models.

In next chapter, We compare different deep learning models for time series forecasting on pollution data. It requires the application of advanced and novel methods that can handle the complexity and uncertainty of the pollution data. It also involves the exploration and analysis of the underlying patterns and factors that affect the pollution levels.

The deep learning models that are compared in this study are gated recurrent units (GRUs) [28], long short-term memory (LSTM) [5] networks, convolutional neural network - long short-term memory [30] networks, and support vector regression (SVR) [9]. These models are chosen because they can capture the complex and nonlinear patterns in the pollution data, and handle variable-length input sequences and produce accurate and robust forecasts. The data used in this study is the Delhi air quality data from 2015 to 2020, which contains various pollutants and environmental factors.

In the model evaluation step, the models are evaluated by plotting the predicted and actual values of pollution levels on the training and testing sets, and by calculating the root mean squared error (RMSE) as a metric of accuracy. We studied that deep learning models are effective and powerful tools for time series forecasting on pollution data. Among them, the SVR model is superior to other models on pollution data-set. The study aims to contribute to the practical solutions for air quality prediction.

1.4 Thesis Organization

The thesis is organized as follows:

In chapter 2 reviews the literature on quantile regression and its applications in air pollution analysis and health outcomes.

In chapter 3 presents the experiment design, data description, statistical methods, linear regression model, quadratic model, kernel function, quantile regression model, coverage probability analysis, and results discussion.

In chapter 4 presents the time series forecasting using deep learning models, GRU model, Vanilla LSTM model, LSTM model, CNN-LSTM model and SVR model description, evaluation metrics, results discussion, and comparison.

In chapter 5 concludes the thesis with a summary of the main findings, limitations, and future work.

CHAPTER 2

Literature Survey

2.1 Pollution Analysis

In this thesis, we use a quantile regression model to analyze the relationship between [24] Air pollution is a major environmental and health problem that affects millions of people worldwide. According to the World Health Organization (WHO), air pollution causes about 7 million premature deaths every year, and is linked to various diseases such as respiratory infections, cardiovascular diseases, stroke, and lung cancer . One of the most polluted cities in the world is Delhi, the capital of India, where the air quality often reaches hazardous levels, especially during the winter months . The main sources of air pollution in Delhi are vehicular emissions, industrial activities, biomass burning, dust storms, and meteorological factors .

The objective of this thesis is to use a quantile regression model to analyze the relationship between air pollution and its predictors, such as temperature or particulate matter (PM_{2.5}), on different levels of the pollution distribution. We hypothesize that quantile regression can provide more comprehensive and nuanced insights into the dynamics and drivers of air pollution than traditional regression models.

Quantile regression is a statistical method that estimates the conditional median or other quantiles of the response variable across values of the predictor variables. Quantile regression can capture the variability and diversity of the response over distinct quantiles or percentiles. This can be useful for analyzing data that has unequal variation, non-linear relationships, or extreme values (Koenker, 2017). For example, quantile regression can reveal how temperature affects high or low levels of PM_{2.5} differently, or how PM_{2.5} affects health outcomes differently across different population groups. Quantile regression has been widely used for air quality analysis in various studies , and has shown promising results in terms of accuracy and robustness.

2.2 Estimating PM2.5 concentration using temperature variables

Air pollution [16] is a major environmental and health problem that affects millions of people worldwide. According to the World Health Organization (WHO), air pollution causes about 7 million premature deaths every year, and is linked to various diseases such as respiratory infections, cardiovascular diseases, stroke, and lung cancer . One of the most polluted countries in the world is India, where the air quality often reaches hazardous levels, especially in the metro cities such as Delhi, Mumbai, Kolkata, and Chennai . The main sources of air pollution in India are vehicular emissions, industrial activities, biomass burning, dust storms, and meteorological factors.

The objective of this thesis is to study the trends of different pollutant gases present in the air of metro cities in India, and to analyze their relationship with temperature variables using quantile regression models. We hypothesize that quantile regression models can provide more comprehensive and nuanced insights into the dynamics and drivers of air pollution than traditional regression models.

Quantile regression is a statistical method that estimates the conditional median or other quantiles of the response variable across values of the predictor variables. Quantile regression can capture the variability and diversity of the response over distinct quantiles or percentiles. This can be useful for analyzing data that has unequal variation, non-linear relationships, or extreme values (Koenker, 2017). For example, quantile regression can reveal how temperature affects high or low levels of PM2.5 differently, or how PM2.5 affects health outcomes differently across different population groups. Quantile regression has been widely used for air quality analysis in various studies , and has shown promising results in terms of accuracy and robustness.

However, according to the data available in 2020, the levels of different pollutant gases in the air of metro cities in India have been consistently high and hazardous to human health . Therefore, there is a need for more detailed and comparative studies for quantile regression models in the field of air pollution control. In this thesis, we aim to fill this gap by thoroughly comparing different quantile regression models, using them to estimate the PM2.5 concentration in an urban location using the temperature variables. We also aim to provide more insights into the factors that influence the extreme values of PM2.5 concentration, such as the 10th or 90th percentile.

2.3 Comparative analysis of pollutant gas levels in metro city

Delhi, for instance, has consistently ranked as one of the most polluted cities in the world, with high levels of particulate matter (PM_{2.5}), nitrogen dioxide, and ozone. Other metro cities such as Mumbai, Kolkata, and Chennai [8] also have high levels of air pollution. However, the levels may vary depending on the time of the year and local sources of pollution. Several factors contribute to these differences in air pollution levels among different metro cities in India. Some of the factors that contribute to higher levels of air pollution in some cities compared to others are:

Vehicular emissions: Cities with higher numbers of vehicles on the road tend to have higher levels of air pollution. Delhi has one of the highest registered vehicles in India, which contributes significantly to air pollution.

Industrial activities: Cities with a high concentration of industries tend to have higher levels of air pollution. Mumbai, for instance, is a hub for industries such as textiles, petrochemicals, and engineering, which contribute to air pollution.

Geography and climate: Cities located in geographically disadvantaged areas, such as valleys or with specific climatic conditions and weather patterns, are more likely to have high levels of air pollution. For example, Delhi is located in a region with relatively low wind speeds, making it difficult for pollutants to disperse.

Construction and demolition: Rapid urbanization and construction activities can lead to high levels of dust and other particulate matter, which can contribute to air pollution.

Agricultural practices: Burning of crop stubble and other agricultural waste can contribute significantly to air pollution levels in some areas of the country.

Overall, multiple factors contribute to the differences in air pollution levels among different metro cities in India, and addressing these factors is crucial to reducing air pollution and improving air quality in these cities.

2.4 Seasonal Trends of air pollution in india: causes and consequences

There are certain times of the year when pollutant levels are higher in metro cities in India, and the reasons for these seasonal trends are mainly related to weather patterns and human activities [11]. One of the most significant contributors to

seasonal changes in air pollution levels in metro cities in India is the occurrence of weather phenomena such as temperature inversions and monsoons. During winter months, temperature inversions often occur in northern India, where the ground is cooler than the air above, causing pollutants to become trapped near the surface, resulting in high levels of pollution.

These temperature inversions, combined with the use of coal for heating, burning of crop stubble, and vehicular emissions, contribute to the high levels of pollution in cities such as Delhi during winter months. During the summer months, the temperature in India can become very high, leading to the formation of ground-level ozone, which can cause respiratory problems. Additionally, the dry and hot weather conditions can lead to an increase in wildfires, which can significantly contribute to air pollution levels in nearby cities.

The monsoon season in India, which generally occurs between June and September, can help to alleviate air pollution levels in some metro cities by washing away pollutants from the atmosphere. However, the onset of the monsoon season can also lead to an increase in humidity levels, which can cause an increase in mould and dust mites, leading to respiratory problems. Human activities, such as festivals and agricultural practices, can also contribute to seasonal trends in air pollution levels. For example, during the festival of Diwali, the burning of firecrackers can lead to a significant increase in particulate matter in the air.

The burning of crop stubble in northern India after the harvest season can also lead to a significant increase in air pollution levels during the winter months. In summary, seasonal trends in air pollution levels in metro cities in India are mainly driven by weather patterns and human activities, and addressing these factors is crucial to reducing air pollution levels and improving air quality in these cities.

2.5 Air pollution levels and compliance with international norms

Pollutant levels in metro cities in India are often higher than international standards and guidelines set by organizations such as the World Health Organization (WHO) [25] and the United States Environmental Protection Agency (EPA). For example, the WHO recommends that the annual average concentration of PM_{2.5} not exceed 10 micro grams per meter cube, while the Indian national ambient air quality standard (NAAQS) sets a limit of 40 micro gram per meter cube. In Delhi, for instance, the average annual concentration of PM_{2.5} is around 100 micro gram per meter cube, which is ten times higher than the WHO guideline and two and

a half times higher than the Indian NAAQS limit. Other pollutants, such as nitrogen dioxide and ozone, also often exceed recommended limits in metro cities in India.

The implications of these differences between actual pollutant levels and recommended standards and guidelines are severe. Exposure to high levels of air pollution can lead to a wide range of health problems, including respiratory and cardiovascular diseases, lung cancer, and stroke. It is estimated that air pollution is responsible for millions of premature deaths each year worldwide, with a significant proportion of these occurring in India. Additionally, high levels of air pollution can have significant economic costs, including lost productivity due to illness, increased healthcare costs, and damage to crops and other natural resources.

In terms of analysis, you may want to use statistical techniques to identify trends and patterns in the data, as well as to explore relationships between pollutant levels and other factors (such as weather patterns, population density, or industrial activity). You may also want to consider using visualization techniques to communicate your findings to a wider audience. Overall, this is an important and complex area of research, but one that has the potential to have a significant impact on public health and environmental policy in India.

Overall, the period from 2015 to 2020 saw a continued struggle to address air pollution in India, with various measures being taken at different levels to reduce emissions and improve air quality. The study finds that particulate pollution is the dominant pollutant in India, with virtually all sites in northern India exceeding the annual average national ambient air quality standards (NAAQS) for PM10 and PM2.5. Southern India also experiences high levels of particulate pollution, exceeding the PM10 standard by 50 percent and the PM2.5 standard by 40

However, SO₂, NO₂, and O₃ generally meet the residential NAAQS across India. The study also finds no significant trend of these pollutants over the five-year period, and the reanalyzed dataset can be useful for evaluating Indian air quality from satellite data, atmospheric models, and low-cost sensors. Overall, this dataset provides a baseline to evaluate the effectiveness of the National Clean Air Programme and future air pollution mitigation policies in India

Investigate the interaction between air pollution and physical activity (PA) on lung function in healthy adults living in high-polluted areas. The study monitored the fine particulate matter (PM_{2.5}), particulate matter less than 10 micro meter (PM₁₀), particulate matter less than 1 micro meter (PM₁), black carbon (BC), nitrogen dioxide (NO₂), and ozone (O₃) continuously during the 2-h exposure.

Lung function was measured at five time points for each visit.

The results showed that PA, compared to rest, alleviated the detrimental effects of air pollutants on lung function. The study highlighted the importance of timing of measurements for capturing associations and suggested that PA might alleviate the associations between various pollutant exposures and lung function. Further research in this area is recommended.

2.6 A Comparative Study and Application

Several studies have used quantile regression to investigate the potential varying effects of air pollution exposure on different health outcomes, such as birth weight, carotid intima-media thickness (CIMT), and mortality [23]. For example, Lamichhane et al. (2020) used quantile regression to examine the socioeconomic inequalities in air pollution and birth weight and found that low maternal education positively modified the association between PM_{2.5} exposure and low birth weight 1. Wang et al. (2021) used quantile regression to examine the association of air pollution with CIMT, a marker of atherosclerosis, and found that PM_{2.5} exposure had a stronger effect on CIMT at higher percentiles 2. Pérez Vasseur and Aznarte (2021) compared 10 state-of-the-art quantile regression models for probabilistic forecasting of NO₂ pollution levels and found that quantile gradient boosted trees showed the best performance 3.

