Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 21 | No. 2 | pp. 691-696 | 2022

Original Research Paper

https://doi.org/10.46488/NEPT.2022.v21i02.029

Open Access Journal

Prediction of $PM_{2.5}$ Over Hyderabad Using Deep Learning Technique

P. Vinay Kumar*, M. C. Ajay Kumar*, B. Anil Kumar and P. Venkateswara Rao***†**
*Department of Physics, Aurora's Technological and Research Institute, Hyderabad, India
**Department of Mechanical Engineering, Aurora's Technological and Research Institute, Hyderabad, India
***Department of Physics, Vasavi College of Engineering, Hyderabad, India
†Corresponding author: P. Venkateswara Rao; kishansetty@gmail.com

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 24-07-2021
Revised: 04-09-2021
Accepted: 13-09-2021

Key Words:
$PM_{2.5}$
Air pollution
Deep learning
Forecasting

ABSTRACT

Urbanization and Industrialization during the last few decades have increased air pollution causing harm to human health. Air pollution in metro cities turns out to be a serious environmental problem, especially in developing countries like India. The major environmental challenge is, to predict accurate air quality from pollutants. Envisaging air quality from pollutants like $PM_{2.5,}$ using the latest deep learning technique (LSTM timer series) has turned out to be a significant research area. The primary goal of this research paper is to forecast near-time pollution using the LSTM time series multivariate regression technique. The air quality data from Central Pollution Control Board over Hyderabad station has been used for the present study. All the processing is done in real-time and the system is found to be functionally very stable and works under all conditions. The Root Mean Square Error (RMSE) and R^2 have been used as evaluation criteria for this regression technique. Further, the time series regression has been used to find the best fit model in terms of processing time to get the lowest error rate. The statistical model based on machine learning established a relevant prediction of $PM_{2.5}$ concentrations from meteorological data.

INTRODUCTION

Air quality in India has declined significantly over the past few decades and today, air pollution has been implicated as a major risk factor for illnesses and premature mortality in India. As per the study based on 2019 data, 21 of the 30 most polluted cities in the world are in India (World Air Quality Report 2019)**.** Rapid growth in industrialization, urbanization, and population critically affect the atmospheric environment. Innumerable studies have reported that industrial and vehicular emissions are the main contributors to atmospheric pollution (Ajay Kumar et al. 2020, Van 2017, Ravindra et al. 2016, Zhang et al. 2019, Zhao et al. 2019, Singh et al. 2020a). It is a well-known fact that air pollution has adverse effects on humans as well as the environment. The most concerning air pollutant is particulate matter $PM_{2.5}$ (particles with an aerodynamic diameter less than 2.5 microns). These tiny particles stay for a longer time in the atmosphere and they can easily bypass the filters of the human nose and throat, which cause serious impact on human health that includes respiratory problems, asthma, chronic bronchitis, and lung diseases.

The sources of $PM_{2.5}$ can be classified as outdoor and indoor sources. The outdoor sources include vehicular exhausts, construction equipment, burning of fuels such as wood, heating oil or coal, and the reaction of gases or droplets in the atmosphere originating from power plants. Indoor sources include tobacco smoke, cooking, burning candles or oil lamps, operating fireplaces and fuel-burning space heaters (e.g., kerosene heaters). Prediction of these pollutants is very important to take precautionary measures. The concentration of $PM_{2.5}$ depends on many factors and it does not relate to the previous repetitive patterns. Several studies have been conducted to forecast air quality using various statistical and machine learning models. Statistical models can predict atmospheric pollution, but their accuracy is always biased as it depends on an updated source list. Hence, a machine learning technique is more reliable to derive a real-time assessment of air quality based on sample data.

Artificial Neural Networks, based on Machine Learning, is the most commonly used technique to predict air quality. In recent years, the large availability of data enables researchers to implement different prediction models using deep learning techniques. Deep learning is the subset or a special kind of Machine learning, that makes the simultaneous process of huge data set in sequential layers, thereby providing more reliable results. The deep learning technique consists of an artificial intelligence system that can obtain data unsupervised, in an unstructured or unlabeled learning approach. Usually, deep learning has been widely used in academic and practical applications like translation, speech recognition, language, and image processing. The advanced features of deep learning make it a suitable method to predict air pollutants (Yasin et al. 2018). Deep neural networks can be used to obtain short-term air quality forecasts.

