Hindawi
Adsorption Science & Technology
Volume 2022, Article ID 5086622, 15 pages
https://doi.org/10.1155/2022/5086622

Research Article

Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning

D. Kothandaraman,[1] **N. Praveena**,[2] **K. Varadarajkumar**,[3] **B. Madhav Rao**,[4] **Dharmesh Dhabliya**,[5] **Shivaprasad Satla**,[6] **and Worku Abera**[7]

[1]*School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India*
[2]*Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India*
[3]*Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, 500043 Telangana, India*
[4]*Department of Computer Science and Engineering, SIR C R Reddy College of Engineering, Eluru, India*
[5]*Department of Computer Engineering, Vishwakarma Institute of Information Technology, India*
[6]*Department of Computer Science and Engineering, Malla Reddy Engineering College, Secunderabad, 500100 Telangana, India*
[7]*Department of Food Process Engineering, College of Engineering and Technology, Wolkite University, Wolkite, Ethiopia*

Correspondence should be addressed to D. Kothandaraman; kothanda_raman_d@srecwarangal.ac.in, N. Praveena; praveena.4u@gmail.com, and Worku Abera; worku.abera@wku.edu.et

Received 28 March 2022; Revised 24 April 2022; Accepted 6 May 2022; Published 26 June 2022

Academic Editor: Lakshmipathy R

Air pollution consists of harmful gases and fine Particulate Matter ($PM_{2.5}$) which affect the quality of air. This has not only become the key issues in scientific research but also turned to be an important social issues of the public's life. Therefore, many experts and scholars at different R&Ds, universities, and abroad are involved in lot of research on $PM_{2.5}$ pollutant predictions. In this scenario, the authors proposed various machine learning models such as linear regression, random forest, KNN, ridge and lasso, XGBoost, and AdaBoost models to predict $PM_{2.5}$ pollutants in polluted cities. This experiment is carried out using Jupyter Notebook in Python 3.7.3. From the results with respect to MAE, MAPE, and RMSE metrics, among the models, XGBoost, AdaBoost, random forest, and KNN models (8.27, 0.40, and 13.85; 9.23, 0.45, and 10.59; 39.84, 1.94, and 54.59; and 49.13, 2.40, and 69.92, respectively) are observed to be more reliable models. The $PM_{2.5}$ pollutant concentration (PC_{low}-PC_{high}) range observed for these models is 0-18.583 $\mu g/m^3$, 18.583-25.023 $\mu g/m^3$, 25.023-28.234$\mu g/m^3$, and 28.234-49.032 $\mu g/m^3$, respectively, so these models can both predict the $PM_{2.5}$ pollutant and can forecast the air quality levels in a better way. On comparison between various existing models and proposed models, it was observed that the proposed models can predict the $PM_{2.5}$ pollutant with a better performance with a reduced error rate than the existing models.

1. Introduction

Nowadays, accurate air pollution prediction and forecast become a challenging and significant task due to increased air pollution which acts as a fundamental problem in many parts of the world. Generally, the pollution is divided into two types: (1) natural pollution because of volcanic eruptions and forest fires resulting in emission of SO_2, CO_2, CO, NO_2, and sulfate as air pollutants and (2) man-made pollution because of some human activities such as burning of oils, discharges from industrial production processes, and transportation emissions that have $PM_{2.5}$ as its major air pollutant [1] which has received much attention due to their destructive effects on human health, other kinds of creatures, and environment [2]. Various studies testify that air pollution leads to respiratory and cardiovascular disease leading to death of animals and plants, acid rain, climate change, global warming, etc. thus making economic loses and the human life of a society difficult to survive in the world [3]. Regarding the effects of $PM_{2.5}$ investigated over the last 25

years using the comparative analysis of ML techniques, Ameer et al. [4] have estimated that approximately 4.2 million people have died due to long-term exposure of $PM_{2.5}$ in the atmosphere, while an additional 250,000 deaths have occurred due to ozone exposure [1]. In worldwide rankings of mortality risk factors, $PM_{2.5}$ was ranked as 5^{th} and accounted for 7.6% of total deaths all over the world. From 1990 to 2015, the number of deaths due to air pollution has increased, especially in China and India with more than 20% of 1.1 million deaths worldwide attributed to respiratory diseases [5]. Hence, worldwide, huge number of research has been carried out on topics like air pollution levels and air quality forecasts to control air pollution more effectively. Extensive research specifies that air pollution forecasting approaches can be imprecisely divided into three traditional classes: (1) statistical forecasting methods, (2) artificial intelligence methods [6], and (3) numerical forecasting methods [4].

$PM_{2.5}$ pollutants are fine particles that are made up of a combination of gases and particles which are hazardous when released into the atmosphere [2]. These pollutants are mainly responsible for causing human respiratory diseases in one way or another, and when severe, it can further lead to the pandemic COVID-19 [7, 8] resulting in increased death level. The present models focus on only the $PM_{2.5}$ pollutant because from the survey, it is obvious that $PM_{2.5}$ causes high issues in human being compared to other pollutants, and it is the one that creates other pollutants. Statistical analysis for $PM_{2.5}$ pollutant prediction is done using historical meteorological datasets. However, existing models are constrained to utilize some basic standard classification techniques; few models are used for forecasting, yet the results showed poor error rate performance.

In this proposed approach, six different machine learning models [9] which include regression models such as linear regression model (LR), random forest model (RF), KNN model, ridge and lasso model (RL), XGBoost model (Xgb), and AdaBoost model (Adab) have been implemented to predict the $PM_{2.5}$ pollutant using meteorological and $PM_{2.5}$ pollutant historical datasets that are downloaded from 1^{st} Jan 2014 to 1^{st} Dec 2019. These data have been monitored continuously for 24 h with a time period of an hour using the following meteorological features such as temperature (T in °C), minimum temperature (Tm in °C), maximum temperature (TM in °C), total train/snowmelt (PP in mm), humidity (H in %), wind speed (V in km/h), visibility (VV in km), and maximum sustained wind speed (VM in km/h). Also, the proposed machine learning models have been evaluated using statistical metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R^2. Results show the achievement of better performance with decreased error rate when compared to traditional prediction models. This paper has been organized as follows. Section 2 discusses the related works, Section 3 introduces machine learning models for predicting $PM_{2.5}$ and forecasting air quality, Section 4 presents model results and analysis, and Section 5 concludes the paper.

2. Related Works

In the recent years, many prediction models were developed for solving $PM_{2.5}$ pollutant issues. Zhang et al. [10] used a light gradient boosting decision tree model to process the high dimensional data to predict $PM_{2.5}$ within 24 h based on the historical datasets and predictive datasets and then compared it with various models using evaluated metrics such as Symmetric Mean Absolute Percentage Error (SMAPE), MAE, and RMSE.

[11]) reported a spatial ensemble model to predict $PM_{2.5}$ for the Beijing railway station, but it is not reliable for other locations. Kim et al. [12] reported effects of the indoor $PM_{2.5}$ pollutant, i.e., asthma attacks in children, based on peak breath flow rates using deep learning methods' rule for predicting respiratory disease risk. Caraka et al. [13] reported prediction of $PM_{2.5}$ using the Markov chain stochastic process and VAR-NN-PSO. Using the $PM_{2.5}$ feature of higher probability to pass through the lower respiratory tract, its range can be categorized into no risk (1-30), medium risk (30-48), and moderate risk (>49) in Chaozhou and Pingtung for the datasets obtained from Jan 2014 to May 2019.

Beelen et al. [14] established a multicenter cohort study in Europe to study the positive correlation between $PM_{2.5}$ concentration and heart disease mortality during a long time exposure period to $PM_{2.5}$ [1, 15]. Tiwari et al. [16] considered an XGBoost model on a building that utilizes atmospheric data of Velachery and database of the central control room collected from a commercial station in Tamil Nadu for predicting air quality management. This model also considers the highly unstable meteorological parameters such as relative humidity, wind speed pressure, temperature, and wind direction of the geographic region.

Bing et al. [17] and Pasha et al. [18] reported a new model for forecasting air quality index in China using support vector regression, and the results showed a decrease in MAPE when there is a robust interaction. Lin et al. [19] proposed a novel system based on a cloud model granulation algorithm for air quality forecasting through data exploration in three monitoring localities in Wuhan City with high accuracy.

