Environmental Research 239 (2023) 117354

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Impact of air pollutants on climate change and prediction of air quality index using machine learning models

Gokulan Ravindiran [a,b,*], Sivarethinamohan Rajamanickam [c], Karthick Kanagarathinam [d,**], Gasim Hayder [a,e], Gorti Janardhan [f], Priya Arunkumar [g], Sivakumar Arunachalam [h], Abeer A. AlObaid [i], Ismail Warad [j,k], Senthil Kumar Muniasamy [l]

[a] Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), 43000, Kajang, Selangor Darul Ehsan, Malaysia
[b] Department of Civil Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, 500090, Telangana, India
[c] Symbiosis Centre for Management Studies (Constituent of Symbiosis International Deemed University), Bengaluru, 560 100, Karnataka, India
[d] Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, 532 127, Andhra Pradesh, India
[e] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000, Kajang, Selangor Darul Ehsan, Malaysia
[f] Department of Mechanical Engineering, GMR Institute of Technology, Rajam, 532 127, Andhra Pradesh, India
[g] Department of Chemical Engineering, KPR Institute of Engineering and Technology, Coimbatore, 641 407, India
[h] Department of Electrical and Electronics Engineering, Panimalar Engineering College, Chennai, India
[i] Department of Chemistry, College of Science, King Saud University, P.O. Box- 2455, Riyadh, 11451, Saudi Arabia
[j] Department of Chemistry, AN- Najah National University, P.O. Box 7, Nablus, Palestine
[k] Research Centre, Manchester Salt & Catalysis, Unit C, 88-90, Chorlton Rd, M154AN, Manchester, United Kingdom
[l] Department of Biotechnology, Karpaga Vinayaga College of Engineering and Technology, Chengalpattu, 603308, Tamilnadu, India

ARTICLE INFO

Keywords:
Climate action
Air pollution
Air quality index
Machine learning

ABSTRACT

The impact of air pollution in Chennai metropolitan city, a southern Indian coastal city was examined to predict the Air Quality Index (AQI). Regular monitoring and prediction of the Air Quality Index (AQI) are critical for combating air pollution. The current study created machine learning models such as XGBoost, Random Forest, BaggingRegressor, and LGBMRegressor for the prediction of the AQI using the historical data available from 2017 to 2022. According to historical data, the AQI is highest in January, with a mean value of 104.6 g/gm, and the lowest in August, with a mean AQI value of 63.87 g/gm. Particulate matter, gaseous pollutants, and meteorological parameters were used to predict AQI, and the heat map generated showed that of all the parameters, $PM_{2.5}$ has the greatest impact on AQI, with a value of 0.91. The log transformation method is used to normalize datasets and determine skewness and kurtosis. The XGBoost model demonstrated strong performance, achieving an R^2 (correlation coefficient) of 0.9935, a mean absolute error (MAE) of 0.02, a mean square error (MSE) of 0.001, and a root mean square error (RMSE) of 0.04. In comparison, the LightGBM model's prediction was less effective, as it attained an R^2 of 0.9748. According to the study, the AQI in Chennai has been increasing over the last two years, and if the same conditions persist, the city's air pollution will worsen in the future. Furthermore, accurate future air quality level predictions can be made using historical data and advanced machine learning algorithms.

1. Introduction

Climate change has a long-term impact on atmospheric temperatures and weather patterns and these could be man-made or natural phenomena. Some of the major sectors that contribute to air pollution are transportation, industrial development, and the combustion of fossil fuels such as coal and gas (Perera, 2018). Greenhouse gas emissions cause global warming, and global warming causes climate change.

* Corresponding author. Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), Kajang, 43000, Selangor Darul Ehsan, Malaysia.
** Corresponding author. Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, 532 127, Andhra Pradesh, India.
E-mail addresses: gokulravi4455@gmail.com, gokulan_r@vnrvjiet.in (G. Ravindiran), mohan.dimat@gmail.com (S. Rajamanickam), karthick.k@gmrit.edu.in (K. Kanagarathinam), gasim@uniten.edu.my (G. Hayder), g.janardhan@hotmail.com (G. Janardhan), a.k.priya@kpriet.ac.in (P. Arunkumar), arunsiva75@gmail.com (S. Arunachalam), aalobaid@ksu.edu.sa (A.A. AlObaid), i.kh.warad@gmail.com (I. Warad), senthilenvtce@gmail.com (S.K. Muniasamy).

https://doi.org/10.1016/j.envres.2023.117354
Received 13 June 2023; Received in revised form 26 September 2023; Accepted 7 October 2023
Available online 12 October 2023