However, there is a lack of comparative studies for quantile regression models in the field of air pollution control. In this thesis, we aim to fill this gap by thoroughly comparing different quantile regression models, using them to estimate the PM_{2.5} concentration in a urban location using the temperature variables. We also aim to provide more insights into the factors that influence the extreme values of PM_{2.5} concentration, such as the 10th or 90th percentile, which are more relevant for public health and environmental regulation.

CHAPTER 3

A statistically study of polluted gases in delhi city:

To study the pollutant gasses present in the air of delhi city, the air quality data in India during 2015-2020 is collected by the Central Pollution Control Board (CPCB) and State Pollution Control Boards (SPCBs) through a network of air quality monitoring stations located in various cities and towns across India. We have downloaded the dataset from Kaggle website which contains the daily recordings of various pollutants such as particulate matter (PM10 and PM2.5), nitrogen oxides (NO₂), sulfur dioxide (SO₂) present in the air of delhi city. Further, we have also collected data for daily temperature and humidity present in the air of Delhi city.

A limitation of AQI values is that they only summarise the average air quality over time. They do not reflect the variability and diversity of air pollution levels over different quantiles or percentiles of the distribution. This means they may miss important information about factors influencing extreme air pollution values, such as the 10th or 90th percentile. These extreme values are more relevant for public health and environmental regulation, as they indicate the potential risks and impacts of air pollution on the most vulnerable or exposed groups. Therefore, there is a need for a more comprehensive and excellent analysis of air pollution levels that can capture the variability and diversity of the distribution.

3.1 Analysis of polluted gases

we present our analysis of pollutant gases for the Delhi city using the collected datasets. We focus on the PM_{2.5} concentration in air, which is one of the most harmful pollutants for human health. We explore the monthly trend of PM_{2.5} concentration in air during 2015-2020 and examine the effect of temperature on PM_{2.5} concentration.

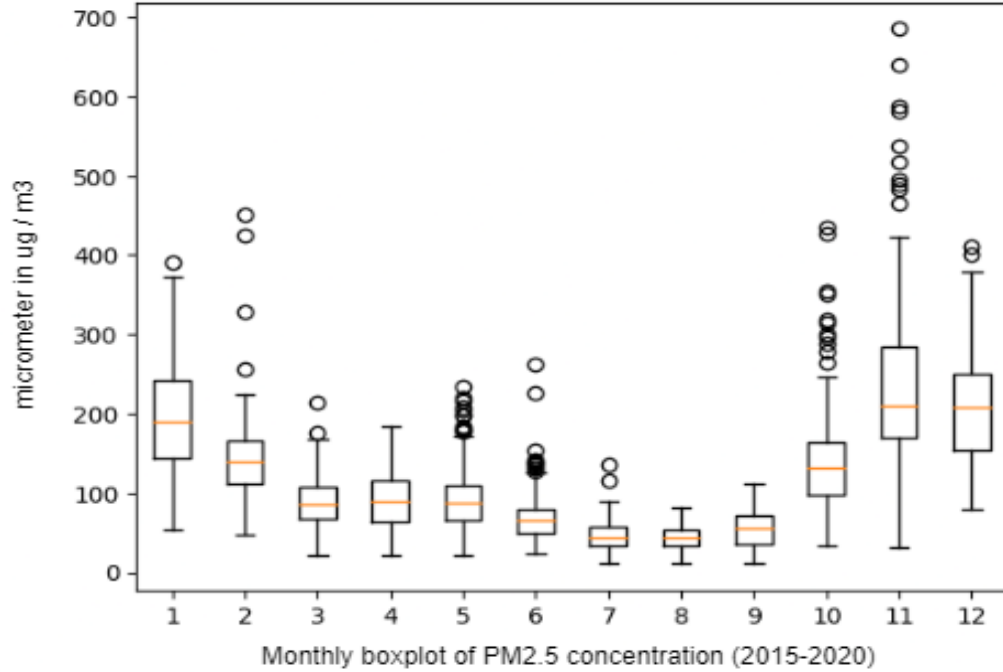


Figure 3.1: Monthly box plot of PM2.5 concentration

Months	Jan	Feb	march	April	May	June	July	Aug	Sep	Oct	Nov	Dec
Var(PM2.5)	4856.55	2800.22	999.95	1324.36	1607.49	1047.36	346.20	212.02	514.27	4753.54	13203.62	4814.29
Var(Temp)	29.96	33.24	55.50	34.40	30.44	33.48	24.87	12.39	19.10	12.68	19.26	43.22
Min Value	60	50	25	25	25	28	20	20	20	50	50	90
Max Value	380	230	180	190	185	140	100	90	120	240	410	390
IO Value	90	45	35	45	40	25	15	15	30	70	100	90
Median	190	140	80	85	80	60	50	50	55	150	220	215

Table 3.1: Monthly Var, Min, Max, IO and Median Values 2015-2020

To study the monthly trend of PM2.5 concentration in air, we obtain the monthly boxplot of PM2.5 concentration in air during 2015-2020, as shown in Fig. 3.1. The boxplot shows the median, quartiles, and outliers of PM2.5 concentration for each month. We can observe that the PM2.5 concentration varies significantly across different months, with higher values in winter and lower values in summer and monsoon.

In the months of November, December and January, the median value of PM 2.5 concentration is higher than 190 AQI . Also in these months, the maximum value of PM2.5 AQI touches 420. According to the National Air Quality Index (QAI) measures, Delhi air during the months of November, December and January can be ranked extremely hazardous and may have a very adversarial health impact on the people . The range and interquartile distances for November, December and January are also large, indicating a high variability and diversity of

cor	Jan	Feb	march	April	May	June	July	Aug	Sep	Oct	Nov	Dec
value	0.354	0.103	0.041	0.050	0.035	0.041	0.036	0.030	0.063	0.280	0.392	0.329

Table 3.2: Correlation between PM2.5 and Temperature

PM2.5 concentration in these months.

After February, the PM2.5 concentration starts decreasing from March and becomes moderate in April, May, June, July, August and September. The median value of PM2.5 concentration in these months is below 100 AQI, which is still unhealthy but less severe than winter months. The maximum value of PM2.5 AQI in these months is below 300 AQI, which is still very unhealthy but not hazardous. The range and interquartile distances for these months are also smaller than winter months, indicating a lower variability and diversity of PM2.5 concentration in these months.

We can observe that in winter months of Delhi, the PM 2.5 concentration is higher than summer and monsoon months. This may be due to various factors, such as lower wind speed, higher humidity, higher emissions from vehicles and industries, and crop burning in nearby states. Taking motivation from this, we attempt to study the effect of the temperature on PM 2.5 concentration in Delhi air. For this, we first compute the correlation of temperature with PM2.5 concentration. The correlation coefficient measures the strength and direction of the linear relationship between two variables. A positive correlation means that the variables tend to increase or decrease together, while a negative correlation means that the variables tend to move in opposite directions. A correlation coefficient close to 1 or -1 indicates a strong relationship, while a correlation coefficient close to 0 indicates a weak or no relationship.

The correlation analysis shows that in table 3.2 there is a positive relationship between temperature and PM2.5 concentration for all months, which means that as the temperature increases, the PM2.5 concentration decreases, and vice versa. This is in line with our observation that winter months have higher PM2.5 concentration than summer and monsoon months. The correlation coefficient is also higher in absolute value for winter months than summer and monsoon months, which means that the relationship between temperature and PM2.5 concentration is stronger in winter months than summer and monsoon months. Based on these results, we decide to use temperature as a predictor variable for estimating the PM2.5 concentration using regression analysis. We expect that temperature can explain some of the variation in PM2.5 concentration across different months and seasons.

3.2 Linear Regression Approach

We first applied a linear regression model analysis to our collected data, which included PM2.5 concentration in Delhi city. We also obtained the temperature data during 2015 to 2020 from a website. The temperature was the independent variable and PM2.5 concentration was the dependent variable. We had total 2010 data points.

$$y = \beta_0 + \beta_1 x + \epsilon \quad (3.1)$$

where y is the dependent variable, x is the independent variable, and β_0 and β_1 are the coefficients.

To extend this equation to a linear regression model with multiple independent variables, we will need to add more terms. We will use the following notation:

The coefficient of linear regression is β_0 and β_1 , and ϵ is the vector of errors.

Slope Equation:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad (3.2)$$

Intercept Equation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3.3)$$

where, $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ are predicted value and y_1, y_2, \dots, y_n are observed value, n is the number of observations.

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3.4)$$

We conducted our regression analysis with a significance level of 0.05. Our null hypothesis was $\beta_1 = 0$ and our alternative hypothesis was $\beta_1 \neq 0$. Since we obtained a p-value less than 0.05, we rejected the null hypothesis. It means that temperature impacts the PM2.5 concentration significantly. Our regression analysis makes sense, but the RMSE value obtained was high, so we decided to work on a nonlinear regression model analysis.

The RMSE for linear regression is 73.85, which means that there is a large average error between the observed and predicted values. The SSE for linear regression is 10958714.52, which means that there is a lot of unexplained variation


```

=====
                        OLS Regression Results
=====
Dep. Variable:          PM      R-squared:                0.211
Model:                 OLS     Adj. R-squared:           0.210
Method:                Least Squares   F-statistic:              535.9
Date:                  Tue, 23 May 2023   Prob (F-statistic):       2.74e-105
Time:                  02:33:11   Log-Likelihood:           -11488.
No. Observations:     2009   AIC:                      2.298e+04
Df Residuals:         2007   BIC:                      2.299e+04
Df Model:              1
Covariance Type:      nonrobust
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept      293.3085      7.787         37.667      0.000      278.037      308.580
temp          -5.4239      0.234        -23.151      0.000      -5.883      -4.964
=====
Omnibus:              792.742   Durbin-Watson:           0.367
Prob(Omnibus):        0.000   Jarque-Bera (JB):        4514.720
Skew:                 1.768   Prob(JB):                 0.00
Kurtosis:             9.437   Cond. No.                 158.
=====

```

Figure 3.2: Shows the plot obtained by linear regression carried out by us

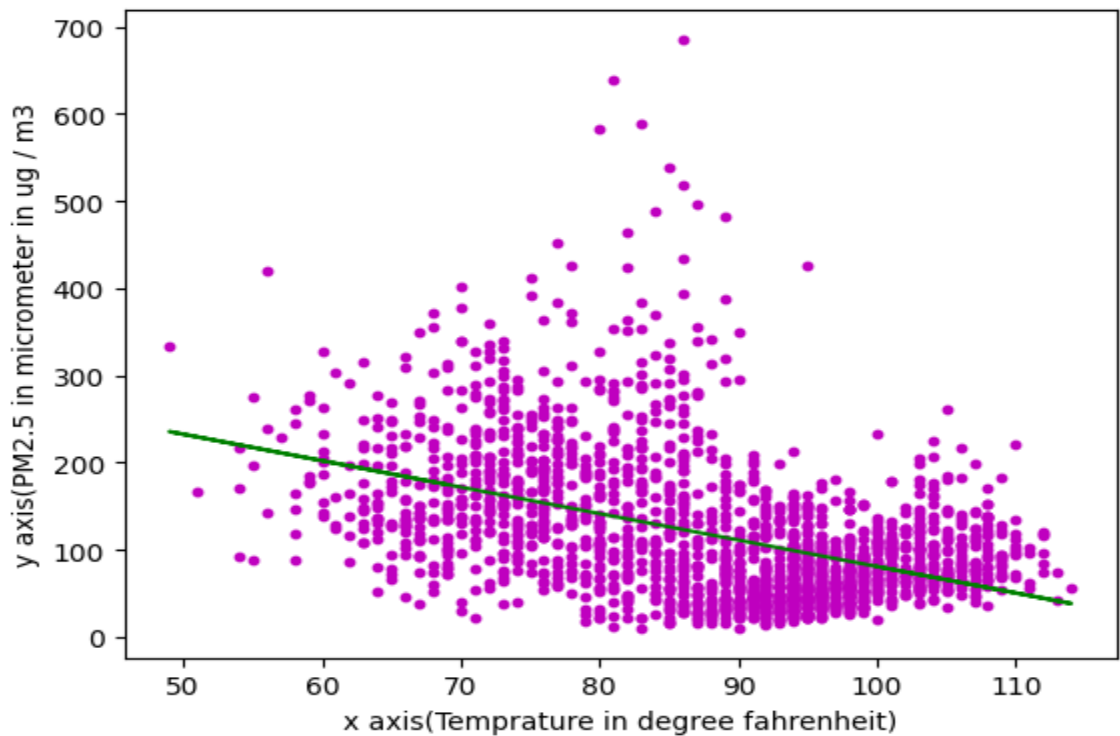


Figure 3.3: Linear Regression Model

in the data by the model. These results suggest that linear regression is not suitable for modeling the relationship between PM2.5 concentration and temperature

variables.

