Application of LSTM Models in Predicting Particulate Matter (PM2.5) Levels for Urban Area

DOI : 10.36909/jer.11781

Sundarambal Balaraman[*], Partheeban Pachaivannan[**], P. Navin Elamparithi[***] and S. Manimozhi[****]

[*] Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, India, sundarambalb@citchennai.net

Orcid: 0000-0002-7098-8363

[**] Department of Civil Engineering, Chennai Institute of Technology, Chennai, India; Email: parthi011@yahoo.co.in, dean.pd@citchennai.net

Orcid: 0000-0003-4345-0741

[***] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirapalli, India, parithi1999@gmail.com

Orcid: 0000-0002-7542-6626

[****] Department of Electrical and Electronics Engineering, National Institute of Technology, Tiruchirapalli, India, manimozhisekar4@gmail.com

Corresponding author's email id: parthi011@yahoo.co.in

ABSTRACT

Air pollution in India poses a big threat to human lives. In 2017, 77% of population of India was subjected to $PM_{2.5}$ (Particulate Matter) exposure resulting in mortality of 6.7 lakh throughout the country. In this study, Long Short-Term Memory (LSTM) model, a powerful deep learning technique is applied for $PM_{2.5}$ prediction. Three variants of LSTM model, LSTM for regression, LSTM for regression using window and LSTM for regression with time steps are developed to predict $PM_{2.5}$ concentration in India. The metrics used to evaluate the performance of the predictive models are root mean square error (RMSE) and coefficient of determination (R^2). The models are applied to continuous ambient air quality data collected from 14 stations in India, for the period from May 01, 2019 to April 30, 2020 at an interval of every 15 minutes. The optimal results are obtained from the models with the tuned parameters of 64 epochs and batch size of 32. All the three variants of LSTM model performed equally well in predicting $PM_{2.5}$ concentration. The experimental results revealed that the value of R^2 is maintained at 0.9 consistently for all the variants of LSTM model. The low values of RMSE and high values of R^2 proved the reliability of the model. Thus, the proposed model gives awareness about the air pollution level in India and alerts the society to take precautionary steps to save their lives. Further the urban planners can have idea of the pollution levels for their planning and decision making.

Keywords: air pollution; deep learning; long short-term memory (LSTM); particulate matter; $PM_{2.5}$

INTRODUCTION

Today, air pollution in India, especially the northern part, is at unbearable levels. Air quality in many parts of Delhi, the capital city of India, has worsened into the toxic category, with the possibility of causing respiratory ailments. According to the World Health Organization (WHO), one-third of the deaths from stroke, lung cancer and heart disease are due to air pollution. Thus, air quality in India poses a severe health issue. In 2019, 30 cities were declared as the most polluted in the world, in which 21 cities are from India [W1]. As per 2019 Air Quality Index (AQI) country ranking, India ranks 5th place globally in air pollution out of 193 countries [W2].

Rapid industrialization in India not only escalates the country's economy significantly but also pollution in quality of air (Patnaik 2018). The major air pollutants are Particulate Matter (PM_{10} and $PM_{2.5}$), Nitrogen oxide, Sulphur dioxide, Carbon Monoxide and Ozone. Out of which, the most critical pollutant is $PM_{2.5}$. $PM_{2.5}$is the name given to tiny particles in the air whose size is smaller than two and a half microns in width. The diameter of the larger $PM_{2.5}$particles will be around thirty times less than that of human hair. Particles in the $PM_{2.5}$range are likely to penetrate directly through the respiratory system to enter the lungs. Exposure to $PM_{2.5}$ leads to increased incidence of chronic bronchitis and malfunctioning of lungs, which strongly associates with a high mortality rate. $PM_{2.5}$is ranked as the sixth-largest risk factor for global premature mortality (Apte et al. 2015). People with breathing trouble and respiratory issues, infants and elders could be easily prone to $PM_{2.5}$.

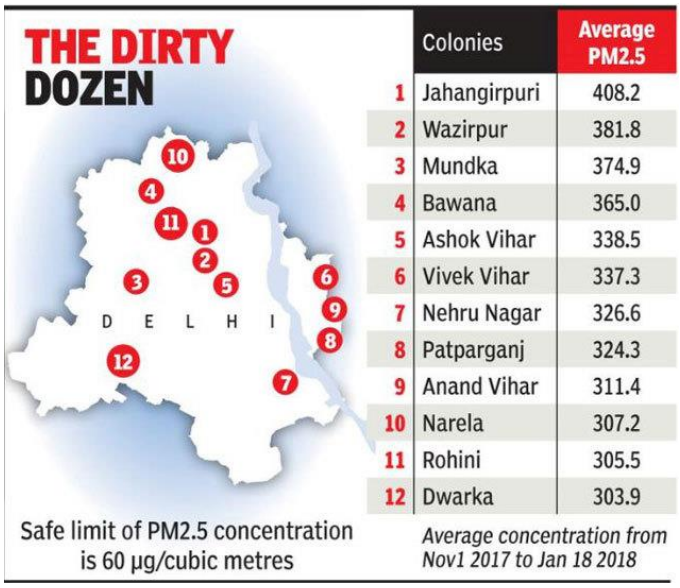
The concentration of $PM_{2.5}$in India in 2019 was five times higher than the WHO recommendation. According to the National Air Quality Index report released by Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change,

Government of India, the threshold values for $PM_{2.5}$as proposed by USEPA (United States Environmental Protection Agency, 2015) are adopted for India as given in Table 1 [W3].

Table 1. Threshold values for PM 2.5

Range of $PM_{2.5}$ (Micrograms per cubic metre)	**AQI category**
0 – 30	Good
31 – 60	Satisfactory
61 – 90	Moderate
91 – 120	Poor
121 – 250	Very Poor
>250	Severe

With the consistent increase in urbanization and industrialization in metropolitan cities of India like Delhi, Mumbai, Kolkata, Bangalore and Chennai, the quality of air also degrades consistently posing a threat to human lives (Garg et al. 1995). In recent years, the rise of $PM_{2.5}$levels above its safe limit 60μg/m³ in the metropolitan city of Delhi is visualized in Figure 1, from an article in Times of India [W4] which reduces the average life span of a human.



	Colonies	Average PM2.5
1	Jahangirpuri	408.2
2	Wazirpur	381.8
3	Mundka	374.9
4	Bawana	365.0
5	Ashok Vihar	338.5
6	Vivek Vihar	337.3
7	Nehru Nagar	326.6
8	Patparganj	324.3
9	Anand Vihar	311.4
10	Narela	307.2
11	Rohini	305.5
12	Dwarka	303.9

Figure 1. The increase in the concentration of $PM_{2.5}$in Delhi

In a nutshell, focusing our efforts on $PM_{2.5}$to identify urban air pollution sources will suffice to address the overall urban air quality scenario in India, without discussing everything under the sun. Consequently, several experiments have been performed by researchers to assess and evaluate air quality based on $PM_{2.5}$.

The Central Pollution Control Board (CPCB) increases the number of stations to be monitored for air quality every year. Thus, a massive amount of data is being generated and maintained at CPCB. Manual exploration of those voluminous data becomes impractical and thus machine learning and data science have been applied for data exploration. The study shows various machine learning and deep learning approaches for air quality prediction, elaborated in section 2.

In this work, Long Short-Term Memory (LSTM) model is developed to predict $PM_{2.5}$ concentration in India. Three variants of LSTM such as LSTM for regression, LSTM for regression with window and LSTM for regression with time steps are implemented and

compared for predicting $PM_{2.5}$ levels in India based on meteorological data collected from five metropolitan cities. The continuous ambient air quality details of one year from May 01, 2019, to April 30, 2020, for every 15 minutes from five major metropolitan cities are collected from the Central Pollution Control Board. The metrics used to assess the performance of the models are root mean square error (RMSE) between the actual observation and the predicted value and coefficient of determination (R^2).