Emerging countries like India are dealing with a slew of issues concerning air pollution and its aggressive effects on human health and the environment (Manisalidis et al., 2020a). Air is the main source for the survival of living beings on Earth. Earth life is not possible without air and it is very much essential for survival. The increase in population, industrial activities, burning of fossil fuels, poor agricultural practices and motor vehicle emissions resulted in degradation of air quality (Ravindra, 2019; Ravindra et al., 2020). Some of the most commonly emitted air pollutants from these activities are particulate matter 2.5 and 10 ($PM_{2.5}$ and PM_{10}), Carbon dioxide (CO_2), Sulfur dioxide (SO_X), Nitrous Oxide (NO_x), Nitrogen dioxide (NO_2), ammonia, benzene, volatile organic compounds (VOCs), Carbon monoxide (CO) and Ozone (Villanueva et al., 2016). These pollutants are emitted directly from the source mentioned above and cause air pollution. Generally, air pollutants are categorized into particulate matter and gaseous pollutants. The extent and magnitude of the pollutants depend on meteorological factors namely rainfall intensity, wind speed, relative humidity, wind direction, solar radiation and temperature(Bodor et al., 2020). These air pollutants are used to calculate the AQI of a region or city. Further, these air pollutants result in air pollution and have several other impacts namely ozone layer depletion, global warming, rise in average earth temperature, climate change and acid rain (Balakrishnan et al., 2019; Manisalidis et al., 2020b).

Every year IQAir publishes the statistics of AQI and the latest was published in the year 2023 for the assessing year 2022("IQAir | First in Air Quality," n.d.). The World Air Quality report is based on $P.M_{2.5}$. The reason for considering $P.M_{2.5}$ concentration is that most aerosols have a size of less than 2.5 μm. $P.M_{2.5}$ is one of the most important air pollutants based on its extent and impact on the environment. The report consolidated the $P.M_{2.5}$ data for 7323 cities across 131 countries worldwide. WHO fixed the guidelines of $P.M_{2.5}$ should be always less than 5 $\mu g/m^3$ (annual) and 15 $\mu g/m^3$ (daily) ("Ambient (outdoor) air pollution," n.d.), but only 13 countries out of 131 countries showed the $P.M_{2.5}$ less than WHO guidelines and remaining all countries have exceeded the recommended values. India ranked at 8th position with an average $P.M_{2.5}$ of 53.3 $\mu g/m^3$ which is 10 times higher than the WHO guidelines. It is also observed that 12 cities out of 15 cities from central and south Asia are from India, indicating that most of the cities in India are severely affected due to air pollution. 60% of the Indian cities have 7 times higher values of $PM_{2.5}$ than WHO guidelines. In the previous year 2021, India ranked 5th with an average $P.M_{2.5}$ of 58.1 $\mu g/m^3$, currently which is 4.8 $\mu g/m^3$ less than the previous year (2021). In India, around 25–30% of the $PM_{2.5}$ was emitted from the transport sector followed by burring of crops increasing $PM_{2.5}$ concentration, particularly in North India (Singh et al., 2021a). Further, in the last quarter of 2022, the Indian government relaxed the environmental rules and regulations for coal and thermal power plants, which resulted in an 11.72% increase in coal production and increased $PM_{2.5}$ concentration (Singh et al., 2021b).

Because of the country's large population and economy, air pollution monitoring and control are major challenges for India. India is facing a major challenge in terms of AQI monitoring and, as a result, reducing pollution at the source. Prediction of AQI is critical for implementing a strategy or developing a policy to control air pollution. AQI prediction or future forecasting will provide policymakers with a solution for developing a protocol for air pollution mitigation measures (Ramírez et al., 2019). Since the last ten decades, India's carbon dioxide emissions have steadily increased, and the country now ranks third in the world in CO_2 emissions, contributing to 7% of global CO_2 emissions. Similarly, India ranks second in methane gas emissions after China, with agricultural emissions accounting for 62% of total emissions. CO_2 and CH_4 are the primary greenhouse gases that contribute significantly to global warming and climate change. To address all of these issues, it is necessary to forecast the AQI of a country/region to develop mitigation measures to reduce future pollution.

AQI is calculated on several input parameters as mentioned above. These input parameters are continuously recorded in all monitoring stations using sensors and the data are recorded for the AQI prediction and future predictions of AQI(Wang et al., 2022). In recent days, AQI forecasting and predictions have been considered a viable process for sustainable development (Rybarczyk et al., 2017). A high value of AQI represents the hazardous nature and it causes several damage to human health and the environment. So, it has become mandatory to predict the futuristic AQI and recommend mitigation measures in the present to overcome the adverse effects of air pollutants (Bekkar et al., 2021). AQI prediction models (ML - Machine Learning and DL - Deep Learning) are developed by many researchers based on the statistics, dynamic and futuristic that could predict the AQI accurately (Maltare and Vahora, 2023). ML techniques are not only based on statistics, they develop a model based on the correlation between every independent variable that is used for the AQI predictions (Bhalgat, 2019; Kumar and Pande, 2022).

