National College of Ireland

Forecasting the air pollution in New Delhi using deep Learning methodology with Covid-19 lockdown focus

MSc Research Project
Data Analytics

Kumar Parakram Singh
Student ID: x20253788

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Kumar Parakram Singh
Student ID:	x20253788
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	15/08/2022
Project Title:	Forecasting the air pollution in New Delhi using deep Learning methodology with Covid-19 lockdown focus
Word Count:	5340
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Kumar Singh
Date:	13th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting the air pollution in New Delhi using deep Learning methodology with Covid-19 lockdown focus

Kumar Parakram Singh
x20253788
MSc Research Project in Data Analytics

15/08/2022

Abstract

New Delhi has witnessed a sharp spike in industrialization within a very short time frame. The primary source of pollution in Delhi is the pollutants coming from meteorological activities, vehicles and heavy industries. In this study, pollutant data from four locations in Delhi - Anand Vihar, DTU, Bawana and Vivek Vihar have been analysed. In this paper, the dataset from the Delhi Air Pollution Control Board has been used.[1] The critical reason for choosing the pollutant Particulate Matter 2.5 is because of its lethal nature. In this paper, multivariate analyses have been done with the help of Long Short Term Memory (LSTM) -deep learning methodologies. Recent versions of LSTM like the Encoder-Decoder-LSTM, Bi-Directional LSTM, and LSTM-Forward Neural Network have also been implemented and analysed. In terms of novelty, this paper implements the technique (ten steps ahead or multi-step ahead) short-term air quality forecasting method. In this paper, one month's future air pollutant Particulate Matter 2.5 predictions have been done. In addition, twelve predictor variables with eighty hours of data have been used. The accuracy of the models has been evaluated with the help of the Root Mean Square Error evaluation matrix. The impact of the Covid-19 lockdown in Delhi has also been investigated, and it was evaluated that the air quality deteriorated once the restrictions were relaxed. It was evaluated that the bi-directional LSTM with the least RMSE error was the best prediction model.

Keywords: Industrial emissions, Multi-Step ahead method, Covid-19, Long Short Term Memory, Root Mean Square Error, Forecasting

1 Introduction

1.1 Background and Motivation

The population of Delhi rose astronomically in the last decade. The World Economic Forum in its latest report on pollution worldwide has picked 10 cities with the most

[1] https://aqicn.org/data-platform/register/)

pollution. In this list, Delhi is also included. Moreover, medical research has revealed that with every increase in the concentration of the pollutant Particulate Matter 2.5 particulate, the risk of lung cancer and heart stroke increases by nearly 6 per cent. In Delhi, the concentration of population in a localised area can be attributed to the places with the least economic development. It has also been studied that the concentration of Particulate Matter (PM) 2.5 in Delhi is higher than in any other capital city of the world. To uplift millions of people from poverty, this challenge has forced the Indian government to industrialize its key cities, Delhi in particular. Industrialization and an increase in vehicle counts have contributed to an increase in PM 2.5. Most of the research papers have established both the environment and the health of humans are co-related. In recent times, the most important research area is in the field of human health. These research works need backing from the soft-wares, statistics, mathematics implementation and deep science to establish any hypothesis in the field of air pollution prediction. The principal difference between simulation and numerical forms of modelling is their accuracy. For instance, there are several strategies to limit air pollution we can reduce air pollution by micro-managing the solid waste pollutants and implementing good engines in the vehicles to name a few. But in this paper, air pollution forecasting uses the air pollutants in the air. The primary pollutant in Delhi is the Particulate Matter 2.5, and Particulate Matter 10. It additionally includes a rising percentage of Nitrogen dioxide, Sulphur dioxide and Ozone. It has also been analysed that air forecasting or air pollution management should not merely be done in a short range but also for a more prolonged duration of time. In the given below the figure, the causes of PM 2.5 Delhi pollution have been shown below the figure. 1

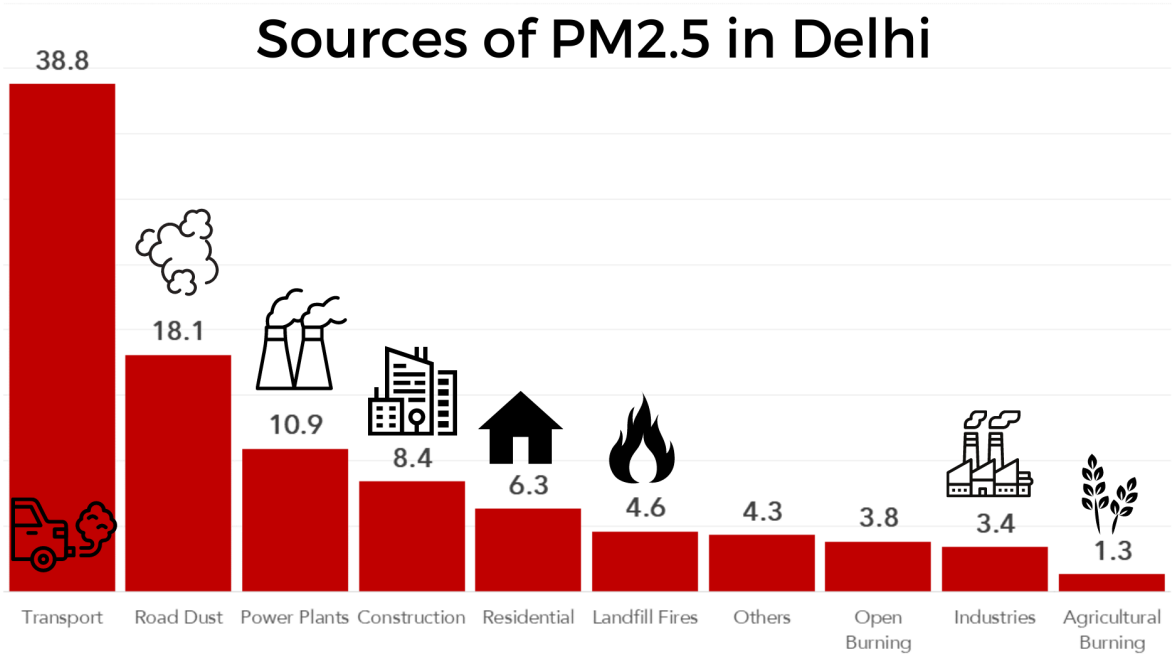


Figure 1: Delhi PM 2.5 Pollution Sources
Figure source : link

1.2 Impact of Corona Virus

The Corona Virus became a world pandemic in the year if 2019. It was contiguous and forced several nations to completely close their borders and lockdown their cities. Similar was the case with New Delhi which has come under strict lockdown to contain the spread of the virus. This lockdown has critically paralysed the global economy. There have been a number of studies done on the impact of COVID -19 in several parts of the world. It was detected that there has been a drastic decrease in air pollution due to less number of vehicles on the road whereas in some cases no significant drop in the pollutant due to meteorological reasons. There has been a significant study on COVID -19 with the help of machine learning methods have been used. But there are several deep learning and advanced versions of LSTM methods that can be used to get better results.

In this research paper, the novel approach of multi-step ahead(10-steps) prediction has been used. Here extended duration forecasting has been done in four locations in New Delhi has been done. In our method, 12 pollutants have been used. These pollutants are Benzene, Toluene, Ozone, Sulphur Dioxide, Carbon Monoxide, Nitric Oxide, Nitrous Oxide, PM 2.5, PM 10, Ammonia Gas, Wind Speed and Nitrogen Oxides. The time frame between two observations takes place 8 hours. At this juncture, our core objective is to provide how the air quality changes after and before lockdown. In this report, one month ahead of data prediction has been done.

1.3 Specification

Question 1: " The after and before the impact of COVID-19 lockdown on the Delhi pollution can be assessed satisfactorily by LSTM methodology or not?"

Question 2: "Whether the novel technique followed in this paper i.e multi-step ahead LSTM methodology can provide more impressive results when compared to other machine learning forecasting techniques or not?"