3.3 Quadratic regression Approach

After that, we have used quadratic regression [29] to model the relationship between PM2.5 and Temperature variables. Quadratic regression is a type of polynomial regression that fits a curve of the form $y = ax^2 + bx + \epsilon$ to the data, where a is not equal to zero. We chose this method because the scatter plot of PM2.5 and Temperature showed a clear parabolic pattern, suggesting that a linear model would not be adequate. In this section, we derived the equation for multiple regression in matrix form and Quadratic regression model is

$$y = ax^2 + bx + \epsilon \quad (3.5)$$

Define the error and the sum of squared errors

$$e_i = y_i - (ax_i^2 + bx_i + \epsilon) \quad (3.6)$$

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - ax_i^2 - bx_i - \epsilon)^2 \quad (3.7)$$

Define the partial derivatives and set them equal to zero

$$\frac{\partial SSE}{\partial a} = -2 \sum_{i=1}^n x_i^2 (y_i - ax_i^2 - bx_i - \epsilon) = 0 \quad (3.8)$$

$$\frac{\partial SSE}{\partial b} = -2 \sum_{i=1}^n x_i (y_i - ax_i^2 - bx_i - \epsilon) = 0 \quad (3.9)$$

$$\frac{\partial SSE}{\partial c} = -2 \sum_{i=1}^n (y_i - ax_i^2 - bx_i - \epsilon) = 0 \quad (3.10)$$

Solve the equations for a , b , and c using matrix algebra that can perform non-linear regression. The solution gave us the estimates of a , b , and c that best fit the data.

The RMSE for quadratic regression is 73.61, which means that there is still a large average error between the observed and predicted values. The SSE for quadratic regression is 10885632.83, which means that there is still a lot of unexplained variation in the data by the model. We still find that the quadratic regression model results are very high so we have used a kernel regression model

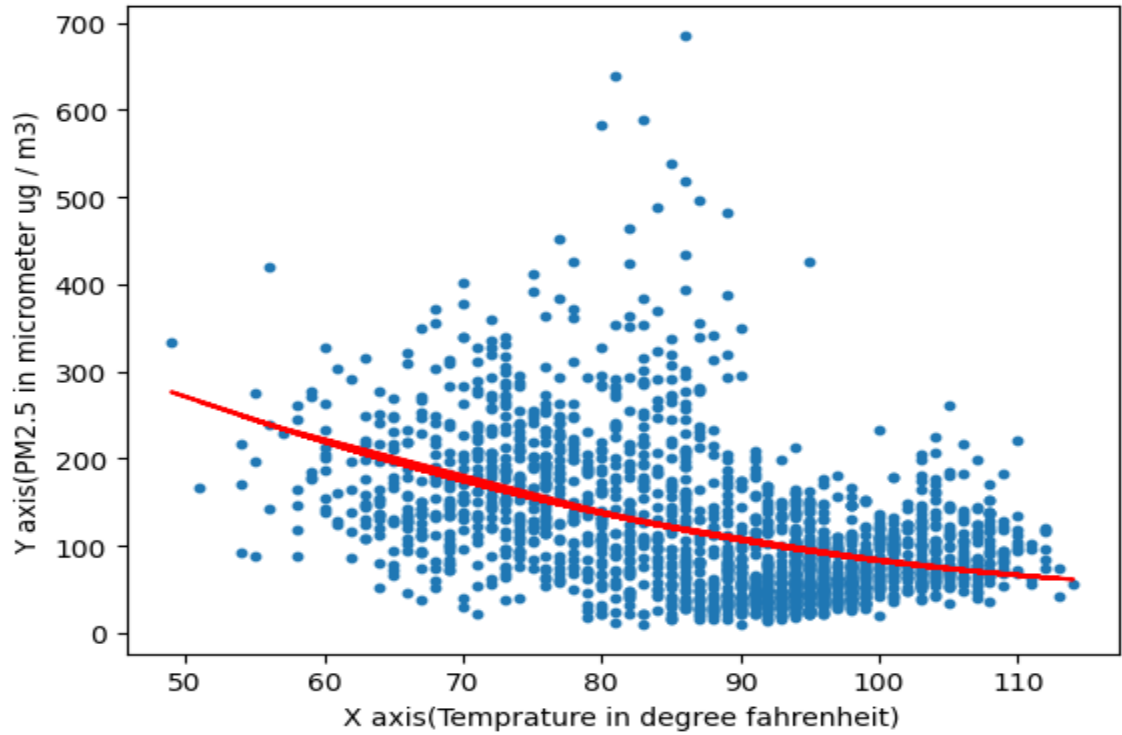


Figure 3.4: Quadratic regression Model

which will be capable of any non linear regression. We have used a gaussian kernel which is in the form the parameter sigma was tuned and opatain from RMSE value.

3.4 Gaussian Kernel Function Approach

In this section, we will derive the equation for the Gaussian kernel function. The Gaussian kernel function is a popular kernel function that is used in machine learning algorithms such as support vector machines and Gaussian process regression.

The Gaussian kernel function is defined as:

$$K(x, y) = \exp\left(-\frac{(x - y)^2}{2\sigma^2}\right) \quad (3.11)$$

where x and y are two data points, and σ is a hyperparameter that controls the smoothness of the kernel function.

The Gaussian kernel function can be derived using the following steps:

1. We start with the equation for the Euclidean distance between two data points:

$$d(x, y) = \sqrt{(x - y)^2} \quad (3.12)$$

2. We then take the exponential of the negative Euclidean distance:

$$K(x, y) = \exp(-d(x, y)^2) \quad (3.13)$$

3. Finally, we replace the Euclidean distance with a Gaussian kernel:

$$K(x, y) = \exp\left(-\frac{(x - y)^2}{2\sigma^2}\right) \quad (3.14)$$

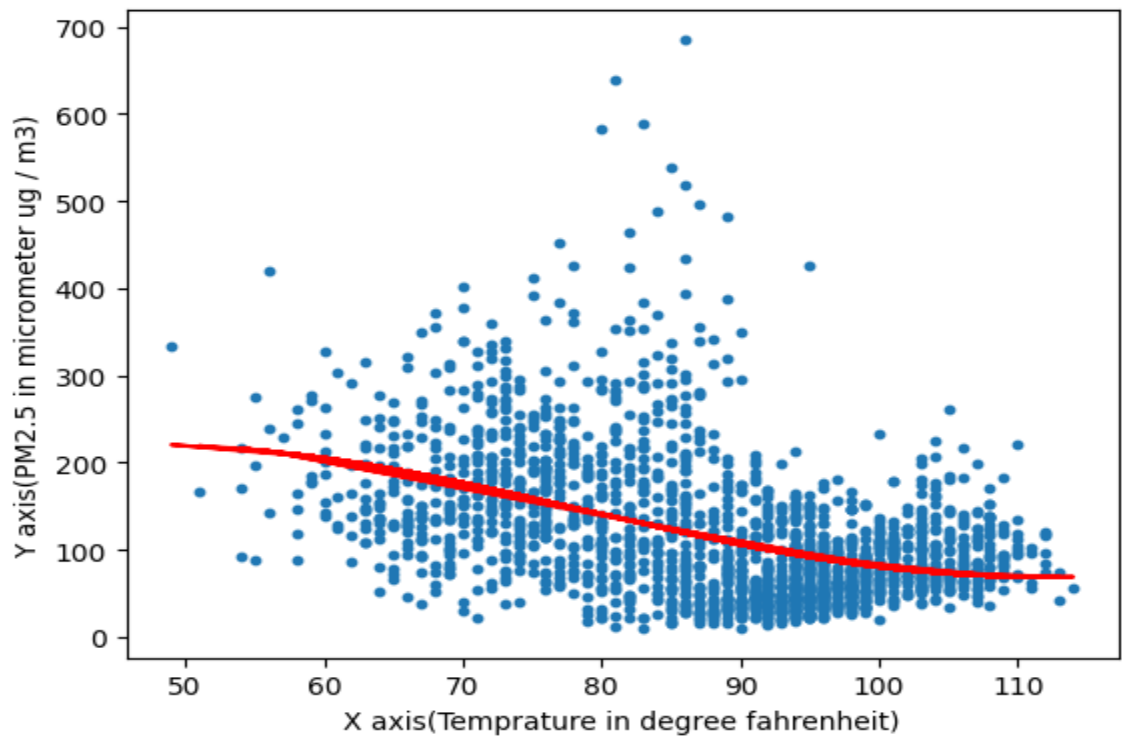


Figure 3.5: Kernel Function

The RMSE for kernel function is 72.94, which means that there is still a large average error between the observed and predicted values. The SSE for kernel function is 10688668.04, We can observe that even in the kernel regression model our prediction involves lots of uncertainty so it is very imperative to model the uncertainty in the relationship with temperature and pm2.5 in detail. We target to estimate the prediction pm2.5 concert given the temp we have used kernel quantile regression model obtain the prediction interval with 80 percent.

Regression Model	RMSE Value	SSE Value
Linear Regression	73.85	10958714.52
Quadratic Regression	73.61	10885632.83
Kernel Function	72.94	10688668.04

Table 3.3: RMSE and SSE value on Regression models.

3.5 Quantile regression model with Pin Loss Function

In this section, we use a quantile regression model based on a kernel function. Here we apply pinball loss function which we can use for the estimation of conditional quantiles. This pinball loss function can be given by

$$L_{\tau}(u) = \begin{cases} \tau u & \text{if } u \geq 0. \\ (\tau - 1)u & \text{otherwise.} \end{cases}$$

We apply the Quantile regression model. We evaluate the coverage probability as an essential criterion for assessing the accuracy and reliability of confidence intervals for quantile regression. Ideally, we want a method that produces confidence intervals with high coverage probability, which means they are likely to contain the actual value.