Many researchers utilized many machine learning models for the AQI predictions namely Adaptive Boosting (Adaboost) (Liang et al., 2020), Artificial Neural Network (ANN) (Liang et al., 2020; Madan et al., 2020), Catboost Regression (CR) (Gupta et al., 2023), Convolution Neural Networks (CNN) (Wang et al., 2022), Decision Tree Regression (DTR) (Abirami et al., 2022; Madan et al., 2020), Gaussian Naive Bayes (GNB) (Kumar and Pande, 2022), Improved Long Short-Term Memory (ILSTM) (Wang et al., 2022), K-Nearest Neighbors Algorithm (KNN) (Kumar and Pande, 2022), Linear Regression (Madan et al., 2020), long short-term memory (LSTM) (Maltare and Vahora, 2023), Multiple Linear Regression (MLR) (Abirami et al., 2022), Random Forest Regression (RFR) (Gupta et al., 2023; Liang et al., 2020), Seasonal Autoregressive Integrated Moving Average (SARIMA) (Maltare and Vahora, 2023), Support Vector Machine (SVM) (Liang et al., 2020; Maltare and Vahora, 2023), Support Vector Regression (SVR) (Abirami et al., 2022; Gupta et al., 2023), and Xgboost (Kumar and Pande, 2022).

The current study examines data collected from the years 2017 and 2022 in Chennai, Tamilnadu, India. Random Forest (RF), XGBoost, Bagging Regressor and LGBM Regressor are the four machine learning algorithms used for the AQI predictions. The specific machine learning methods, such as Random Forest (RF), XGBoost, Bagging Regressor, and LGBM Regressor, were adopted in the study based on their unique characteristics and advantages. In Random Forest multiple decision trees are used to make predictions and it has several advantages namely overfitting, the capacity to handle high-dimensional data, and automatic feature selection. XGBoost (Extreme Gradient Boosting) is a powerful machine-learning technique. XGBoost is known for its high efficiency, scalability, and accuracy. It utilizes gradient boosting to create an ensemble of weak learners and sequentially trains models on weighted versions of the data. Bagging is another ensemble learning technique that combines multiple models trained on different subsets of the raw datasets. The Bagging Regressor is a variant of the bagging approach that applies bagging to regression tasks. LightGBM is scalable and can handle large datasets efficiently. It has built-in support for categorical features without requiring one-hot encoding. It can be used for a wide range of data types, including numerical, categorical, and even sparse data. LightGBM supports L1 and L2 regularization, allowing for better control of model complexity and preventing overfitting.

2. Methods and materials

2.1. Study area

Chennai is the capital city of Tamil Nadu, South India located along the east coastal region of the Bay of Bengal. The economic development of the city started from the year 1990. Software development and electronic manufacturing were the two major sectors in early 1990. But now, many other industries namely automobile, rubber, fertilizer manufacturing, leather, iron and ore, and cotton developed (Mehta and Rajan, 2017). These industrial development activities not only increased the economy of the city but also increased air pollution. Currently, Chennai is reporting a minimum of 5 times higher values of $PM_{2.5}$ as per

the WHO guidelines value. As illustrated in Fig. 2 Chennai is located with latitudes ranging from 12°51′00″ N to 13°09′00″ N and longitudes ranging from 80°08′36″ E to 80°18′17″ E (Saravanan et al., 2019). With an area of 1189 square kilometers, it is Tamil Nadu's largest city and primarily an industrial city.

Chennai has a tropical climate, with most of the year being hot and humid. Because the city is located near the equator and on the coast, seasonal temperature variations are minimal. The lowest temperature in January is around 18–20 °C, and the extreme temperature in May and June is around 38–42 °C. From mid-October to mid-December, the northeast monsoon winds bring rain to the city. The annual rainfall averages around 140 cm. The most common winds are south-westerly from the end of May to the end of September and north-easterly the rest of the year. Chennai is also one of the top 100 cities in the world's fastest-growing cities. Fig. 1 illustrated the location map of Chennai city.

2.2. Machine learning technique in the prediction of AQI

Deep learning models have shown their effectiveness in many applications, for a relatively small dataset such as the AQI dataset used in this study, traditional machine learning models may be more suitable and easier to interpret (Ayus et al., 2023). In this study, we started by dividing the dataset into train and test sets and then standardized the data. Next, we developed an ML model along with a hyperparameter grid. To find the best hyperparameters, a cross-validation grid search using the GridSearchCV function was used. The model was trained using the fit() function, and the best parameters and scores were reported. Finally, we evaluated the model's efficiency on the test set. To ensure the robustness of the model, we used 5-fold cross-validation, computed MAE, MSE, RMSE, and R^2 scores for each fold, and reported the mean and standard deviation of each metric across all folds. When a model has a higher R^2 score and lower MAE and MSE scores, it is generally regarded as performing better (Sekeroglu et al., 2022).