1.4 Contribution

The main contribution of this paper is:

1. Accurate and elaborate forecasting of the air pollution at four locations in New Delhi. This paper might help the local health authorities in Delhi to incorporate health measures to mitigate the air pollution hazards at some key venues.

2. State of the art comparison of various machine learning models to predict the Delhi air pollution.

1.5 Outline

In the following sections, all the key attributes of the paper have been organised as follows. First of all, in section 2, literature reviews on the previous work on Delhi air pollution have been studied. The subsequent section (section 3) elaborates on the methodologies implemented while undertaking Covid-19 analysis. Then in the section, the results obtained after the experiments have been analysed. In the end, the conclusion and future work is presented and discussed.

1.6 Air pollution Study Area

The city of New Delhi has been selected for this paper. The sensor data from the geographical area of four locations are chosen to obtain the prediction of air quality. There are 20 air pollutants monitoring stations are there in Delhi. The reason for preferring the below-given locations is that the traffic and population density are highest in those four locations. In addition, most of the iconic places of India and places of international importance are also there in those four areas. Any suggestions to curtail air pollution can significantly help the concerned authorities. The locations selected for this paper are given below:

1. Bawana
2. Ashok Vihar
3. Vivek Vihar and
4. DTU

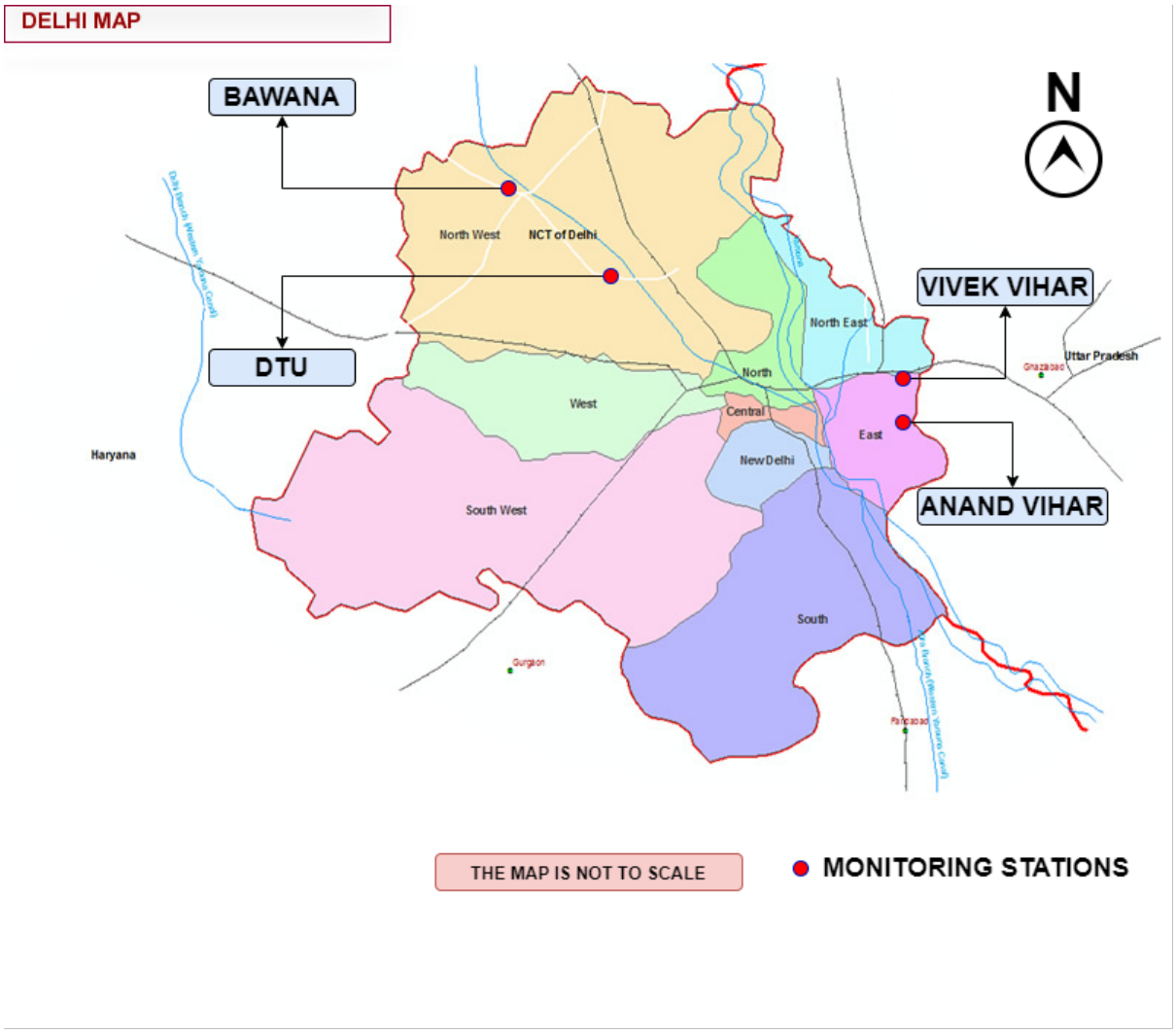


Figure 2: Delhi PM 2.5 Pollution Monitoring Stations
Figure source : link

2 Related Work

The situation of pollution in New Delhi is in dire condition. There are numerous studies that have already been done on air pollution and the root cause analysis has also been evaluated by researchers from almost all parts of the world **(Sinnott and Guan; 2018)**. From the year 2000 - 2010, there has been a tremendous increase in the number of vehicles in Delhi causing enormous air pollution. Even the government of India also funded millions of dollars to tackle this issue. The main idea behind this report is to predict the future value based on the historical values of the Delhi Air Pollution Dataset. For evaluation purposes, machine learning, deep learning methodology and statistical methods are used.

2.1 Work Done in Machine learning field and Covid -19 analysis.

Schmidhuber et al. (1997) In this paper bi-directional methodology variant of LSTM has been used. The author evaluated that the LSTM methods used indeed improved the results. The key reason for implementing the bi-directional LSTM is that the output layer in the bi-directional LSTM can extract the information from future and past results in the same instance. Accuracy in the result is obtained when the RNN-Canonical are connected to the opposite hidden layers which in turn are connected to the same output layer.**Hochreiter and Schmidhuber (1997)**.

Salcedo et al. (1999) The researcher evaluated the air pollution in the region of Oporto area of Spain. He has used the basic time series analysis methodology. They have used the SATSA machine learning method. Here, the step-wise method has been used. In this method, black smoke, and acidity concentrations were analysed. In the observation, the researchers have evaluated that there has been a long-term was identified in the cyclic air pollutant components. They have tested is for if there is any trend is there in the dataset or if there are any seasonality elements are there. They used the step-wise implementation and hence in each step they searched for residual correlation analysis was done. The most important implication of the step-wise method was that they had found the white noise which is used in the time series forecasting analysis.

In the paper **(Taieb et al.; 2012)** evaluated the Direct H step methodology used in LSTM modelling technique. In this model, a number of models say n, has been made to forecast n given steps. In this method, forecasting any other point in say f(time + 2) needs to be analysed and calculated from the value of f(time +1) first. It was also evaluated that the second forecasting point was independent of the first forecasting point.

Researcher **Sindhwani and Goyal (2014)** evaluated that there have been cyclic changes in the pollution of New Delhi. The researcher had divided the Delhi geography into several locations. In their research, they used the dataset from the year 2000 - 2015. One of the key findings in their research paper was that they witnessed a 90 per cent increase in particulate matter 2.5 over the years. But on another side, they have failed to provide credible results due to their limited air pollutants analysis and also neglecting other environmental factors.

The Researchers like the **Zheng et al. (2016)** evaluated the correlation between lung cancer and the concentration of PM 2.5 and PM 10. They have used time series analysis on the particulate matters PM 10 and PM 2.5.

In the year 2016**Li et al. (2016)** in this paper, the author evaluated various PM 2.5

models based on the aerosol optical depth methodologies. The author's work was based on the correlation between traffic pollution and the meteorological data for the city of Madrid, Spain. He evaluated various other machine learning models like the random forest method, Deep Learning Methodologies, GBRT - Gradient Boosting regression technique and the extra tree models. In his evaluation, the researchers found that the RMSE error for the extra tree model that is used to evaluate the machine learning model was the least and hence, was able to answer the pollution question effectively.