RELATED RESEARCH

This study focused on three variants of LSTM approach to estimate $PM_{2.5}$ concentrations in India based on the detailed survey on existing methodologies for $PM_{2.5}$ prediction as below.

A study by Abdelmoula (Rihab et al. 2021) on Multiobjective optimisation of a series hybrid electric vehicle using DIRECT algorithm. They have studied about the effect of various parameters that affects environment in electric vehicle technology. They optimised the use of fuel consumption and emissions (HC, CO, and NOx) of the vehicle engine. Further they analysed the driving performance requirements of environment pollution. Ahmed et.al, (Alharbi et al. 2020), examined effect of adding hydrogen-rich synthesis gas and ethanol on NOx with gasoline at different fuel mixers. They have modified the engine and plasma converter system for feeding the same type of fuel. They have concluded that the modification of fuel and engine design reduce the pollution level.

Shuyue Zhang et al. (Zhang et al. 2020) used LSTM model to predict $PM_{2.5}$ concentrations in five Chinese cities. They obtained correlation coefficient as 0.86724, 0.80070, 0.78225, 0.72147 and 0.64118 for the cities, Wuhan, Chengdu, Shenzhen, Shanghai and Beijing respectively. Hyun S. Kim et al (Kim et al. 2019)built a deep recurrent neural network based on the LSTM model for daily PM_{10} and $PM_{2.5}$ predictions. They examined the efficiency of the

system by comparing its PM_{10} and $PM_{2.5}$ predictions with the observed and CMAQ (Community Multiscale Air Quality) predicted rates. They found that an LSTM-based PM prediction outperforms CMAQ-based PM predictions.

Xueling Wu et al. (Wu et al. 2020) also used LSTM model to predict the ratio of $PM_{2.5}$ / PM_{10} based on time, space, and random patterns observed in aerosol optical depth, meteorological data, and gaseous pollutant data. They proved LSTM as a dynamic model that understands past data and relates it to current output perfectly than other models. Mei Yang et al. (Yang, Fan, and Zhao 2019) proposed a long short-term memory-convolutional neural network based on dynamic-wind field distance (LSTM-CNN-DWFD) to predict the $PM_{2.5}$ concentration of a specific site for the next 24 hours. The model is proved as the best with low RMSE and high R^2.

Mohit Bansal et al. (Bansal, Aggarwal, and Verma 2019)] established an efficient model for predicting air quality index (AQI) in Delhi, India. They proposed an RNN – LSTM model that predicts pollutant concentration for every hour. They obtained RMSE of 12.79, MAE of 7.84 and R^2 of 0.99. Xiang Li et al. (X. Li et al. 2017) collected $PM_{2.5}$ concentration data from 12 stations in Beijing, China from January 2014 to May 2016 and applied various models such as LSTM model, spatiotemporal deep learning model, time-delay neural network model, autoregressive moving average model and support vector regression model on data. Experimental results revealed that LSTM outperforms the other statistical models.

Klymet Kaya et al. (Kaya and Gündüz Öğüdücü 2020) outlined PM_{10} as target pollutant and proposed a deep flexible sequential model composed of CNN, LSTM and Dropout layer. Their study uses hourly data from Istanbul, Turkey between 2014 and 2018 to predict the air pollution before 4, 12 and 24 hours. Yves Rybarczyk et al. (-Galvan et al. 2016) demonstrated

that the $PM_{2.5}$ predictive performance is improved with a rich set of data. The data from multiple sources such as time, traffic, weather and atmospheric pollutant concentrations gave a clear predictive picture of air quality with just two months data.

Jan Kleine Deters et al. (Kleine Deters et al. 2017) proposed a machine learning approach on six years of meteorological data of two air quality monitoring sites, namely, Cotocollao and Belisario, located in Quito, the capital city of Ecuador and predicted $PM_{2.5}$ concentration. They used Matlab toolbox for implementation. Ping-Wei Soh et al. (Soh, Chang, and Huang 2018) forecasted air quality of Taiwan and Beijing before 48 hours using a combination of various neural networks such as artificial neural network, convolutional neural network and LSTM.

Nandigala Venkat Anurag et al. (Anurag et al. 2019) deployed an XGBoost model that uses meteorological data of Velachery, a fast-growing station in South India and predicted AQI. Experimental results proved XGBoost had shown a decline in error rate comparing with other models such as neural networks, Decision tree and multiple linear regression. Thanongsak Xayasouk et al. (Xayasouk and Lee 2018) proposed a stacked Autoencoders model to predict the quality of air in South Korea. The performance of the model is evaluated using the metric RMSE and the results are predicted for eight areas in South Korea such as Busan, Daegu, Daejeon, Gwangju, Incheon, Sejong, Seoul and Ulsan.

To address the temporal and spatial dependencies of $PM_{2.5}$ concentration simultaneously, Songzhou Li et al.(S. Li et al. 2020) proposed a deep learning model AC-LSTM which includes CNN, LSTM and attention-based network. This hybrid model is applied to air quality data of the city Taiyuan, China and predicted $PM_{2.5}$ concentration over the next 24 hours. Many reported works did not investigate the factors which influence $PM_{2.5}$. The significance of features on $PM_{2.5}$ is studied by Mehdi Zamani Joharestani et al. (Zamani Joharestani et al. 2019)

by implementing random forest, XGBoost and machine learning approach on Tehran, the capital of Iran. After eliminating unnecessary features, XGBoost gave the best results with $R^2 = 0.81$ and RMSE = 13.58 $\mu g/m^3$.

Mangayarkarasi et al. (Mangayarkarasi et al. 2021) proposed forecasting model to predict annual PM2.5 and AQI using Seasonal Autoregressive Integrated Moving Average and Facebook's Prophet Library with the samples collected from 23 Indian cities during the period January 2015 to July 2020. Adil Masood et al. (Masood and Ahmad 2020) built two models using SVM and ANN on the meteorological inputs from the metropolitan city, Delhi of two year periods from 2016-18 and proved that PM2.5 prediction is better with ANN. Since the daily lives at Delhi is affected worse, Chinmay Jena et al. (Jena et al. 2021) addressed the problem by developing a very high resolution operational air quality forecasting system.

Chiou-Jye Huang et al. (C. J. Huang and Kuo 2018) predicted $PM_{2.5}$ concentration for Beijing by employing the combined CNN and LSTM, called APNet. The performance metrics used to evaluate the work are MAE, RMSE, Pearson correlation coefficient and Index of Agreement. Qian Di et al. (Di et al. 2019) used an ensemble model integrating neural network, random forest and gradient boosting to predict $PM_{2.5}$ from 2000 to 2015 for the entire contiguous United States. They proved that a single machine learning algorithm might underperform at a specific year, season and location. In contrast, the ensemble model combining the outputs of all machine learning algorithms would improve the predictive performance. Thanongsak Xayasouk et al. (Xayasouk, Lee, and Lee 2020) applied LSTM and deep autoencoder models on air quality data obtained from 25 stations in Seoul, South Korea. They predicted $PM_{2.5}$ concentration for the next ten days. Their study concludes that the performance of the LSTM model is superior to deep autoencoder model.

The exhaustive literature review reveals that LSTM gives better results for $PM_{2.5}$ prediction in many works. Thus, we proposed to apply three variants of LSTM, namely, LSTM for regression, LSTM for regression using window method and LSTM for regression with time steps for $PM_{2.5}$ prediction in India.

MATERIALS

Study Areas

In this work, the data is collected for five metropolitan cities of India, namely, Delhi, Mumbai, Kolkata, Bangalore and Chennai from CPCB. The various stations at Delhi considered in this study are Anand Vihar, Ashok Vihar, Bawana, Dr.Karni Singh Shooting Range, Dwarka-Sector 8, Jawaharlal Nehru Stadium and Shadipur. Similarly, the stations considered at Chennai are Alandur, Manali and Velachery. The station in Bangalore, BWSSB Kadabesanahalli and the station in Mumbai, Bandra and finally, the two stations from Kolkata, Rabindra Bharati University and Victoria are taken into consideration for analyzing the pollutant levels.