The present study aims to build a model for hourly $PM_{2.5}$ forecasting over Bollaram, Hyderabad, Telangana State, India, using one of the most powerful existing deep learning methods, namely Long Short-Term Memory (LSTM).

RELATED RESEARCH WORK

Recent studies have used many machine learning techniques to overcome the problem related to the prediction of air pollutants like $PM_{2.5}$. Hoek et al. (2008) reported that annual mean $PM_{2.5}$ concentrations can be predicted using land-use regression methods. Using a neural network algorithm, the prediction of O_3 and PM_{10} concentrations was made by Corani (2005). He addressed the problem of the prediction of these two pollutants using feed-forward neural networks (FNs), pruned neural networks (PNNs), and lazy learning (LL). All three approaches are tested in the prediction of pollutants. LSTM neural network has been extensively used to process time-series data as it is capable of stimulating simultaneously (Hochreiter & Schmidhuder 1997). Li (2017) developed a predictive method by using concentrations from different sites as an input to the LSTM layer. Experiments were performed using the spatiotemporal deep learning (STDL) model, the time-delay neural network (TDNN) model, the autoregressive moving average (ARMA) model, the support vector regression (SVR) model, and the traditional LSTM NN model, and a comparison of the results demonstrated that the LSTME model is superior to the other statistics-based models. Further, numerous forecasting models have been developed by using various models based on LSTM. One such study is a prediction of hourly fine PM concentration at 25 target locations in Seoul (Kim et al. 2019). Results showed that the mean squared errors between predicted concentrations of $PM_{2.5}$ and PM_{10} for all 25 locations are well below 45 x 10^{-5}. Qin (2019) proposed a novel model based on a convolutional neural network (CNN) and LSTM to predict urban PM concentration. The model predicts future particulate matter concentration as a time series. Results are compared with results of numerical models and it shows an improved prediction performance. Xiao et al. (2020) recently proposed an improved deep learning model to predict daily PM concentration by collecting three years of data from 2015 to 2017 and evaluating the performance of different models. In this study, the author compared three models namely GWR, LSTME, and STSVR with the proposed WLSTME (weighted LSTME) model. Results showed that the proposed WLSTME model has the lowest RMSE (40.67) and MAE (26.10) with the highest p (0.59).

DATA DETAILS

To develop a predictive model using deep learning or machine learning, it is required to have adequate previous data. In the present study, we used the data set extracted from the Central Pollution Control Board (CPCB), India. Hyderabad, popularly known as Pearl City, is the capital of the Indian state of Telangana. It is one of the most polluted and populated cities in India. IDA Bollaram, located in Hyderabad is categorized by CPCB as an industrial hub of Hyderabad. It is home to 100 polluting units, about half of which are pharma and drug companies. Due to a high level of pollution, IDA Bollaram was chosen as a sample area in our study.

The hourly data, separated by different meteorological variables and $PM_{2.5,}$ was downloaded from the CPCB website for this location from 1st March 2017 to 31st August 2020. Table 1 reports the summary of the site and observed variables.

MATERIALS AND METHODS

To develop the predictive model using deep learning technique, the following steps were implemented:

- Identification of the relevant data set
- Pre-processing of data for analysis
- Modeling
- Training of the model in the test data sets and running the model to generate test score
- Evaluation of the performance of the proposed model.

Identification of the Relevant Data Set and Basic Statistical Analysis

The variations in the concentration of $PM_{2.5}$ greatly depend on meteorological and weather conditions. Considering that,

Table 1: Summary of site location and variables used in the present study.