In the year 2017 **Saeed et al. (2017)** the researcher worked on the implication of the LSTM machine learning method in forecasting the pollution in the air. He used open-source datasets. His main objective was to find the air pollution in the area of the middle east. He worked on the hourly concentration of air pollutants. He further used the meteorological datasets to predict over the next day. He used the multi-task learning methodology to adhere to this problem.

In the research paper, **Ye (2019)** the researcher has evaluated that in New Delhi, the average life of an individual can be increased by up to 10 years, if we decrease the particulate matter PM 2.5 to 15 ccm.[2]In addition to this the global footprint of death caused by air pollution death is nearly 4.2 million people.

Vu et al. (2019) also worked on the air pollution methodology. He worked on the LUR method. In the Land Use Regression methodology, he evaluated and used its boundary conditions. In addition, he also worked on the gradient boost method and the Artificial Neural Network. He evaluated that the LUR method performed better than the later methodologies. He further suggested that the LUR method can be improved by better understanding the explanatory variables used in the machine learning methodology. His work provided the key insight that PM 2.5 was the primary pollutant.

In the year of 2020, **Kalajdjieski et al. (2020)** the researcher worked on installing cameras to capture photographic images of the air pollutants in the concerned area. In addition to cameras, they have also installed sensors to get the real-time values of the air pollutants like Sulphur dioxide, ozone, PM 2.5 and PM 10. He further enhanced the results by combining the images and the sensor data. The machine learning methodologies which were used were time multi-layer perceptron and linear regression methods. The author fails to evaluate properly the time series analysis and hence his work does not answer specific questions.

in the year 2020**Castelli et al. (2020)** researcher worked on the support vector regression machine methodology to forecast air pollution. The kernel employed in this technique was the RBF method (the radial basis function was used.). Furthermore, he worked on the hourly pollutants like PM 2.5, sulphur dioxide and carbon monoxide. He worked on a comparative study of ANN (Artificial Neural Network), GWR (Geographical Weighted regression), SVR (Support Vector Regression) and NARX (Non -linear Auto regressive Exogenous Model) methodologies to predict air pollution. In his results, he evaluated that the NARA method performed most optimally.

Stephens et al. (2020) the researcher has conducted his work on the model negative bi-model to study the exposure of pollutants like PM 2.5, Sulphur dioxide to the environment. The researcher also evaluated the impact of the Australian wildfire on the environment. He evaluated that the current 2020 wildfire causes the most havoc on the health of the Australian citizens compared to the previous twenty recorded wildfire pollutant impacts. In addition, the a significant increase in premature deaths related to wildfire pollutants. Also, during this time, there has been a significant increase in the

[2]`https://www.news-medical.net/news/20220616/`

patients related to cardiovascular and respiratory disorders patients in the Australian medical units have been witnessed.

In addition to PM 10 and PM 2.5 **Usmani et al. (2020)** researchers also evaluated other pollutants like the impact of ozone, nitrogen dioxide and sulphur dioxide. Their work was carried out in Malaysia. Malaysia is also densely populated and hence, the particulate emissions have a similar pattern to that of Delhi. The Researchers have used a dataset to comprise emissions from the industries. He also has used meteorological data and traffic sensor data. They have used deep learning methodology like the Artificial Neural Network for forecasting and analysing air pollution. They used data on 4-hour intervals. The researchers have also used the backpropagation method. For the backpropagation, they have used four layers of perceptron layers have been used. The most limiting factor for their methodology was the lack of other machine learning methods which makes their findings less trustworthy. The result obtained in their research was merely 0.62 R 2 co-efficient which makes their work less authentic.

Sur et al. (2020) In this paper, the researcher worked on the implementation of the Fb-Prophet prediction model on the Delhi air pollution dataset has been implemented. This model has never been implemented in the research on air pollution in New Delhi. The dataset from four stations has been selected from the observatory stations in Delhi. The prophet model has been implemented to predict the coming 1 month's quality of the air pollution in Delhi. The basic reason to adopt this model was that it provided good predictions while enduring holidays and seasonality defects in the datasets[3]. In addition, it uses intuitive and simple parameters in configuration. When the government tried to replace the old model vehicles cars with CNG (Compressed Natural Gas) fitted automobiles, there has been a significant drop of PM 2.5 and PM 10 particulates have been observed. While some measures were remarkable but few promising methods like the implementation of the odd-even number plate care model failed in Delhi. In the early winters of 2016, the government implemented the ODD - Even models. Here, the vehicle owners with odd number plate were allowed for one day and even number plate was allowed during the next day. But this method proved fruitless while causing inconvenience to Delhi residents.

For forecasting analysis, a number of deep learning methods have been researched for this paper. To begin with, RNN has been studied. RNN is a Recurrent Neural Network. The RNN technique is used for long-term dependency-modelling of the temporal sequences of the data.**Chang et al. (2020)** RNN was studied because in this paper a one-month prediction of PM 2.5 needs to be done and hence all methodologies which work well for long-term forecasting analysis were studied. Secondly, in this paper, the advanced version of RNN, LSTM is studied and implemented which addresses the long-term dependencies faced by the RNN. In the LSTM method, particular sub-forms of LSTMs like the bi-direction LSTM and LSTM (Encoder - Decoder) have also been studied.

There have been multiple studies that have been carried out to relate the impact of Covid-19 lockdown on local area pollutant levels. In China, the researcher **(Zhu et al.; 2020)** worked on the demography of China and found that there is a co-relation between Covid-19 infection and air pollution in China. In addition to this researcher **(Kerimray et al.; 2020)** worked on the dataset of Kazakhstan. According to his findings, the covid-19 lockdown played no role in the reduction of air pollutants. Meteorological variations played a significant role in increasing the air pollutants whereas there was minute temporal reduction was recorded in the Kazakh cities. There are several coal plants were

[3]`https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python`

operational in Kazakh and China to supply energy during the lockdown period. Hence, all these coal power plants played a significant role in maintaining the pollutant level decreased by reducing cars on roads. In addition, to these, another researcher **(Li et al.; 2020)** worked on the effect of the Covid lockdown on the Yangtze River flowing through the northern part of the China delta. The researcher has found that there had been no decrease in the amount of Ozone in that area. It was also noted that the residual pollution mainly arising from the industry source was an all-time high.

2.2 Related Work Gaps Identified

Most of the research workers have worked on simple time series analysis. Most of the research works lacked hyper-tuning or other advanced machine learning techniques for parameter tuning. In addition, the researchers have used visual graphics like meteorological pictures for air quality analysis. For overcoming this scenario, in this paper, we have used eight minutes analyses of particulate matters PM 2.5 analyses. In the previous paper, the researchers have not been able to handle the data set seasonality and the effect of holidays. By introducing the LSTM in this paper, the missing values, seasonality and outlier problems in the datasets are tried to be answered and the providing accurate 1 month ahead PM 2.5 predictions is also been addressed.

2.3 Summarizing Related Work

After an exhaustive literature review, it has been identified that the two most important pollutant for Delhi air pollution is PM 10 and PM 2.5. The main reason for using PM 10 and PM 2.5 is their size. Both these pollutants have the ability to enter the lungs because their size is minuscule. LSTM methodology was used by several researchers and the most important LSTM variant was the bi-directional LSTM method. Bi-LSTM in most of the papers produced better results. One of the key positive aspects of using the LSTM - multi-step-ahead method is that it has the ability able to predict a long period of time. The model can critically handle the missing values and also have resistance towards the outliers.

3 Methodology

CRISP-DM methodology has been used in this paper. This methodology is widely used by almost all contemporary researchers. The key aspect of this method is that it can be altered or manipulated according to the business needs. For example in this paper, the codes and methodology has been continuously has been updated at any time depending upon the requirements.

3.1 Business - Understanding

The business understanding for this research is to be able to make a model which can forecast air quality in Delhi by using data from the four source locations in Delhi. The findings can be shared with the concerned authorities which can help them to make future planning regarding air pollution.