Fig. 2 depicts the Flowchart describing the Machine Learning technique in the prediction of AQI. R^2 values will be between 0 and 1 and close to 1 indicates the datasets are similar and model performance is good. MAE is typically used when measuring performance using continuous variable data. The lower the value, the better the performance of the model. The utility of RMSE increases significantly when there are large errors that significantly affect the performance of the model. The absolute value of the error, which is important in many mathematical calculations, is not taken into account. In this metric as well, a lower value indicates better model performance.

2.3. Air pollutants and meteorological parameters

The datasets were obtained from the Central Pollution Control Board's - Central Control Room for Air (CPCBCRR). They include measurements of air pollutants and meteorological parameters, as summarized in Table 1, which were used to calculate the AQI. The daily mean value of each pollutant was considered for the study for the selected durations. The data sets contained a total of 2099 observations spanning the years 01-01-2017 to 30-9-2022. The study utilized an original dataset consisting of 2099 instances, which included 14 variables (08 Air Pollutants attributes, 05 Meteorological Factors, and 1 AQI). The AQI represents the variable under investigation. Conventionally, the AQI is computed using the sub-index ranges of each air pollutant and categorized based on the AQI values as summarized in Table 2. For example, AQI values are calculated in a three-step process. First, the concentration of air pollutants needs to be measured, followed by converting all these concentrations into sub-index as mentioned in Table 2 and finally arriving at AQI values (Maximum Sub-Index of Pollutants). Sample AQI calculation is included in supplementary materials (Table S1 & S2) for better understanding. Table 1 summarizes the list of datasets used for the prediction of AQI using machine learning models.