Xiao et al. [20] identified a novel hybrid model by combining air mass trajectory analysis and wavelet transformation to improve the artificial neural network for forecasting the daily average concentrations of $PM_{2.5}$. Soh et al. [21] recognized the data-driven model ST-DNN to predict $PM_{2.5}$ time series data and other pollutants in seven locations for only 48 h using real-time Taiwan and Beijing datasets. Heni et al. [22] and Li et al. [23] used multivariate multistep time series prediction with random forest models to improve the performance and to reduce the time complexity of the air pollutant prediction models.

Regarding the effects of $PM_{2.5}$ over the last 25 years, Ameer et al. [4] discussed comparison among various regression techniques such as decision tree, random forest gradient boosting, and ANN [24] multilayer perceptron regression with respect to error rate and processing time for forecasting air quality in smart cities. In [25], a deep

learning model consisting of a recurrent neural network with long-short-term memory is used to predict local 8 h averaged surface ozone concentrations for 72 h based on hourly air quality and also used meteorological data measurements as a tool to forecast air pollution values with decreased error rate.

Deters et al. [26] and Sallauddin et al. [27] considered a machine learning method based on six years of meteorological and pollution data analyses in Belisario and Cotocollao to predict the concentrations of $PM_{2.5}$ using wind direction, its speed, and rainfall levels and then compared it to various ML algorithms such as BT, L-SVM [28], and ANN regression models. The high correlation between estimated and real data for a time series analysis during the wet season confirms a better prediction of $PM_{2.5}$ when the climatic conditions are getting more dangerous or there are high-level conditions of precipitation or strong winds. Zhao et al. [29] and Ni et al. [30] introduced a multivariate linear regression model to achieve short period prediction of $PM_{2.5}$, and the parameters included are data on aerosol optical depth obtained through remote sensing and meteorological factors from ground monitoring temperature, relative humidity, and wind velocity.

The present paper investigated different prediction models related to the $PM_{2.5}$ pollutant which are statistically analyzed. All existing approaches have mostly implemented so many prediction models such as NN [31], L-SVM (Linear Support Vector Machines), BT (Boosted Trees), CGM, and NN (neural network) [26]; deep learning consisting of a recurrent neural network with long-short-term memory [25]; decision tree, gradient boosting, random forest, ANN multilayer perceptron regression [4, 15], and multivariate linear regression model [29]; AdaBoost, XGBoost, GBDT, LightGBM, and DNN [10]; and predictive data feature exploration-based air quality prediction approach. In the proposed $PM_{2.5}$ pollutant prediction, six different machine learning models have been used, and the results were compared with those of the above-mentioned existing models.

3. Machine Learning Models for Predicting $PM_{2.5}$ and Forecasting Air Quality

In these proposed machine learning models to predict the $PM_{2.5}$ pollutant, meteorological datasets were collected for 24 hours of the day from 1st Jan 2014 to 31st Dec 2019. The main objective of the proposed models is to apply various machine learning models to predict the $PM_{2.5}$ pollutant range and its level of air quality in any polluted cities. Though not more than three or four techniques in existing models have predicted the $PM_{2.5}$ pollutant [4, 10, 25, 26, 29], here six different machine learning models such as LR, RF, KNN, RL, Xgb, and Adab models were implemented to predict the $PM_{2.5}$ pollutant with different hyperparameter tuning to increase the accuracy with reduced error rate. The present models were initially preprocessed with various meteorological and $PM_{2.5}$ pollutant datasets. During the model creation, the datasets were split as training sets of 70% and testing sets of 30%. When compared with existing models' performance, machine learning models achieve a better performance with minimum error rates.

3.1. Architecture for Machine Learning Models. Figure 1 represents the machine learning model for predicting the $PM_{2.5}$ pollutant in the affected cities. Figure 1 consists of three layers: (1) the first layer is an input layer which has the $PM_{2.5}$ pollutant and meteorological datasets for preprocessing and feature extraction, (2) the second layer contains six different machine learning models which are used to predict the $PM_{2.5}$ pollutant along with its working principle, and (3) the output layer consists of certain steps like training models and testing models and then the final step to predict the $PM_{2.5}$ pollutant range and to forecast its air quality level among the various categories.

3.2. Flowchart Representation. Figure 2 represents the flowchart for predicting the $PM_{2.5}$ pollutant with the assistance of machine learning models. Here, the prediction process was started using real-time meteorological and its $PM_{2.5}$ pollutant historical datasets. Then, the data were preprocessed and then feature extracted to remove unwanted data to obtain cleaned datasets for training models. Then, six different models were integrated for training and testing with real-time data. Then, finally check the prediction of the $PM_{2.5}$ pollutant range and then proceed further to forecast whether air quality levels are good or satisfied in order to deploy the models; otherwise, the models and datasets should be enhanced again.

3.3. Implementation of $PM_{2.5}$ Pollutant Prediction Models. For all the models, performances of training and testing models were evaluated using metrics such as R^2 (equation (1)), Mean Absolute Error (MAE) (equation (2)), Mean Absolute Percentage Error (MAPE) (equation (3)), Mean Square Error (MSE) (equation (4)), and Root Mean Square Error (RMSE) (equation (5)), and similarly the $PM_{2.5}$ pollutant was also evaluated.

$$R^2 = \left(\frac{1/m\sum_{i=1}^{m}\left(x_{\text{observed}}(i) - \bar{x}_{\text{observed}}\right)\left(x_{\text{predicted}}(i)\bar{x}_{\text{predicted}}\right)}{\sqrt{1/m\sum_{i=1}^{m}\left(x_{\text{observed}}(i) - \bar{x}_{\text{observed}}\right)^2}\sqrt{1/m\sum_{i=1}^{m}\left(x_{\text{predicted}}(i)\bar{x}_{\text{predicted}}\right)^2}} \right)^2, \tag{1}$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|x_{\text{observed}}(i) - x_{\text{predicted}}(i)\right|, \tag{2}$$

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\frac{x_{\text{observed}}(i) - x_{\text{predicted}}(i)}{x_{\text{observed}}(i)} \times 100, \tag{3}$$

$$\text{RME} = \frac{1}{m}\sum_{i=1}^{m}\left(x_{\text{observed}}(i) - x_{\text{predicted}}(i)\right), \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_{\text{observed}}(i) - x_{\text{predicted}}(i)\right)^2}. \tag{5}$$

3.4. Model Deployment for Forecasting Air Quality. To evaluate the $PM_{2.5}$ pollutant concentration for forecasting air

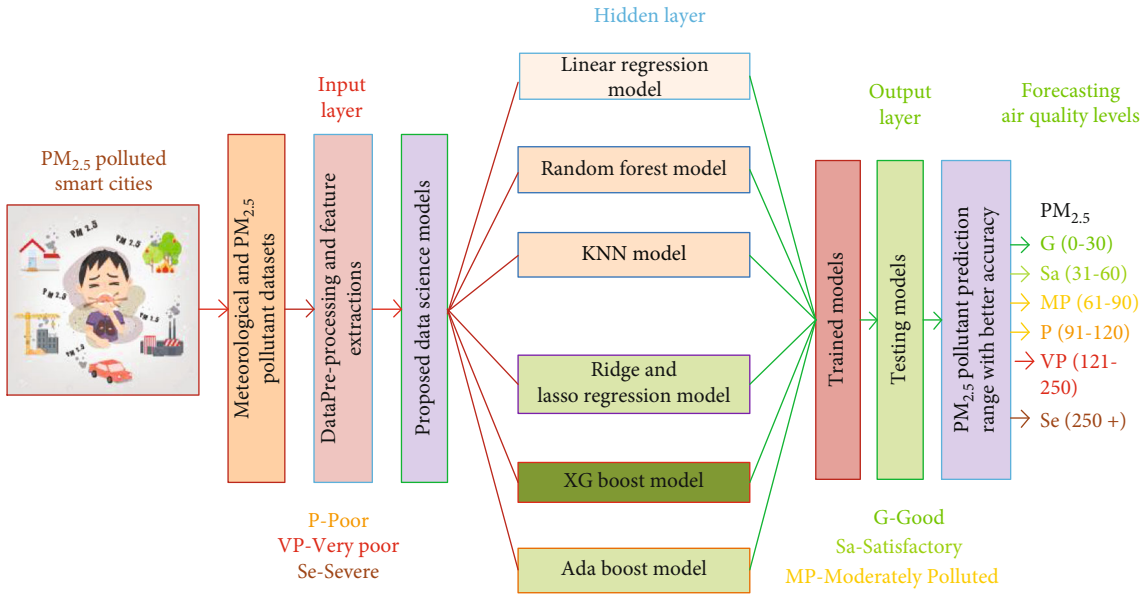


Figure 1: Machine learning model for $PM_{2.5}$ pollutant prediction and air quality forecasting.