We find that quantile regression analysis between PM2.5 and temperature variables has high coverage probability using different methods of constructing confidence intervals. This means that we can estimate the relationship between PM2.5 and temperature variables on different levels of pollution distribution with high accuracy and reliability. We also find that quantile regression can capture the variability and diversity of PM2.5 over distinct quantiles or percentiles, which can provide more comprehensive and nuanced insights into the dynamics and drivers of PM2.5 pollution than traditional regression models.

For the training, 80 per cent of the total data, and for testing, 20 per cent of the total data set at τ value 0.1 at tuning parameter $s = 2^1$, $c1 = 2^2$, $\tau = 0.1$, $v1 = 0.01$. The coverage probability that we calculate at $\tau = 0.1$ is 0.1015

For the training and testing data set at τ value 0.9 at tuning parameter $s = 2^9$, $c1 = 2^2$, $\tau = 0.9$, $v1 = 0.01$. The coverage probability that we calculate at $\tau = 0.9$ is 0.9014

We perform this procedure 500 times and compute the mean of the outcomes. We use a kernel-based quantile regression model to estimate the conditional quantiles of the PM2.5 levels given the temperature. The diagram shows that 80 percent

τ (Training Data)	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
CP	0.1015	0.2573	0.5027	0.7551	0.9014

Table 3.4: Coverage Probability Result

of our data lies between two regression lines corresponding to the 10th and 90th percentiles. The x-axis represents temperature, and the y-axis represents PM2.5 levels. For instance, at a temperature of 80 degrees Fahrenheit, the 80 percent data interval ranges from 51 to 230 on the y-axis.

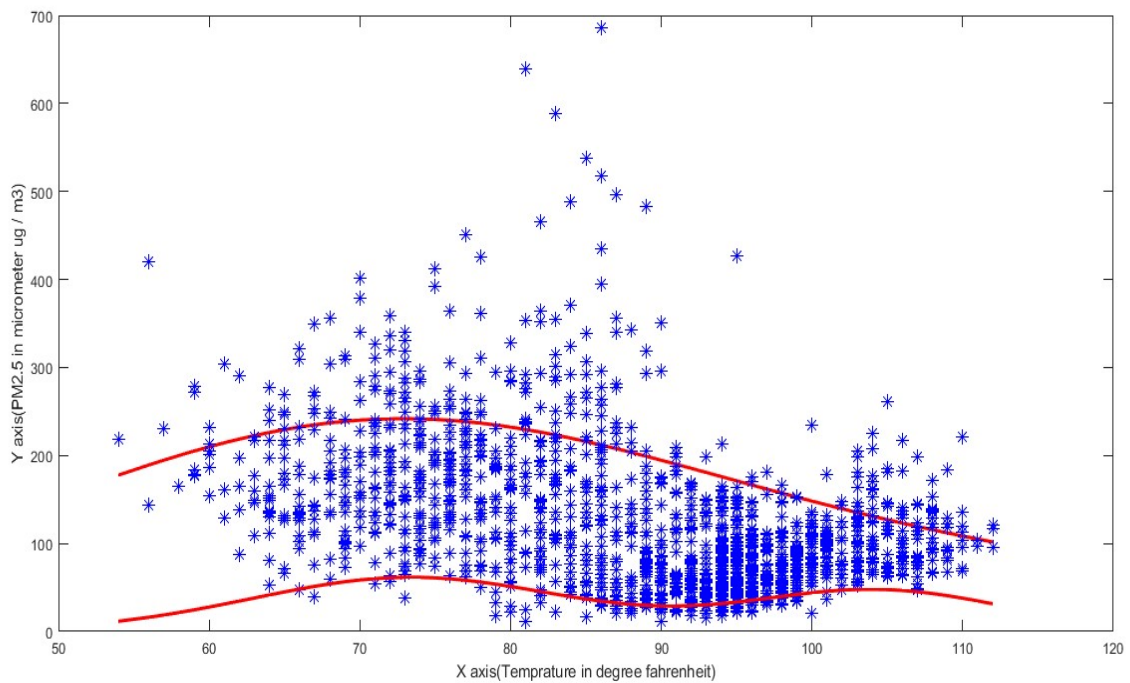


Figure 3.6: Training data with a 0.10 τ and 0.90 τ value

CHAPTER 4

Time Series Forecasting using Deep Learning Models

To study the pollutant gasses present in the air of delhi city, the air quality data in India during 2015-2020 is collected by the Central Pollution Control Board (CPCB) and State Pollution Control Boards (SPCBs) through a network of air quality monitoring stations located in various cities and towns across India. We have downloaded the dataset from Kaggle website which contains the daily recordings of various pollutants such as particulate matter (PM10 and PM2.5), nitrogen oxides (NO₂), sulfur dioxide (SO₂) present in the air of delhi city.

Pollution is a challenging and essential problem that significantly impacts human health and the environment. It requires the application of advanced and novel methods that can handle the complexity and uncertainty of pollution data. It also involves exploring and analysing the underlying patterns and factors that affect pollution levels. It contributes to the scientific knowledge and practical solutions for air quality management and policy making. It also provides opportunities for further research and improvement in this field.

Time series forecasting using a deep learning model is a novel and effective approach that can capture the complex and nonlinear patterns in the pollution data. It requires the application of advanced and powerful deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), that can learn from sequential and high-dimensional data. It also involves exploring and analysing the impact of various factors, such as environmental factors, on pollution levels. It contributes to the scientific knowledge and practical solutions for air quality prediction and management.

In this chapter, we compare different deep-learning models for sequential data processing. These models include gated recurrent units (GRUs), long short-term memory (LSTM) networks, and convolutional neural network-long short-term memory (CNN-LSTM) networks. GRUs and LSTMs are recurrent neural network

(RNN) types that can capture long-term dependencies in sequential data using different gating mechanisms. CNN-LSTMs are hybrid models that combine a convolutional neural network (CNN) for feature extraction from pollution data and an LSTM for sequence generation, such as captions or labels. At last, we use time series forecasting to predict pollution levels based on historical data. One of the methods we employ for time series forecasting is support vector regression (SVR), a machine learning model that can learn nonlinear relationships between the input and output variables.

4.1 GRU Model

Time series forecasting is the task of predicting the future values of a series based on past observations. A GRU model [28] is a recurrent neural network with a more straightforward structure that can learn from sequential data. A GRU model has two gates: reset and update gates. These gates control how information flows in and out of the hidden state. A brief description of time series forecasting using the GRU model is as follows:

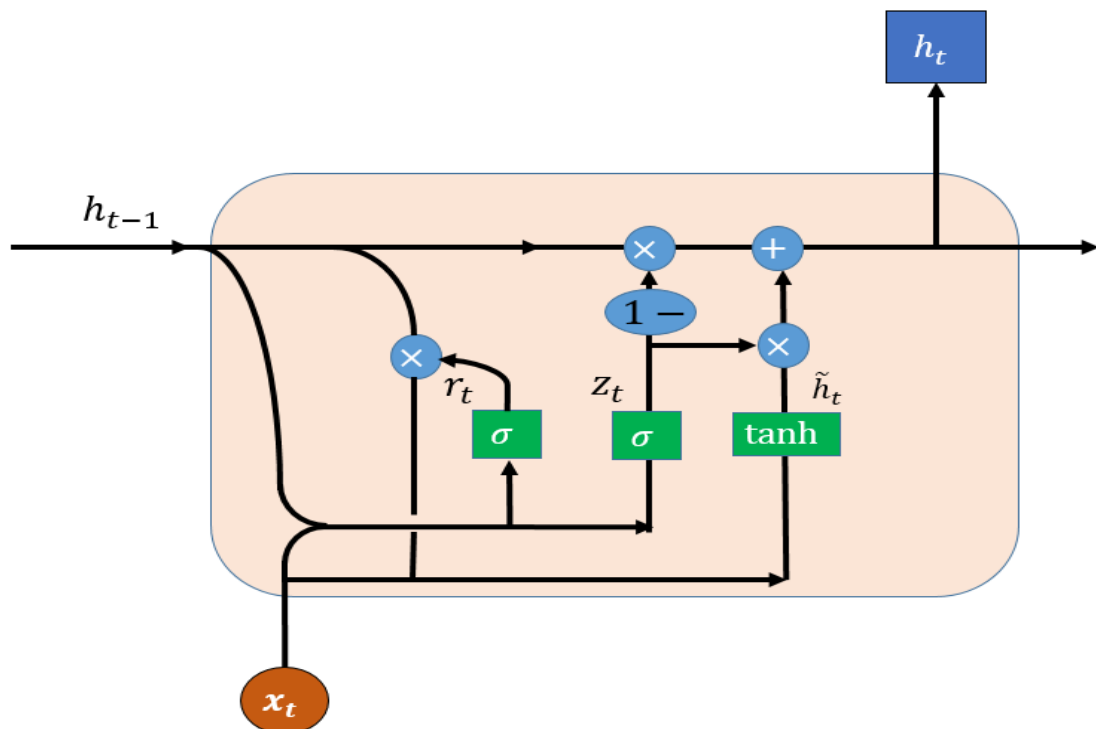


Figure 4.1: GRU Model Architecture

Reset gate

$$\mathbf{r}_t = \sigma(\mathbf{W}_{xr}\mathbf{x}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1} + \mathbf{b}_r) \quad (4.1)$$

Update gate

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{W}_{hz}\mathbf{h}_{t-1} + \mathbf{b}_z) \quad (4.2)$$

Candidate hidden state

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (4.3)$$

Hidden state

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \quad (4.4)$$

In this study, we aimed to forecast the pollution levels for the winter months (October to January) from 2015 to 2020 using a recurrent neural network (RNN) model. We chose a gated recurrent unit (GRU) as the RNN architecture, which has been shown to perform well on time series forecasting. We used a sliding window approach to create the input and output sequences, with a window size of 5. We split the data into three sets: training, validation and testing. The training set contained data from 2015 to 2018, the validation set contained data from 2019, and the testing set contained data from 2020. We built the GRU model with 5 input units and 1 output unit, and optimized it using the Adam algorithm with a learning rate of 0.0001. We minimized the mean squared error (MSE) as the loss function and trained the model for 1000 epochs. We evaluated the model by plotting the predicted and actual values of pollution levels on the training and testing sets, and by calculating the root mean squared error (RMSE) as a metric of accuracy.

4.2 Vanilla LSTM Model

A Vanilla LSTM model [5] is a simple type of recurrent neural network that can learn from sequential data and remember long-term dependencies. The Vanilla LSTM model has a single hidden layer of LSTM units and an output layer used to make a prediction. The Vanilla LSTM model can handle variable-length input sequences and produce accurate and robust forecasts.