3.2 New Delhi data and report - Situation Interpretation

According to the detail study conducted by the United Nation, Delhi would continue to remain the second moist populated city in year 2035. With this surge in population would put tremendous pressure on natural resource and leading to sharp increase in pollution. With increase in pollution, it would make it difficult for the Delhi residents to sustain their lives in Delhi. The number of vehicles would increase four fold and amount PM 2.5 would also increase drastically. During winter season Delhi is always crippled with very low visibility. The low visibility can be attributed to high concentration PM 10 and PM 2.5 pollutants. In this report, the data is imported from the Delhi Air Pollution Control website. We had chosen four suburbs of Delhi, which are Bawana, Anand Vihar, Vikash Vihar and (Delhi Technological University)DTU. The main reason behind using these four monitoring station was its air quality. Bawana was the most polluted area followed by the DTU, Anand Vihar and Vivek Vihar according to 2020 survey.[4]. The machine learning model formulation depends upon the quality of data and missing data poses a serious risk in data model formulation. The larger the missing values in the dataset, the higher the prediction errors of the models. There are serious concerns about the data missing or NAN values which we have tried to fix in the coming sections. In the given below the figure, a snapshot of data from the Anand Vihar and DTU has been presented. There is a total of 12 parameters have been selected for multivariate analysis. The total interval between two sampled data is eight hours. Some cursory findings were that the Bawana had the highest values both in terms of the mean and highest value for PM 2.5. In other words, the air in Bawana was the most polluted.

3.3 Data Evaluating

For evaluation of the models, two techniques have been used which are RMSE and MSE. In this report, plotting techniques like the PACF and ACF is also been used. ACF and PACF are used to evaluate the residual values.

1. RMSE (Root Mean Square Error) - RMSE is defined as the difference between the population/ sample values and the values which we get after observation in the experiment. In our experiment, the prediction performance has been evaluated by the RMSE values. The equation used is given below.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2} \tag{1}$$

4 Design and Implementation

4.1 LSTM - Method

In this paper, the LSTM methodology has been used. The principle of LSTM relies on RNN advanced architecture. RNN - Recurrent Neural Network, nowadays has far more popular than the feed-forward method. RNN is a simple methodology that is significantly trained by the backward propagation method. It uses the gradient descent

[4] https://www.hindustantimes.com/environment/delhi-had-7-of-india-s-10-bad-air-hot-spots-last-yea
html

method. The LSTM technology was formed on the basis of all deep learning methods. It works smartly in remembering long-term dependencies. These dependencies are also referred to as architectures. Temporal sequences, it has gates and memory cells. BPTT (Back Propagation Through Time) - is used to train the memory cells in a supervised learning methodology. In this report, an optimiser for LSTM like ADAM has been used to get accurate results. ADAM optimiser implements feature like adaptive learning rate which smooths the learning process of the training data.

4.1.1 Bi - Directional LSTM

RNN demonstrates some flaws like it is only able to contemplate the previous sequences to provide predictions for future values/states. The primary purpose of using - Bidirectional LSTM is to have the sequence information of both back propagation knowledge of previous past and future values. It has two separate hidden layers in the neural network, where both these hidden layers are feed-forward connected to the output same layer. In addition, to it, BD-LSTM is broadly used when the input sequence and the output sequence are known beforehand. The application of BD-LSTM in the case of sentence classification and phoneme classification is studied in detail to understand our objective in this report.

4.1.2 Encoder - Decoder LSTM

In this methodology, it uses variable length input to map variable length output. ED-LSTM was used in this paper because of its ability to address sequence-to-sequence issues. It generally works by first encoding a given sample of input say of length n which has a vector representation (latent) on sequence with m length output. Since it was effectively used in the text simplification process and moreover, its implementation was studied in speech recognition methodology also. In this paper, both the given methods were studied, and their implication was tried to replicate in the current problem.

4.2 Implemented Framework

To begin with the air pollution datasets have been downloaded from the Delhi Air Control Pollution website[5]. Here, data from four of the prominent stations in New Delhi namely Anand-vihar, DTU, Bawana and Vivek Vihar are used. In the given below figure3, the framework has been defined. In this diagram, four base stations have been chosen. In the next step, the data set has been cleaned. In the data wrangling methodology, the null values obtained in the datasets have been replaced with median value imputation. An explanation can be given for the null values regards the inconsistency in data recording from the government employees. Moving forward, for training the dataset in the LSTM methodology the feature and temporal sequences in the dataset have been converted from 2-d to the 3-d model. It has been explained in the figure 4. The data in the dataset have an 8-hour sequence recording. In the LSTM methodology, I have used both uni-variate analysis and multivariate analysis. In the multivariate analysis, five pollutants have been used with an interval of eight hours. After that data pre-processing has been done. In the multivariate analysis deep learning methodology like LSTM and its modern versions like ED-LSTM, BD-LSTM and FNN-LSTM have been implemented. In this process, COVID lockdown change in air pollutant concentration has also been visualized.

[5] (https://aqicn.org/data-platform/register/)