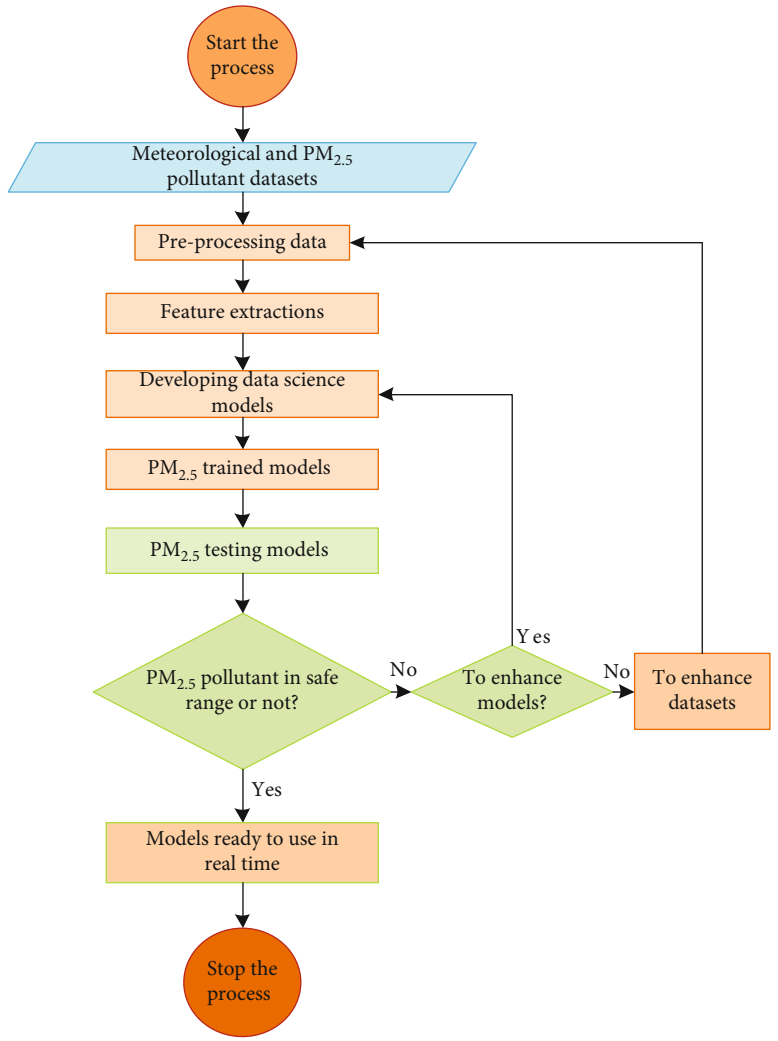
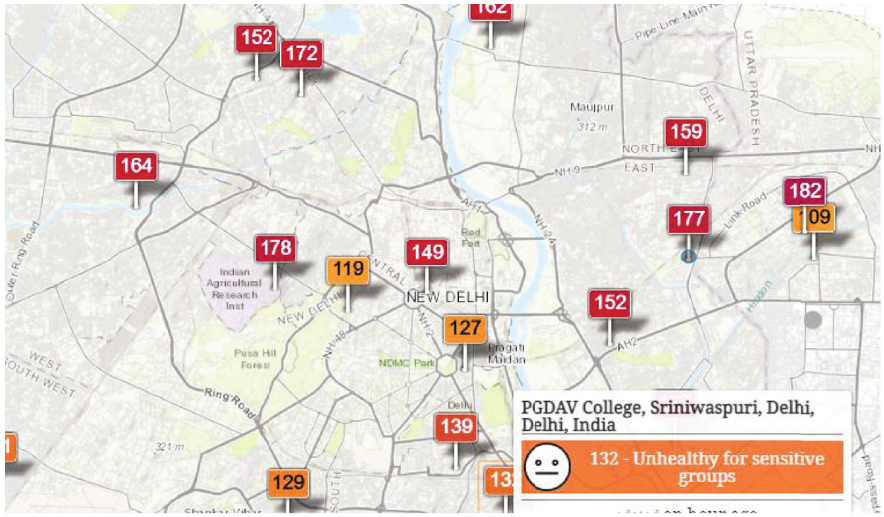
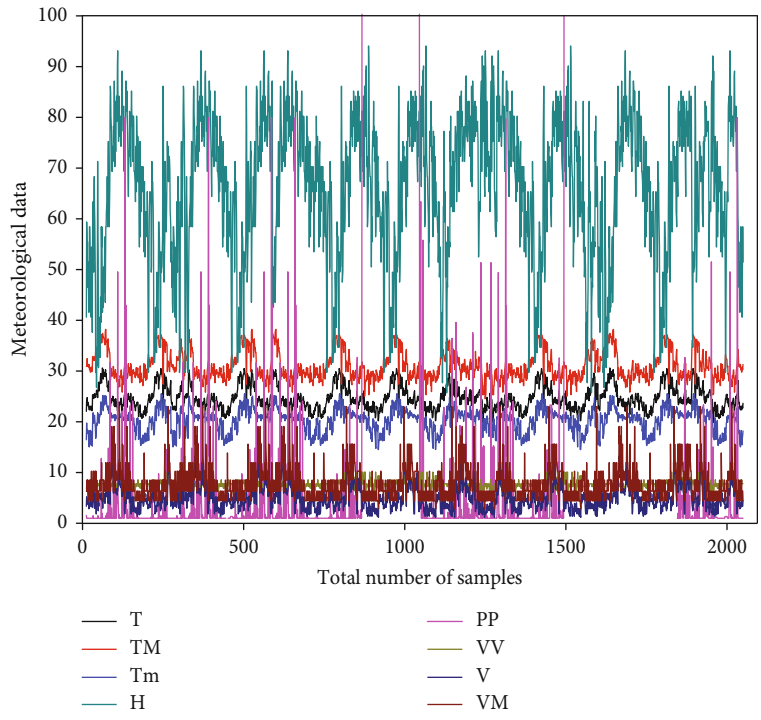


Figure 2: Flowchart representations for predicting $PM_{2.5}$ and air quality forecasting.



(a)



(b)

Figure 3: (a) Sample study area map for experimental purpose. (b) Variation among meteorological data.

quality level, equation (6) is used [4].

$$\mathrm{AQR} = \frac{\mathrm{AQR}_{\mathrm{high}-}\mathrm{AQR}_{\mathrm{low}}}{\mathrm{PC}_{\mathrm{high}-}\mathrm{PC}_{\mathrm{low}}}(\mathrm{PC} - \mathrm{PC}_{\mathrm{low}}) + \mathrm{AQR}_{\mathrm{low}}, \quad (6)$$

where AQR is the air quality range, PC is the pollutant concentration, $\mathrm{PC}_{\mathrm{low}}$ is the concentration break point $\leq$ PC, $\mathrm{PC}_{\mathrm{high}}$ is the concentration break point $\geq$ PC, $\mathrm{AQR}_{\mathrm{low}}$ is the AQR break point corresponding to $\mathrm{PC}_{\mathrm{low}}$, and $\mathrm{AQR}_{\mathrm{high}}$ is the AQR break point corresponding to $\mathrm{PC}_{\mathrm{high}}$.

4. Results and Analysis

4.1. Experiment Setup. This experiment was carried out using Jupyter Notebook and a computing system which has a processor speed of Intel(R) Core(TM) i5-2450M CPU@2.50 GHz and RAM of 12 GB. The proposed machine learning models are exposed to data cleaning and feature extraction for training and testing models using Python 3.7.3.