Input gate

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \quad (4.5)$$

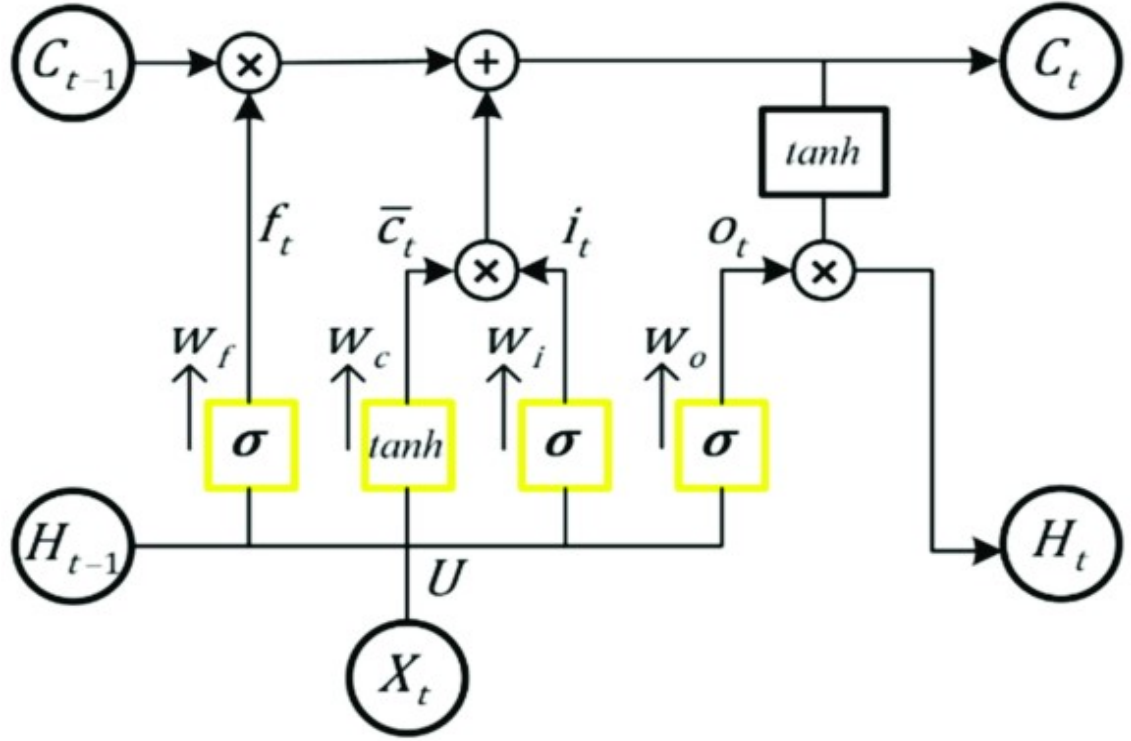


Figure 4.2: Vanilla LSTM Model Architecture

Forget gate

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (4.6)$$

Output gate

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \quad (4.7)$$

Candidate cell state

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (4.8)$$

Hidden state

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (4.9)$$

We used a recurrent neural network (RNN) model to forecast pollution levels over the winter months (October to January) from 2015 to 2020. We chose a vanilla long short-term memory (LSTM) as the RNN architecture since it is the most basic form of LSTM and uses a hyperbolic tangent (tanh) activation function. We built the input and output sequences using a sliding window method with a window size of 5. We divided the data into three sets: training, validation, and testing. The training set had data from 2015 to 2018, the validation set contained data from 2019, and the testing set contained data from 2020. We used the Adam algorithm

with a learning rate 0.0001 to optimize the vanilla LSTM model, which included five input units and one output unit. As the loss function, we minimized the mean squared error (MSE) and trained the model for 1000 epochs. We evaluated the model by graphing projected and actual pollution levels on the training and testing sets and computed the root mean squared error (RMSE) as an accuracy metric.

4.3 Simple LSTM Model

Time series forecasting is the task of predicting the future values of a series based on past observations. A simple LSTM model is a recurrent neural network that can learn from sequential data and remember long-term dependencies. The LSTM model has a unique structure that consists of a cell state and three gates: an input gate, an output gate, and a forget gate. The LSTM model can handle variable-length input sequences and produce accurate and robust forecasts.

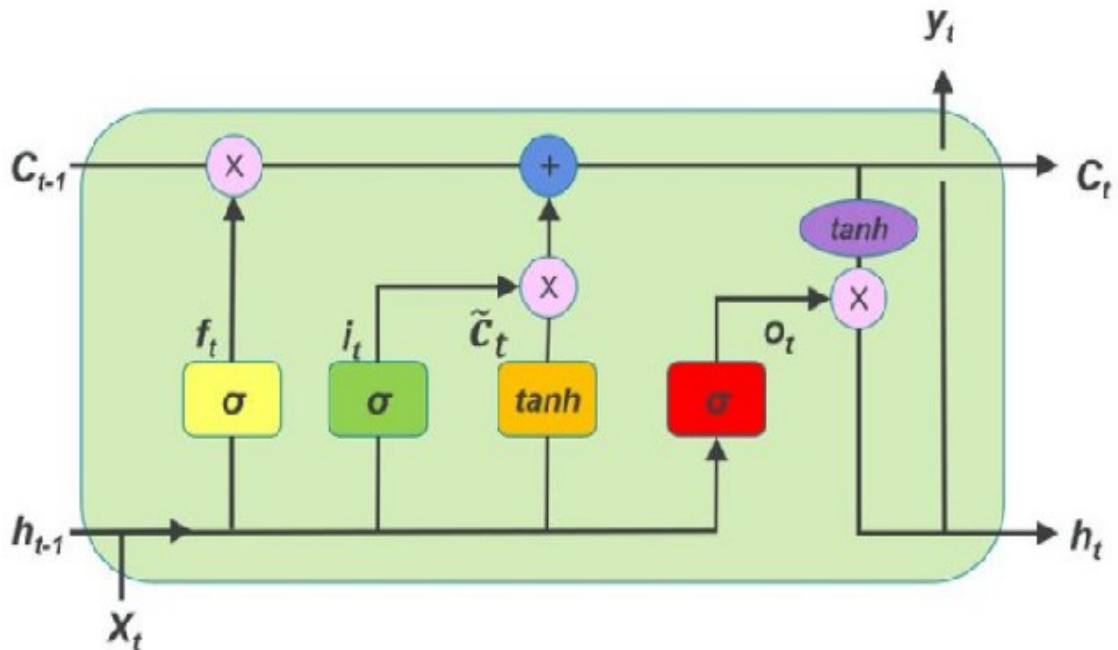


Figure 4.3: LSTM Model Architecture

Input gate

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \quad (4.10)$$

Forget gate

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (4.11)$$

Output gate

$$\mathbf{O}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \quad (4.12)$$

Candidate cell state

$$\tilde{\mathbf{O}}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (4.13)$$

Cell state

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad (4.14)$$

Hidden state

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (4.15)$$

This study aimed to anticipate pollution levels for the winter months (October to January) from 2015 to 2020 using a recurrent neural network (RNN) model. We picked a long short-term memory (LSTM) architecture as the RNN architecture since it has been proven to handle long-term dependencies and avoid vanishing gradients in time series data. The input and output sequences were built using a sliding window method with a window size of 5. We divided the data into three groups: training, validation, and testing. The training set included data from 2015 to 2018, the validation set included data from 2019, and the testing set included data from 2020. The LSTM model was implemented using five input units and one output unit, and it was optimized using the Adam method with a learning rate of 0.0001. As the loss function, we minimized the mean squared error (MSE) and trained the model for 1000 epochs. We evaluated the model by plotting showed and actual pollution levels on the training and testing sets and calculated the root mean squared error (RMSE) as an accuracy metric.

4.4 CNN-LSTM Model

A CNN-LSTM hybrid model combines convolutional neural networks (CNNs) and extended short-term memory networks (LSTMs) to process sequential data. The CNN model extracts features from sub-sequences of the input series, while the LSTM model captures the temporal dependencies among the features. The CNN-LSTM model can handle long input sequences and produce accurate and robust forecasts.

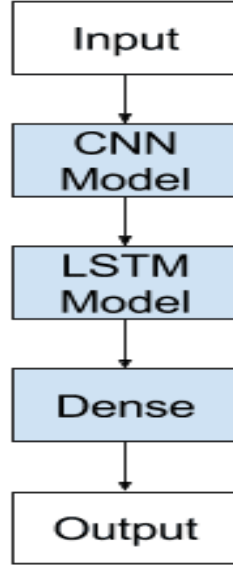


Figure 4.4: LSTM Model Architecture

CNN layer

$$\mathbf{x}_t = \text{CNN}(\mathbf{x}_{t-1}) \quad (4.16)$$

Input gate

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \quad (4.17)$$

Forget gate

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (4.18)$$

Output gate

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \quad (4.19)$$

Candidate cell state

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (4.20)$$

Cell state

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad (4.21)$$

Hidden state

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (4.22)$$

In this study, We used a hybrid neural network model to anticipate pollution levels over the winter months (October to January) from 2015 to 2020. We used a convolutional neural network (CNN) and a long short-term memory (LSTM) network to capture the time series data's spatial and temporal features. We employed

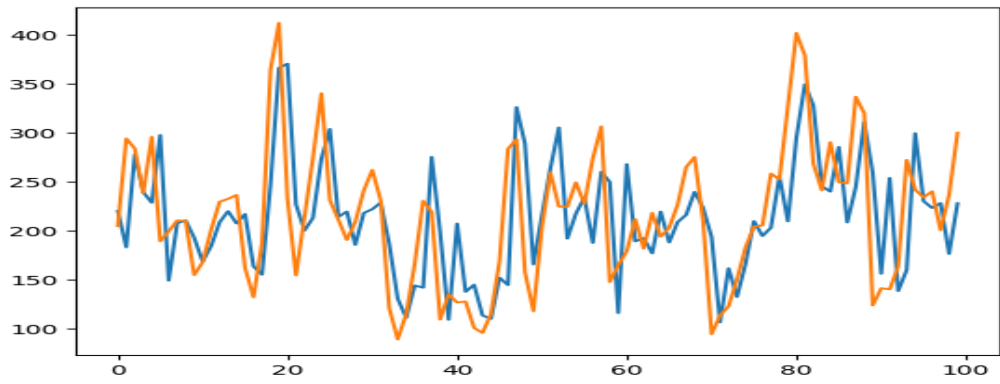
a sliding window technique to build the input and output sequences with a window size of 5. The data was divided into three categories: training, validation, and testing. The training set included information from 2015 to 2018, the validation set included information from 2019, and the testing set included information from 2020. We created the CNN-LSTM model with five input units and one output unit and then optimized it with the Adam algorithm with a learning rate 0.0001. As the loss function, we minimized the mean squared error (MSE) and trained the model for 1000 epochs. We assessed the model by plotting showed and actual pollution levels on the training and testing sets and computing the root mean squared error (RMSE) as an accuracy metric.

4.5 SVR Model

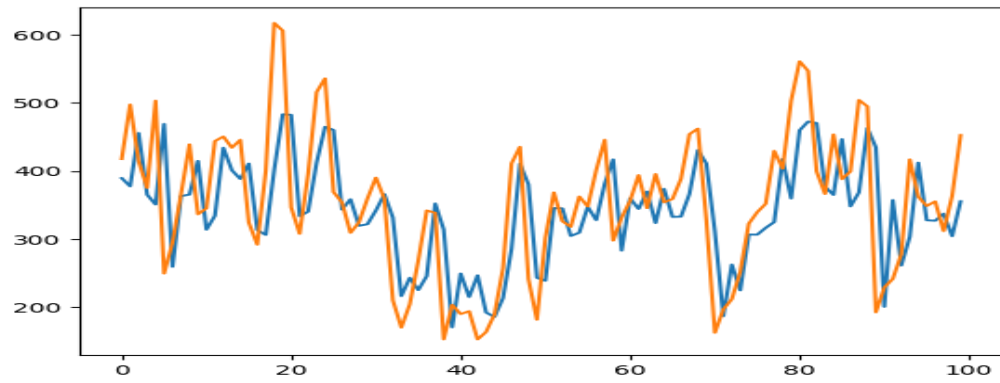
We collected our time series data from the Delhi air quality data, which contained daily measurements of four pollutants: particulate matter (PM10 and PM2.5), nitrogen dioxide (NO₂), sulfur dioxide (SO₂). The data spanned from 2015 to 2020, covering the winter months (October to January) when the air quality was the worst. We divided the data into two sets: training and testing. The training set had 80 per cent of the data, and the testing set contained 20 per cent of the total data. We normalized the data using min-max scaling to avoid numerical instability and improve the performance of the model.