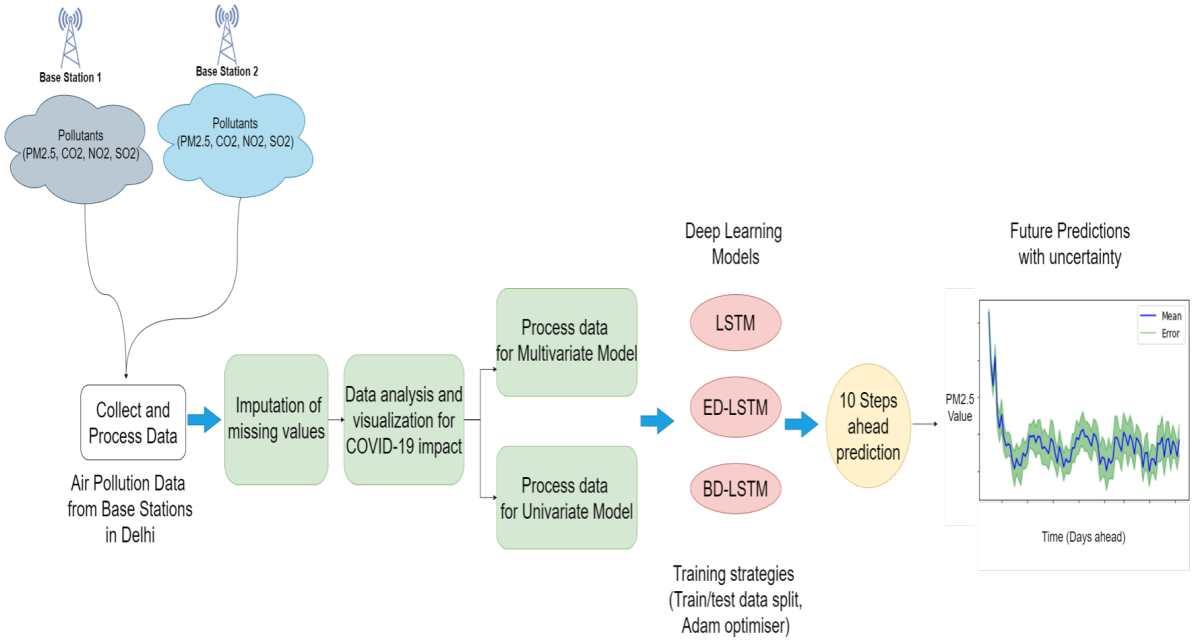


Figure 3: Multivariate Methodology Implemented

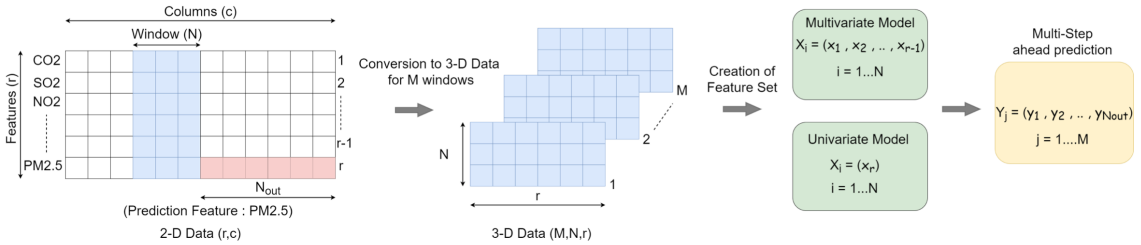


Figure 4: Conversion of 2D -3D Methodology
The figure source link

In the above figure, the concept of two dimensional data to three dimensional data has been explained. In this method, window size of N in dimension and M number of windows have been used. These windows are used for both the uni-variate and multivariate analysis. In the given figure, the window blue in colour is two dimensional and the window in pink colour is the predicted PM 2.5 values. In the framework purpose, LSTM models has used 11 features as N as 5 windows (here n takes into account 5 steps in the past 5 window)and N(f) excluding PM 2.5 in case for multivariate analysis. The concept of window is how much times the models used in LSTM methodology would unfold. Neurons stands for number of feature which is being used at that step of model. N(out) is defined as the number neurons or the steps ahead count. N(out) 10 means 10 step ahead has been used in this paper to predict the PM2.5 concentration. Step ahead 10 means 8 hours multiplied by steps or 80 hours ahead prediction. If we are using the uni-variate analysis then the input neuron is only 1 or 1 step ahead and similarly output neurons would be 10 steps ahead. This concept helps to understand the table 5.1.1.

5 Result and Experiments

In the experiment which I have performed uses three models based on LSTM machine learning predicting models. In the first step, Covid -19 impact analysis was done. In the next step, all three LSTM method's multi-step method has been used to evaluate the air pollutant (PM 2.5). The data has been setup in two ways, one before Covid lockdown and other after Covid to access the impact of Covid 19 lockdown.

In the given below table, the results obtained from the LSTM and its extended versions of deep learning methods have be shown. Here, RMSE value has been used to evaluate model accuracy.

Model	Pollutant	RMSE	Time - Execution(secs)
LSTM	All pollutants	2017.92	1254
LSTM - Bidirectional	All pollutants	1808.92	1298
LSTM - FNN	All pollutants	1821.29	1343
LSTM - ED	All pollutants	1871.44	1232

It can be inferred for the multivariate analysis Bi Directional-LSTM model performed well. Though the execution time was similar for other multivariate analysis. In addition, to it the BD-LSTM performed well on both the training and test datasets.

5.1 Experiment Steps

The process of experiments and visualisations are given below :

1. Visualisation has been done on the PM2.5 concentration for the four air quality stations along with Covid -19 lockdown air quality. In addition, different time intervals of data has been chosen to get granular forecasting result.

2. Performance of the deep learning methods have been evaluated with the help help of Adam optimiser method.

3. The training data has been made in such a way that it has values before and after Covid 19 time frame. After that models have been compared

4. Multivariate and uni-variate analysis of pollutants have been done by using the LSTM models.

5. In last step, one month prediction of the pollutant has been done using the LSTM model which has the best performance. Here, numerous runs of data have been done.

5.1.1 Technical details

In our trial runs several learning rates (hyper parameter tuning) and hidden neuron numbers have been selected. I have used Rectifier liner Unit (ReLU) as an activation function along with used ADAM optimiser which has 20 as batch size and has 200 as epoch units. In the given below table, the selected hidden layers and other values have been described.

Model	Dimensions - Input	Hidden Layers	Output
LSTM	(5,11)	1	10
LSTM - Bidirectional	(5,11)	1	10
LSTM - FNN	55	5	10
LSTM - ED	(5,11)	4	10

Moving forward the time stamp of before Covid is January 2019 to May 2020 whereas post covid time frame chosen is June - Dec 2020. The data chosen in the pre covid test dataset by shuffling technique. In the shuffling technique random data belonging to that time frame window has been chosen. Rather than choosing consecutive weeks, different weeks have been chosen by the shuffling technique.

Best model has been tested in two different methodology. Firstly, by providing them training on the data which is seasonal in nature and in time frame from February 2019 till September 2019. This time frame does not take into consideration of the Covid-19 lockdown period. In our best LSTM model 50 epoch period has been taken. In second testing method, uni-variate PM 2.5 has been tested using the multi step methodology. Here, 1000 epoch period has been selected. Moreover, for further enhancing the tuning capability the testing dataset has been further divided into validation datasets. For maintaining the homogeneity the testing dataset has been used in 1:1 ratio.

In our experiment the total number of experiments conducted are 30, each having different initialisation of weights. In addition to it, in my results standard deviation and mean.

5.2 Visualisation of result

In the given below figures visualisation has been given on the important aspects of the experiment. We have analyse the COVID-19 lockdown effect in New Delhi while comparing air quality before and after the lockdown. In given below figure the PM 2.5 values at four stations - Bawana, Anand Vihar, Vivek Vihar and DTU have been given before and after lockdown. The time frame is from 1 January 2018 till December.

In the above given heat-map of four stations data many conclusions can be drawn. Heat map is a visual representations of the features used in experiments. Here, a value close to zero means independent features. While more positive value means that both the features are highly correlated to each other while negative values means features are highly unrelated to each other. Positive values are indicated by the dark red colour where as negatively correlated values are shown by blue colour.

1. In all the heat maps have high correlation features, like the PM2.5 and PM10 are highly co-related values for all the monitoring stations.

2. It was also observed that the nitrogen oxides have very high correlation between nitrogen oxide and nitrogen dioxide . It can be explained as all three of them are nitrogen components.