4.2. Details about Meteorological and $PM_{2.5}$ Datasets. Meteorological and $PM_{2.5}$ historical datasets were collected (anand-vihar, delhi-air-quality) from the Delhi Pollution Control Committee (http://aqicn.org) for experimental

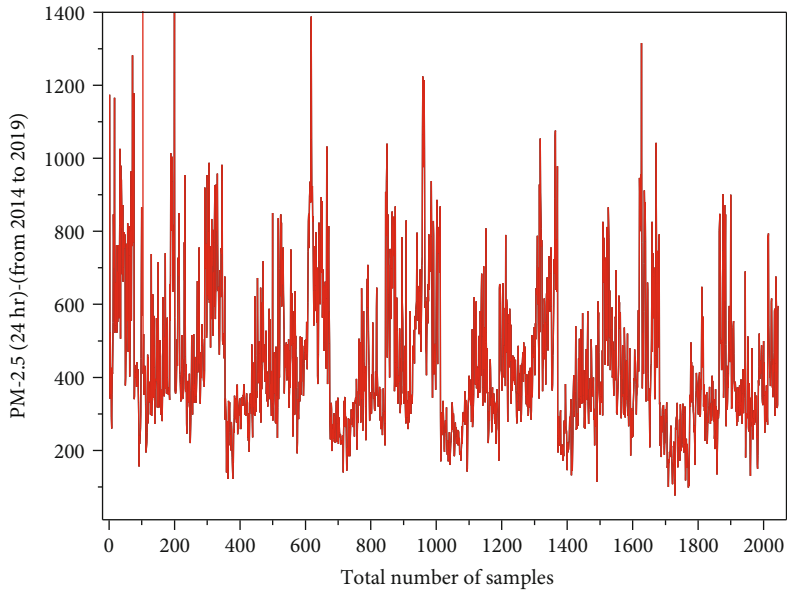


Figure 4: Overall $PM_{2.5}$ variation with respect to time series.

Table 1: Meteorological and $PM_{2.5}$ dataset analysis.

Observed datasets (years)	Samples obtained (from and to, months)	Samples not obtained (from and to, months)	Total samples obtained 24 h per day	Mean of $PM_{2.5}$ per year in $\mu g/m^3$	SD of $PM_{2.5}$ per year
2014	01-01-2014; 1:00 AM and 01-12-2014; 24:00 PM	Nil	6360	258	119.3437
2015	01-01-2015; 1:00 AM and 01-12-2015; 24:00 PM	Nil	7584	228	90.30255
2016	01-01-2016; 1:00 AM and 01-12-2016; 24:00 PM	Nil	8136	229	107.5823
2017	01-01-2017; 1:00 AM and 01-12-2017; 24:00 PM	Nil	8616	221	94.87083
2018	Data of all months are available except for the 7^{th} month	01-07-2018; 1:00 AM and 31-07-2018; 24 PM	7536	215	88.63759
2019	01-01-2016; 1:00 AM and 01-12-2016; 24:00 PM	Nil	8664	261	92.81299

purpose only as shown in Figures 3(a), 3(b), and 4. These datasets include various climatic conditions based on T (°C), Tm (°C), TM (°C), PP (mm), H (%), V (km/h), VV (km), and VM (km/h) (Figure 3). The $PM_{2.5}$ pollutant is shown in Figure 4. The data was obtained for 24 hours with a time period of an hour from 1^{st} Jan 2014 (1:00 AM) to 31^{st} Dec 2019 (24:00 PM), and data sources are stored in CSV file format. Average $PM_{2.5}$ samples are stored in the file $2044 * 24 = 49056$. For a year, 8176 samples (approximately) are observed, and for an hour, a maximum of two samples (approximately) is appended depending on climatic conditions. The remaining data are considered to be null values or improper data which are removed by using data preprocessing techniques. Further information about datasets has been presented in Table 1.

Using datasets in Table 1, variation of $PM_{2.5}i^{th}$ daily concentration was measured in terms of statistical features such as mean and standard deviation as shown in Figure 5, where "N" is the number of samples and "i" is a single sample in the i^{th} $PM_{2.5}$ range.

4.3. Statistical Information about Datasets. Table 2 represents the statistical analysis of both meteorological and $PM_{2.5}$ datasets that are considered with various features such as T, TM, Tm, H, PP, VV, V, VM, and $PM_{2.5}$. Datasets are evaluated using statistical features such as the count, mean, SD, MIN, 25%, 50%, 75%, and MAX. The overall $PM_{2.5}$ varies from 78 to 824 ($\mu g/m^2$) for 2014, from 61 to 494 ($\mu g/m^2$) for 2015, from 70 to 694 ($\mu g/m^2$) for 2016, from 71 to 612 ($\mu g/m^2$) for 2017, from 57 to 538 ($\mu g/m^2$) for 2018, from 38 to 658 ($\mu g/m^2$) for 2019, and from 38 to 824 ($\mu g/m^2$) for 2014-2019. Based on statistics, the maximum $PM_{2.5}$ pollutant range is exceeding the default air quality forecasting limit levels, and this is indicated as "severe" in Table 2. So in this work, six different machine learning models were

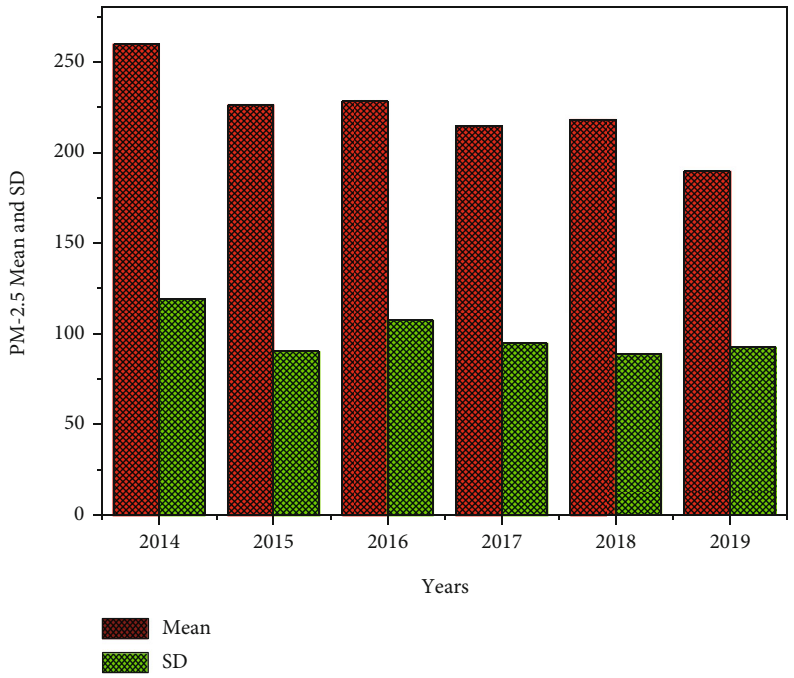


Figure 5: Years vs. $PM_{2.5}$ mean and SD.

Table 2: Statistical analysis of both meteorological and PM2.5 datasets (2014 to 2019).

Statistical features	*T*	TM	Tm	*H*	PP	VV	*V*	VM	$PM_{2.5}$
Count	2044	2044	2044	2044	2038	2044	2044	2044	2044
Mean	23.98728	30.4362	19.60274	66.01761	3.085113	6.75093	4.114335	7.037818	219.8787
SD	2.318939	2.879207	2.268557	14.38204	10.13789	0.637014	2.324433	3.311582	100.0151
MIN	19.1	23.8	13.7	25	0	4	0.2	1.9	38
25%	22.43359	28.50713	18.08281	56.38164	-3.70727	6.32413	2.556964	4.819058	152.8685
50%	22.48728	28.9362	18.10274	64.51761	1.585113	5.25093	2.614335	5.537818	218.3787
75%	25.54097	32.36527	21.12267	75.65358	9.877499	7.177729	5.671705	9.256578	286.8888
MAX	29.9	37.6	24.8	94	132.33	9.2	12.4	22.2	824

applied to minimize the $PM_{2.5}$ pollutant range and are observed to predict air quality levels in a better way.

4.4. Feature Extraction. Figure 6 represents the pair plot of feature extraction for meteorological and $PM_{2.5}$ pollutant datasets which clear the null values using preprocessing mean and SD. *x*- and *y*-axes represent eight different meteorological features such as *T*, TM, Tm, *H*, PP, VV, *V*, and VM and the $PM_{2.5}$ pollutant. Figure 7 represents the feature extraction using regression.

4.4.1. Heat Map for Correlating Coefficient between Features. Figure 8 represents the heat map to find the cross-correlation between different meteorological and $PM_{2.5}$ pollutant features; if values come nearby 1, then it shows a strong positive correlation; if values come nearby -1, then it shows a negative correlation; and if values come nearby 0 meaning neutral, it is an independent correlation. Thus, the heat map is used to remove the unwanted features in $PM_{2.5}$ pollutant datasets (i.e., strongly correlated).

