Particulate matter forecasting using artificial neural network and support vector machine based models

Adil Masood[1], Kafeel Ahmad[1]

{adil169375@st.jmi.ac.in, kahmad2@jmi.ac.in}

Department of Civil Engineering, Jamia Millia Islamia University, New Delhi, 110025, India[1].

Abstract. Particulate matter (PM_{10}) remains the most important air pollutant that plays a dominant negative role in deteriorating the air quality of Delhi and its surrounding regions. Accurate forecasting of this criteria pollutant is crucial for managing the air status and providing valuable information to the sensitive population groups. In this study, two machine learning-based techniques, i.e., Artificial neural network (ANN) and Support vector machine (SVM) were applied to predict PM_{10} using air quality and meteorological parameters as inputs. The data corresponding to a period of two years (2015-16) having 730 observations was used in this study. The model performances were assessed by R^2, RMSE, and IA values. The results suggested that the ANN model with $R^2 = 0.896$, RMSE = 46.6 and IA = 0.987, outperformed the SVM based model in terms of PM_{10} forecasting. Overall, the findings of this work highlighted that ANN is an accurate and reliable technique for forecasting PM_{10} concentrations.

Keywords: Artificial neural network (ANN); Support vector machine (SVM); PM_{10}; Roughness coefficient.

1 Introduction

Particulate matter (PM_{10}) based pollution has become one of the most pressing challenges in the city of Delhi and has gained widespread public attention, especially due to its adverse environmental impact and associated health effects. Therefore, accurate prediction of PM_{10} concentration is required to help planners and sensitive population groups in this region to reduce the effect of long-term exposure to it. In recent years, machine learning (ML) based forecasting techniques such as ANN and SVM have demonstrated great success in predicting high PM_{10} episodes. Because of its immense potential to manage and prevent the destructive implications of different air pollutants on human health, these techniques have gained substantial attention in the field of atmospheric sciences. Several

ICIDSSD 2022, May 24-25, New Delhi, India

DOI 10.4108/eai.24-5-2022.2318563

researchers have used various forecasting techniques to perform short- and long-term forecasting studies [1], [2], [3]. These forecasting methods can be broadly grouped into two main categories: Deterministic and Non-deterministic techniques (Statistical and Machine learning-based methods). It is noteworthy that between these two categories, the non-deterministic approach, more specifically the ML-based techniques are currently more popular than the traditional deterministic techniques due to their ability to handle complex and non-linear relationships that exist between air quality variables. Machine learning (ML) has experienced a resurgence of interest in forecasting ambient PM_{10} concentrations in the last decade. With rapid technological advances in big data analytics, e.g., improved computing platforms, scalable storage systems, and high-speed parallel processing machines, ML has drawn the attention of researchers for developing advanced and accurate PM_{10} forecasting systems. Previous studies have effectively developed numerous data-driven models based on ANN and SVM techniques for forecasting PM_{10} concentrations. For example, Gualtieri et al. [4] developed an ANN model to predict PM_{10} concentrations for an urban region and compared the accuracy of the ANN model with a linear model. It was concluded that the ANN model was better suited for PM_{10} forecasting in terms of accuracy and robustness than the linear model. Bozdağ et al. [5] designed an ANN model for forecasting PM_{10} concentrations by applying air quality data as input . Based on the results, it was observed that the ANN model presented good overall accuracy for PM_{10} forecasting. Akhtar et al. [6] investigated the performance of ANN and other machine learning models for the prediction of PM_{10} concentrations in an urban airshed . The results revealed that the ANN model was able to predict more reliable and accurate PM_{10} concentrations compared to other machine learning models. Li and Tao, [7] implemented a Support Vector Machine model based on wavelet transform (W-SVM) for the prediction of PM_{10} concentrations using air quality and meteorological variables as inputs. The results suggested that the proposed SVM model successfully predicted the PM_{10} concentration levels with improved accuracy. Weizhen et al. [8] reported the development of an SVM-based forecasting model on the inputs of air quality and other influential meteorological parameters such as aerosol optical depth (AOD) . It was concluded that the inclusion of AOD as an input significantly improved the PM_{10} forecasting performance of the SVM model. All above-cited works present deep indispensability for a precise and reliable study with reference to PM_{10} forecasting and also highlight that both ANN and SVM are considered as powerful state-of-the-art techniques for predicting high PM_{10} episodes. Therefore, keeping the above in mind, the study lays out the following objectives.