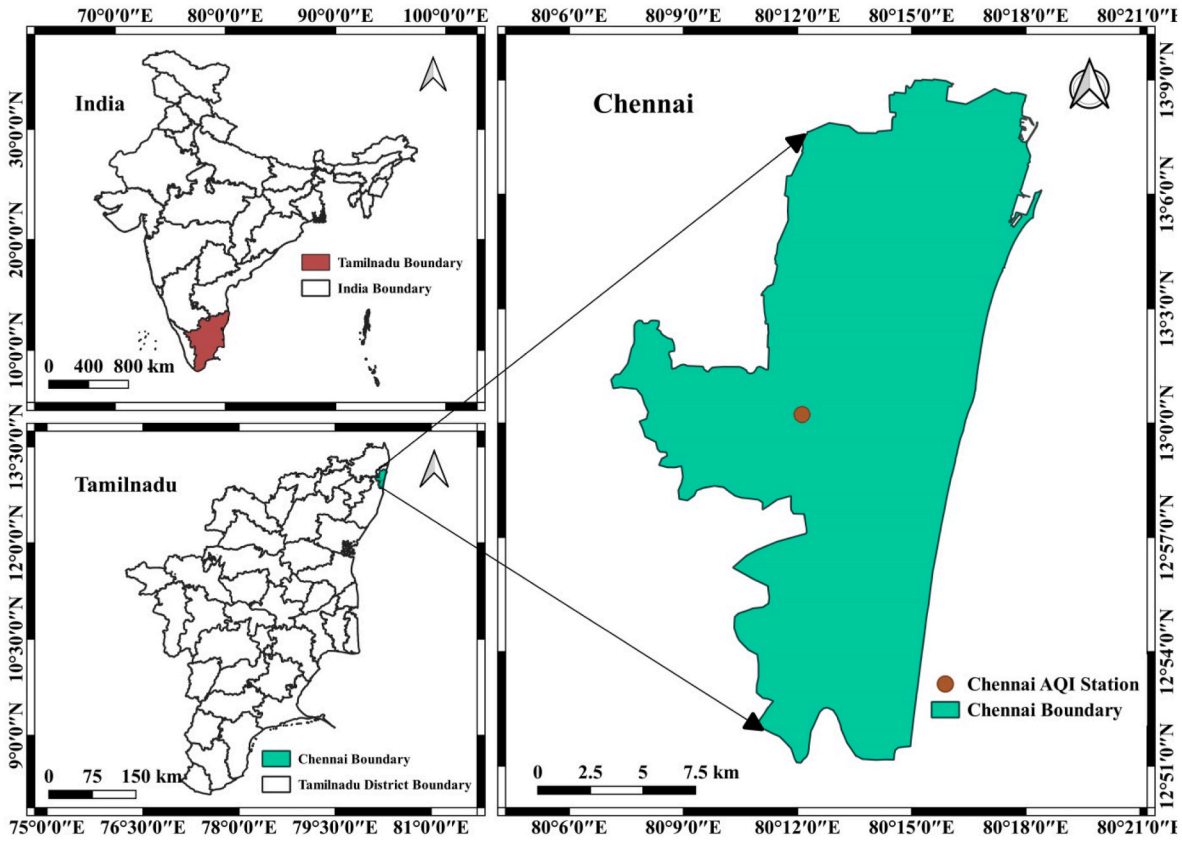

Fig. 1. Location map of Chennai City, Tamil Nadu.

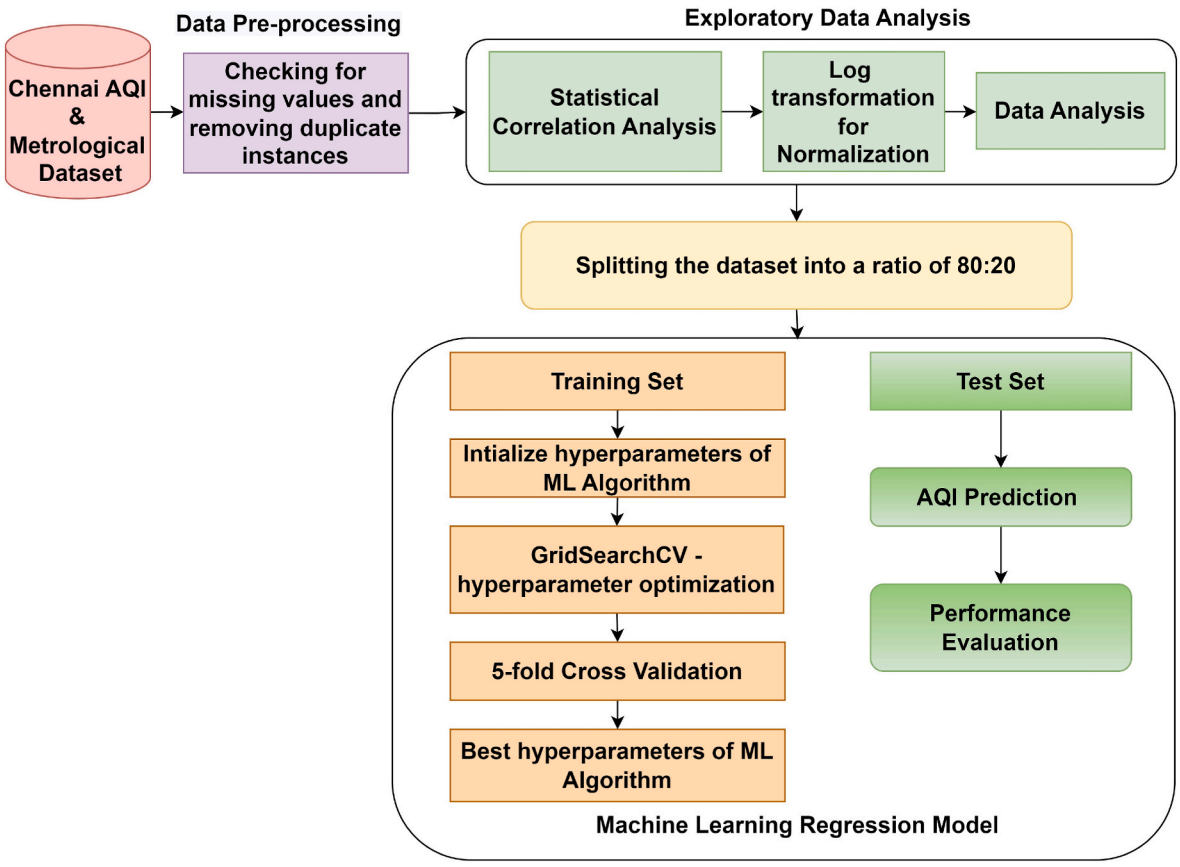

Fig. 2. Flowchart describing machine learning technique in the prediction of AQI.

2.4. Preliminary processing of datasets

A total of 2099 instances were available in the raw dataset. The missing values of the characteristics were initially removed. After removing the missing values, there are a total of 1840 instances remaining, with 14 characteristics. The removal of missing values helps maintain consistency in the dataset, as it ensures that each instance has values for all the specified characteristics. This consistency is crucial for training machine learning models effectively and making reliable predictions. Type conversion from object to float data type was performed on AQI. Table 3 summarizes the statistical data from the datasets for air pollutants and meteorological parameters.