Name of the site	Location	Type	Variables
IDA Bollaram, Hyderabad, Telangana State, India	17.38 °N, 78.48 °E	Particulates	$PM_{2.5}$
		Meteorological parameters	Temperature (Temp), Solar Radiation (SR), Relative Humidity (RH), Wind Speed (WS), and Wind Direction (WD)

we collected hourly meteorological data which included temperature (^oC), relative humidity (%), solar radiation ($W.m^{-2}$), wind speed ($m.s^{-1}$), rainfall (mm), and wind direction (degree) in addition to the $PM_{2.5}$ from CPCB website for the same location and period. The data matrix was prepared with Year, Month, Day, Hour, $PM_{2.5,}$ and meteorological parameters.

To find the interdependency of $PM_{2.5}$ with other meteorological parameters, we have computed correlation and depicted the same in Fig. 1. We observed that $PM_{2.5}$ has a negative correlation with all meteorological parameters, especially high values with temperature and relative humidity, which depicts that lowering the temperature or relative humidity increases the amount of $PM_{2.5}$ concentrations. Based on these results, it was decided to retain all pollutants in the data set except rainfall due to insufficient data.

Data Pre-Processing

Pre-processing helps to transform the raw data into a smooth information set. The data we downloaded may contain missing and inconsistent values. To get better prediction results, the raw data should be cleaned; missing values or null data must be filled with appropriate mean or median values. Superfluous, insufficient, or extreme data must be removed to avoid biasing the results so that the output is more accurate.

In this section, the method adopted for pre-processing the data to forecast the concentration of $PM_{2.5}$ was presented. It included refining the raw data, removing the outliers, normalization, and standardization (data scaling) of the data for visualization and understanding the data.

Data Refining

In this study, we processed three years five months of data with 30510 total number of records, mentioned in data details. In our data set, we have found missing values that represent 10% of the total data for the study site. We have used means to handle the missing data.

Outlier Detection Process

Generally, outliers present in the data can skew and misinform the procedure used in the algorithms which leads to inaccurate results giving poor output. Table 2 shows the basic statistical analysis of $PM_{2.5}$ concentrations with mean, standard deviation, maxima, minima, and quartiles before the removal of outliers.

To summarize the distribution of variables, we have represented the data in the form of a box plot as illustrated in Fig. 2. This illustration helps us to find out the skewness and outliers. Mean, first, third quartiles, maxima, and minima values of $PM_{2.5}$ before and after removal of outliers are reported in Fig. 2. The outliers with z values greater than 3 are discarded from the data set.

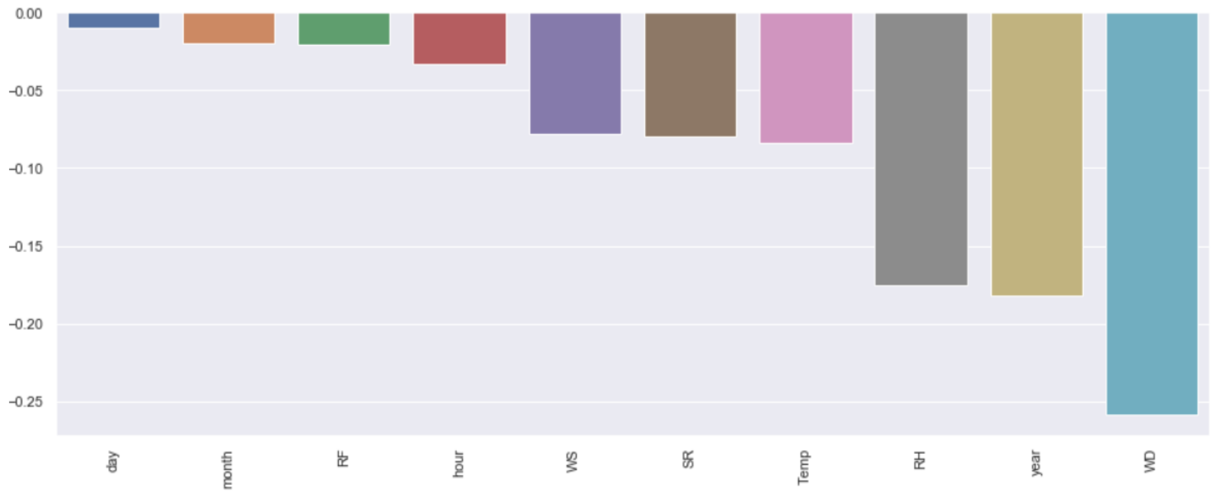


Fig. 1: Correlation between $PM_{2.5}$ with various parameters.