- To develop two high-precision ML models i.e. ANN and SVM for prediction of PM_{10} concentrations.
- To compare the forecasting performances of these models and identify the best model in terms of accuracy for PM_{10} prediction.

2 Design of Experiment

2.1 Study area

With the fast-growing economy and population, air quality deterioration in Delhi has become a topic of concern in recent years. The capital city of India is the center of power, trade, and commerce and covers an area of 1483 km^2 with a population of 29 million in 2020[9]. The unique topography and poor dispersion conditions often result in high PM episodes in the region. The city experiences semi-arid sub-tropical climatic conditions for most of the year and has a mean annual precipitation of around 610 mm. The GIS-based map of Delhi showing the monitoring site locations has been presented in **Figure 1** below.

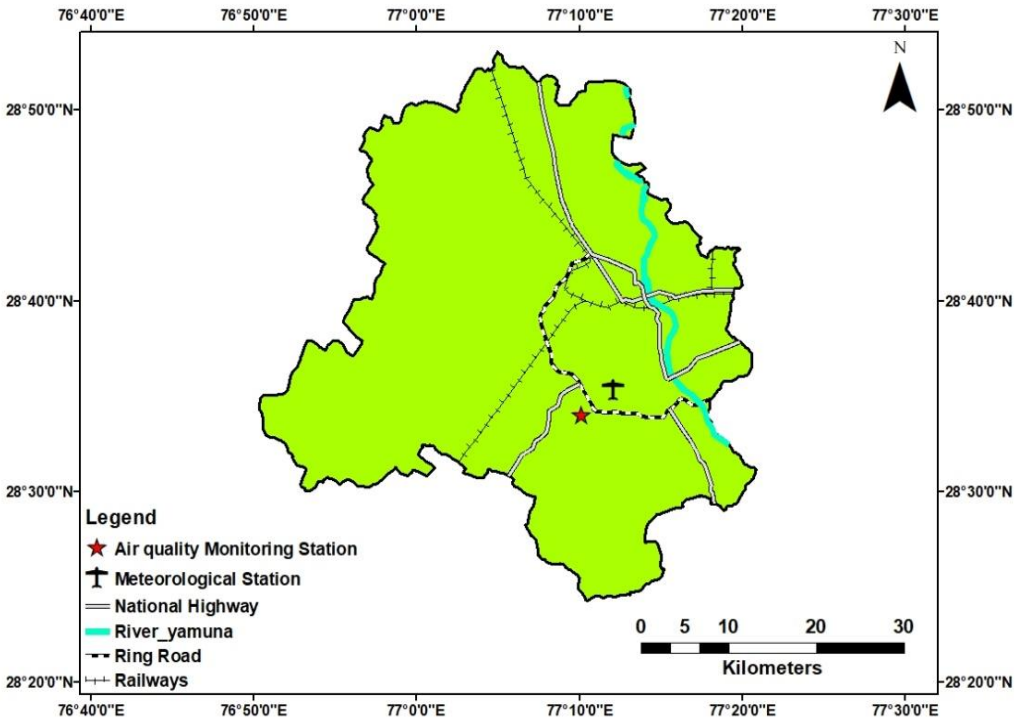

Fig.1. Map of the study area

2.2 Dataset

Daily averaged air quality and meteorological data used in this study has been procured from the air quality monitoring station at R K Puram, governed by the Delhi pollution control committee (DPCC) and meteorological station at Safdarjung airport operated by the Indian meteorological department (IMD). The data belongs to a period of two years (2015-16) and includes 730 observations. The entire dataset was preprocessed using the min-max normalization technique and was further divided in such a way that 80% of the data was used for training and 20% for testing. The characteristics of both the training and testing data sets have been presented in **Table 1**. A total of twenty-one parameters (wind speed, wind direction, temperature, atmospheric pressure, etc.), including the novel aerodynamic roughness coefficient (Z_0), were considered as inputs, and PM_{10} was adopted as the output for the ML models.

Table 1. Characteristics of the dataset used in the study

Parameters	**Units**	**Minimum**		**Maximum**		**Mean**		**Standard Deviation**	
		Training	Testing	Training	Testing	Training	Testing	Training	Testing
