Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 2101–2110

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches

Adil Masood[a]* , Kafeel Ahmad[b]

[a,b]*Department of Civil Engineering, Jamia Millia Islamia, New Delhi – 110025, INDIA*

Abstract

Particulate matter ($PM_{2.5}$) remains one of the most dominant contributors to air pollution in Delhi and its acute or chronic exposures have exerted serious health implications. Hence, it is necessary to accurately predict the magnitude of $PM_{2.5}$ concentrations in order to develop emission reduction strategies for air quality management. In regard to this, few machine learning techniques have been applied to predict daily $PM_{2.5}$ concentrations in Delhi. Two Different models i.e. SVM and ANN, were built on the inputs of various meteorological and pollutant parameters corresponding to 2-year period from 2016-18. Performance evaluation of the models for $PM_{2.5}$ prediction has been executed and the results have been discussed. The results of this simulation exercise indicate that the ANN shows better prediction accuracy than SVM for $PM_{2.5}$ prediction.

Keywords: Particulate Matter ($PM_{2.5}$); Machine Learning; Artificial Neural Network; Support Vector Machines.

1. Introduction

In recent years, Delhi, India's vast central metropolis has been subjected to the impact of urban anthropological activities which has resulted in an unprecedented environmental crisis [1]. One of the main features of this crisis has been the degraded air quality in this region. In the context of Delhi, particulate matter ($PM_{2.5}$) plays a dominant role in altering air quality, human health, and climate change. It has been reported that the average annual concentration

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: adil.engg.cvl@gmail.com

10.1016/j.procs.2020.03.258

of $PM_{2.5}$ in Delhi exceeds the air quality standard (NAAQS) of 40µg/m^3 by more than 200%, highlighting a serious health risk [2]. Vehicular emissions, construction activities, re-suspension of dust on roads, power plants, landfill fires and refuse burnings have been some of the major sources which contribute significantly to total $PM_{2.5}$ load in ambient air of Delhi. $PM_{2.5}$ is considered as a criteria pollutant and its exposure is associated with serious health ailments, including cardiovascular and respiratory issues, premature deaths and reduced birth weight in infants [3-5]. According to a study conducted by Kandlikar and Ramachandaran [6], people living in Delhi are 12 times more susceptible to the above mentioned health problems as compared to the people living in other parts of the country. Hence, there is a need to develop a reliable air quality forecasting model in order to improve the public awareness towards extreme air pollution events and also to assist vulnerable population groups in accessing useful information that would enable them to limit their exposure time.

However, despite the pivotal importance of these models in air pollution forecasting, it is difficult to characterize non-linear natural phenomena effectively. On the other hand, Machine learning approaches such as ANN and SVM provide faster prediction, better accuracy, and ease in performing multidimensional data operations. A number of studies have successfully applied machine learning approaches in air pollution forecasting. For example, Moisan et al. [7] compared the performances of a Dynamic multiple equation (DME) model in forecasting $PM_{2.5}$ concentrations with an ANN model and an ARIMAX model. They concluded that, although ANN on very few instances showed more significant and accurate results than the DME model, the overall performance of the DME model was slightly better than the ANN.

In another study, Suleiman et al. [8] assessed and compared three air quality control strategies, including Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Boosted Regression Trees (BRT) for forecasting and controlling roadside PM_{10} and $PM_{2.5}$. Interestingly, they reported that the neural network and regression tree based models exhibit better prediction performance as compared to the SVM model for PM_{10} forecasting while these models tend to slightly underperform for $PM_{2.5}$ forecasting.

Further Biancofiore et al. [9] applied three different approaches i.e. Recursive Neural Network, Feed Forward Neural Network, and multiple linear regression to estimate the daily average PM_{10} and $PM_{2.5}$ concentrations. Meteorological parameters together with the pollutant concentration data were considered as the input variables to the proposed models and the $PM_{2.5}$ concentrations were obtained as the output. The authors concluded that their recursive model achieved better results for the mentioned problem.

Contemporarily, Hoshyaripour et al. [10] compared the prediction capabilities of a deterministic model (WRF–Chem) with a neural network model to estimate the ozone level concentration in Sao Paulo, Brazil. Their ANN model was a feed-forward neural network with one hidden layer, containing two hidden neurons. The backpropagation learning algorithm was applied for training the network. Their results indicated that WRF-Chem model shows higher prediction accuracy, better convergence, and faster generalization than the ANN model.