We developed a support vector regression (SVR) model [9] to forecast the pollution levels for each pollutant separately. SVR is a machine learning technique that can perform nonlinear regression by mapping the input data into a high-dimensional feature space using a kernel function. We used a radial basis function (RBF) kernel, which is a common choice for SVR models. We tuned the hyperparameters of the model using grid search and cross-validation, and selected the optimal values as follows: a gamma value of 0.5, a regularization parameter of 10, and an epsilon parameter of 0.05. We trained the model on the training set and tested it on the testing set. We evaluated the model by computing the root mean squared error (RMSE) for each pollutant as an accuracy metric. The RMSE measures the average deviation between the projected and actual pollution levels, with lower values indicating better fit.

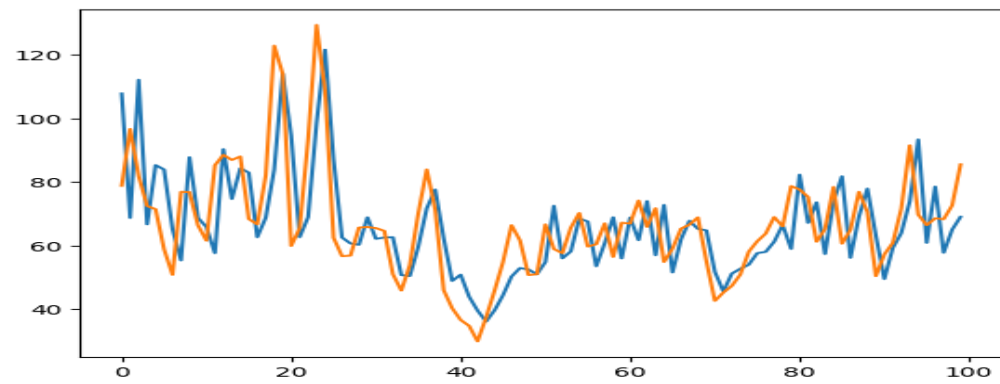
The results of models are as following:



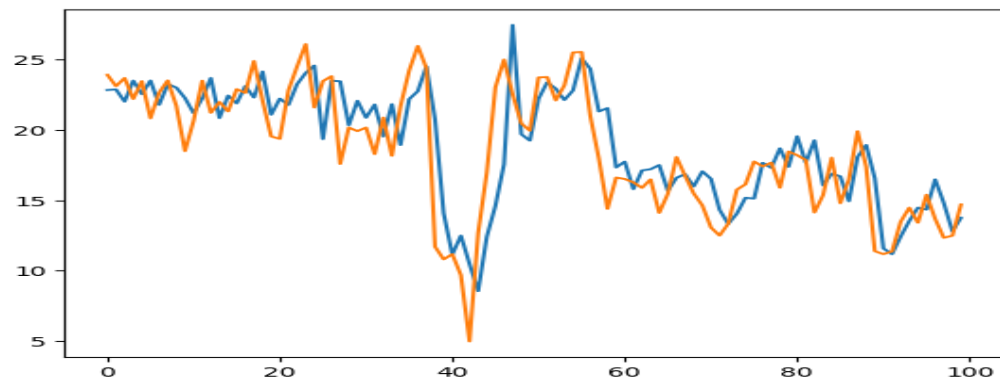
((a)) PM2.5 Actual (Blue) vs Predicted(Orange)



((b)) PM 10 Actual (Blue) vs Predicted(Orange)

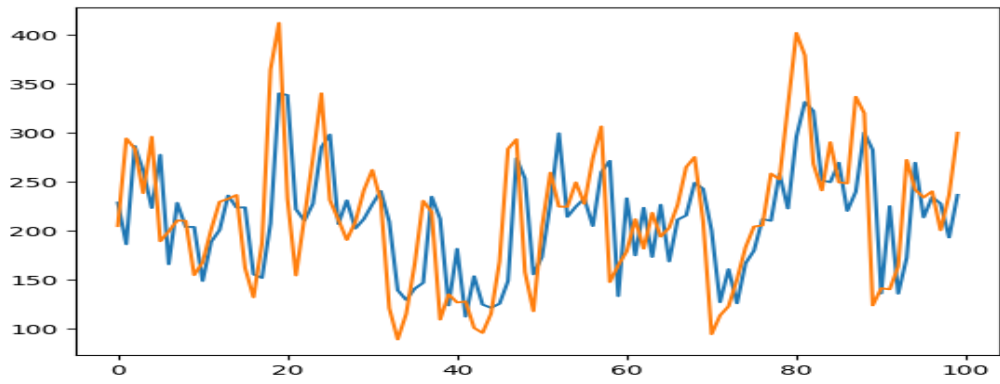


((c)) NO2 Actual (Blue) vs Predicted(Orange)

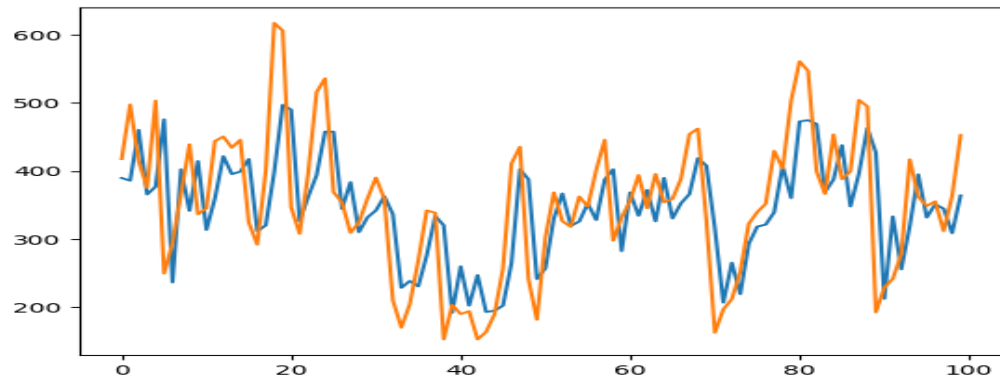


((d)) SO2 Actual (Blue) vs Predicted(Orange)

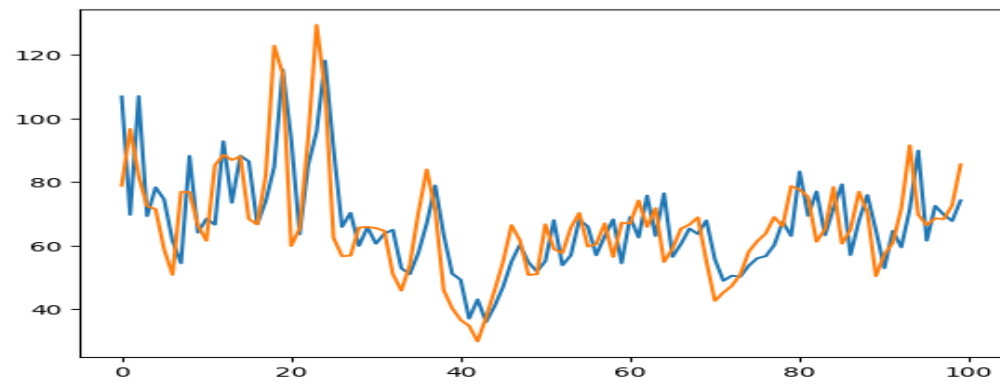
Figure 4.5: Time Series Forecasting using GRU Model



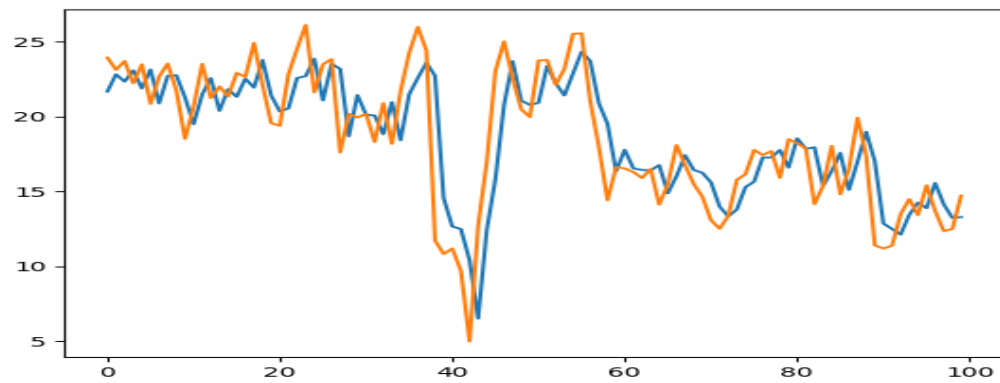
((a)) PM2.5 Actual (Blue) vs Predicted(Orange)



((b)) PM 10 Actual (Blue) vs Predicted(Orange)

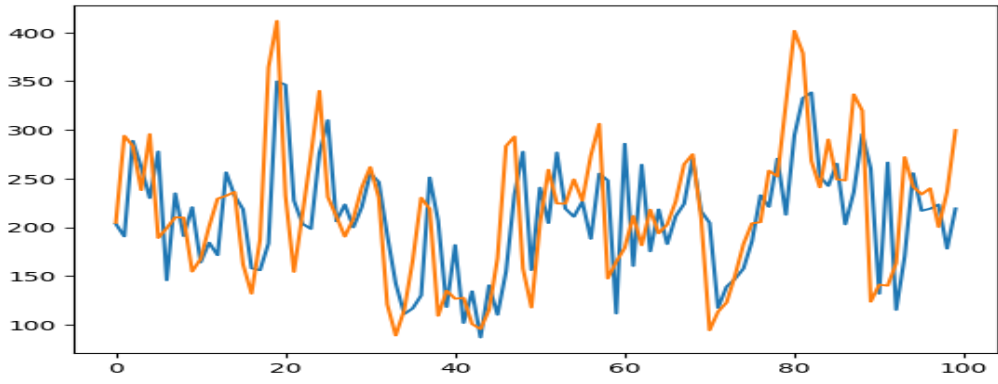


((c)) NO2 Actual (Blue) vs Predicted(Orange)

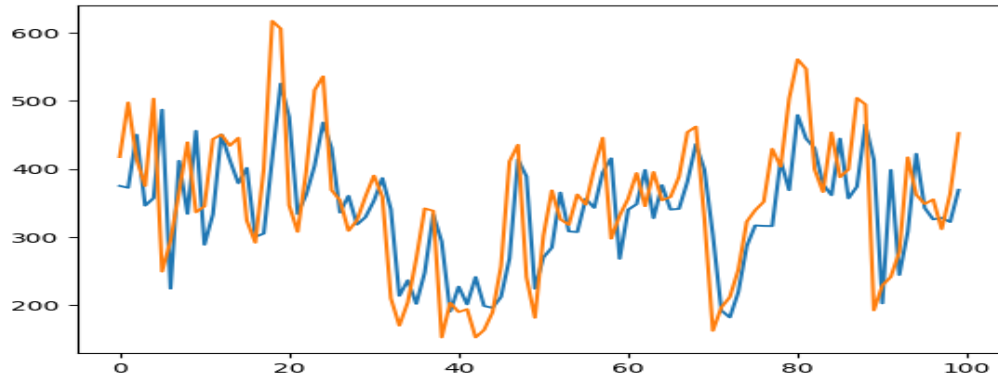


((d)) SO2 Actual (Blue) vs Predicted(Orange)