The study areas in this work are as shown in Figure 2. The air quality of nation's capital city Delhi is deteriorated to the worst in global level. Kolkata in the east has declined to Moderate Zone with 152 index value. Bengaluru, Chennai and Mumbai remain in satisfactory zone. According to WHO, out of the 20 most polluted cities in the world, 13 are in India [W6].

Figure 2a. Study area comprising five metropolitan cities of India

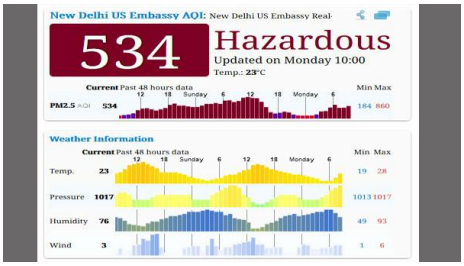


Figure 2b. Hazardous pollution level in Delhi **(Source :** Pollution crisis in India: Before you breathe, check out air quality index of your city today - Oneindia News)

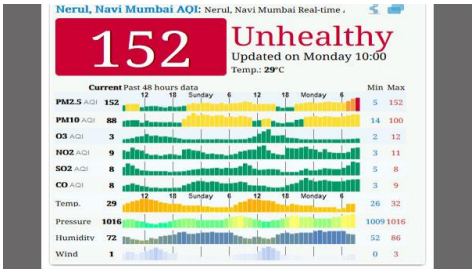


Figure 2c. Unhealthy pollution level in Mumbai **(Source :** Pollution crisis in India: Before you breathe, check out air quality index of your city today - Oneindia News)

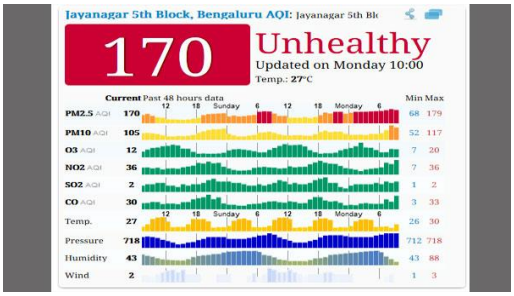


Figure 2d. Unhealthy pollution level in Bengaluru **(Source :** Pollution crisis in India: Before you breathe, check out air quality index of your city today - Oneindia News)

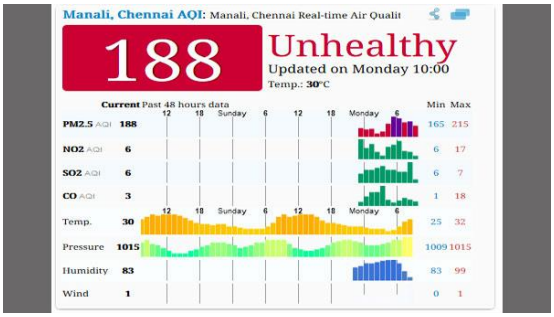


Figure 2e. Unhealthy pollution level in Chennai **(Source :** Pollution crisis in India: Before you breathe, check out air quality index of your city today - Oneindia News)

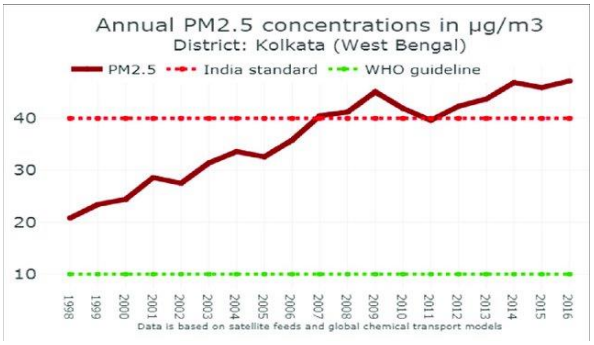


Figure 2f. The results of satellite data derived Surface PM2.5 Concentrations
(Source: West Bengal Pollution Control Board. Accessed from: http://www.wbpcb.gov.in/writereaddata/files/comprehensive%20air%20quality%20action%20plan%20(3).pdf)

Figure 2. Details of PM2.5 concentration in study areas

Dataset

The meteorological features such as Barometric Pressure (BP), Relative Humidity (RH), Wind speed (WS) and wind direction (WD) are collected in addition to $PM_{2.5}$ for every 15