In one of the examples, Dragomir et al. [11] proposed an experimental analysis to approximate the value of air quality index based on the data of various pollutant parameters such as Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Ozone (O_3) and Carbon monoxide (CO). In their study, WEKA, a machine learning application was used for analysis and other predictive modelling operations. The authors concluded that the proposed methodology was able to forecast the air quality index with less prediction error. Martin et al. [12] used two modelling approaches i.e. ANN and KNN to predict carbon monoxide for an urban area. The models were based on the meteorological and pollutant data observed for a period of three years (1999-2001). It was concluded that the prediction results can further be fine-tuned by incorporating the exogenous information of other parameters and the cyclic behaviour of Carbon monoxide. It is worth mentioning that the above-mentioned studies share a common drawback with respect to machine learning approach based on air pollution forecasting. Most of the studies reported that the techniques like ANN show a tendency towards under predicting the pollutant concentrations. Therefore, this work tries to address this issue by fine tuning the performance of ANN and SVM in particulate matter forecasting. The present work was conducted to evaluate the potential of various machine learning approaches such as ANN and SVM in making reliable and accurate predictions of $PM_{2.5}$ forecasting for New Delhi. The aim of this paper is twofold. Firstly, to develop a $PM_{2.5}$ based prediction model with meteorological parameters and pollutant concentrations as inputs using ANN and SVM techniques and secondly, to comparatively examine the prediction performance of these developed models.

2. Design of Experiment

2.1. Study Area

Delhi, city and national capital territory (NCT) of India has earned the distinction of being one of the most polluted cities in the world [13]. Being a landlocked region there are limited avenues for dispersion of pollutants that often results in high pollution episodes [1]. The region has sub-tropical semi-arid (steppe) climatic conditions with an annual average temperature of 31.5°C. Prevailing wind direction is westerly or north-westerly. The average annual wind speed varies from 0.9-2 m/s [14]. Delhi gets most of its rainfall during the monsoon season and the total average rainfall received is 611.8 mm/year. The location of the study area is presented in the Fig. 1. below.

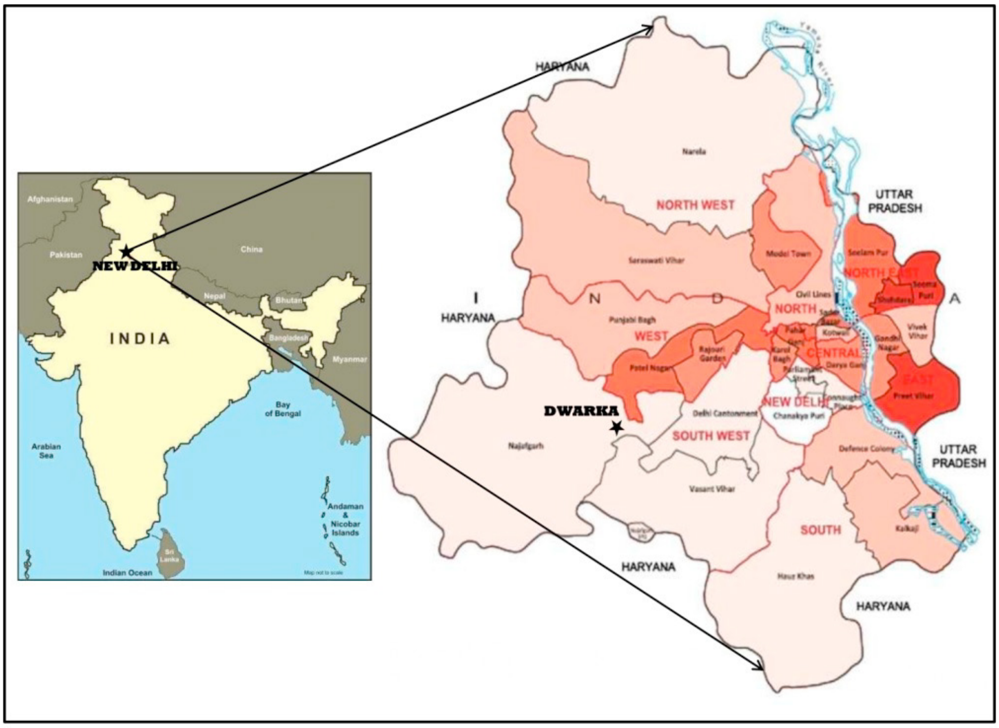

Fig. 1. Delhi map showing the CPCB monitoring station–Dwarka (Map not to scale)