4.4.2. Normal Distribution Curve Fitting (NDCF) for $PM_{2.5}$. Figure 9 represents the curve fitting using normal distribution for $PM_{2.5}$ pollutant datasets. Perfect fit range for the normal distribution curve is observed to be 0.0085, and this value can be satisfactorily considered near to 0.01. The *x*-axis shows the correlation coefficient features, and the *y*-axis shows the dependent feature of $PM_{2.5}$.

4.5. Comparing NDCF among Machine Learning Models. Figure 10(a) represents the LR model curve fitting showing a value of about 0.0085 with the correlation coefficient in the *x*-axis and the dependent feature of $PM_{2.5}$ in the *y*-axis. Figure 10(b) represents the KNN model without hyperparameter tuning which shows overfit of the curve while the curve fitting value is 0.0095 for the KNN model using hyperparameter tuning and is shown in Figure 10(c). Figure 10(d) represents RF models without hyperparameter tuning which shows overfit of the curve while the curve fitting value is 0.0094 for the RF model using hyperparameter tuning and is shown in Figure 10(e). Figure 10(f) represents RL models

Figure 6: Feature extraction of $PM_{2.5}$.

without hyperparameter tuning which otherwise represents overfit of the curve while the curve fitting value is 0.0075 for RL models using hyperparameter tuning and is shown in Figure 10(g). Figure 10(h) represents Xgb models without hyperparameter tuning which otherwise represents overfit of the curve while the curve fitting value is 0.0086 for Xgb models using hyperparameter tuning and is shown in Figure 10(i). Figure 10(j) represents the curve fitting for the Adab model with tuning which is observed to have 0.0095 which is a perfect fit model.

4.6. Performance Measures. Table 3 represents the performance results of different machine learning models which are used to predict the $PM_{2.5}$ pollutant. The results of LR, RF, KNN, RL, Xgb, and Adab for various performance metrics are as follows: for MAE, their values are 55.12, 39.84, 49.13, 55.12, 8.27, and 9.23, respectively; for MAPE, their values are 2.69, 1.94, 2.40, 2.69, 0.40, and 0.45, respectively; for MSE, their values are 5157.17, 2980.71, 4889.74, 5157.17, 192.08, and 112.15, respectively; and for RMSE, their values are 71.81, 54.59, 69.92, 71.81, 13.85, and 10.59, respectively. From the above results, Xgb, Adab, RF, and KNN models are considered to achieve better performance results in all means and then compared to the other models.

Table 4 represents the correlation coefficient determination in terms of R^2 using LR, RF, KNN, RL, Xgb, and Adab. From Table 4, when the performance results of the training set value is nearer to one, it is considered to be the better

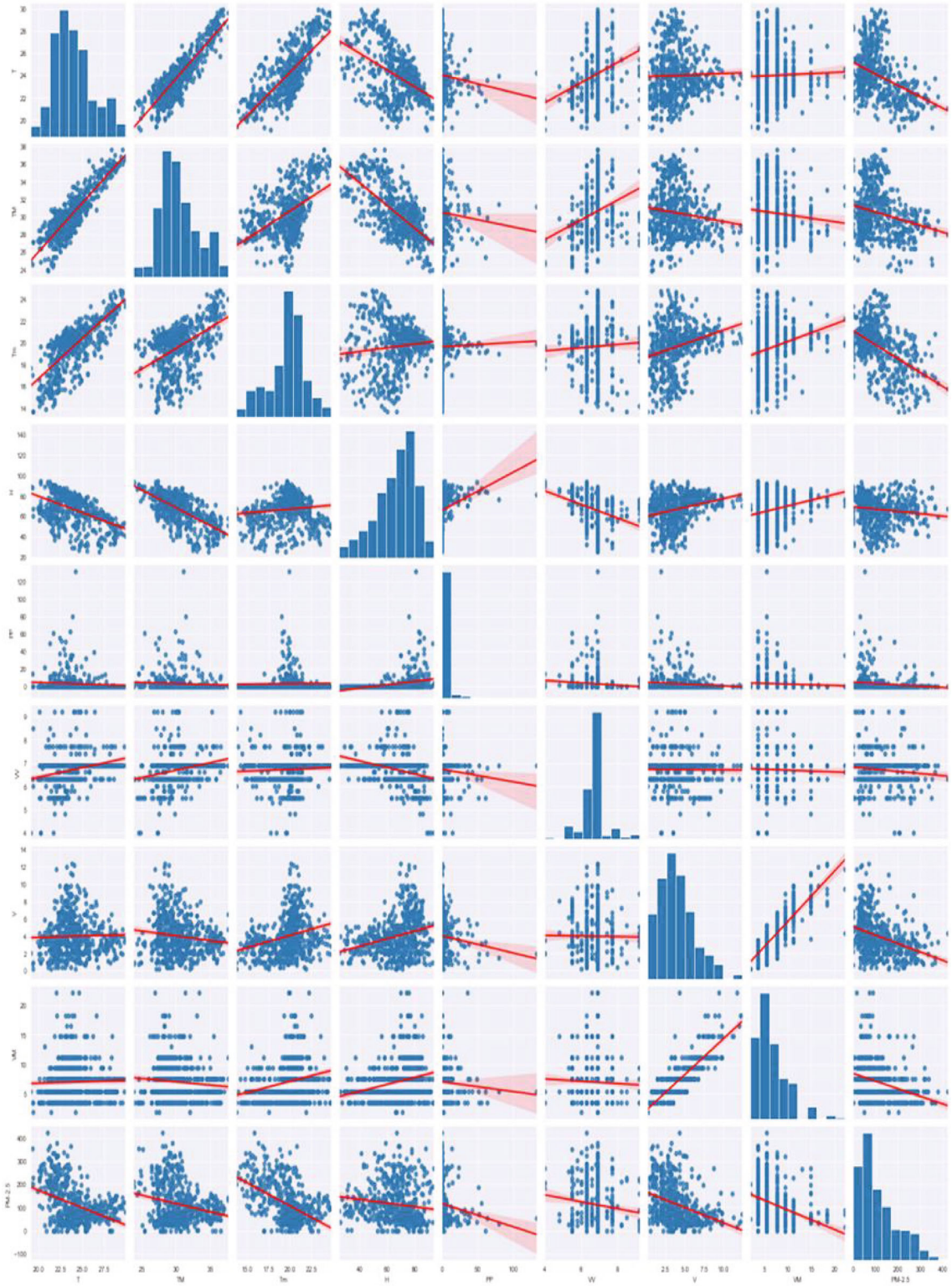


Figure 7: Feature extraction using regression.

performance. So the better performance results are KNN train set and test set values of 1.0 and -0.228, respectively; Xgb train set and test set values of 0.999 and 0.3072, respectively; and RF train set and test set values of 0.904 and 0.382, respectively.

4.7. Comparative Analysis

4.7.1. Comparison in Terms of RMSE and MAE. Among all pollutants, only the $PM_{2.5}$ pollutant is considered in the existing Xgb and Adab models [10] for comparison with

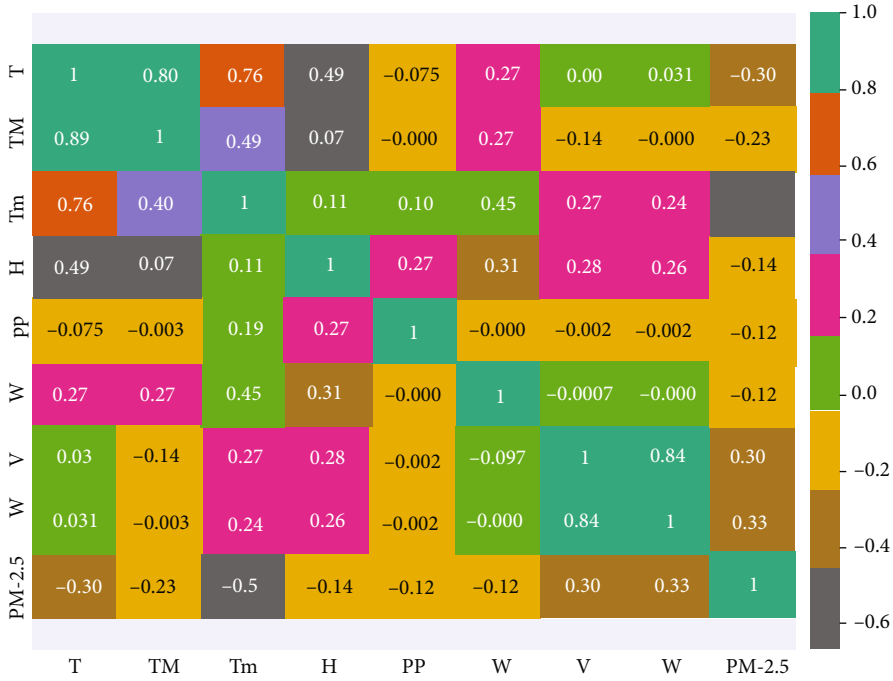


Figure 8: Correlation coefficient matrix of $PM_{2.5}$.