PM_{10}	$\mu g/m^3$	35.25	39.30	880.64	652.12	266.17	259.78	138.21	141.21
NO	$\mu g/m^3$	1.01	1.90	410.12	249.46	62.82	62.55	58.11	58.98
NO_2	$\mu g/m^3$	21.44	25.70	153.26	145.78	73.17	72.91	25.07	27.53
NO_x	ppb	24.06	25.91	315.23	284.55	120.46	127.54	57.81	66.55
NH_3	$\mu g/m^3$	7.71	5.52	144.55	102.09	39.95	41.15	21.91	19.12
CO	mg/m^3	0.55	0.54	11.38	21.84	2.58	2.43	1.64	1.98
SO_2	$\mu g/m^3$	1.90	3.22	114.20	107.30	26.27	24.50	17.12	16.14
Ozone	$\mu g/m^3$	15.34	21.08	86.42	85.26	55.14	56.64	15.23	14.05
Benzene	$\mu g/m^3$	0.19	0.25	32.01	25.13	7.58	6.96	5.23	5.90
Toluene	$\mu g/m^3$	0.25	1.10	69.06	62.22	18.21	17.57	12.20	13.08
$PM_{2.5}$	$\mu g/m^3$	18.09	20.30	720.11	460.5	133.20	131.39	93.20	86.68
Temp.(max.)	°C	13.00	13.21	45.20	42.40	34.29	31.38	6.54	7.08
Temp.(min.)	°C	4.20	4.09	30.90	30.60	18.90	19.36	7.30	7.60
WS(avg.)	m/s	1.29	1.19	8.96	9.07	5.20	5.38	2.02	2.10
WD	Degree	0	0	98	81	19.25	18.41	14.56	15.25
Rainfall	mm	0	0	67.60	92.90	1.20	2.83	5.10	10.46
Evaporation	mm	0.91	0	8.40	7.50	4.14	4.38	1.30	1.40
DOR	hours	0	0	12	12	0.13	0.47	0.65	2.60
Humidity	g/m^3	13.12	22.86	93.75	97	58.18	61.95	16.29	15.84
Atm. Pressure	millibar	12.04	312.83	7680.33	1341.61	1017.12	1008.02	298.04	58.16
Z_0	meter	0.000442	0.003646	10	10	2.80	3.35	2.10	2.27
AQI	-	90	90	720	560	290.61	298.67	106.10	104.15

3 Methodology

3.1 Artificial Neural Networks

Neural networks are a branch of machine learning created in the 1950s for the purpose of replicating the cognitive functioning and behavior of a biological brain. They are systems that comprise many interconnected non-linear processing units, called neurons. The network architecture of an ANN is composed of many nodes or neurons that are arranged into a sequence of layers [10]. In general, the network incorporates an input layer, hidden layers, and an output layer. In this study, a Multilayer feed-forward neural network (MLFFNN) trained using the Levenburg-Marquardt algorithm with two hidden layers was applied to forecast the PM_{10} concentrations (**Figure 2**).