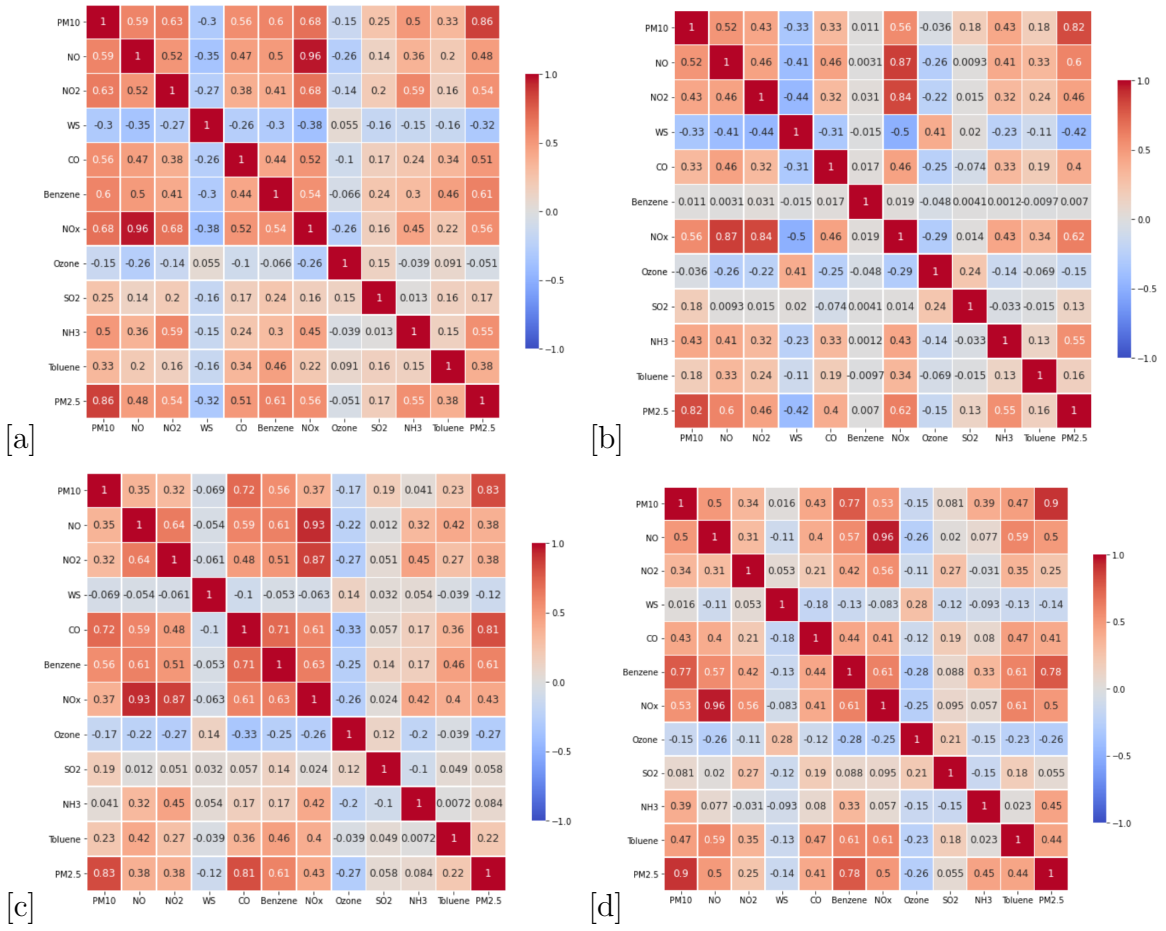


Figure 5: LSTM - Deep Learning Analysis Heat-map (a) Anand Vihar - Deep Learning Analysis Topology(b) Bawana - Deep Learning Analysis Topology (c) DTU - Deep Learning Analysis Topology(d) Vivek Vihar - Deep Learning Analysis Topology

Time Interval (PM2.5)	Bawana (Mean, Interval)	Anand Vihar(Mean, Interval)	DTU(Mean, Interval)	Vivek Vihar(Mean, Interval)
Mar-Jun (2018)	– –	104.59 10.47	189.55 7.39	85.98 8.05
Mar-Jun (2019)	99.38 7.41	97.62 8.48	77.65 8.48	84.81 6.13
Mar-Jun (2020)	64.31 5.44	49.11 6.45	52.78 4.08	57.94 4.74

Table : March-June PM2.5 concentration over 4 stations.

5.3 Forecasting and Modelling - Results

In this paper, four machine learning methods have been used and evaluated for predicting the Delhi's air pollution. First method is LSTM and rest three are the LSTM extensions. These models are LSTM - Bidirectional, Encoder Decoder -LSTM and Forward Neural Network -LSTM. The air pollution dataset have been taken from the Delhi government website. The original data source is Delhi Air Pollution Control board. The Data are

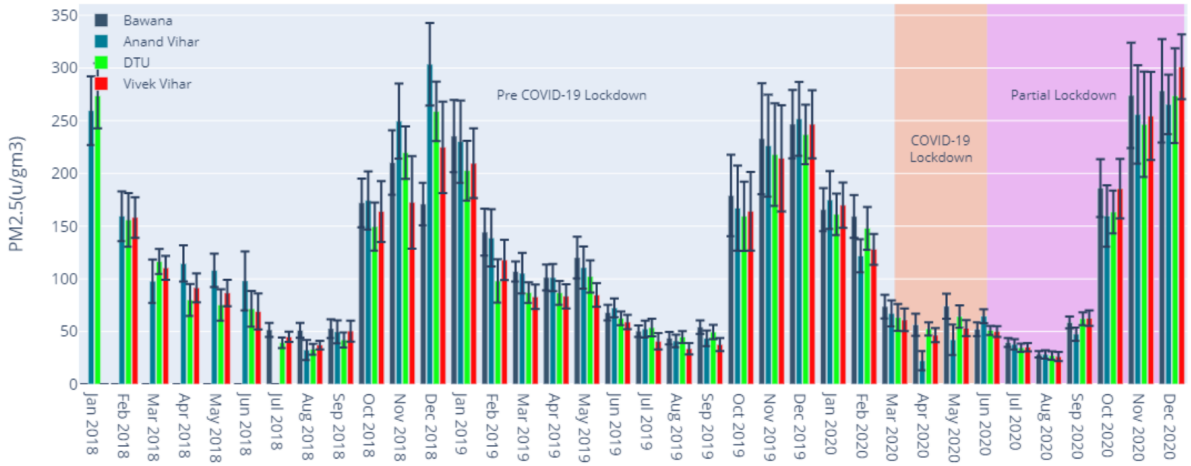


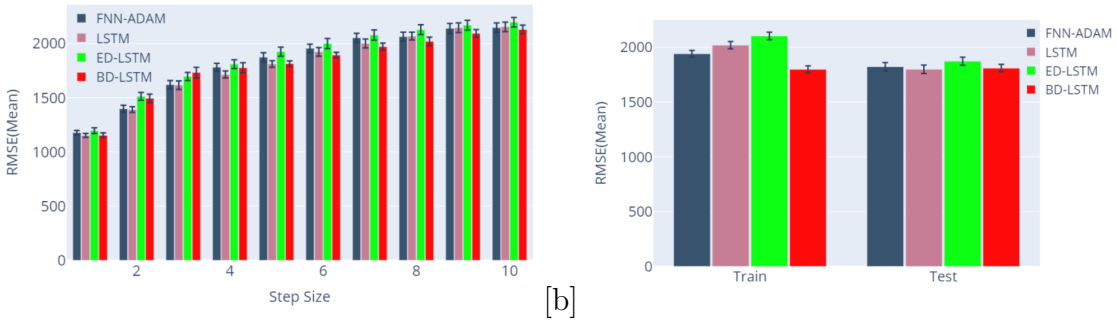
Figure 6: **PM2.5 concentration over 4 stations from time period 1 Jan - Dec 2020**

obtained from four location centres Anand - Vihar, DTU, BAWANA and Vivek Vihar. In this report, the performance of the testing dataset and training dataset mean with 95 percent interval has been described in the figure **??**. 30 experimental results have been run for each model. The year which has been used are from the 2018,2019 and 2020. Few key points inferred from the results are :

1. There has been a significant decrease in PM 2.5 during the 2020 when we have compared with the 2019. This reduction can be attributed to the COVID-19 curfew imposed in the Delhi and hence a decreed in pollutants have also been seen.

2. In our primary set of I have used the ADAM optimiser technique for the LSTM methods. The multivariate analysis have been done on the Anand Vihar Dataset, followed by the Bawana, DTU and Vivek Vihar.

In the given below figure : 10 step ahead method prediction has been shown in the figure a, whereas RMSE value using the entire test and train sets have been used.The dataset here used is from Anand Vihar. Similarly for rest three stations have been implemented.



[a] [b]

Figure 7: LSTM- Evaluation (a) Anand Vihar - Diff. Prediction Horizons (b) Anand Vihar - RMSE Different Models.

In the given below figure 8 reflects one month prediction of the PM2.5 concentration for the four different monitoring stations. The data frame chosen is from 11-Dec till 9-Jan

2021. It was obtained by the BD-LSTM model. Total 720 hours have been used, with 30 experiments run everyday with 8 hour gap. The plot has 95(+-) percentage confidence interval.