Table 2: Basic statistical analysis of data set.

Parameter	Minimum	Maximum	STD	25%	50%	75%	Mean
$PM_{2.5}$ ($\mu g.m^{-3}$)	1.0	544	25.827	24.250	37.750	58.000	42.489

Data Scaling

Data scaling is the technique used to normalize or standardize the independent variables present in the data set in a fixed range. In deep learning neural network models, data scaling is always recommended as one of the pre-processing steps. In the present study, we used data scaling that normalizes the real values of the input variables within a fixed range [0, 1]. This makes all features equally weighted between 0 and 1.

Modelling

Multivariate linear regression emphasizes the relationship between one dependent variable and multiple independent variables. The long Short-Term Memory model, part of standard Recurrent Neural Networks (RNN), is a trustable technique to predict air quality using pollution and meteorological information of time series data. It basically consists of different gates, including input, forgets, and output gates, which provide the information to flow between, in, and out of a cell. The block system of LSTM is depicted in Fig. 3.

Training of the Model

We have used around 21357 training samples and 9153 testing samples in the ratio of 70:30. The data has been split into train and test, to ensure that the model is not overfitting the data. According to the loss graph (shown below in section 5), the train and test data converged smoothly, which indicates that the model can be properly generalized. In the process of training the libraries used for LSTM implementation: the number of LSTM units used is 25; the number of epochs 50; dense layers 1; dropout 0.; batch size 80. The hardware and software specifications are Memory 8GB; 128 – bit LPDDR4x@1600MHz 512 GB/s. Compute 6-core NVIDIA Carmel ARM@v8.2 64-bit CPU 6MB L2 + 4MB L3; OS Ubuntu 20.04. The developing environment python 3.6 and TensorFlow 2.1.0.

Evaluation Criteria

The performance of the predictive model has been evaluated using the standard measure, Root Mean Square Error (RMSE). It is used to calculate an average magnitude of the error between measured and predicted $PM_{2.5}$ concentration values using the formula

$$RMSE = \sqrt{\sum \left(P_m - P_r\right)^2 / N} \qquad \ldots(1)$$

Where P_m denotes measured air pollution, P_r denotes predicted air pollution and N is the number of measured data.

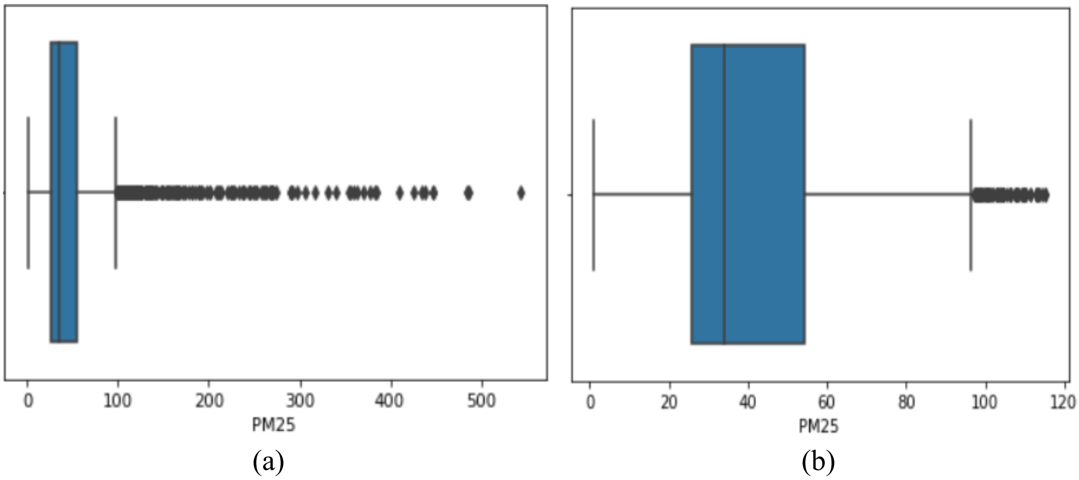
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2} \qquad \ldots(2)$$

Where y_i and $\hat{y}_i$ are original and predicted values and μ is the sample mean.