minutes from 14 stations throughout India. The data is collected for one-year duration from the period May 01, 2019, to April 30, 2020, with each station contributing 35039 rows of information and hence totally 490546 rows in the dataset. The dataset is preprocessed to remove the null values and the rows of information with $PM_{2.5}$ values not exceeding 250 µg/m³ are considered resulting in 209216 rows. The details of the dataset are as shown in Figure 3.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209216 entries, 0 to 209215
Data columns (total 5 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   PM2.5   209216 non-null  float64
 1   BP      209216 non-null  float64
 2   RH      209216 non-null  float64
 3   WS      209216 non-null  float64
 4   WD      209216 non-null  float64
dtypes: float64(5)
memory usage: 8.0 MB
```

Figure 3. Details of the Dataset used in the study

The sample data from the dataset is shown in Table 2 for visualization.

Table 2. First five rows of the dataset

	$PM_{2.5}$	**BP**	**RH**	**WS**	**WD**
0	8.00	1032.0	67.07	0.97	243.77
1	60.85	1031.9	73.63	1.14	69.07
2	35.34	1031.8	56.73	0.14	327.98
3	71.13	1031.7	36.79	1.49	173.51
4	35.68	1031.3	0.00	1.11	114.31

The statistical summary of the dataset such as count of nun null observations, mean of the values, standard deviations of the observations, maximum value, minimum value and percentiles (lower 25%, upper 75% and median 50%) are as shown in Table 3.

Table 3. Statistical summary of the variables in the dataset.

	$PM_{2.5}$	BP	RH	WS	WD
count	209216.000000	209216.000000	209216.000000	209216.000000	209216.000000
mean	50.811938	929.804218	62.606710	0.994761	173.055102
std	35.535513	99.727319	22.164266	0.833194	99.593779
min	0.010000	700.010000	0.000000	0.010000	0.000000
25%	23.000000	912.400000	48.100000	0.400000	82.700000
50%	42.370000	980.600000	65.130000	0.830000	170.800000
75%	70.900000	993.800000	80.000000	1.300000	259.160000
max	150.000000	1032.000000	100.000000	10.330000	360.000000

PROPOSED METHODS

LSTM Model

The major drawback with recurrent neural networks is its inability to retain memory. For lengthy sequences, they would have a tough time bringing knowledge from the earlier phases to apply to the later ones. This is because of the vanishing gradient problem faced by recurrent neural networks. The vanishing gradient problem is observed when the neural network learns through backpropagation based on gradient. During this learning phase, the weights of the network are updated in relation to the partial derivative of the error function concerning the current weight. This weight update at some iteration where the gradient may be extremely small is prevented from doing so. This forced the network to stop learning further. Thus, recurrent neural network loses its memory.

The drawback with a recurrent neural network is overcome by LSTM. LSTM models are very powerful, particularly for long short-term memory retention. It is an effective machine learning algorithm that can look at the past of the data series and accurately forecast what the

upcoming elements of the series will be. Applications of LSTM are many and to mention few, robot control (Mayer et al. 2006), time series prediction (Munich 2001), speech recognition (Graves, n.d.), handwriting recognition (Graves, n.d.), airport passenger management(Orsini et al. 2019), sign language translation (J. Huang et al. 2017) etc.

LSTM remembers the sequential data in which data at time t depends on data at time t-1. The three gates, such as input gate, forget gate and output gate are available in each neuron of LSTM. The previous message neurons are connected to current message neurons and thus the LSTM gates are used to solve the long-term dependency on the data. LSTM has memory blocks that are connected with each other through layers. The block contains gates which decide the state of the block. The gates are responsible for remembering or forgetting the information during training. It is accomplished using a sigmoid function. The value of this function squishes between 0 and 1. When the data multiplied by 0, it is forgotten and when multiplied by 1, it is remembered. There are three types of gates in the block as explained.

- Forget gate: This gate is responsible for forgetting or retaining the information. It considers h_{t-1} and x_t and produces the output with a value between 0 and 1. A 1 indicates to maintain the data and 0 means to get rid of the information.

$$f_t=\sigma^{\cdot} W_f[h_{t-1},x_t]+b_f \quad (1)$$

  where f_t represents the forget gate at time t, h_{t-1} is a hidden vector in previous time step, x_t is the input at the current time step, b is the bias, σ is the sigmoid function. W is the weight matrix used to transform the information as vectors.