2.2. Data Observation

In the present study, data set pertaining to CPCB monitoring station at Dwarka, New Delhi (28° 35' 31.70"N, 77° 2' 45.78"E) has been procured from the online data monitoring portal of CPCB (http://cpcb.nic.in/) for a period of two years (May 2016 - May 2018). The daily concentration data of various pollutants, such as $PM_{2.5}$ (particulate matter with aerodynamic diameter less than 2.5μm), SO_2 (Sulphur dioxide), CO (Carbon monoxide), NO (Nitric oxide), NOx (oxides of Nitrogen), C_7H_8 (Toluene), NO_2 (Nitrogen dioxide) and daily meteorological data including vertical wind speed, wind speed, wind direction, temperature, relative humidity and solar radiation for the same period have been utilized as primary variables to develop the ML models. In order to avoid any bias with a particular missing value in the dataset that may affect the forecasting model results, a linear interpolation technique has been applied. Out of the total dataset which included, 687 values, the missing data for $PM_{2.5}$, SO_2 and C_7H_8 were recorded as 25, 7 and 13 (days), respectively. Moreover, the general statistics of the experimental observations considered in this work from May 2016-May 2018 has been shown in Table 1.

Table 1. General statistics of the experimental data considered in this study

Parameter	Unit	Range	Mean	St. Dev.	Min.	Max.
$PM_{2.5}$	µg m^{-3}	1454.93	137.77	109.40	1.47	1456.40
CO	mg m^{-3}	3.47	0.69	0.35	0.01	3.48
SO_2	µg m^{-3}	121.20	9.69	6.59	1.44	122.64
NO_x	ppb	1386.80	37.13	71.21	1.06	1387.86
NO	µg m^{-3}	466.54	20.15	25.22	3	469.54
C_7H_8	µg m^{-3}	69.20	10.17	8.87	0.04	69.24
NO_2	µg m^{-3}	283.61	33.74	19.80	1.92	285.53
Wind Speed (WS)	ms^{-1}	3.08	0.98	0.43	0.06	3.14
Wind Direction (WD)	degrees	320.83	189.91	54.39	4.55	325.38
Vertical Wind Speed (VWS)	ms^{-1}	12.10	0.11	0.86	-10	2.10
Relative Humidity (RH)	%	140.40	54.10	20.42	1.69	142.09
Temperature	°C	58.80	28.15	6.90	1.17	59.97
Solar Radiation	w/m^2	787.67	159.63	112.74	1.50	789.17

3. Methodology

3.1. Artificial Neural Networks

ANN's are tools for applying machine learning that mathematically model the functioning of a human brain and have been used to build solutions for a number of problems from diverse domains [19-21]. These networks based on connectionist architectures are interconnected by non-linear processing elements called nodes. These nodes are normally stacked in a fully-connected system of three or more layers namely, the input layer, hidden layer(s), and the output layer. In the present work, we applied a Feed Forward single hidden layer neural network (FFNN) to predict the $PM_{2.5}$ concentrations (Fig. 2). This network was trained using the Resilient back propagation algorithm which has performed effectively in our case and has been successfully implemented by other researchers in the field of air pollution forecasting [18]. The input data were normalized by applying the min-max normalization technique as given by the following equation (1):

$$Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Where,
Z_i= ith normalized value
x_i= ith observed value for the variable x
min(x)= minimum value in the dataset
max(x)= maximum value in the dataset

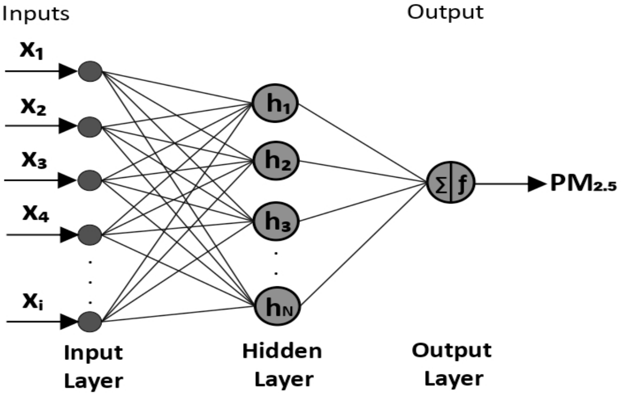

Fig. 2. Feed forward neural network for $PM_{2.5}$ prediction

3.2. Support Vector Machines

SVMs are discriminative classifier techniques that convert the input space into a multi-dimensional characteristic space [22]. These supervised machine learning algorithms find their application in both classification and regression based problems [23]. In this work the SVM was implemented with a linear kernel, with kernel scale value of 1 and a Sequential Minimal optimization as the solver.