RESULTS AND DISCUSSION

Multivariate regression analysis was performed on the data set collected over Hyderabad (IDA Bollaram). 70% of the data was used to train the model, to obtain the $PM_{2.5}$ concentration values and predictions, for the next hour. Fig. 4 illustrates the loss graph of the tested and trained data set. From Fig. 4**,** it is evident that multivariate regression analysis performs well in predicting $PM_{2.5}$ values.

The RMSE value of training and test data as per the above graph is 0.0051 and the R^2 value is found to be 0.9862580518554221. This indicates that regression analysis gives better results in predicting the $PM_{2.5}$ levels in terms of RMSE. The R^2 value is the amount of the variation in



(a) (b)

Fig. 2: Box plot of $PM_{2.5}$ before (a) and after (b) removal of outliers.

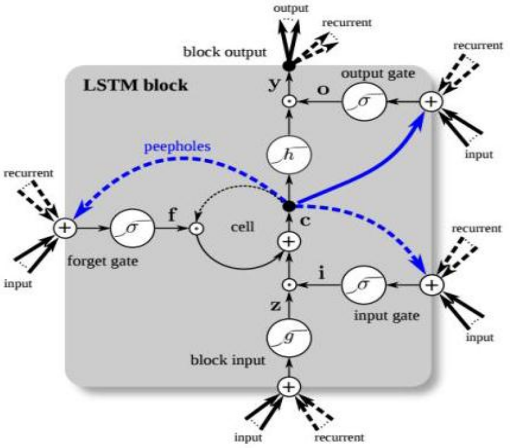


Fig. 3: Block System of LSTM model (Kök et al. 2017).

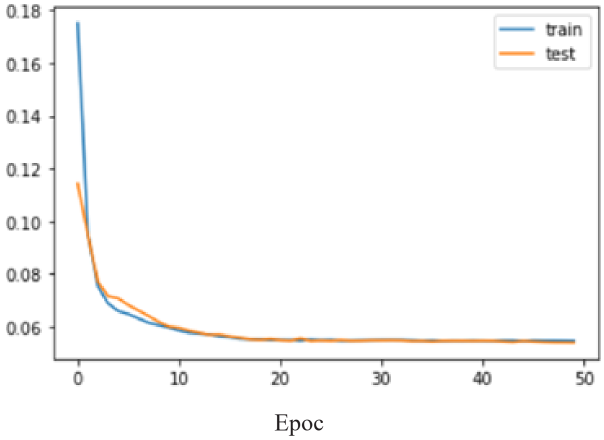


Fig. 4: Loss graph of the tested and trained data set.

the $PM_{2.5}$, which is predictable from the input independent variables, it indicates 98% of changes in $PM_{2.5}$ are due to the variations in the input variables.

The relation between $PM_{2.5}$ and other meteorological parameters was reported in the previous section. It was found that $PM_{2.5}$ has a negative correlation with temperature, relative humidity, and wind speed. This indicates that the low value of city temperature results in high $PM_{2.5}$ concentrations. As a result, there will be an increase in the potentiality of suspended particles, which causes the density of air to increase. Thereby, $PM_{2.5}$ records high concentration values at low temperatures due to a long stay of air in the atmosphere. Several studies reported that the urban $PM_{2.5}$ concentration decreases with an increase in relative humidity (Li et al. 2015). Similarly, the concentration of $PM_{2.5}$ is lower when the wind speed is high. This may be because the particles may have been washed away from the atmosphere due to the higher wind speed at that particular location. From the result obtained, the model converges with a very low value of mean squared error and gives a more accurate prediction for all ideal conditions.