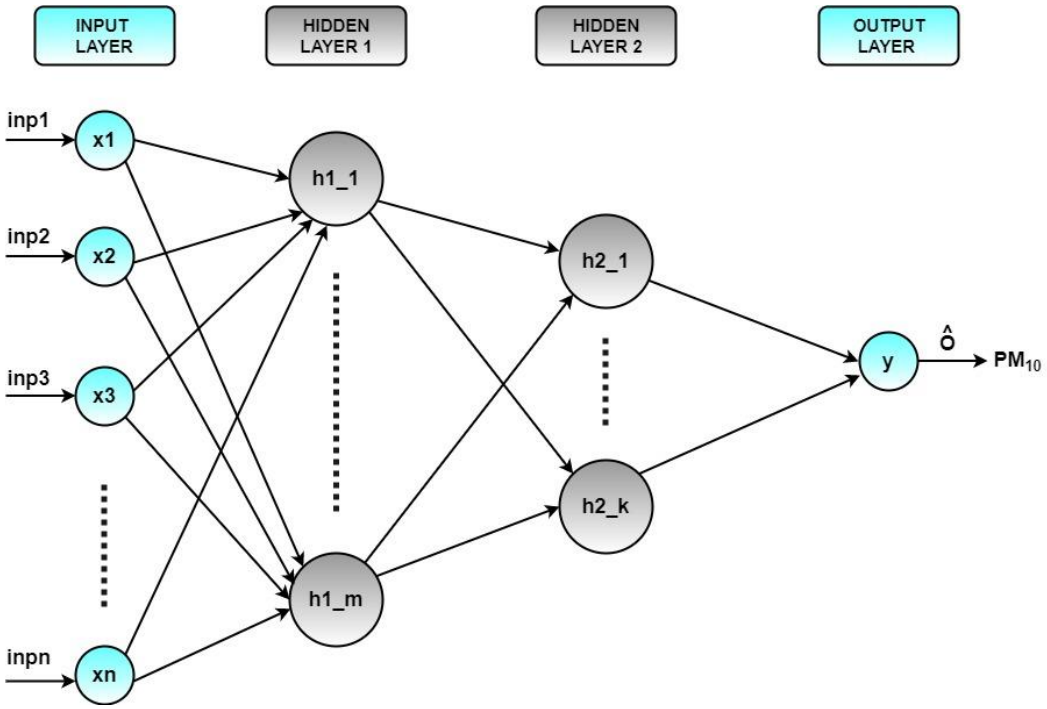

Fig.2. A multilayer feed forward neural network

3.2 Support Vector Machine

Support Vector Machine (SVM) is a type of machine learning technique that operates on the principle of creating a hyperplane in an n-dimensional space which allows the separation of data into groups or classes [11][12]. The technique finds application for both classification and regression-based problems. In this study, two SVM models (SVM_lin and SVM_pol) were developed for PM_{10} forecasting using the linear and polynomial kernels. Sequential minimal optimization (SMO) and Iterative Single Data Algorithm (ISDA) were chosen as the solvers. The kernel scale and polynomial order values were set as 1 for the model development.

4 Results

4.1 Results of the MLFFNN model

The MLFFNN model proposed for this work was an outcome of the iterative procedure which led to the development of an optimal model for PM_{10} forecasting. The dataset shown in **Table 1** was selected for the model development. The scatter plot between the observed and the simulated PM_{10} concentrations has been shown in **Figure 3a** for the testing stage. The figure suggests that the predicted values from the MLFFNN model are in close proximity to the observed values and almost follow the same trend. The statistical analysis results for both training and testing phases have been presented in **Table 2**. It is evident from these figures that MLFFNN model with an R^2= 0.950, RMSE = 30.831, IA = 0.965 for training and R^2 = 0.873, RMSE = 51.125, IA = 0.962 for testing has performed well for PM_{10} forecasting.