4. Results

In order to avoid poor generalization by the model, and produce better simulations, 80% of the observed data was applied for training and the remaining 20% was utilized for the validation process. A number of experiments were performed to decide the best combination of hidden layers and their neurons, learning algorithm and an optimum learning rate. Accordingly, the ANN model with a topology of 12:20:1 and at 600 epochs incorporating a learning rate of 0.2 was found to be optimal for predicting the $PM_{2.5}$ concentrations. The accuracy of the ANN model was evaluated by performing a regression analysis between the resulted output and the test data. In the training stage, the simulated $PM_{2.5}$ values and the target data showed the best fit with a high correlation coefficient (R) value of 0.856 as presented in Fig. 3.

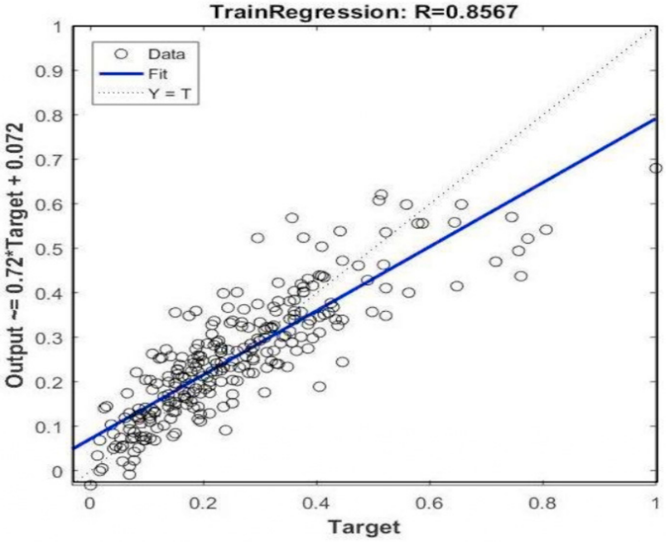

Fig. 3. Correlation plot for training phase of the ANN based model

The high R-value indicated that the simulated data was consistent with the input data and the proposed model is validated to be applied as a $PM_{2.5}$ prediction model. The more the line of linear regression (solid line; labelled as Fit) converges with the dashed line (labelled as Y=T, signifying that the simulated ANN results match the desired values), the more efficient the model is expected (Fig. 3 & Fig. 4).

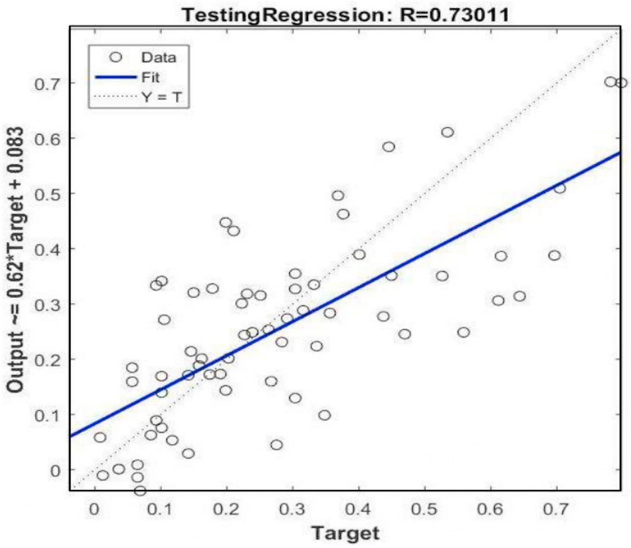

Fig. 4. Correlation plot for the testing phase of the ANN based model

Furthermore, the MSE (mean square error) was used as a statistical measures to analyse the prediction accuracy of the ANN model. As shown in Fig. 5, the training performance of the ANN model improves considerably as there is a reduction in MSE and the best training performance was observed as 0.00698 at 600 epochs.