- Input gate: This gate is responsible for updating the state of the block. The sigmoid function with previous hidden state and the current input transforms the value

between 0 and 1. Thus it conditionally decides which input value can update the state of the block.

$$i_t=\sigma^{\cdot} W_f[h_{t-1},x_t]+b_i \quad (2)$$

where i_t represents the input gate.

- Output gate: This gate is responsible for determining the next hidden state with the current input and previous hidden state.

$$o_t=\sigma^{\cdot} W_o[h_{t-1},x_t]+b_o \quad (3)$$

where o_t represents the output gate.

Thus, each unit in LSTM acts like a state machine maintaining current state, previous state, the current input and next state. The weights for forget, input and output gate, are learned during the training phase. The architecture of LSTM network is shown in Figure 4.

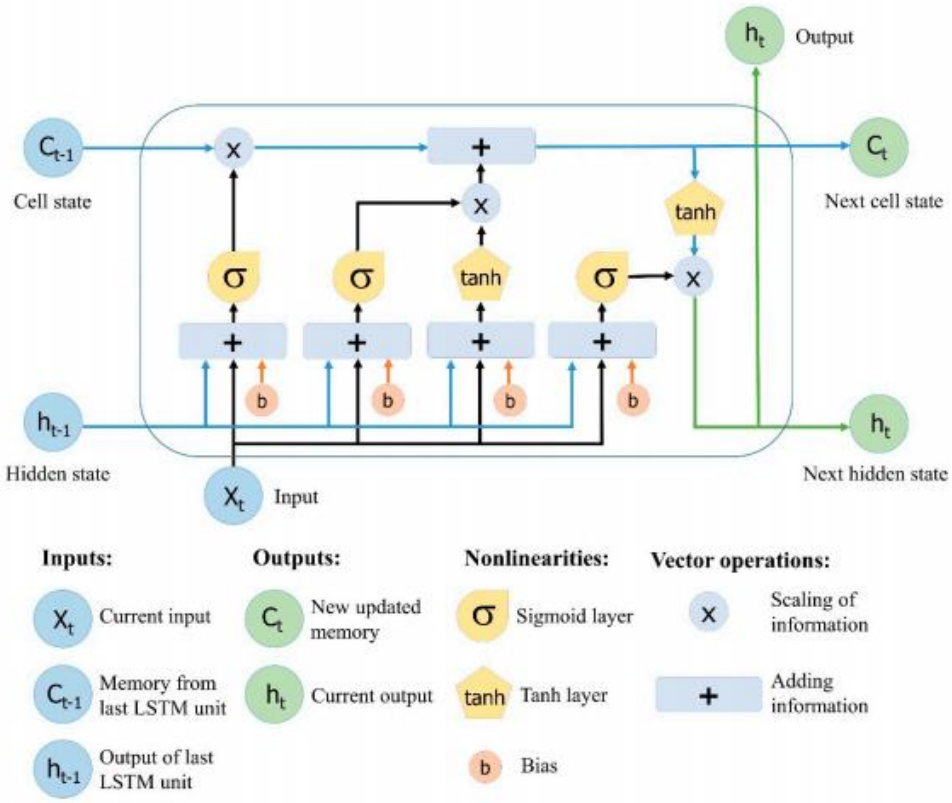


Figure 4. Architecture of LSTM network (reproduced from [W5])

The LSTM algorithm used in our proposed system is described in Table 4.

Table 4. Training the LSTM algorithm.

Step	Description
1	Preprocessing of particulate matter and meteorological data · Inspect, visualize and clean the dataset · Normalize the dataset and set the look back
2	LSTM training • Construct LSTM network with one input, four hidden layers and an output layer with single value prediction · Apply the sigmoid function for LSTM layer · Train the network with epochs = 64 and batch size = 32
3	Obtain prediction results for test dataset using the trained model

Model Performance Evaluation

The metric used to evaluate the performance of the proposed model is the root mean square error (RMSE)between measured $PM_{2.5}$ values and predicted $PM_{2.5}$ values. The formula for calculating RMSE is as follows:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(P_m - P_r)^2}{N}} \quad (4)$$

where P_m and P_r are the measured and predicted PM concentrations, respectively, and N is the number of measured values.

The other metric used to evaluate the performance of the proposed model is the coefficient of determination (R^2), a statistical measure indicating the closeness of data fitting the regression line. The interpretation of RMSE and R^2 is that the high value of R^2 and the low value of RMSE indicate the best fit.

RESULTS AND DISCUSSIONS

LSTM for Regression

In this study, the input data considered are $PM_{2.5}$ concentration data along with meteorological data consisting of barometric pressure, relative humidity, wind speed and direction collected from five metropolitan cities of India. The output is the prediction of $PM_{2.5}$ concentration.

The required Keras deep learning library is imported using Python code and Google Colab environment is used for implementation. As there is no uniformity in the magnitude of the values of available independent features, feature scaling must be employed to standardize the values at a uniform scale. In machine learning algorithms, scaling is done so that a feature with high

magnitude does not over dominate features with low magnitude. The contribution of all the features of the model can then be compared at the same level. In this research work, a min-max scalar is employed which translates each feature to have the values between 0 and 1.

The sequence of data in the training set is very crucial for time series prediction. Thus, the dataset is carefully inspected for the sequential ordering of information. The dataset is then split to consider the first 90% of observations as the training set and the remaining 10% of observations as the test set. Here the parameter lookback is set as 1 which indicates the number of previous steps to consider for predicting the next step.

LSTM network has one input, four hidden LSTM layers and an output layer. The various steps adopted in this work for $PM_{2.5}$ prediction is shown in Figure 5. The performance of the three LSTM models is compared in terms of RMSE and R^2. The different combinations of epoch and batch size are tried, and the optimal results are obtained with 64 epochs and batch size as 32.

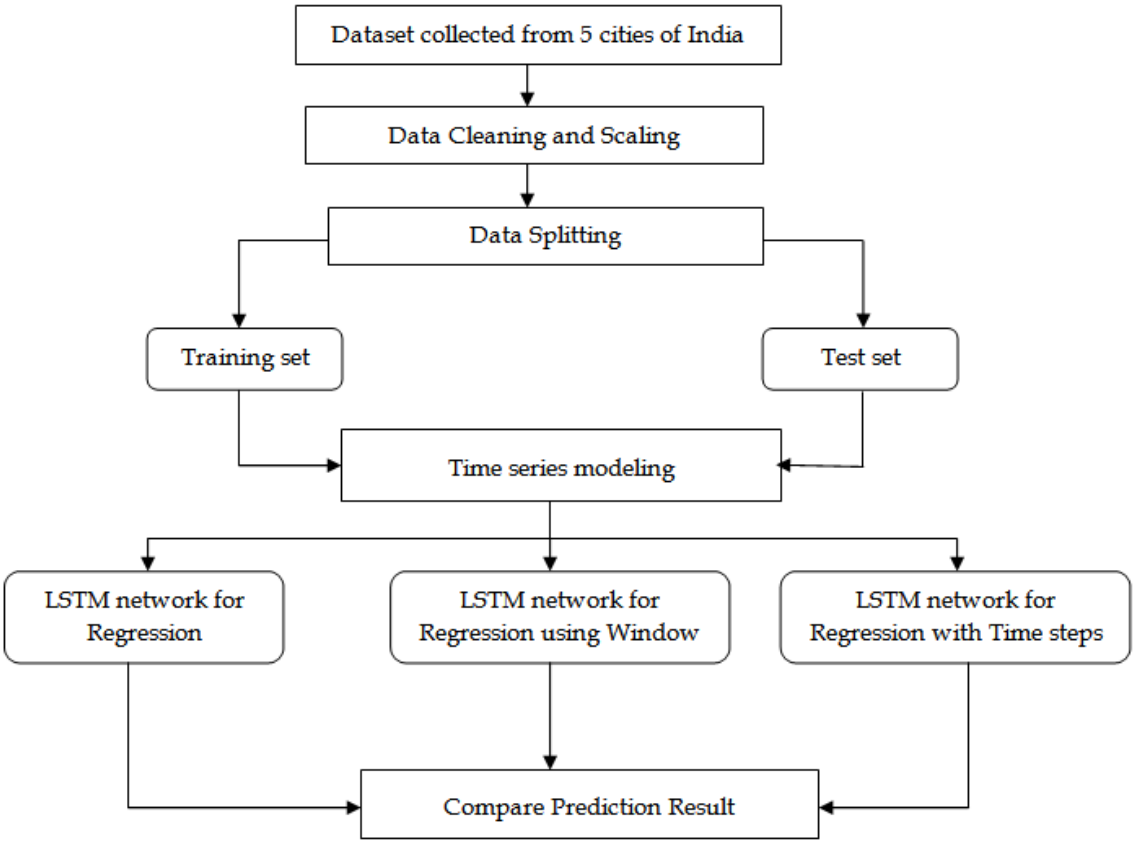


Figure 5. Workflow for predicting $PM_{2.5}$ concentration using three variants of the LSTM model.

The predictions are generated using the LSTM model for both the training and testing data to assess the performance of the model, as shown in Figure 6. The graph is plotted using the tools like numpy, pandas and matplotlib in python.

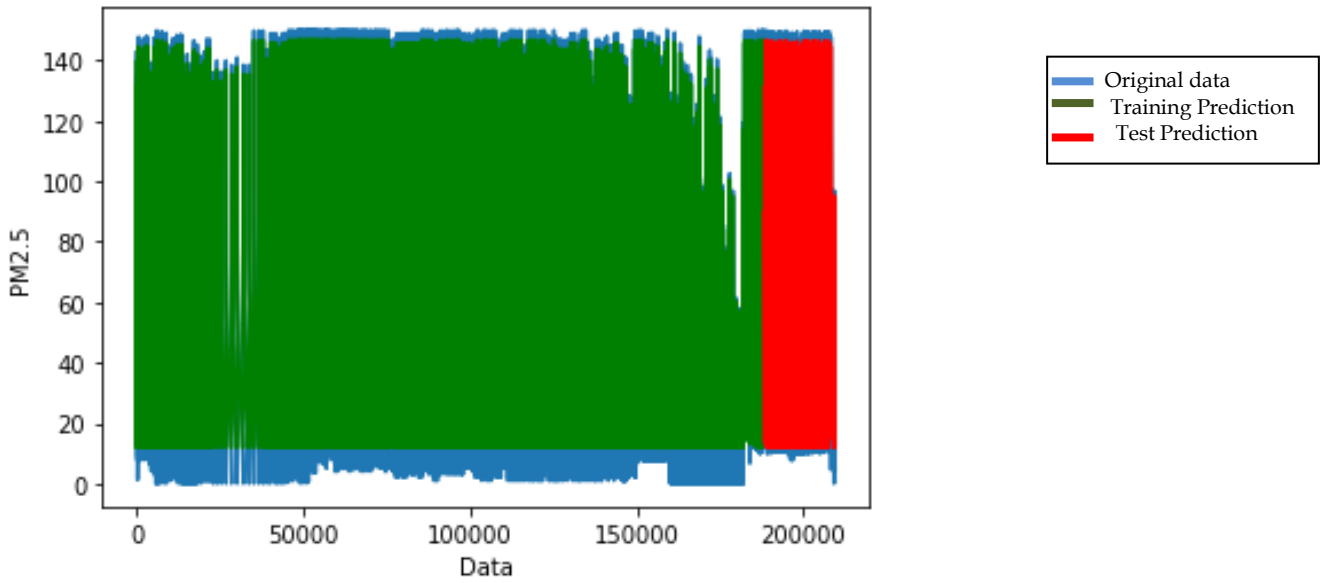


Figure 6. $PM_{2.5}$ prediction using LSTM model for Regression.

In the data plot, the original dataset is shown in blue color, the prediction of the training set in green color and the prediction of the unseen test set in red color. It is seen that the model performs a good job in fitting both training and test data. The RMSE and R^2 values obtained with LSTM model for Regression is shown in Figure 7. The result shows that the model has an average error of 10.758µg/m^3 for the training set and 11.472µg/$m^3$for the test set. The value of R^2 obtained with training and testing set is0.907 and 0.902, respectively, which indicates that the model is the best fit for $PM_{2.5}$ prediction.