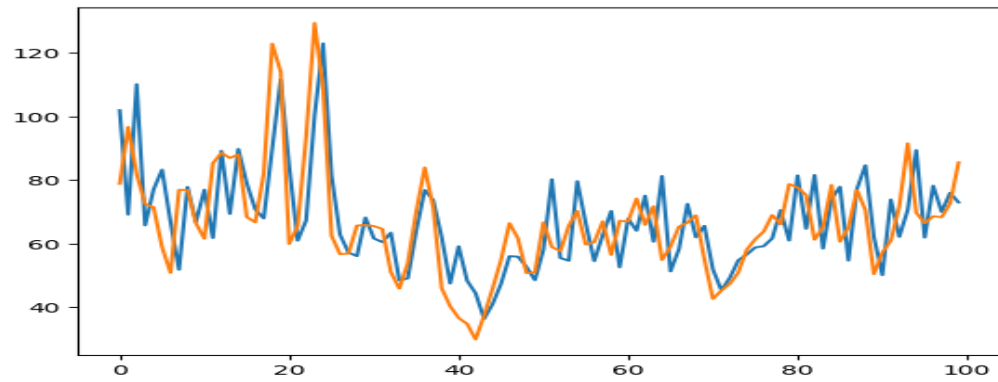
Figure 4.6: Time Series Forecasting using LSTM Model



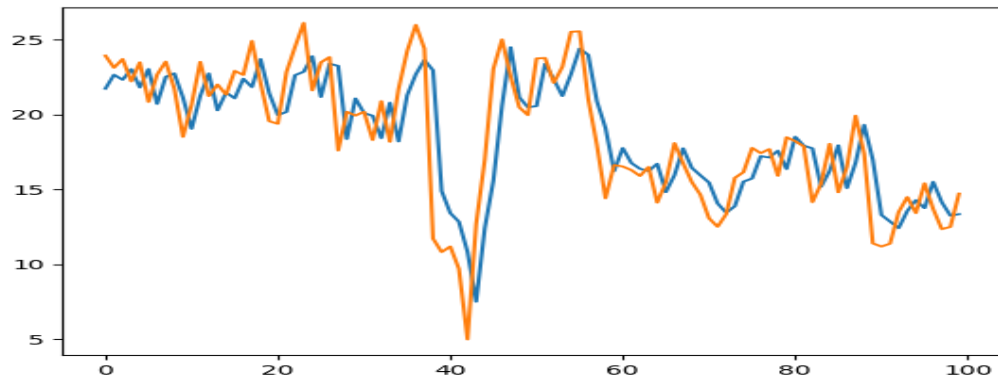
((a)) PM2.5 Actual (Blue) vs Predicted(Orange)



((b)) PM 10 Actual (Blue) vs Predicted(Orange)

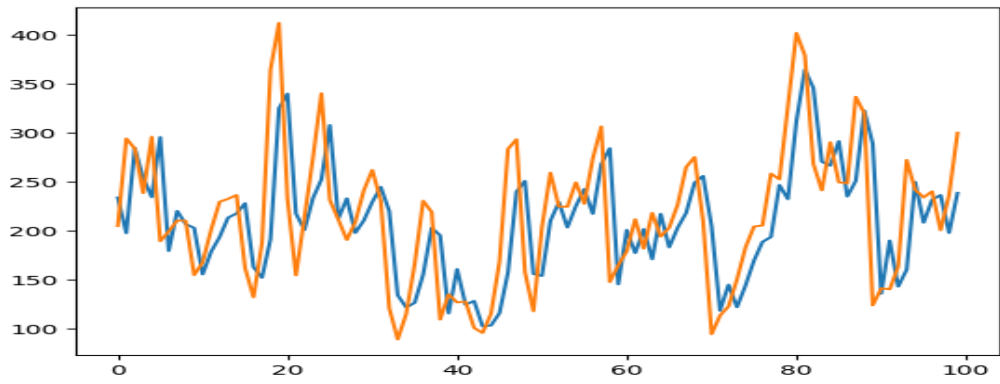


((c)) NO2 Actual (Blue) vs Predicted(Orange)

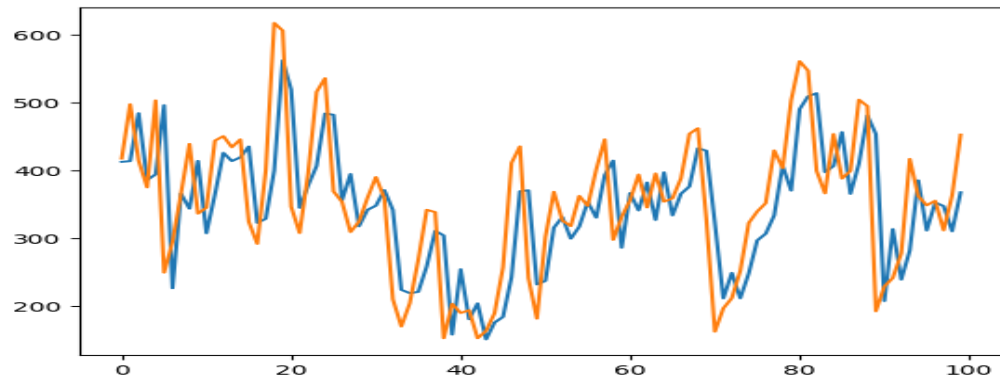


((d)) SO2 Actual (Blue) vs Predicted(Orange)

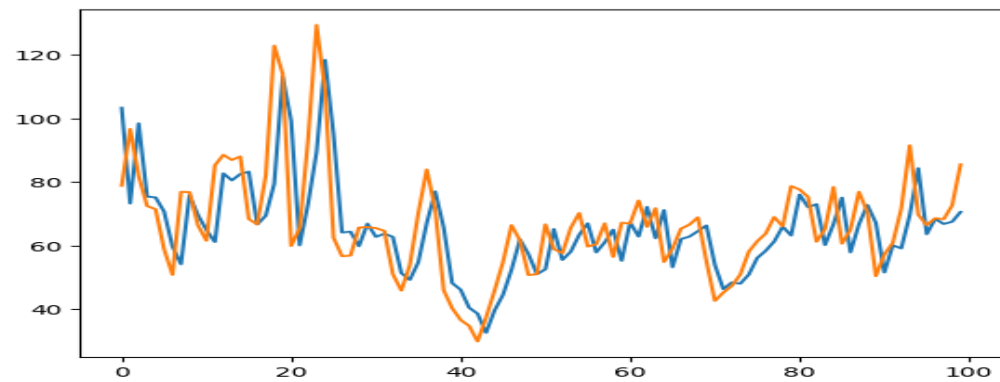
Figure 4.7: Time Series Forecasting using Vanilla LSTM Model



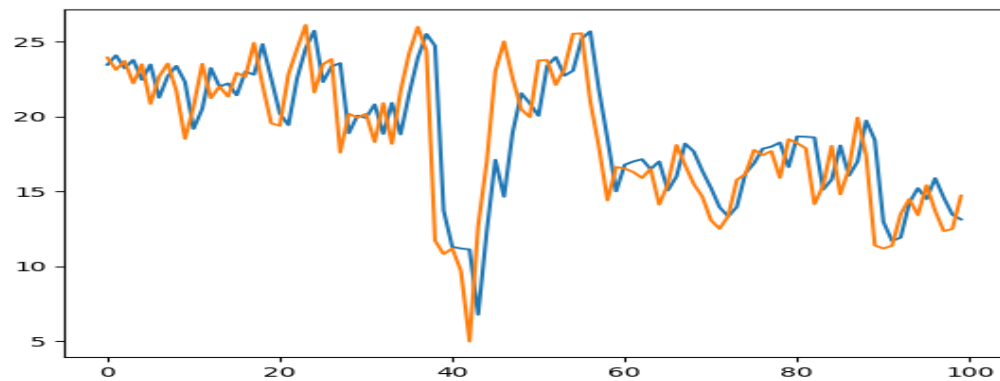
((a)) PM2.5 Actual (Blue) vs Predicted(Orange)



((b)) PM 10 Actual (Blue) vs Predicted(Orange)

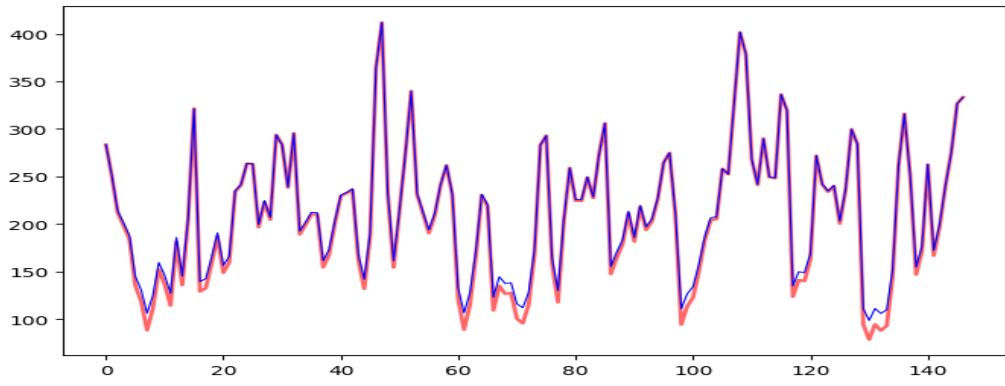


((c)) NO2 Actual (Blue) vs Predicted(Orange)

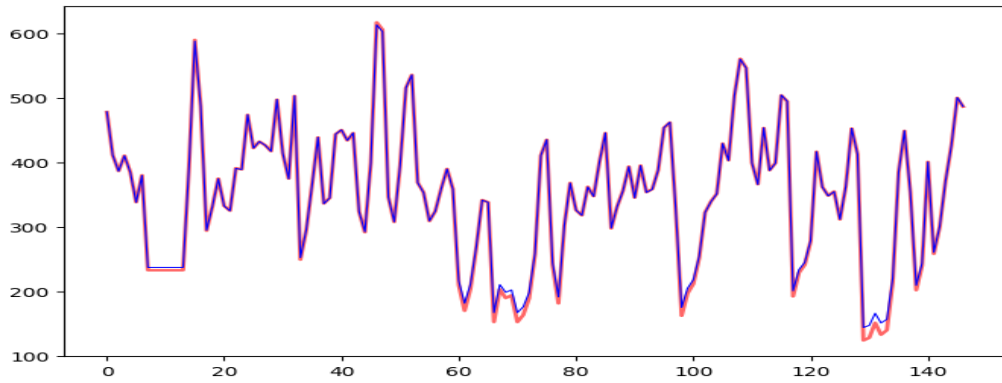


((d)) SO2 Actual (Blue) vs Predicted(Orange)

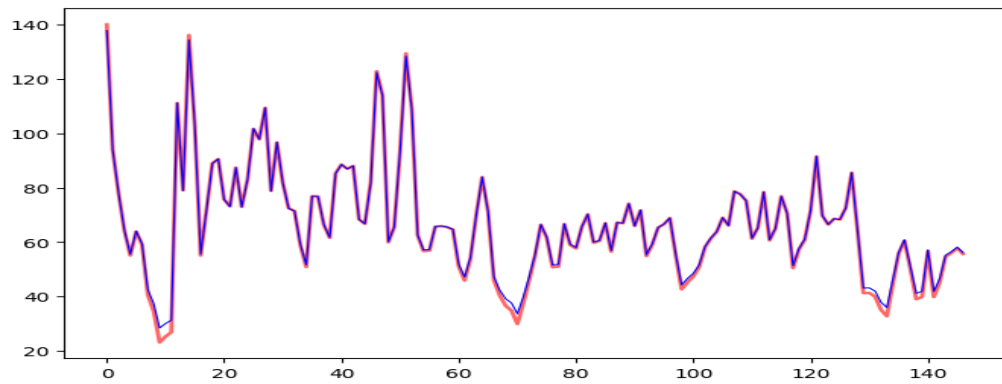
Figure 4.8: Time Series Forecasting using CNN-LSTM Model