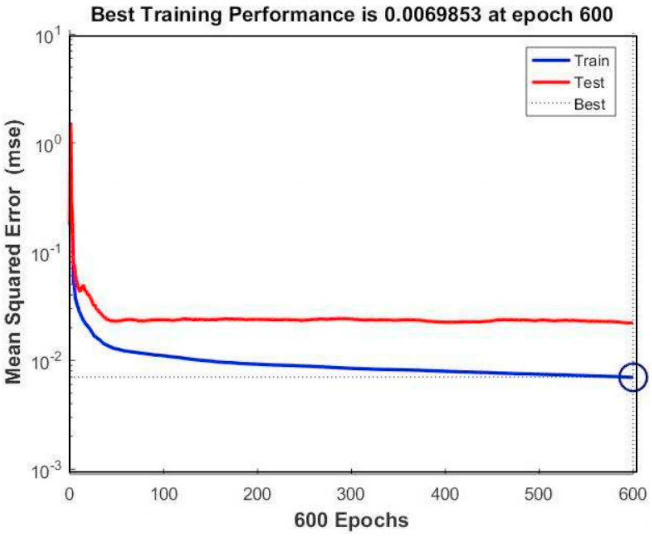

Fig. 5. Training performance of ANN for simulating $PM_{2.5}$

The MSE observed during the training phase was 0.0070. The ANN prediction performance was also found to be satisfactory for the testing phase as well. Fig. 4 specifies a good correlation of 0.7301 between the model predictions and the observed data. It was observed that the ANN model has an MSE of 0.0191 during the testing phase. The error histogram, illustrated in Fig. 6 was used to verify the prediction performance of the ANN model.

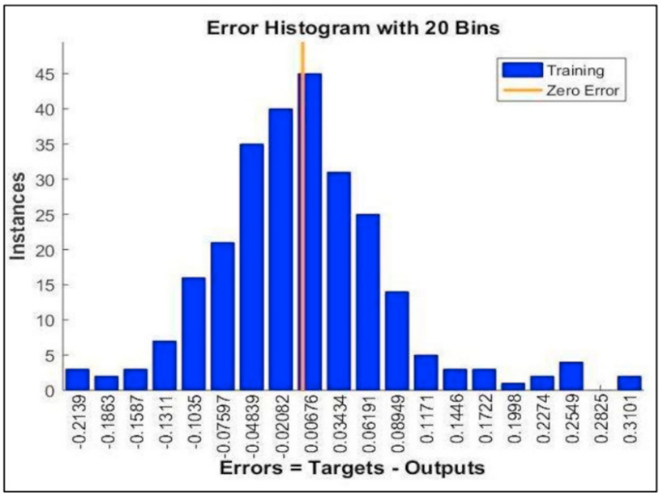

Fig. 6. Error histogram for the ANN model

The total error range of the ANN model has been divided into 20 bins or vertical bars. As shown in Fig. 6 the total error for our model ranges from -0.2139 (leftmost bin) to 0.3101 (rightmost bin). It is worth mentioning that, around 45 samples of the training dataset lie in the error range of (-.00634, .01986), 40 samples lie in the error range of (-0.03392, -0.0072), 35 samples lie in the range of (-.06149, -0.03529), 30 samples have an error range of (0.02124, 0.04744), 25 samples have an error range of (0.04881, 0.07501) and for 20 samples the error varies from (-0.08907, - 0.06287), respectively.

Table 2. Observed values of MSE for each applied technique

Model	MSE on Train data	MSE on Test data
ANN	0.00669	0.0191
SVM	0.0171	0.0314

It may be observed from Table 2, that the ANN based model was able to achieve considerably lower values of the mean square error over the train and test dataset selected, which signifies that the differences between the observed and the expected target values in the modelling exercise, were minimal.

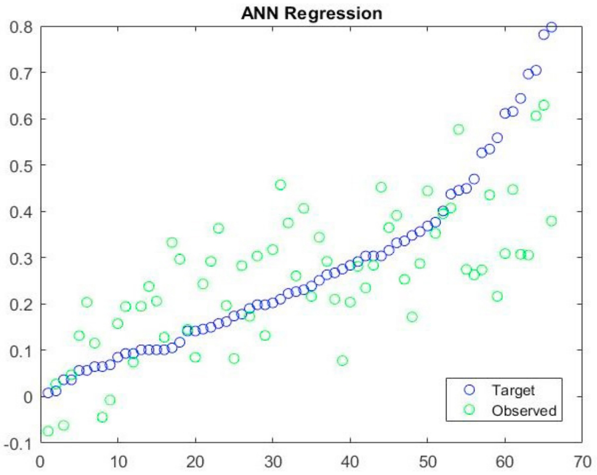

Fig. 7. Plot representing the Target and observed value from the ANN based regression model

Fig. 7 presents a plot of the test set target values against the observed predicted values from the ANN based regression model implemented in this work. It is evident from this figure that the predicted values were very close to the actual test set target values, suggesting the fact that ANN based model was very much efficient in performing this prediction for most of the data points.