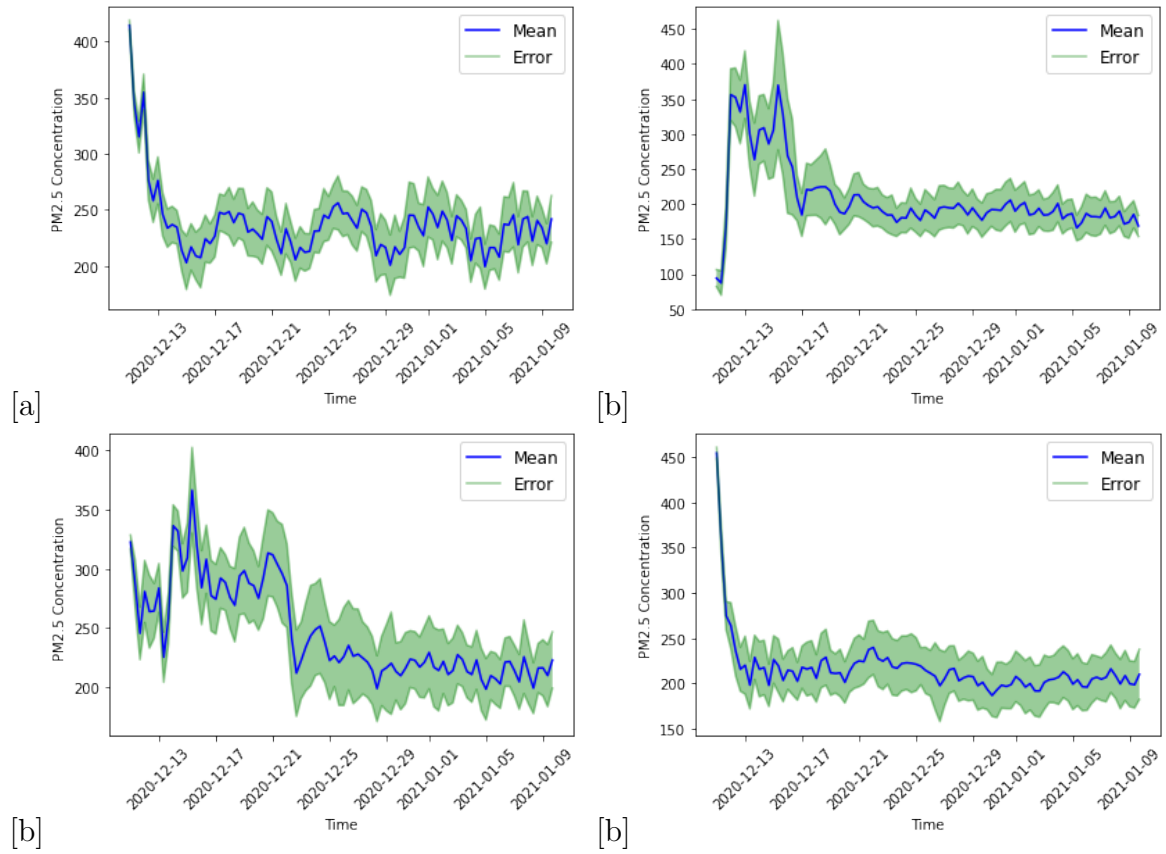


Figure 8: (a) Anand Vihar - PM2.5 Concentration Prediction(b) DTU - PM2.5 Concentration Prediction (c) Bawana - PM2.5 Concentration Prediction(d) Vivek Vihar - PM2.5 Concentration Prediction

In the given below figure 10 Predicted and actual figure has have been displayed for PM 2.5 prediction for the month of September 2020. It has 30 experimental values runs are executed with mean and confidence interval of (+-95) percent.

In this section, we try to give brief findings about the project. Given below are key points.

1. In case of multivariate analysis LSTM - Bidirectional performed well with 1808.92 RMSE value. This is evident from the figure 7. In bi directional architecture with two layers of LSTM performs better for for larger period of time from steps 6 to step 10. Here, we see that rest of the methods does not perform that well.

2. For different monitoring stations, models like BD LSTM was able to predict the PM 2.5 values significantly. This can be inferred that the model was not affected due to the lockdown imposed during Covid-19 lockdown.

6 Conclusion and Future Work

In conclusion, the core objective of this report was to analyse the application of the novel model multi-step ahead model in LSTM models for multi-variate air pollution analysis.For

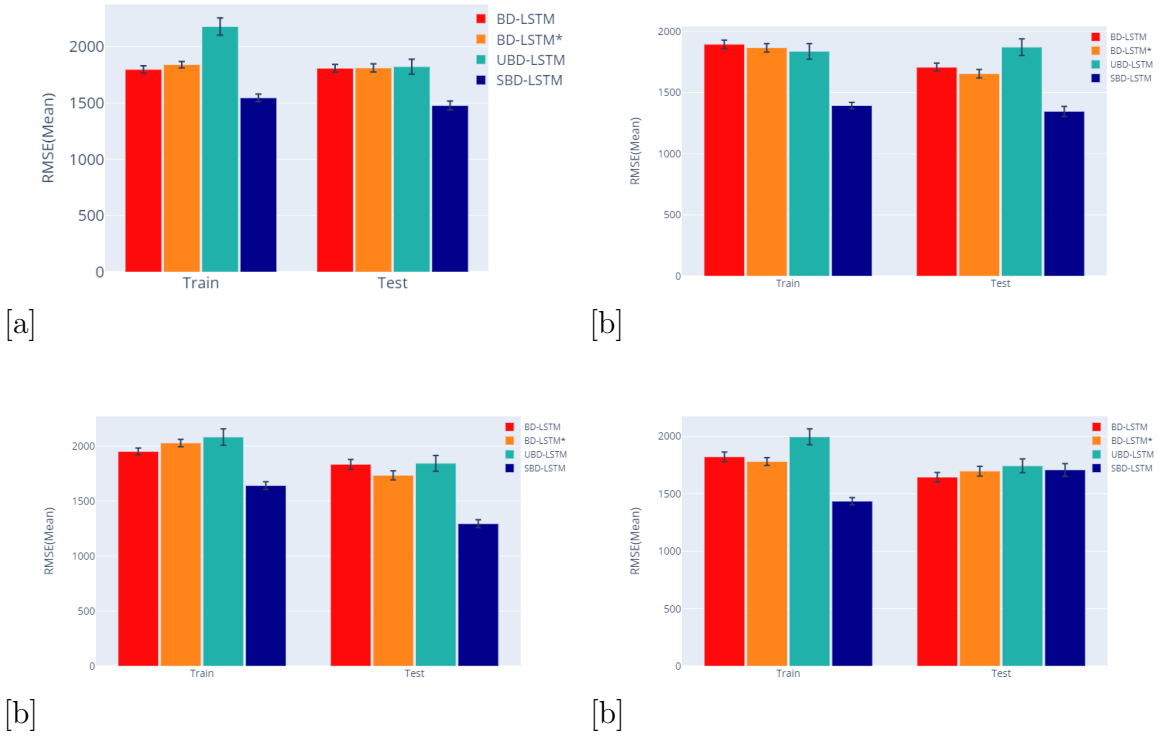


Figure 9: RMSE value of several models (a) Anand Vihar - RMSE value(b) DTU - RMSE value (c) Bawana - RMSE value(d) Vivek Vihar - RMSE value

multivariate air pollution prediction, twelve pollutant data have been used and analysed from the four air quality stations in Delhi. These stations are Vivek Vihar, Anand Vihar, DTU and Bawana. To get better results a number of deep learning methodologies have also been used like LSTM, LSTM-BD, LSTM -ED and LSTM-FNN methods. The criteria to analyse the models are their Root Mean Square Error values and their compilation time. The lesser the RMSE value is the better the performance of the model. It was analysed for the multivariate ait pollutant analysis LSTM - BD method was best. In this paper, a study of the impact of the Covid -19 lockdown on the air quality of New Delhi was also analysed. It was also analysed that the air quality has a seasonal trend with the current year having more air pollutants than the previous company.

For future work, there are several machine learning models and modelling techniques are there to be evaluated in this area of research. For instance, using deep learning methodology like the auto-encoders, ensemble methods can in addition be used. These methods help in the prediction elongation(Increasing the forecasting the date period) of the methodology to improve the accuracy of the models. Moreover, the Delhi government is trying to implement various structural changes like building Nuclear power plants to replace thermal power plants which are key sources of pollutants. In addition, they are planning to roll out cheap electric vehicles. A web-based approach where week;y manage of traffic systems can also be evaluated in the city of Delhi. Therefore, in future, the researchers might have to deal with other sources of pollutants and have to focus on other measures that are challenging to get the results in the field of air pollution detection. Since the dataset used in this research is extremely volatile and traces of volatility can be detected. For addressing such issues GARCH, the method can also be