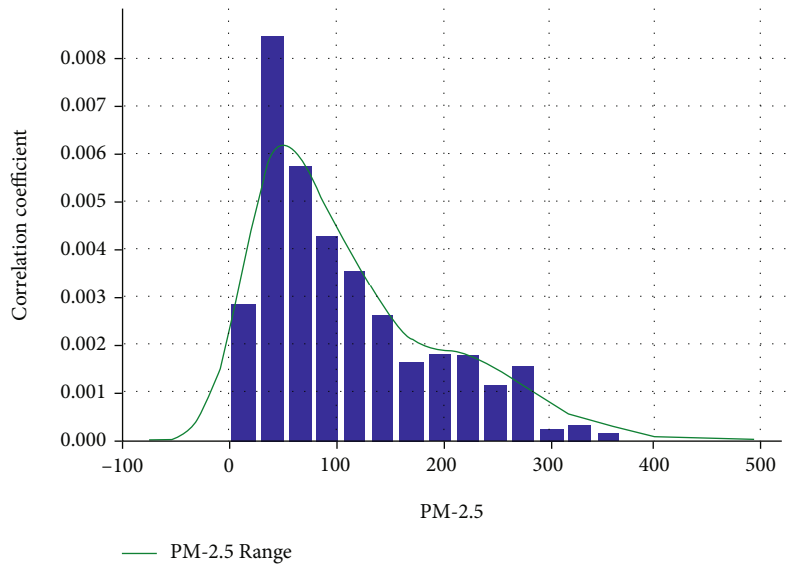


Figure 9: Normal distribution curve fitting for the $PM_{2.5}$ pollutant.

proposed models in terms of performance metrics like RMSE and MAE because other types of models were not reported in the existing work. In the case of the existing work, RMSE for Xgb and Adab is observed to be 38.8253 and 38.825, respectively, while MAE for Xgb and Adab is 27.054 and 32.957, respectively; in the case of proposed models, RMSE for Xgb and Adab is 13.85 and 10.59, respectively, while MAE for Xgb and Adab is 8.27 and 9.23, respectively. On comparing these two data, proposed models represent better results than the existing work, and regarding error rate, the existing model shows increased error rates compared to the proposed model which is represented in Table 5(a).

In the case of the existing work especially that use the trajectory model and trajectory with wavelet model to predict the $PM_{2.5}$ pollutant [20], 2 days for each monitoring station (a, b, c, and d) are considered with RMSE and MAE as evaluating metrics. But for comparison with the present model, only one station with one day is considered because the error rate for the remaining days for other stations is higher than the proposed value. On comparing these two data, proposed models (Xgb and Adab) represent better

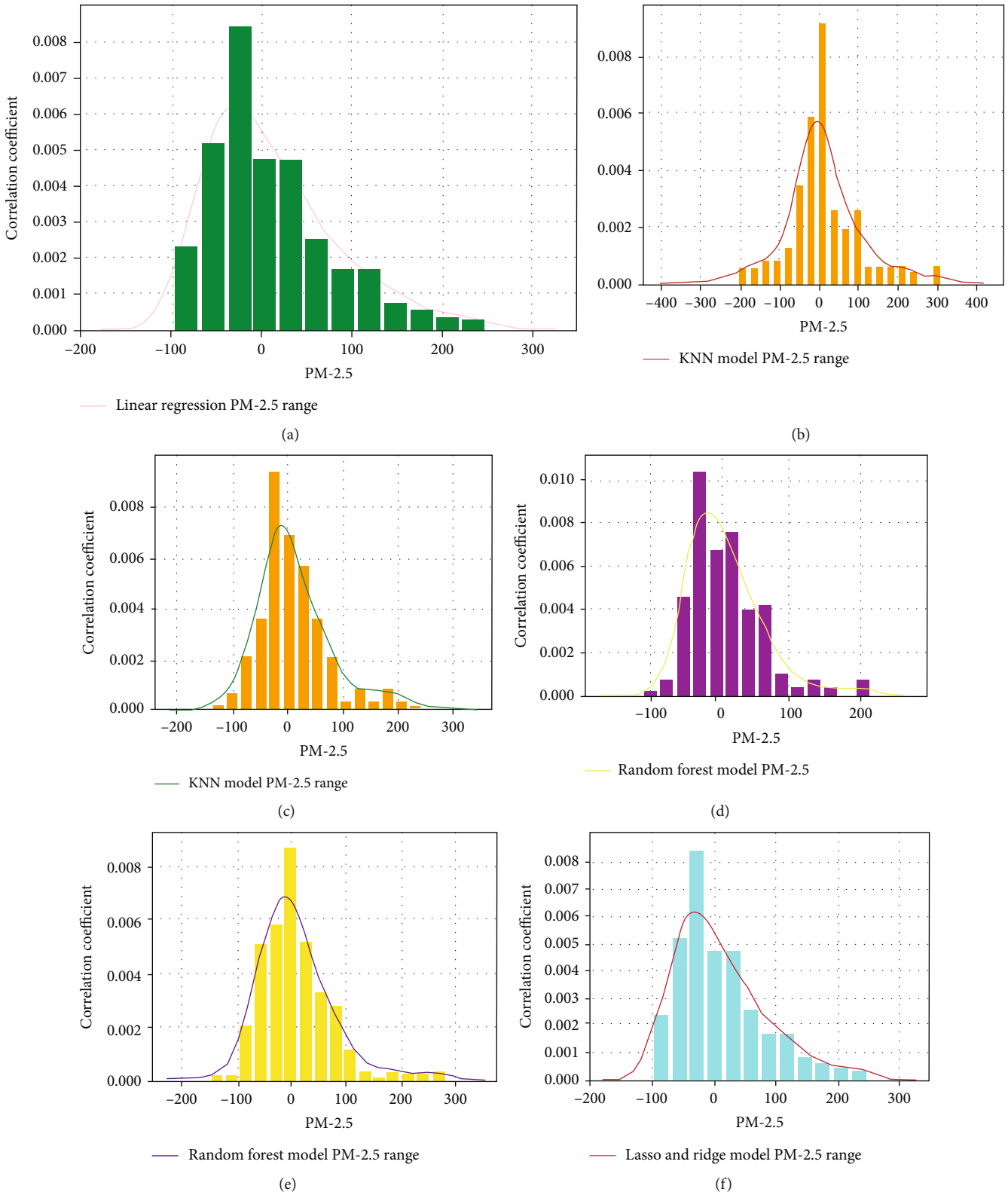


Figure 10: Continued.