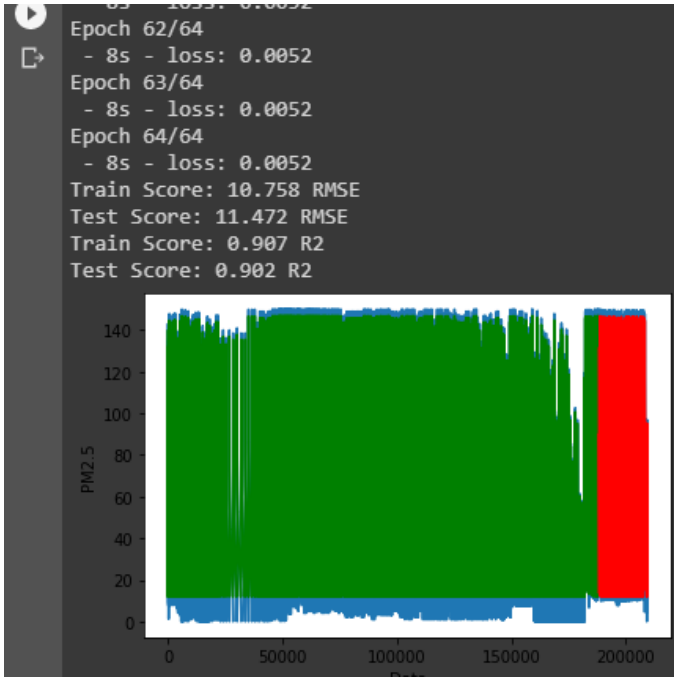


Figure 7. RMSE and R^2 values obtained with LSTM model for Regression.

LSTM for Regression using Window

Another variant of LSTM is built by increasing the size of the Window, in another way, looking back multiple recent time steps to obtain the prediction for the next time step. In this study, the size of the Window, that is, the parameter lookback is set as 3. This implies observations at t-2, t-1 and t are considered to predict the output at t+1.

Again, the predictions are generated using the LSTM model with window size as 3 for both the training and test data to assess the performance of the model as shown in Figure 8.

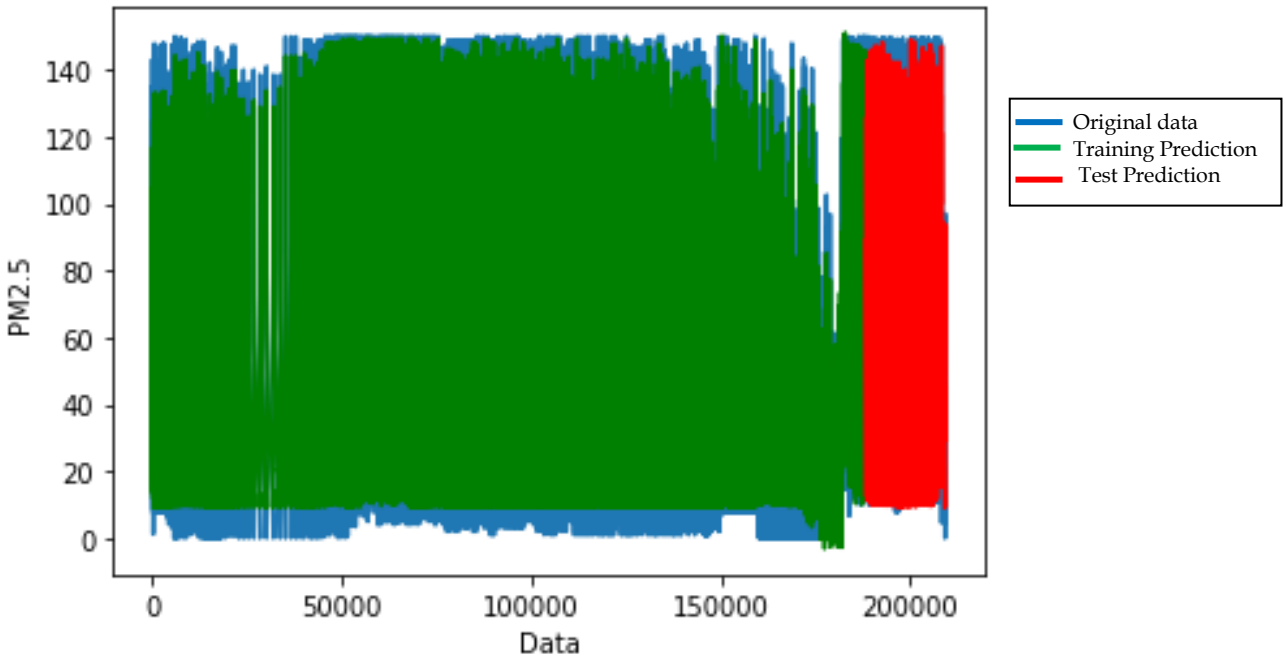


Figure 8. $PM_{2.5}$ prediction using LSTM model for Regression with increased window size.

The RMSE and R^2 values obtained using the LSTM model for Regression with window size as 3 is shown in Figure 9.

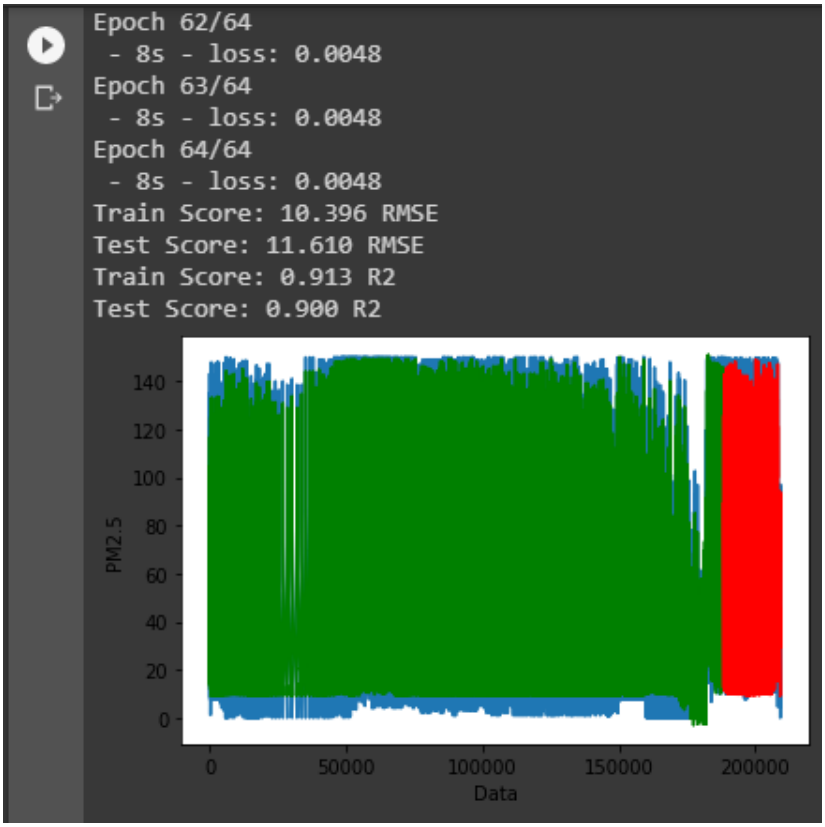


Figure 9. RMSE and R^2 values obtained using the LSTM model for Regression with increased window size.

The result shows that the model has an average error of 10.396µg/m^3 for the training set and 11.610µg/$m^3$for the test set. The score of RMSE for the training set is improved when compared with the previous variant but the error with the test set is increased. Even then, the value of R^2 as 0.913 with the training set and 0.900 with test set indicates that the model can be used for $PM_{2.5}$ prediction.

LSTM for Regression with Time steps

Another variant of LSTM is built using a time step as one of the input features. This is accomplished by reshaping the input to accommodate the time step.

Again, the predictions are generated using the LSTM models with time steps for both the training and test data to assess the performance of the model as shown in Figure 10.

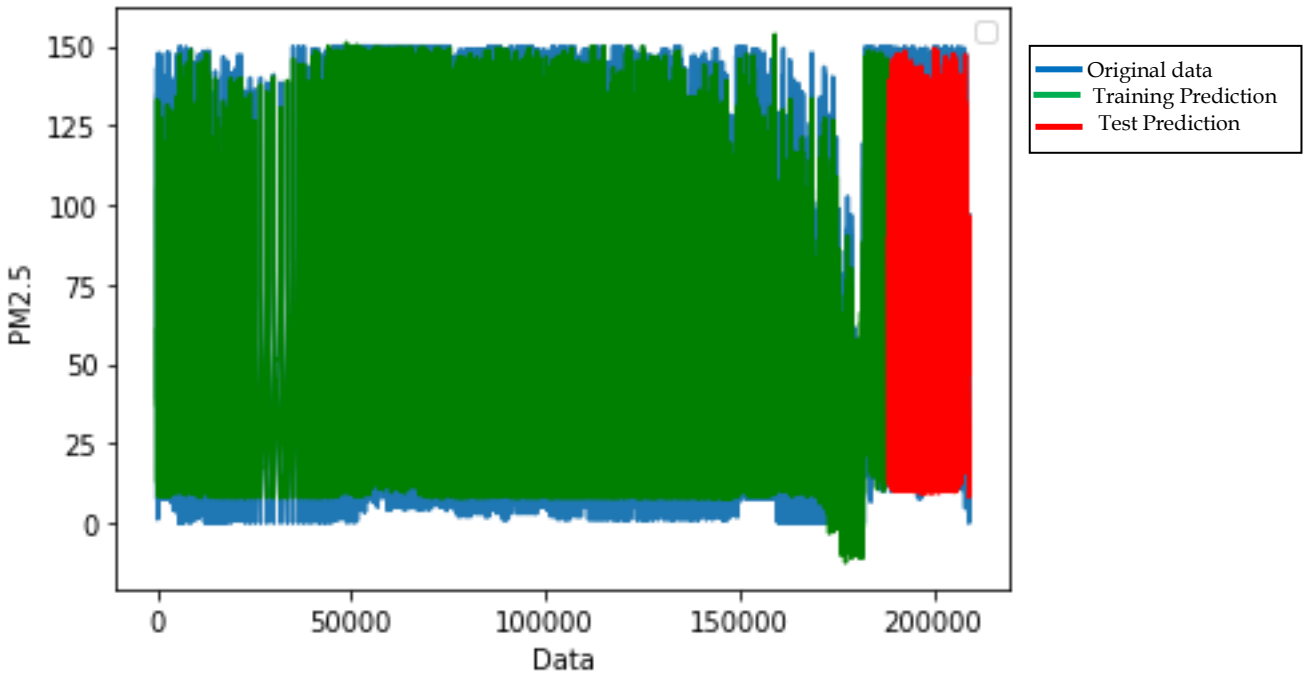


Figure 10. $PM_{2.5}$ prediction using LSTM model for Regressionwith time steps.

The RMSE and R^2 values obtained using theLSTM model for Regression with time steps are shown in Figure 11.

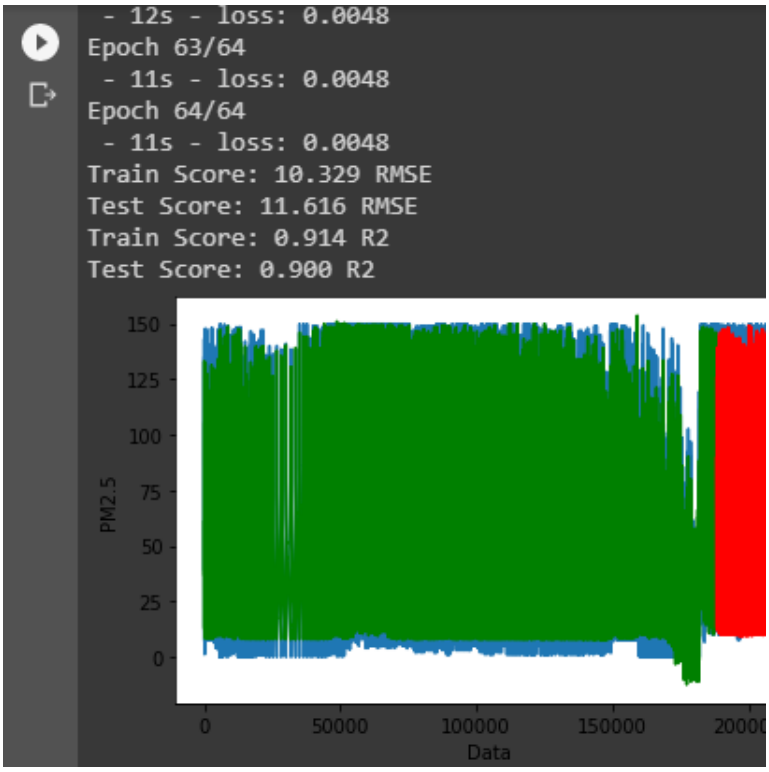


Figure 11. RMSE and R^2 values obtained using the LSTM model for Regression with time steps.

The result shows that the model has an average error of 10.329µg/m^3 for the training set and 11.616µg/$m^3$for the test set. The score of RMSE for the training set is improved when compared with the previous variant and the score with test set remains the same. Again, the value of R^2 obtained with training and test set is0.914 and 0.900, respectively, which indicates that the model is the best fit for $PM_{2.5}$ prediction.

The RMSE and R^2 values are used to compare the results obtained using the LSTM for Regression, LSTM for Regression with Window and LSTM for Regression with time steps, as shown in Table 5.Table 5clearly shows that all the three models performed well for $PM_{2.5}$ predictions and no model is superior to the other. However, the experiments are carried out using different combinations of epoch and batch size and obtained the best results with 64 epochs and batch size as 32. Once the values are tuned, increasing window size or inclusion of

time steps to LSTM model does not show improvements in the performance.

Table 5. Comparison performance for predicting $PM_{2.5}$ concentrations using three variants of LSTM models.

Type of LSTM Models	**RMSE**		**R^2**	
	Training	**Testing**	**Training**	**Testing**
LSTM Network for Regression	10.758	11.472	0.907	0.902
LSTM for Regression Using the Window Method	10.396	11.610	0.913	0.900
LSTM for Regression with Time Steps	10.329	11.615	0.914	0.900

The worst RMSE value which we obtained in this study is 11.472μg/m^3 but comparatively far better than others who carried out researches with air quality data of India. To cite an example, Mohit Bansal et al. (Bansal, Aggarwal, and Verma 2019) used LSTM model to predict $PM_{2.5}$ of Delhi, a capital city of India and they obtained RMSE value of 24.55 μg/m^3. Similarly in another work by Anurag et al. (Anurag et al. 2019), XGBoost is used to predict air quality of Velachery, a fast developing station in Chennai, South India. They obtained RMSE value of 15.97μg/m^3. The lowest RMSE value obtained by Songzhou Li et al. (S. Li et al. 2020) for predicting PM2.5 in Taiyuan city, China is 13.01. The comparison with the existing works on air quality prediction in India proved that our results are superior to others.

CONCLUSIONS