Table 2. Statistical analysis based results for the machine learning techniques

Techniques	**Training**			**Testing**		
	IA	R^2	RMSE	IA	R^2	RMSE
ANN	0.987	0.950	30.831	0.970	0.896	46.600
SVM_lin	0.968	0.882	47.513	0.967	0.875	50.120
SVM_pol	0.965	0.880	47.859	0.962	0.873	51.125

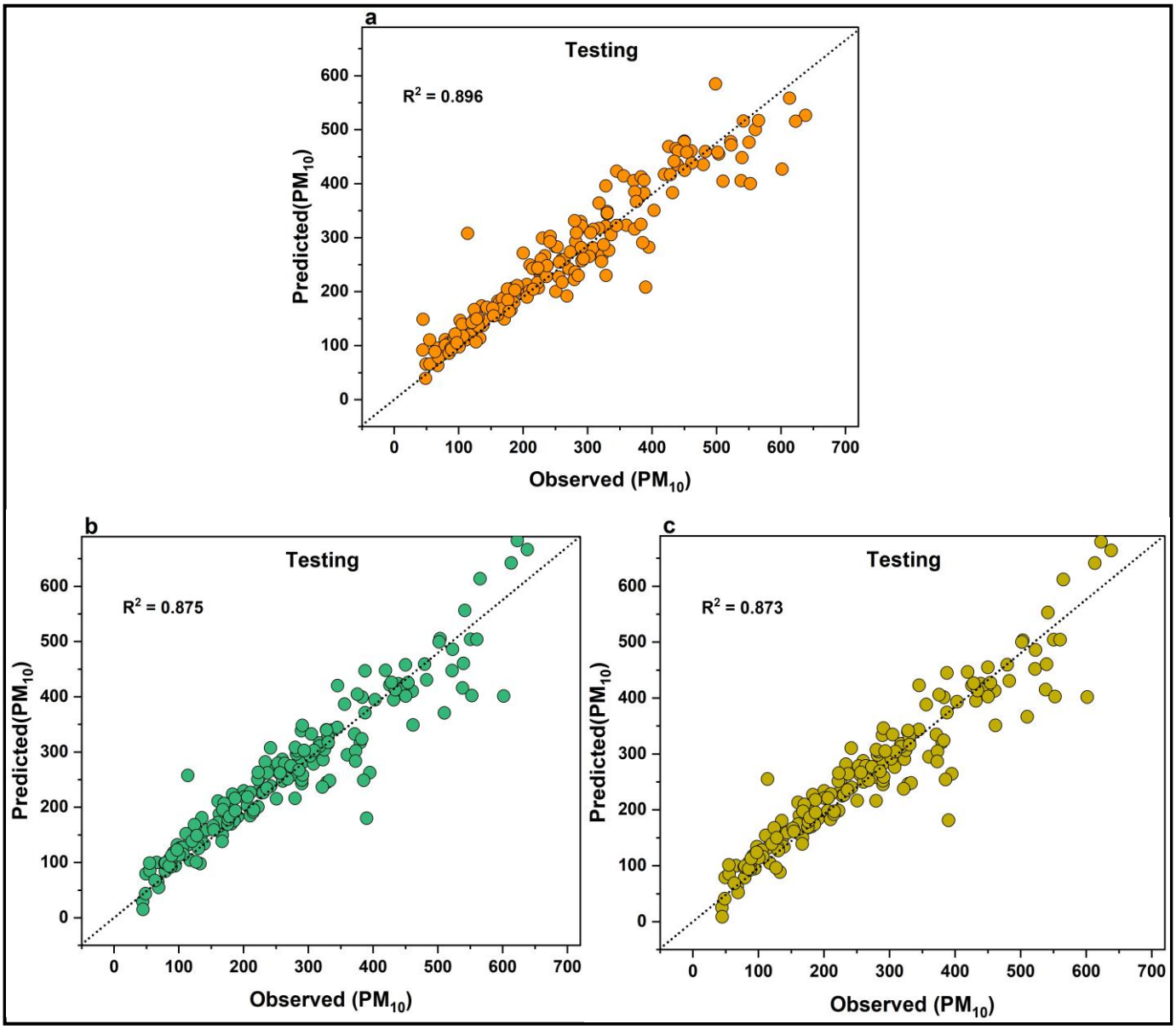

Fig.3 The PM_{10} forecasting performance of (**a)** MLFFNN model, (**b)** SVM_lin model, and (**c**) SVM_pol model