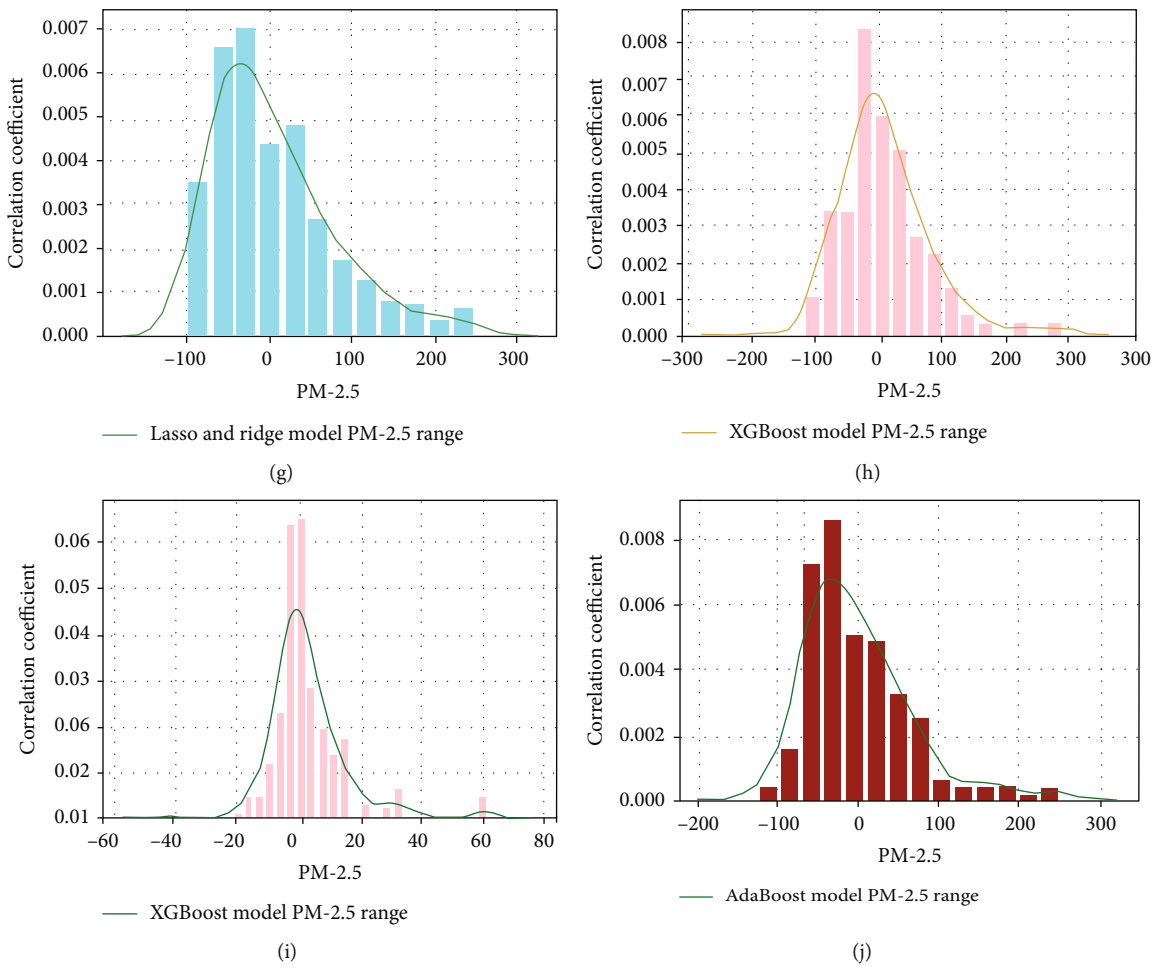


Figure 10: (a) LR model curve fitting. (b) KNN model without hyperparameter tuning. (c) KNN model using hyperparameter tuning. (d) RF models without hyperparameter tuning. (e) RF model using hyperparameter tuning. (f) RL models without hyperparameter tuning. (g) RL models using hyperparameter tuning. (h) Xgb models without hyperparameter tuning. (i) Xgb models using hyperparameter tuning. (j) Curve fitting for the Adab model with tuning.

Table 3: Statistical validation for proposed models using the following metrics.

S. no	Proposed models	MAE	MAPE	MSE	RMSE
1.	LR	55.12	2.69	5157.17	71.81
2.	RF	39.84	1.94	2980.71	54.59
3.	KNN	49.13	2.40	4889.74	69.92
4.	RL	55.12	2.69	5157.17	71.81
5.	Xgb	8.27	0.40	192.08	13.85
6.	Adab	9.23	0.45	112.15	10.59

Table 4: Statistical validation in terms of correlation coefficient R^2.

S. no	Proposed models	R^2 train set	R^2 test set
1.	LR	0.401	0.320
2.	RF	0.904	0.382
3.	KNN	1.0	-0.228
4.	RL	0.4013	0.320
5.	Xgb	0.999	0.3072
6.	Adab	0.6055	0.4290

results than the existing work, and also regarding error rate, the existing model shows increased error rates compared to the proposed model which is represented in Table 5(b).

4.7.2. Comparison in Terms of MAPE. In the case of the existing paper, MAPE values for Linear-Support Vector Machines (L-SVM), Boosted Trees (BT), Convolutional Generalization Model (CGM), and neural networks (NN) are observed to be 41.8, 44.4, 15.0, and 40.7, respectively [26], while in the case of proposed models, MAPE values for LR, RF, KNN, RL, Xgb, and Adab are observed to be 2.69, 1.94, 2.40, 2.69, 0.40, and 0.45, respectively. This result clearly shows that the proposed models represent better MAPE with decreased error rates for all the six models when compared with existing models and is shown in Table 6(a).

The proposed models use 2190 days data for predicting $PM_{2.5}$ with better results while the existing VAR-NN-PSO model [13] shows a MAPE value of 3.57% for 180 days $PM_{2.5}$ data in Pingtung and a MAPE value of 4.87% in Chaozhou. This is shown in Table 6(b).

Table 5

(a) Comparison in terms of RMSE and MAE

Proposed models	Present RMSE	Present MAE	Existing RMSE	Existing MAE
Xgb	13.85	8.27	33.0947	27.054
Adab	10.59	9.23	38.825	32.957

(b) Comparison in terms of RMSE and MAE

Proposed models	Present RMSE	Present MAE	Existing models	Existing RMSE value for 1 day	Existing MAE value for 1 day
Xgb	13.85	8.27	Trajectory	28.98	21.52
Adab	10.59	9.23	Trajectory with wavelet	19.75	11.58

Table 6

(a) Comparison in terms of MAPE

Proposed models	Present MAPE	Existing models	Existing MAPE
LR	2.69	L-SVM	41.8
RF	1.94	BT	44.4
KNN	2.40	CGM	15.0
RL	2.69	NN	40.7
Xgb	0.40		
Adab	0.45		

(b) Comparison in terms of MAPE

Proposed models	Present MAPE	Existing MAPE
LR	2.69	
RF	1.94	3.57
KNN	2.40	
RL	2.69	
Xgb	0.40	4.87
Adab	0.45	

(c) Comparison in terms of MAPE

Proposed models	Present MAPE	Existing model	Existing MAPE
LR	2.69		5.70
RF	1.94	Spatial ensemble model	13.90
KNN	2.40		28.78
RL	2.69		9.80
Xgb	0.40		
Adab	2.55		

In the case of the existing spatial ensemble model [11], one location with 4 quadrants is considered for $PM_{2.5}$ data, and MAPE values obtained for the 1st, 2nd, 3rd, and 4th quarter are 5.7034%, 13.9070%, 28.7859%, and 9.8086%, respectively. But in the case of the proposed models, data from all polluted locations are considered for predicting $PM_{2.5}$, and it is in a better way than the existing models as shown in Table 6(c).

4.8. Deployment of the Models. In proposed models for testing, various meteorological data are randomly selected from datasets like *T* (25.3), TM (31.6), Tm (22.4), *H* (74), PP (0), VV (6.3), *V* (3.9), and VM (9.4) to predict the $PM_{2.5}$ pollutant range. For Xgb, KNN, and Adab, the results obtained are 0-18.583 $\mu g/m^3$, 18.583-25.023 $\mu g/m^3$, and 25.023-28.234 $\mu g/m^3$, respectively, which fall in the category of "good" air quality levels. Similarly, RF of 28.234-49.032 $\mu g/m^3$ and RL

Table 7: Forecasting air quality levels.

S. no	Deployment models	Predicted $PM_{2.5}$ range (PC_{low}-PC_{high})	Default $PM_{2.5}$ range (AQR_{low}-AQR_{high})	Air quality levels	Impact on health
1.	Xgb	0-18.583			
2.	KNN	18.583-25.023	0~30.0	Good	Air is good for health
3.	Adab	25.023-28.234			
4.	RF	28.234-49.032	31.0~60.0	Satisfactory	Air is acceptable
5.	RL	49.032-51.334			
6.	LR	51.334-65.345	61.0~90.0	Moderately polluted	Irritation symptoms occur
			91.0~120.0	Poor	
	No models were found	Not in predicted range	121.0~250.0	Very poor	Cause respiratory diseases
			250+	Severe	

of 49.032-51.334 $\mu g/m^3$ value fall in the category of "satisfactory" air quality levels. In the case of "moderately pollutant," air quality levels of 51.334-65.345 $\mu g/m^3$ in LR agree to this. In the remaining default $PM_{2.5}$ pollutant ranges like 91-120, 121-250, and 250+, none of the proposed machine learning models is forecasting air quality levels. Comparing the models regarding the category of "good" air quality levels, Xgb comes first followed by KNN and then Adab, which is shown in Table 7.