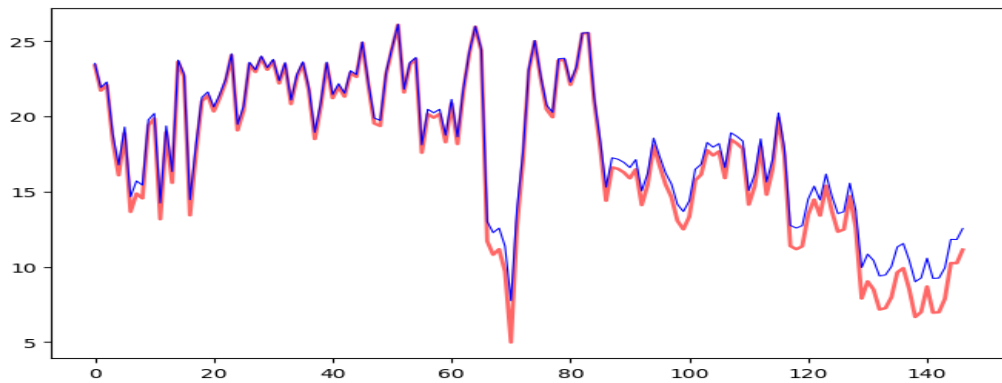
((a)) PM2.5 Actual (Blue) vs Predicted(Red)



((b)) PM 10 Actual (Blue) vs Predicted(Red)



((c)) NO2 Actual (Blue) vs Predicted(Red)



((d)) SO2 Actual (Blue) vs Predicted(Red)

Figure 4.9: Time Series Forecasting using SVR Model

Models	RMSE	Exe-Time (sec)	Accuracy	R-Squared
GRU Models	58.708	3,103	76.25	0.594
CNN-LSTM Models	57.547	3,756	76.95	0.671
LSTM Models	58.532	4,326	75.73	0.591
Vanilla LSTM Models	58.437	1,305	76.04	0.573
SVR Models	06.874	78	96.40	0.873

Table 4.1: Result on Pollution Dataset (PM2.5)

Models	RMSE	Exe-Time (sec)	Accuracy	R-Squared
GRU Models	81.283	3,658	78.94	0.445
CNN-LSTM Models	74.067	3478	78.27	0.734
LSTM Models	80.457	3,324	78.69	0.695
Vanilla LSTM Models	81.960	1,825	78.16	0.564
SVR Models	04.889	69	98.70	0.936

Table 4.2: Result on Pollution Dataset (PM10)

Models	RMSE	Exe-Time (sec)	Accurecy	R-Squared
GRU Models	12.009	2,285	84.40	0.780
CNN-LSTM Models	11.152	248	85.46	0.757
LSTM Models	11.877	915	85.17	0.770
Vanilla LSTM Models	13.671	652	82.72	0.995
SVR Models	01.162	115	98.47	0.919

Table 4.3: Result on Pollution Dataset (NO2)

Models	RMSE	Exe-Time (sec)	Accuracy	R-Squared
GRU Models	2.517	906	86.70	0.775
CNN-LSTM Models	1.818	353	85.86	0.836
LSTM Models	2.530	828	86.61	0.822
Vanilla LSTM Models	2.522	506	86.45	0.743
SVR Models	0.941	87	93.70	0.773

Table 4.4: Result on Pollution Dataset (SO2)

CHAPTER 5

Conclusions

This study aims to examine the relationship between air pollution and temperature variables using quantile regression, a statistical technique that allows for estimating the conditional median and other quantiles of the response variable. Unlike ordinary least squares regression, which only captures the average effect of the explanatory variables, quantile regression can reveal how the effect varies across different levels of the pollution distribution, such as the 10th or 90th percentile. This method can provide more insights into the variability and diversity of PM_{2.5}, PM₁₀, NO₂, and SO₂ concentrations, and help identify the changing trends and causes of air pollution in different regions and seasons.

Apply deep learning models for time series forecasting on pollution data, and to compare their performance with other conventional models. Deep learning models are effective and powerful tools that can handle the complexity and uncertainty of the pollution data, and capture the complex and nonlinear patterns in the pollution levels. Among the deep learning models, the SVR model is selected as the best model based on its accuracy and robustness. The SVR model is able to forecast the pollution levels of PM_{2.5}, PM₁₀, NO₂, and SO₂ with high accuracy and low error. The results of this study can provide valuable information for policy makers and environmental managers to monitor and control the air quality in different regions and seasons.

This study has significant implications for the practical solutions of air quality prediction and analysis. By applying advanced statistical and machine learning techniques, this study can help humans to understand the trends and factors of air pollution in India's metro cities, and to design and implement appropriate measures to mitigate its adverse impacts on human health and environment. This study can also provide useful insights challenges of air pollution and its consequences.

References

- [1] P. Anand, R. Rastogi, and S. Chandra. A pinball loss function based support vector quantile regression model. *arXiv preprint arXiv:1908.06923*, 2019.
- [2] P. Anand, R. Rastogi, and S. Chandra. A quantile regression model with automatic accuracy control. *arXiv preprint arXiv:1910.09168*, 2019.
- [3] P. Anand, R. Rastogi, and S. Chandra. A new asymmetric -insensitive pinball loss function based support vector quantile regression model. *Applied Soft Computing*, 94:106473, 2020.
- [4] M.-A. C. Bind, B. A. Coull, A. Peters, A. A. Baccarelli, L. Tarantini, L. Cantone, P. S. Vokonas, P. Koutrakis, and J. D. Schwartz. Beyond the mean: quantile regression to explore the association of air pollution with gene-specific methylation in the normative aging study. *Environmental health perspectives*, 123(8):759–765, 2015.
- [5] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin. An lstm-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, 11(8):1451–1463, 2020.
- [6] Y. Cheng, X. Li, Z. Li, S. Jiang, and X. Jiang. Fine-grained air quality monitoring based on gaussian process regression. In *Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014. Proceedings, Part II 21*, pages 126–134. Springer, 2014.
- [7] S. K. Dhaka, G. Longiany, V. Panwar, V. Kumar, S. Malik, N. Singh, A. Dimri, Y. Matsumi, T. Nakayama, S. Hayashida, et al. Trends and variability of pm_{2.5} at different time scales over delhi: Long-term analysis 2007-2021. *Aerosol and Air Quality Research*, 22:220191, 2022.
- [8] S. Dutta, S. Ghosh, and S. Dinda. Urban air-quality assessment and inferring the association between different factors: A comparative study among delhi, kolkata and chennai megacity of india. *Aerosol Science and Engineering*, 5:93–111, 2021.

- [9] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, and A. Rahman. Svr based dense air pollution estimation model using static and wireless sensor network. In *2016 IEEE SENSORS*, pages 1–3. IEEE, 2016.
- [10] S. K. Jindal, A. N. Aggarwal, and A. Jindal. Household air pollution in india and respiratory diseases: current status and future directions. *Current Opinion in Pulmonary Medicine*, 26(2):128–134, 2020.
- [11] P. Joshi, N. J. Raju, N. S. Siddaiah, and D. Karunanidhi. Environmental pollution of potentially toxic elements (ptes) and its human health risk assessment in delhi urban environs, india. *Urban Climate*, 46:101309, 2022.
- [12] P. Kumar. A critical evaluation of air quality index models (1960–2021). *Environmental Monitoring and Assessment*, 194(5):324, 2022.
- [13] S. Kumar and S. Dwivedi. Impact on particulate matters in india’s most polluted cities due to long-term restriction on anthropogenic activities. *Environmental Research*, 200:111754, 2021.
- [14] Y. Liu, J. Tian, W. Zheng, and L. Yin. Spatial and temporal distribution characteristics of haze and pollution particles in china based on spatial statistics. *Urban Climate*, 41:101031, 2022.
- [15] A. Loganathan, P. Sumithra, and V. Deneshkumar. Estimation of air quality index using multiple linear regression. *AEES*, 10:717–722, 2022.
- [16] S. Mandal, K. K. Madhipatla, S. Guttikunda, I. Kloog, D. Prabhakaran, J. D. Schwartz, and G. H. I. Team. Ensemble averaging based assessment of spatiotemporal variations in ambient pm_{2.5} concentrations over delhi, india, during 2010–2016. *Atmospheric Environment*, 224:117309, 2020.
- [17] S. A. Meo, S. A. Alqahtani, R. A. AlRasheed, G. M. Aljedaie, R. M. Albarrak, et al. Effect of environmental pollutants pm_{2.5}, co, o₃ and no₂, on the incidence and mortality of sars-cov-2 in largest metropolitan cities, delhi, mumbai and kolkata, india. *Journal of King Saud University-Science*, 34(1):101687, 2022.
- [18] A. Mhawish, C. Sarangi, P. Babu, M. Kumar, M. Bilal, and Z. Qiu. Observational evidence of elevated smoke layers during crop residue burning season over delhi: Potential implications on associated heterogeneous pm_{2.5} enhancements. *Remote Sensing of Environment*, 280:113167, 2022.

- [19] A. Mishra, Z. M. Jalaluddin, and C. V. Mahamuni. Air quality analysis and smog detection in smart cities for safer transport using machine learning (ml) regression models. In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 200–206. IEEE, 2022.
- [20] A. S. Mohan and L. Abraham. An ensemble deep learning model for forecasting hourly pm_{2.5} concentrations. *IETE Journal of Research*, pages 1–14, 2022.
- [21] Z. Mushtaq, P. S. Bangotra, S. Sajad, A. S. Gautam, M. Sharma, K. Singh, Y. Kumar, P. Jain, S. Gautam, et al. Comparative analysis of particulate matter (pm_{2.5}, pm₁₀) and trace gases (so₂, no₂, o₃) in between satellite derived data and ground based instruments. 2023.
- [22] S. Nenavath. Impact of fintech and green finance on environmental quality protection in india: By applying the semi-parametric difference-in-differences (sdid). *Renewable Energy*, 193:913–919, 2022.
- [23] S. Pandya, T. R. Gadekallu, P. K. R. Maddikunta, and R. Sharma. A study of the impacts of air pollution on the agricultural community and yield crops (indian context). *Sustainability*, 14(20):13098, 2022.
- [24] A. K. Sharma, P. Baliyan, and P. Kumar. Air pollution and public health: the challenges for delhi, india. *Reviews on environmental health*, 33(1):77–86, 2018.
- [25] G. K. Sharma, A. Tewani, and P. Gargava. Comprehensive analysis of ambient air quality during second lockdown in national capital territory of delhi. *Journal of Hazardous Materials Advances*, 6:100078, 2022.
- [26] M. Sharma, E. Gupta, and D. Viji. Air quality index (aqi) prediction using automated machine learning with tpot-ann. In *2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, pages 1–9. IEEE, 2023.
- [27] S. Swain, S. Sahoo, and A. K. Taloor. Groundwater quality assessment using geospatial and statistical approaches over faridabad and gurgaon districts of national capital region, india. *Applied Water Science*, 12(4):75, 2022.
- [28] Q. Tao, F. Liu, Y. Li, and D. Sidorov. Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. *IEEE access*, 7:76690–76698, 2019.

- [29] F. Yao and H.-G. Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010.
- [30] Q. Zhang, J. C. Lam, V. O. Li, and Y. Han. Deep-air: A hybrid cnn-lstm framework for fine-grained air pollution forecast. *arXiv preprint arXiv:2001.11957*, 2020.