4.2 Results of the SVM model

The SVM models proposed for this work were developed with a similar trial and error procedure. A total of two models, i.e., SVM_lin and SVM_pol, were generated and used for PM_{10} forecasting. The dataset shown in **Table 1** was selected for both the model development. The scatter plot between the observed and the simulated PM_{10} concentrations for the testing stage in the case of SVM_lin and SVM_pol have been shown in **Figure 3b** and **Figure 3c**. The figure suggests that the predicted values from the SVM model are in close proximity to the observed values and agree closely with the line of perfect agreement (shown in dotted). As presented in **Table2**, the IA, R^2, and RMSE values for the SVM_lin model were 0.968, 0.882, and 47.513 for the training stage and 0.967, 0.875, and 50.120 for the testing phase. Moreover, the IA, R^2, and RMSE values for the SVM_pol were 0.965, 0.880, and 47.859 for the training phase and 0.962, 0.873, and 51.125 for the testing stage. It was observed from these results that the performance of the SVM_lin model was better than the SVM_pol and thus making it more suitable for the prediction of PM_{10} concentrations.

4.3 Comparison of models

The comparative analysis for the ML models was carried out on the basis of scatter plots and Taylor diagram, as shown in **Figures 3** and **4**. According to the scatter plot, the predicted PM concentrations generated by the MLFFNN were closest to the line of perfect agreement (R^2=0.896) in comparison to the SVM models. Similarly, from the taylor plot (**Figure 4**) it can be ascertained that the MLFFNN model (red square) is placed closest to the observed value showing higher correlation and lower standard deviation compared to the other ML models and thus may be considered as the best and the most accurate model for PM_{10} forecasting.