2.5. Exploratory data analysis

Before implementing machine learning techniques, exploratory data analysis is employed for initial interpretation. It is used to comprehend the relationship between various air pollutants in the prediction of AQI (Kumar and Pande, 2022). The values in the heat map typically extend from +1 to −1. Fig. 3 depicts a heat map for the data input. The pollutants with the highest heat map values had a significant impact on AQI predictions. The heat map among the various variables was created using a correlation matrix method.

Fig. 3 shows that the correlation was positive for some parameters and negative for others. When compared to the negative or inversion correlation, the positive correlation plays a major role in AQI prediction. As a result, 0.5 is used as the threshold value, and values less than 0.5 are ignored because they have less impact on AQI predictions. With a correlation coefficient of 0.91, the highest among all parameters, the relationship between $PM_{2.5}$ and AQI is significant, indicating that $PM_{2.5}$ plays a crucial role in determining the AQI. In addition to $PM_{2.5}$, Ozone and NO_2 have correlation coefficients of 0.26 and 0.24, respectively. The remaining parameters have very little correlation, indicating that those pollutants have little impact on the AQI calculation.

2.6. Data transformation

Table 4 summarizes the skewness and kurtosis values of raw data and transformed data. Skewness is a parameter for assessing symmetry or asymmetry. If the datasets are equally spaced from the center (Left to Right), the datasets are said to be symmetrical. Kurtosis is used to determine the normal distribution of the data that is tailed strongly or weakly. The current study datasets are not normally distributed, so the normal distribution of the datasets needs to be assured by utilizing any data transformation technique. The most commonly used data transformation techniques are Box-Cox transformation, log transformation and square root transformation (Feng et al., 2014).

Numerical variables may exhibit high skewness and non-normal distribution due to the presence of outliers or highly exponential distributions. To address this issue, data transformation is commonly employed. One popular data transformation technique is Log transformation, which involves replacing each variable x with its logarithm, using base 10, base 2, or natural logarithm (Feng et al., 2013). Log transformation is a powerful tool to reduce skewness in data. It works by compressing larger values and spreading out smaller ones. This technique is particularly useful for data with a right-skewed distribution, where most of the data is clustered at lower values but there are a few high values that can significantly affect the results. Log transformation was used to change the data to make it more normal. Table 2 shows that the skewness and kurtosis values for $PM_{2.5}$, Benzene and AQI were quite high.

Table 1
Datasets used in the ML models to predict AQI.

S. No	Features	Description
1	City	Chennai, Tamil Nadu India
2	Date	Day-wise air quality data for five years from 01 to 01–2017 to 30-9-2022
3	$PM_{2.5}$	mixing of ultra-fine particles with liquid droplets in the air known as Particulate Matter with the size of 2.5 μm or smaller size
4	PM_{10}	mixing of fine particles in the air is known as Particulate Matter with the size of 10 μm
5	NO	Nitrogen monoxide. It is released due to the industrial combustion process, motor vehicles, and power stations
6	NO_2	Nitrogen dioxide. It is released through the oxidation of NO in the combustion process
7	NO_x	It is a group of highly reactive gasses that include NO, NO2 and other forms of Nitrogen
8	NH_3	It is released from agricultural activities, animal husbandry, fertilizers, etc
9	CO	Carbon Monoxide is a colorless gas released from fires, industrial processes, kitchen chimneys, etc.
10	SO_2	Sulfur dioxide released from automobiles, chemical industries, etc.
11	O_3	Ozone consists of 3 atoms. It is mainly released from industries
12	Benzene	Coal and oil burning and Tobacco smoking are causes of air pollutant Benzene
13	Toluene	Motor vehicles are the main emission resources for air pollutants Toluene
14	Xylene	The burning of coal, wood, and petroleum products is the main source of air pollutant Xylene
15	Meteorological Factors	Air Pressure (BP), Ambient Temperature (AT), Rainfall (RF), Relative Humidity (RH), Solar Radiation (SR), Temperature, Total Rainfall (TOT-RF), Wind Direction (WD) and Wind Speed (WS)
16	AQI	It is calculated based on available air pollutants. AQI calculation needs at least three pollutants of which either $PM_{2.5}$ or PM_{10} should be one.

2.7. Model development- dataset split and standardization

Data splitting refers to the process of dividing a dataset into two mutually exclusive subsets for different purposes, typically for training and testing machine learning models. The primary goal of data splitting is to create subsets that are representative of the overall dataset while ensuring that there is no overlap between them (Joseph and Vakayil, 2022). In the realm of machine learning, it is crucial to assess how well a model performs on data that it has not previously encountered. To accomplish this, a portion of the available data is used for testing, while the remaining data is used to train the model. During the training process, the model is constructed, whereas, during the testing phase, its performance is evaluated.