Accurate air pollution forecasting helps to create awareness among people and Government to take the necessary actions to curtail the pollution thereby leading to healthy lifestyle. As an illustration, the decision taken by the Supreme Court of India in 2017 to prohibit the selling of firecrackers seems to be a move in the right direction to minimize serious health consequences

in the country. Thus, it became an inevitable endeavor for researchers to analyze the pollutant levels of a country periodically. In this study, the $PM_{2.5}$ concentration of India is predicted using three variants of LSTM model based on the meteorological data collected from five metropolitan cities. The data is collected for one-year duration from the period May 01, 2019 to April 30, 2020 at every 15 minutes interval. The optimal parameters of the model such as number of epochs as 64 and batch size as 32 are determined for the training set used in this study. The prediction of $PM_{2.5}$ concentration is done for the next 15 minutes. The results comparison shows that there exists no major deviation between the predicted results and actual observations. The coefficient correlation of the three proposed LSTM models are almost same which is around 0.9 which indicates the reliability of the model. This work can still be extended into a more generalized model for $PM_{2.5}$ prediction of India in future by considering the data from a greater number of stations throughout India. Amid rising levels of air pollutants in India, the number of measurement stations is still insufficient to analyze accurate PM levels across the country. Also, this study does not consider the increased emissions from vehicles and industries, which can be addressed in future.

REFERENCES

Galvan, Federico R., Violeta Barranco, Juan C. Galvan, Santiago Batlle, Sebastian FeliuFajardo, and García. 2016. "We Are IntechOpen , the World ' s Leading Publisher of Open Access Books Built by Scientists , for Scientists TOP 1 %." *Intech* i (tourism): 13. https://doi.org/http://dx.doi.org/10.5772/57353.

Alharbi, Dr. Ahmed Awadh, Dr. Feraih Sh. Aenazey, Dr. Saud A. Binjuwair, Dr. Ibrahim A. Alshunaifi, Dr. Abdullah M. Alkhedair, Dr. Abdullah J. Alabduly, Mohammed S. Almurat, and Miqad S. Albishi. 2020. "Reducing NOx Emissions by Adding Hydrogen-

Rich Synthesis Gas Generated by a Plasma-Assisted Fuel Reformer Using Saudi Arabian Market Gasoline and Ethanol for Different Air/Fuel Mixtures." *Journal of Engineering Research* 8 (1): 1–16. https://doi.org/10.36909/jer.v8i1.7400.

Anurag, Nandigala Venkat, Yagnavalk Burra, S. Sharanya, and M. G. Gireeshan. 2019. "Air Quality Index Prediction Using Meteorological Data Using Featured Based Weighted Xgboost." *International Journal of Innovative Technology and Exploring Engineering* 8 (11 Special Issue): 1026–29. https://doi.org/10.35940/ijitee.K1211.09811S19.

Apte, Joshua S., Julian D. Marshall, Aaron J. Cohen, and Michael Brauer. 2015. "Addressing Global Mortality from Ambient PM 2.5." *Environmental Science & Technology* 49 (13): 8057–66. https://doi.org/10.1021/acs.est.5b01236.

Bansal, Mohit, Anirudh Aggarwal, and Tanishq Verma. 2019. "Air Quality Index Prediction of Delhi Using LSTM." *International Journal of Emerging Trends & Technology in Computer Science* 8 (5): 59–68. https://doi.org/10.13140/rg.2.2.26885.70884.

Di, Qian, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M. Benjamin Sabath, et al. 2019. "An Ensemble-Based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution." *Environment International* 130 (January): 104909. https://doi.org/10.1016/j.envint.2019.104909.

Garg, A. N., N. L. Chutke, M. N. Ambulkar, and A. L. Aggarwal. 1995. "An Environmental Pollution Study of Indian Metropolitan Cities and Industrial Surroundings by INAA." *Journal of Radioanalytical and Nuclear Chemistry Articles* 192 (2): 307–20. https://doi.org/10.1007/BF02041735.

Graves, Alex. n.d. "Framewise Phoneme Classification with Bidirectional LSTM Networks."

"Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks," 1–8.

Huang, Chiou Jye, and Ping Huan Kuo. 2018. "A Deep Cnn-Lstm Model for Particulate Matter (Pm2.5) Forecasting in Smart Cities." *Sensors (Switzerland)* 18 (7). https://doi.org/10.3390/s18072220.

Huang, Jie, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2017. "Video-Based Sign Language Recognition without Temporal Segmentation."

J**ena, Chinmay, Sachin D Ghude, Rajesh Kumar, Sreyashi Debnath, Gaurav Govardhan, Vijay K Soni, Santosh H Kulkarni, G Beig, Ravi S Nanjundiah, and M. Rajeevan**. 2021. "Performance of High Resolution (400 m) PM2.5 Forecast over Delhi." *Scientific Reports* 11 (1): 4104. https://doi.org/10.1038/s41598-021-83467-8.

Kaya, Kıymet, and Şule Gündüz Öğüdücü. 2020. "Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting." *Scientific Reports* 10 (1): 3346. https://doi.org/10.1038/s41598-020-60102-6.