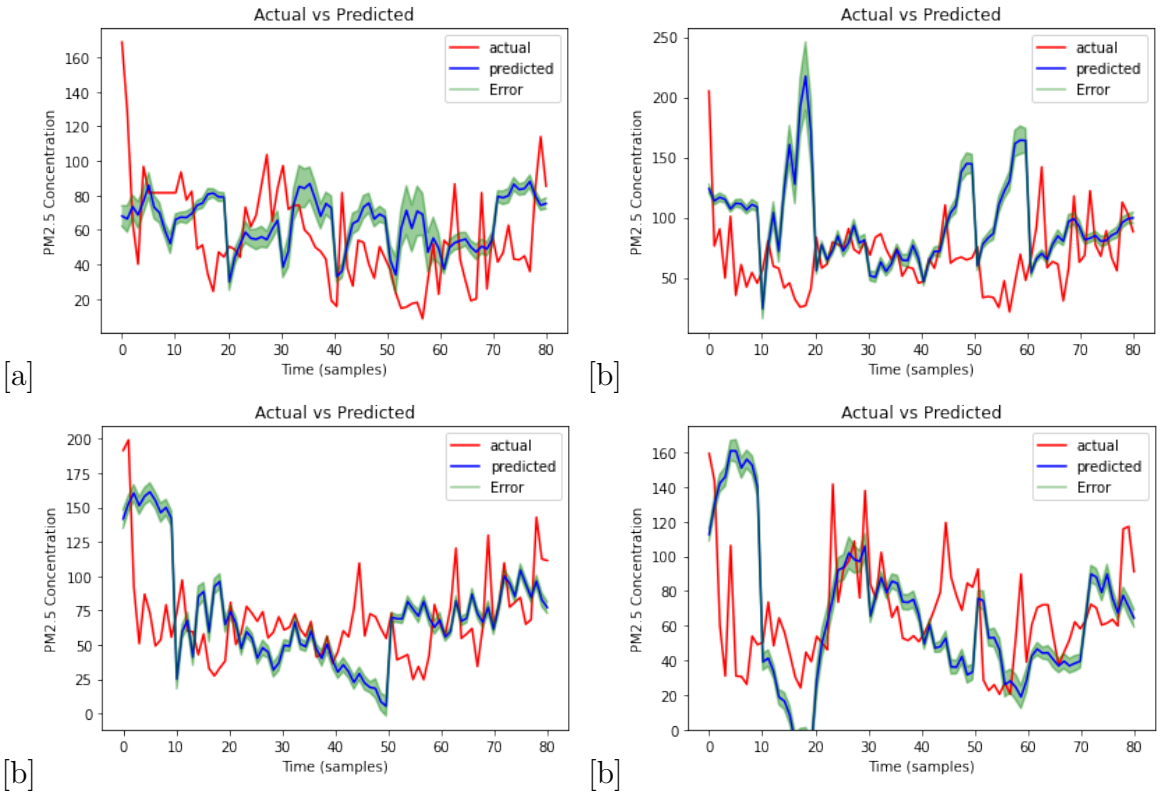


Figure 10: PM2.5 Concentration Prediction actual and prediction for month of Sept 202(a) Anand Vihar (actual and prediction)- PM2.5 Concentration Prediction sept 2020(b) DTU - (actual and prediction)- PM2.5 Concentration Prediction sept 2020(c) Bawana - (actual and prediction)- PM2.5 Concentration Prediction sept 2020(d) Vivek Vihar -(actual and prediction)- PM2.5 Concentration Prediction sept 2020

implemented. For time constraints, the traffic and weather data have been unused in this report. Hence, in future, these data can in addition be employed to get better air value prediction measures. In addition to GARCH, the Bayesian deep learning method can also be implemented. In several papers, it was found that the Bayesian method has a robust prediction of uncertain quantification datasets.

7 Acknowledgement

I am highly humbled and express my deep gratitude to my supervisor Qurrat Ul Ain. She has always advised and guided me from the first week of my program. She assisted me with report writing by guiding me to choose appropriate words and even data models to make meaningful sentences into highly technical sentences. During this short span of time, she guided me immensely in my research domain. In addition, I would like to thank my program coordinator, Dr Anu Sahni, for her throughout guidance in my Data Analytic program. Lastly, I would thank my parents and friends for their constant source of guidance and inspiration to help to achieve my goal.

References

Castelli, M., Clemente, F. M., Popovič, A., Silva, S. and Vanneschi, L. (2020). A machine learning approach to predict air quality in california, *Complexity* **2020**.

Chang, Y.-S., Chiao, H.-T., Abimannan, S., Huang, Y.-P., Tsai, Y.-T. and Lin, K.-M. (2020). An lstm-based aggregated model for air pollution forecasting, *Atmospheric Pollution Research* **11**(8): 1451–1463.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural computation* **9**(8): 1735–1780.

Kalajdjieski, J., Stojkoska, B. R. and Trivodaliev, K. (2020). Iot based framework for air pollution monitoring in smart cities, *2020 28th Telecommunications Forum (TELFOR)*, IEEE, pp. 1–4.

Kerimray, A., Baimatova, N., Ibragimova, O. P., Bukenov, B., Kenessov, B., Plotitsyn, P. and Karaca, F. (2020). Assessing air quality changes in large cities during covid-19 lockdowns: The impacts of traffic-free urban conditions in almaty, kazakhstan, *Science of the Total Environment* **730**: 139179.

Li, L., Li, Q., Huang, L., Wang, Q., Zhu, A., Xu, J., Liu, Z., Li, H., Shi, L., Li, R. et al. (2020). Air quality changes during the covid-19 lockdown over the yangtze river delta region: An insight into the impact of human activity pattern changes on air pollution variation, *Science of the Total Environment* **732**: 139282.

Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. (2016). Deep learning architecture for air quality predictions, *Environmental Science and Pollution Research* **23**(22): 22408–22417.

Saeed, S., Hussain, L., Awan, I. A. and Idris, A. (2017). Comparative analysis of different statistical methods for prediction of pm2. 5 and pm10 concentrations in advance

for several hours, *International Journal of Computer Science and Network Security* **17**(11): 45–52.

Salcedo, R., Ferraz, M. A., Alves, C. and Martins, F. (1999). Time-series analysis of air pollution data, *Atmospheric Environment* **33**(15): 2361–2372.

Schmidhuber, J., Hochreiter, S. et al. (1997). Long short-term memory, *Neural Comput* **9**(8): 1735–1780.

Sindhwani, R. and Goyal, P. (2014). Assessment of traffic-generated gaseous and particulate matter emissions and trends over delhi (2000–2010), *Atmospheric pollution research* **5**(3): 438–446.

Sinnott, R. O. and Guan, Z. (2018). Prediction of air pollution through machine learning approaches on the cloud, *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, IEEE, pp. 51–60.

Stephens, S. L., Westerling, A. L., Hurteau, M. D., Peery, M. Z., Schultz, C. A. and Thompson, S. (2020). Fire and climate change: conserving seasonally dry forests is still possible, *Frontiers in Ecology and the Environment* **18**(6): 354–360.

Sur, S., Ghosal, R. and Mondal, R. (2020). Air pollution hotspot identification and pollution level prediction in the city of delhi, *2020 IEEE 1st International Conference for Convergence in Engineering (ICCE)*, IEEE, pp. 290–294.

Taieb, S. B., Bontempi, G., Atiya, A. F. and Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition, *Expert systems with applications* **39**(8): 7067–7083.

Usmani, R. S. A., Saeed, A., Abdullahi, A. M., Pillai, T. R., Jhanjhi, N. Z. and Hashem, I. A. T. (2020). Air pollution and its health impacts in malaysia: a review, *Air Quality, Atmosphere & Health* **13**(9): 1093–1118.

Vu, T. V., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S. and Harrison, R. M. (2019). Assessing the impact of clean air action on air quality trends in beijing using a machine learning technique, *Atmospheric Chemistry and Physics* **19**(17): 11303–11314.

Ye, Z. (2019). Air pollutants prediction in shenzhen based on arima and prophet method, *E3S Web of Conferences*, Vol. 136, EDP Sciences, p. 05001.

Zheng, K., Zhao, S., Yang, Z., Xiong, X. and Xiang, W. (2016). Design and implementation of lpwa-based air quality monitoring system, *IEEE Access* **4**: 3238–3245.

Zhu, Y., Xie, J., Huang, F. and Cao, L. (2020). Association between short-term exposure to air pollution and covid-19 infection: Evidence from china, *Science of the total environment* **727**: 138704.