Randomly dividing the data into subsets is a common approach, and the allocation of data to each subset varies depending on the problem and the dataset's size. For instance, a larger portion of data is usually allocated to the training set in cases where the dataset is small, whereas a smaller portion may be used for larger datasets. Common split ratios are 70/30, 80/20, and 90/10. The division of data into subsets is a critical stage in machine learning, as it can significantly impact the model's performance. An inadequately selected split can lead to the model being overfit or underfit, which can result in inadequate predictions on unseen data. Therefore, it is crucial to select the split carefully and assess the model's performance on the test data to ensure that it can generalize well to new data.

In this study, we assign the target variable AQI to the variable y and all other features except AQI to the variable X as explanatory variables. We utilize the sklearn.model_selection module to split the data into two distinct sets: training and testing, using the train_test_split() function. To ensure the reproducibility of the results, we set the random state to 0 and the size of the testing set to 20% of the data. The training set is employed to train the model, while the test set is entirely separate and used as unseen data for evaluation purposes. By evaluating the model on the test set, the model's performances have been evaluated on data it has not seen during training. Finally, to normalize the features of the training and testing sets, we use the StandardScaler() function from the sklearn. preprocessing module. We standardize the features of the training set by executing the fit_transform() method, and subsequently, employ the transform() method on the testing set to ensure that the same scaling is applied as was executed on the training set.

2.7.1. XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm that combines multiple weak learners into a powerful model by training them sequentially on weighted versions of the dataset (Mahesh et al., 2022). It is an enhanced implementation of gradient boosting,

Table 2
AQI Ranges for different air pollutants.

AQI Category	PM_{10} 24 h	$PM_{2.5}$ 24 h	NO_2 24 h	O_3 24 h	CO 8 h	SO_2 24 h	NH_3 24 h	Pb 24 h
Good (0–50)	0–50	0–30	0–40	0–50	0–1	0–40	0–200	0.05
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.6–1.0
Moderate (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10.1–17	381–800	801–1200	2.1–3.0
Very Poor (301–400)	351–430	121–250	281–400	209–748	17.1–34	801–1600	1201–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Table 3
Dataset statistical information.

index	PM2.5 (ug/m^3)	NO (ug/m^3)	NO_2 (ug/m^3)	NOx (ppb)	SO_2 (ug/m^3)	Ozone (ug/m^3)	Benzene (ug/m^3)	Toluene (ug/m^3)	WS (m/s)	WD (deg)	SR (W/mt^2)	BP (mmHg)	VWS (m/s)	AQI
count	1840	1840	1840	1840	1840	1840	1840	1840	1840	1840	1840	1840	1840	1840
mean	42.03	7.43	14.42	18.67	11.47	25.73	0.36	0.45	1.25	176.76	83.31	805.95	−0.13	83.73
std	42.19	5.38	7.97	11.23	12.11	19.54	0.87	1.05	1.55	67.50	63.11	101.33	0.41	63.84
min	0.02	0.9	0.02	1.23	0.01	0.21	0	0	0.05	3.83	0.73	742.78	−0.52	15.5
25%	22.71	4.04	9.81	10.47	4.52	12.82	0.00	0.00	0.55	120.16	35.87	752.11	−0.48	51.50
50%	34.68	5.96	12.88	15.53	7.37	21.04	0.07	0.08	1.02	174.61	73.20	759.47	−0.44	68.00
75%	49.38	9.40	17.39	23.54	13.76	33.66	0.31	0.41	1.54	231.25	119.81	763.29	0.37	91.00
max	999.99	59.71	83.09	87.89	144.81	182.94	11.68	20.20	21.09	353.18	765.83	1020.36	1.86	976.92