Kim, Hyun S., Inyoung Park, Chul H. Song, Kyunghwa Lee, Jae W. Yun, Hong K. Kim, Moongu Jeon, and Jiwon Lee. 2019. "Development of Daily PM_{10} and $PM_{2.5}$ Prediction System Using a Deep Long Short-Term Memory Neural Network Model." *Atmospheric Chemistry and Physics Discussions*, 1–27. https://doi.org/10.5194/acp-2019-268.

Kleine Deters, Jan, Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk. 2017. "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters." *Journal of Electrical and Computer Engineering* 2017. https://doi.org/10.1155/2017/5106045.

Li, Songzhou, Gang Xie, Jinchang Ren, Lei Guo, Yunyun Yang, and Xinying Xu. 2020. "Urban PM2.5 Concentration Prediction via Attention-Based CNN–LSTM." *Applied Sciences* 10 (6): 1953. https://doi.org/10.3390/app10061953.

Li, Xiang, Ling Peng, Xiaojing Yao, Shaolong Cui, Yuan Hu, Chengzeng You, and Tianhe Chi. 2017. "Long Short-Term Memory Neural Network for Air Pollutant Concentration Predictions: Method Development and Evaluation." *Environmental Pollution* 231 (December): 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114.

Mangayarkarasi, R., C Vanmathi, Mohammad Zubair Khan, Abdulfattah Noorwali, Rachit Jain, and Priyansh Agarwal. 2021. "COVID19: Forecasting Air Quality Index and Particulate Matter (PM2.5)." *Computers, Materials & Continua* 67 (3): 3363–80. https://doi.org/10.32604/cmc.2021.014991.

Masood, Adil, and Kafeel Ahmad. 2020. "A Model for Particulate Matter (PM2.5) Prediction for Delhi Based on Machine Learning Approaches." *Procedia Computer Science* 167 (2019): 2101–10. https://doi.org/10.1016/j.procs.2020.03.258.

Mayer, Hermann, Faustino Gomez, Daan Wierstra, Istvan Nagy, and Alois Knoll. 2006. "A System for Robotic Heart Surgery That Learns to Tie Knots Using Recurrent Neural Networks," 543–48.

Munich, T U. 2001. "Evolino : Hybrid Neuroevolution / Optimal Linear Search for Sequence Learning Recurrent Neural Network."

Orsini, Federico, Riccardo Rossi, Massimiliano Gastaldi, and Luca Mantecchini. 2019. "Neural Networks Trained with WiFi Traces to Predict Airport Passenger Behavior." *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 1–7.

Patnaik, Rasmi. 2018. "Impact of Industrialization on Environment and Sustainable Solutions – Reflections from a South Indian Region." *IOP Conference Series: Earth and Environmental Science* 120 (1): 012016. https://doi.org/10.1088/1755-1315/120/1/012016.

Rihab, Abdelmoula, Ben Hadj Naourez, Chaieb Mohamed, and Neji Rafik. 2021. "Multiobjective Optimisation of a Series Hybrid Electric Vehicle Using DIRECT Algorithm." *Journal of Engineering Research* 9 (1): 151–67. https://doi.org/10.36909/jer.v9i1.8366.

Soh, Ping Wei, Jia Wei Chang, and Jen Wei Huang. 2018. "Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations." *IEEE Access* 6: 38186–99. https://doi.org/10.1109/ACCESS.2018.2849820.

Wu, Xueling, Ying Wang, Siyuan He, and Zhongfang Wu. 2020. "PM2.5/PM10 Ratio Prediction Based on a Long Short-Term Memory Neural Network in Wuhan, China." *Geoscientific Model Development* 13 (3): 1499–1511. https://doi.org/10.5194/gmd-13-1499-2020.

Xayasouk, Thanongsak, Hwa Min Lee, and Giyeol Lee. 2020. "Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models." *Sustainability (Switzerland)* 12 (6). https://doi.org/10.3390/su12062570.

Xayasouk, Thanongsak, and Hwamin Lee. 2018. "Air Pollution Prediction System Using Deep Learning." *WIT Transactions on Ecology and the Environment* 230: 71–79. https://doi.org/10.2495/AIR180071.

Yang, Mei, Hong Fan, and Kang Zhao. 2019. "PM2.5 Prediction with a Novel Multi-Step-Ahead Forecasting Model Based on Dynamic Wind Field Distance." *International Journal of Environmental Research and Public Health* 16 (22): 4482. https://doi.org/10.3390/ijerph16224482.

Zamani Joharestani, Mehdi, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani. 2019. "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep

Learning Using Multisource Remote Sensing Data." *Atmosphere* 10 (7): 373. https://doi.org/10.3390/atmos10070373.

Zhang, Shuyue, Minfeng Lin, Xiuguo Zou, Steven Su, Wentian Zhang, Xuhui Zhang, and Zijie Guo. 2020. "LSTM-Based Air Quality Predicted Model for Large Cities in China." *Nature Environment and Pollution Technology* 19 (1): 229–36.

Website references

W1. Regan, Helen. "21 of the world's 30 cities with the worst air pollution are in India". *CNN*. Retrieved 2020-02-26.

W2. https://www.iqair.com/india

W3. https://app.cpcbccr.com/ccr_docs/FINAL-REPORT_AQI_.pdf

W4. https://timesofindia.indiatimes.com/city/delhi/twelve-areas-in-delhi-where-you-can-never-breathe-clean-air/articleshow/62677188.cms

W5. Yan, S. Understanding LSTM and Its Diagrams. Available online: https://medium.com/mlreview/ understanding-lstm-and-its-diagrams-37e2f46f1714

W6. https://www.mapsofindia.com/liveblog/india/barometer-with-air-quality-index/

Declarations

Funding : Not applicable

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.