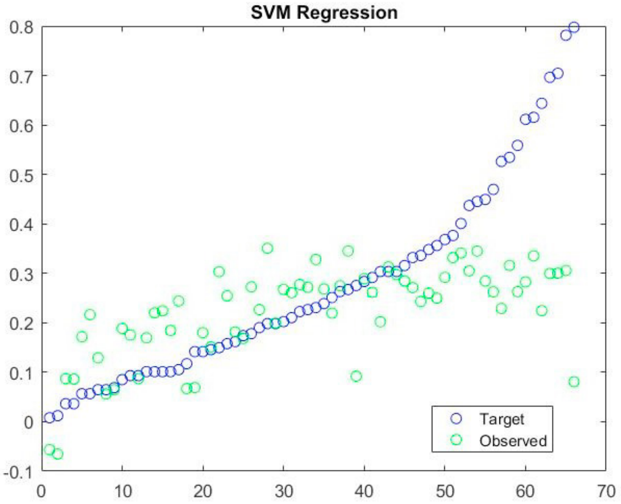

Fig. 8. Plot representing the Target and observed value from the SVM based regression model

However, the SVM based model was not that much efficient in the task of $PM_{2.5}$ prediction as seen in Fig. 8 as its predicted values were not in close ranges to the actual target test set values. Also, for some data points, the SVM based model abberated significantly from the target data points, which eventually led to a decline in its test MSE performance presented in Table 2.

In the light of the above results, it can be argued that the proposed ANN model introduces a good accurate prediction for $PM_{2.5}$ when compared with the experimental dataset. In addition to this, the ANN model outperforms the SVM and provided better accuracy and less error, thus thought as more efficient technique for air pollution prediction studies.

5. Conclusions

In this work, a couple of machine learning techniques, namely the ANN and SVM have been used for $PM_{2.5}$ forecasting application in Sub-tropical semi-arid (steppe) climatic conditions of Delhi. It was observed that the ANN-based $PM_{2.5}$ prediction model involving the air pollution and meteorological data exhibited better performance for the available test dataset. The results indicated that ANN has merit in the prediction of $PM_{2.5}$ concentrations and is able to outperform other machine learning approach used in this study. In the case of the ANN model, the correlation values for training and testing were found to be 0.856 and 0.730 respectively, showing the suitability of the model for $PM_{2.5}$ prediction. The results also indicated that, with improvements in model topography, data quality and computational power, the prediction accuracy of the ANN model can further be fine-tuned for practical application. Hence it may be concluded that the ANN model with better generalization capabilities may be considered as an optimal candidate technique, for model development, for multi-dimensional complex problems like air pollution.

References

[1] Masood, Adil, Ahmad, Kafeel, Ahmad, Shamshad, (2018) “Urban Roadside Monitoring, Modeling and Mapping of Air Pollution.” Applied Journal of Environmental Engineering Science ; **3(2)**:3–2.

[2] Chowdhury, Sourangsu, Sagnik Dey, Larry Di Girolamo, Kirk R. Smith, Ajay Pillarisetti, and Alexei Lyapustin. (2019) "Tracking ambient PM2. 5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset." Atmospheric Environment **204**: 142-150.

[3] Coker, Eric, Silvia Liverani, Jo Kay Ghosh, Michael Jerrett, Bernardo Beckerman, Arthur Li, Beate Ritz, and John Molitor. (2016) "Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County." Environment international **91**: 1-13.

[4] Brook, Robert D., Sanjay Rajagopalan, C. Arden Pope III, Jeffrey R. Brook, Aruni Bhatnagar, Ana V. Diez-Roux, Fernando Holguin et al. (2010) "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association." Circulation **121(21)**: 2331-2378.

[5] Lippmann, Morton. (2014) "Toxicological and epidemiological studies of cardiovascular effects of ambient air fine particulate matter (PM2. 5) and its chemical components: coherence and public health implications." Critical reviews in toxicology **44(4)**: 299-347.