CONCLUSION

There is a lot of costs involved in setting up the pollution sensing system and gathering $PM_{2.5}$ data across the city. We not only need to invest in technical assets but also invest in human capital as well. To address this issue, we may use a mobile technical environment, collect data for a particular period and train our model. In this work, the problem of forecasting hourly $PM_{2.5}$ has been addressed using a deep learning-based model called LSTM with an evaluation criterion of RMSE. We showed the experimental results of prediction with a low value of mean square error. Thus, we propose this method as an accurate prediction approach in all ideal conditions. Our system based on deep learning technology, using a fine-tuned LSTM time series regression model can precisely identify the pollution measure $PM_{2.5}$ on an hourly basis, thus helping the Pollution Control Board to understand the patterns of pollution in different areas across the city, without setting up a permanent sensing system at multiple locations in the city. It can be further ameliorated by training on a more extensive dataset and changing the backbone transfer learning model for latency improvement.

ACKNOWLEDGEMENT

The authors acknowledge Telangana State Pollution Control Board (TSPCB) and Central Pollution Control Board (CPCB) for making the data available to users. The authors would like to thank the management of Vasavi College of Engineering, Hyderabad, India, and Aurora's Technological and Research Institute, Hyderabad, India for their active support and encouragement.

REFERENCES

Ajay Kumar, M.C., Vinay Kumar, P. and Venkateswara Rao, P. 2020. Temporal variations of $PM_{2.5}$ and PM_{10} concentration over Hyderabad. Nat. Environ. Pollut. Technol., 19: 421-428.

Corani, G. 2005. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks, and lazy learning. Ecol. Model., 185: 513-529.

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

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. and Briggs, D. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ., 42: 7561-7578.

Kim, S., Lee, J.M., Lee, J. and Seo, J. 2019. Deep-dust: Predicting concentrations of fine dust in Seoul using LSTM. Mach. Learn., 3: 319.

Kök, I., Şimşek, M.U. and Özdemir, S. 2017. A deep learning model for air quality prediction in smart cities. IEEE Int. Conf. Big Data, 601: 1973-1980.

Li, X. 2017, Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ. Pollut., 231(1): 997-1004.

Li, Y., Chen, Q., Zhao, H., Wang, L. and Tao, R. 2015. Variations in PM_{10}, $PM_{2.5,}$ and $PM_{1.0}$ in an urban area of the Sichuan Basin and their relation to meteorological factors. Atmosphere, 6: 150-163.

Qin, D. 2019. A novel combined prediction scheme based on CNN and LSTM for urban PM25 concentration. IEEE Access, 7: 20050-20059.

Ravindra, K., Sidhu, M.K., Mor, S., John, S. and Pyne, S. 2016. Air pollution in India: bridging the gap between science and policy. J. Hazard. Toxic Radioact. Waste, 20: A4015003.

Singh, V., Biswal, A., Kesarkar, A.P., Mor, S. and Ravindra, K. 2020. High resolution vehicular PM_{10} emissions over megacity Delhi: Relative contributions of exhaust and non-exhaust sources. Sci. Total Environ., 699: 134273.

Van, V.E. 2017. Energy, land-use, and greenhouse gas emissions trajectories under a green growth paradigm. Glob. Environ. Change, 42: 237-250.

World Air Quality Report 2019. https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf

Xiao, F., Yang, M. and Fan, H. 2020. An improved deep learning model for predicting daily PM2.5 concentration. Sci. Rep., 10: 20988.

Yasin, A.A., Zeynep, C.A. and Hüseyin, O.A. 2018. Air pollution modeling with deep learning: A review. Int. J. Environ. Pollut. Environ. Model., 1(3): 58-62.

Zhang, K., Zhao, C., Fan, H., Yang, Y. and Sun, Y. 2019. Toward understanding the differences of $PM_{2.5}$ characteristics among five China urban cities. Asia-Pacific J. Atmos. Sci., 56: 493-502.

Zhao, C., Wang, Y., Shi, X., Zhang, D., Wang, C., Jiang, J.H., Zhang, Q. and Fan, H. 2019. Estimating the contribution of local primary emissions to particulate pollution using high-density station observations. J. Geophys. Res. Atmos., 124: 1648-1661.