5. Conclusions

Air pollution is harmful to both the environment and human existence. When some substances in the atmosphere exceed a certain concentration, it results in air pollution. One of the effective pollution control measures is to predict $PM_{2.5}$ and to forecast the air quality. In the proposed models, the $PM_{2.5}$ pollutant is predicted using meteorological datasets and six different models (LR, RF, KNN, RL, Xgb, and Adab models) are used for forecasting air quality levels. The results were evaluated using statistical metrics such as MAE, MAPE, MSE, RMSE, and R^2. The better performance results for correlation coefficient determination in terms of R^2 are KNN train set and test set values of 1.0 and -0.228, respectively; Xgb train set and test set values of 0.999 and 0.3072, respectively; and RF train set and test set values of 0.904 and 0.382, respectively. Among those proposed models from the results with respect to MAE, MAPE, and RMSE metrics (8.27, 0.40, and 13.85; 9.23, 0.45, and 10.59; 39.84, 1.94, and 54.59; and 49.13, 2.40, and 69.92, respectively, for Xgb, Adab, RF, and KNN), it could be obvious that Xgb, Adab, KNN, and RF are reliable models when compared to the existing models. The $PM_{2.5}$ pollutant (PC_{low}-PC_{high}) range observed for these models is 0-18.583 $\mu g/m^3$, 25.023-28.234 $\mu g/m^3$, 18.583-25.023 $\mu g/m^3$, and 28.234-49.032 $\mu g/m^3$, respectively. It can be concluded that by using the proposed models, the $PM_{2.5}$ pollutant can be predicted; thereby, it can forecast the air quality levels also in a better way. Finally, it is obvious that this research is very useful for the society since forecasting air quality levels acts as an important tool to prevent air pollution by taking necessary actions and steps to control the pollutants.

Data Availability

The data used to support the findings of this study are included within the article. Should further data or information be required, these are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest. The study was performed as a part of the employment.

Acknowledgments

The authors acknowledged the characterization support to complete this research work.

References

[1] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: an overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, 2018.

[2] A. C. Kemp, B. P. Horton, J. P. Donnelly, M. E. Mann, M. Vermeer, and S. Rahmstorf, "Climate related sea-level variations over the past two millennia," *Proceedings of the National Academy of Sciences*, vol. 108, no. 27, pp. 11017–11022, 2011.

[3] J. Wang, H. Jiang, Q. Zhou, J. Wu, and S. Qin, "China's natural gas production and consumption analysis based on the multi-cycle Hubbert model and rolling Grey model," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 1149–1167, 2016.

[4] S. Ameer, M. Ali Shah, A. Khan et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.

[5] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: model regularization and optimization," *Big Data Cognitive Computing*, vol. 2, no. 1, pp. 5–15, 2018.

[6] D. Ramesh, "Enhancements of artificial intelligence and machine learning," *International Journal of Advanced Science and Technology*, vol. 28, no. 17, pp. 16–23, 2019.

[7] M. G. H. David, R. Faner, O. Sibila, J. R. Badia, and A. Agusti, *Do Chronic Respiratory Diseases or Their Treatment Affect the Risk of SARS-CoV-2 Infection*, Elsevier Ltd. Science Direct, 2020.

[8] Y. Ying, L. Chang, and L. Wang, "Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): current status, challenges, and countermeasures," *Reviews in Medical Virology*, vol. 30, no. 3, article e2106, 2020.

[9] M. Chitty, *Artificial Intelligence, Machine Learning & Machine Learning Glossary & Taxonomy*, Cambridge Health Institute, 2020.

[10] Y. Zhang, Y. Wang, M. Gao et al., "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, no. 2019, pp. 30732–30743, 2019.

[11] Y. Xu and H. Liu, "Spatial ensemble prediction of hourly PM2.5 concentrations around Beijing railway station in China," *Air Quality, Atmosphere and Health*, vol. 13, no. 5, pp. 563–573, 2020.

[12] D. Kim, S. Cho, L. Tamil, D. J. Song, and A. S. Seo, "Predicting asthma attacks: effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children," *IEEE Access*, vol. 8, pp. 8791–8797, 2020.

[13] R. E. Caraka, R. C. Chen, T. Toharudin, B. Pardamean, H. Yasin, and S. H. Wu, "Prediction of status particulate matter 2.5 using state Markov chain stochastic process and HYBRID VAR-NN-PSO," *IEEE Access*, vol. 2, pp. 161654–161665, 2019.

[14] R. Beelen, O. Raaschounielsen, M. Stafoggia et al., "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project," *The Lancet*, vol. 383, no. 9919, pp. 785–795, 2014.

[15] Y. Bai, Y. Li, X. Wang, J. Xie, and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," *Atmospheric Pollution Research*, vol. 7, no. 3, pp. 557–566, 2016.

[16] R. Tiwari, S. Upadhyay, P. Singhal, U. Garg, and S. Bisht, "Air pollution level prediction system," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 6C, 2019.

[17] C. L. Bing, B. Arihant, C. Pei-Chann, K. T. Manoj, and T. Cheng-Chin, "Urban air quality forecasting based on multi-dimensional collaborative support vector regression (SVR): a case study of Beijing-Tianjin-Shijiazhuang," *PloS One*, vol. 12, no. 7, article e0179763, 2017.

[18] S. N. Pasha, A. Harshavardhan, D. Ramesh, and S. S. Md, "Variation analysis of artificial intelligence, machine learning and advantages of deep architectures," *International Journal of Advanced Science and Technology*, vol. 28, no. 17, pp. 488–495, 2019.

[19] Y. Lin, L. Zhao, L. Haiyan, and Y. Sun, "Air quality forecasting based on cloud model granulation," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, 10 pages, 2018.

[20] F. Xiao, Y. Li, J. Zhu, L. Hou, and J. W. Jin, "Artificial neural networks forecasting of $PM_{2.5}$ pollution using air masstrajectory based geographic model and wavelet transformation," *Atmospheric Environmen*, vol. 107, pp. 118–128, 2015.

[21] C. J. Soh and J. Huang, "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations," *IEEE Access*, vol. 6, pp. 38186–38199, 2018.

[22] P. Heni and S. Saket, "Air pollution prediction system for smart city using data mining technique: a survey," *Health*, vol. 6, no. 12, pp. 990–999, 2019.

[23] J. Li, L. Xiaoli, and K. Wang, "Atmospheric $PM_{2.5}$ concentration prediction based on time series and interactive multiple model approach," *Advances in Meterology*, vol. 2019, article 1279565, pp. 1–11, 2019.

[24] V. Veeramsetty and R. Deshmukh, "Electric power load forecasting on a 33/11 kV substation using artificial neural networks," *SN Applied Sciences*, vol. 2, no. 5, pp. 1–10, 2020.

[25] B. S. Freeman, G. Taylor, B. J. Gharabaghi, and J. Thé, "forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, no. 8, pp. 866–886, 2018.

[26] J. K. Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 5106045, 14 pages, 2017.

[27] M. Sallauddin, D. Ramesh, A. Harshavardhan, and S. N. S. Pasha, "A comprehensive study on traditional AI and ANN architecture," *International Journal of Advanced Science and Technology*, vol. 28, no. 17, pp. 479–487, 2019.

[28] A. Harshavardhan and B. Suresh, "An improved brain tumor segmentation and classification method using SVM with various kernels," *Journal of International Pharmaceutical Research*, vol. 46, no. 2, pp. 489–495, 2019.

[29] R. Zhao, X. Gu, B. Xue, J. Zhang, and W. Ren, "Short period $PM_{2.5}$ prediction based on multivariate linear regression model," *PloS One*, vol. 13, no. 7, article e0201011, 2018.

[30] X. Ni, H. Huang, and W. Du, "Relevance analysis and short-term prediction of $PM_{2.5}$ concentrations in Beijing based on multi-source data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.

[31] N. H. A. Rahman, M. H. Lee, Suhartono, and M. T. Latif, "Artificial neural networks and fuzzy time series forecasting: an application to air quality," *Quality and Quantity*, vol. 49, no. 6, pp. 2633–2647, 2015.