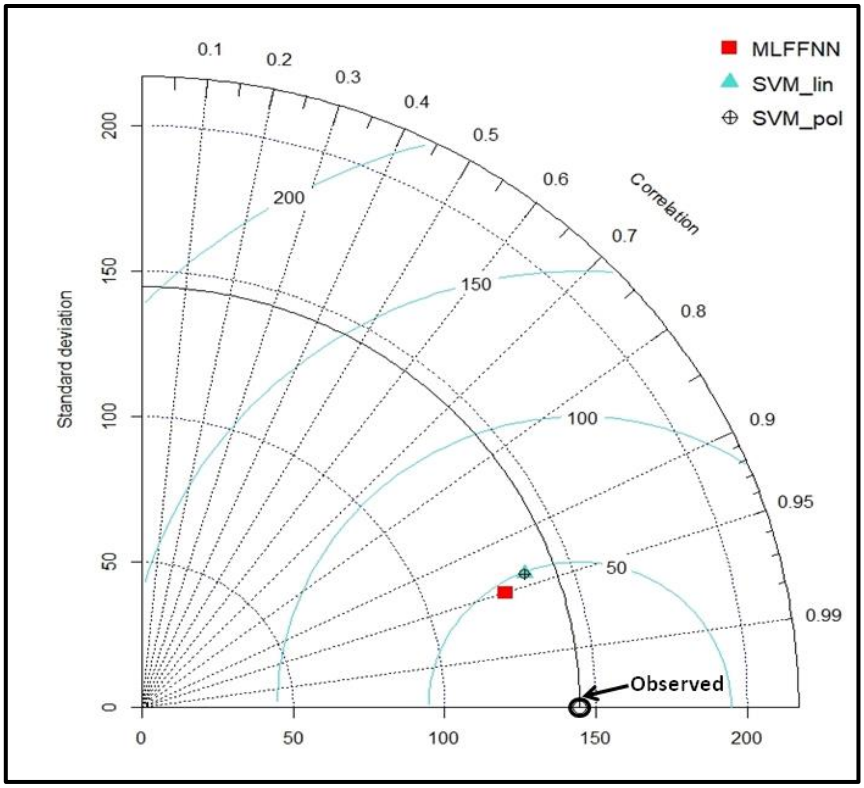

Fig.4. Taylor diagram depicting the PM_{10} forecasting performance of MLFFNN, SVM_lin, and SVM_pol models based on the testing dataset.

5 Conclusion

This work evaluated the performances of two ML models, i.e., ANN and SVM, for predicting the PM_{10} concentrations in the city of Delhi. The air quality and meteorological data corresponding to a period of two years (2015-16) and having 730 observations were used as inputs for the model development. The results of this study signified that the MLFFNN model demonstrates a greater ability to simulate PM_{10} concentrations. A comparative analysis of these models showed that the MLFFNN model presented better forecasting accuracy with R^2= 0.873, RMSE = 51.125, IA = 0.962 during the testing phase compared to its competitive counterparts. Overall, based on the results presented in this work, it may be concluded that the ANN model proves to be an accurate and feasible technique for dealing with non-linear forecasting problems like particulate pollution.

References

[1] Masood, A., Ahmad, K.: A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. Procedia Computer Science. Vol. 167, 2101-2110 (2020).

[2] Rahimi, A.: Short-term prediction of NO2 and NO x concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Ecological Processes. Vol. *6*(1), 1-9 (2017).

[3] Masood, A., Kafeel, A., Shamshad, A.: Urban roadside monitoring, modeling and mapping of air pollution. Applied Journal of Environmental Engineering Science. Vol. *3*(2), 3-2 (2017).

[4] Gualtieri, G., Carotenuto, F., Finardi, S., Tartaglia, M., Toscano, P., Gioli, B.: Forecasting PM10 hourly concentrations in northern Italy: insights on models performance and PM10 drivers through self-organizing maps. Atmospheric Pollution Research. 9(6), 1204-1213 (2018).

[5] Bozdağ, A., Dokuz, Y., Gökçek, Ö. B.: Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. Environmental Pollution. Vol. 263, 114635 (2020).

[6] Akhtar, A., Masood, S., Gupta, C., Masood, A.: Prediction and analysis of pollution levels in Delhi using multilayer perceptron. Data engineering and intelligent computing, Springer, Singapore. 563-572 (2018).

[7] Li, Y., Tao, Y.: PM10 Concentration Forecast Based on Wavelet Support Vector Machine. IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC). 383-386, Shanghai, China (2017).

[8] Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L.. Yan, M.: Using support vector regression to predict PM10 and PM2.5. IOP conference series: earth and environmental science. 35th International symposium on remote sensing of Environment (ISRSE35). 22-26,Beijing, China (2014).

[9] Balha, A., Vishwakarma, B. D., Pandey, S., Singh, C. K.: Predicting impact of urbanization on water resources in megacity Delhi. Remote Sensing Applications: Society and Environment. Vol. 20, 100361 (2020).

[10] Masood, A., Ahmad, K.: A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. Journal of Cleaner Production. Vol. 322, 129072 (2021).

[11] Wang, W., Men, C., Lu, W.: Online prediction model based on support vector machine. Neurocomputing. Vol. 71(4-6), 550-558 (2008).

[12] Alomar MK, Khaleel F, Aljumaily MM, Masood A, Razali SF, AlSaadi MA, Al-Ansari N, Hameed MM. Data-driven models for atmospheric air temperature forecasting at a continental climate region. Plos one. 17(11):e0277079 (2022).