[6] Kandlikar, Milind, and Gurumurthy Ramachandran. (2000) "The causes and consequences of particulate air pollution in urban India: a synthesis of the science." Annual review of energy and the environment, **25 (1)**: 629-684.

[7] Singh, Vikas, Claudio Carnevale, Giovanna Finzi, Enrico Pisoni, and Marialuisa Volta. (2011) "A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations." Environmental Modelling & Software **26(6)**: 778-786.

[8] Konovalov, I. B., Matthias Beekmann, Frédérik Meleux, A. Dutot, and Gilles Foret. (2009) "Combining deterministic and statistical approaches for PM10 forecasting in Europe." Atmospheric Environment **43(40)**: 6425-6434.

[9] Honoré, Cécile, Laurence Rouil, Robert Vautard, Matthias Beekmann, Bertrand Bessagnet, Anne Dufour, Christian Elichegaray et al. (2008) "Predictability of European air quality: Assessment of 3 years of operational forecasts and analyses by the PREV'AIR system." Journal of Geophysical Research: Atmospheres **113(D4).**

[10] Moisan, Stella, Rodrigo Herrera, and Adam Clements. (2018) "A dynamic multiple equation approach for forecasting PM2. 5 pollution in Santiago, Chile." International Journal of Forecasting **34(4)**: 566-581.

[11] Suleiman, A., M. R. Tight, and A. D. Quinn. (2019) "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2. 5)." Atmospheric Pollution Research **10(1)**: 134-144.

[12] Biancofiore, Fabio, Marcella Busilacchio, Marco Verdecchia, Barbara Tomassetti, Eleonora Aruffo, Sebastiano Bianco, Sinibaldo Di Tommaso, Carlo Colangeli, Gianluigi Rosatelli, and Piero Di Carlo. (2017) "Recursive neural network model for analysis and forecast of PM10 and PM2. 5." Atmospheric Pollution Research **8(4)**: 652-659.

[13] Hoshyaripour, Gholamali, G. Brasseur, Maria de Fátima Andrade, M. Gavidia-Calderón, Idir Bouarar, and Rita Yuri Ynoue. (2016) "Prediction of ground-level ozone concentration in São Paulo, Brazil: deterministic versus statistic models." Atmospheric environment **145**: 365-375.

[14] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique. (2010) " Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, **LXII 1(2010)**: 103-108.

[15] Martin, M. L., I. J. Turias, F. J. Gonzalez, P. L. Galindo, F. J. Trujillo, C. G. Puntonet, and J. M. Gorriz. (2008) "Prediction of CO

maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks." Chemosphere **70(7)**: 1190-1195.

[16] Akhtar, Aly, Sarfaraz Masood, Chaitanya Gupta, and Adil Masood. (2018) "Prediction and analysis of pollution levels in Delhi using multilayer perceptron." In Data Engineering and Intelligent Computing, pp. 563-572. Springer, Singapore.

[17] Srivastava, Arun, and V. K. Jain. (2007) "Size distribution and source identification of total suspended particulate matter and associated heavy metals in the urban atmosphere of Delhi." Chemosphere **68(3)**: 579-589.

[18] Athanasiadis, Ioannis N., Vassilis G. Kaburlasos, Pericles A. Mitkas, and Vassilios Petridis. (2003) "Applying machine learning techniques on air quality data for real-time decision support." In First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland.

[19] Masood, Sarfaraz, Shubham Gupta, Abdul Wajid, Suhani Gupta, and Musheer Ahmed. (2018) "Prediction of human ethnicity from facial images using neural networks." In Data Engineering and Intelligent Computing, pp. 217-226. Springer, Singapore.

[20] Goel, Anshuman, Mohd Sheezan, Sarfaraz Masood, and Aadam Saleem. (2014) "Genre classification of songs using neural network." In 2014 International Conference on Computer and Communication Technology (ICCCT), pp. 285-289. IEEE.

[21] Masood, Sarfaraz, Madhav Mehta, and Danish Raza Rizvi. (2015) "Isolated word recognition using neural network." In 2015 Annual IEEE India Conference (INDICON), pp. 1-5. IEEE.

[22] Kecman, Vojislav, and Lipo Wang. (2005) "Support vector machines: theory and applications.”

[23] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: A library for support vector machines. (2011) "ACM transactions on intelligent systems and technology (TIST) **2(3)**: 27.