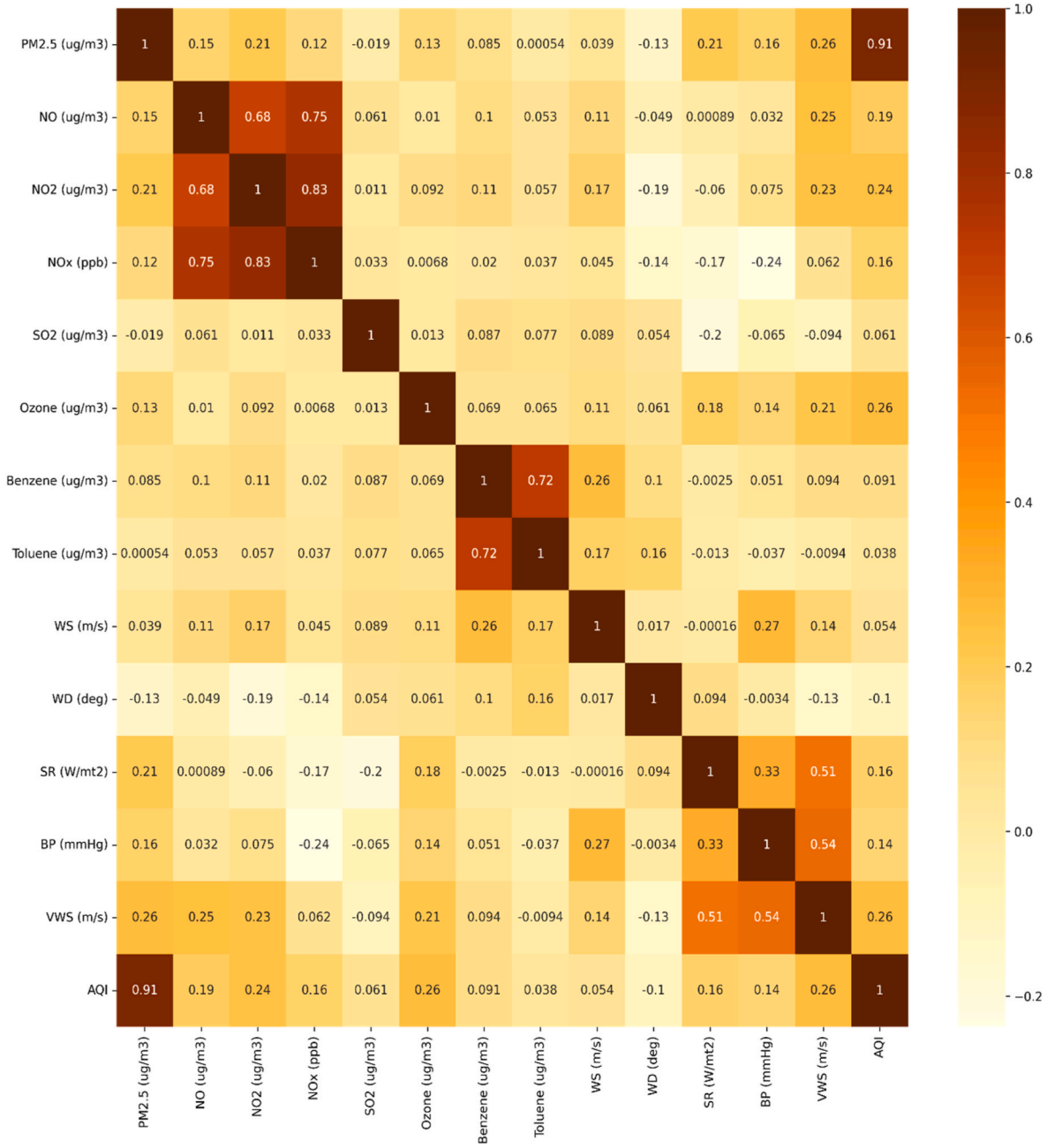

Fig. 3. Correlation matrix of input parameters with AQI- Heat Map.

Table 4
Skew & Kurtosis of datasets before and after log transformation.

Parameters	Before		After	
	Skew	Kurtosis	Skew	Kurtosis
$PM_{2.5}$ ($\mu g/m^3$)	9.73	169.23	−0.105	2.276
NO ($\mu g/m^3$)	2.73	13.98	0.370	0.044
NO_2 ($\mu g/m^3$)	1.87	7.93	1.869	7.930
NOx (ppb)	1.53	2.88	1.529	2.881
Ozone ($\mu g/m^3$)	2.61	12.26	−0.418	0.895
Benzene ($\mu g/m^3$)	6.05	53.47	2.555	7.779
WD (deg)	0.04	−0.89	0.035	−0.890
SR (W/mt^2)	3.72	32.16	−0.540	0.460
BP (mmHg)	1.53	0.36	1.531	0.365
VWS (m/s)	0.52	−1.02	0.521	−1.020
AQI	4.31	32.86	0.705	1.498

known for its efficiency, scalability, and accuracy. In the model development, we imported the XGBRegressor class from the XGBoost library and utilized the GridSearchCV and k-fold classes from the scikit-learn library. To begin, an instance of the XGBRegressor class was created with specific hyperparameters, such as random_state and objective. For optimizing the model's performance, we defined a grid of hyperparameters for tuning, including various values for n_estimators, max_depth, and learning_rate. Specifically, the hyperparameters were set as follows: 'n_estimators': [100, 200], 'max_depth': [5, 10], and 'learning_rate': [0.1, 0.01].

Next, a GridSearchCV object was constructed with 5-fold cross-validation, using the R-squared (R^2) metric for scoring. To expedite the computation process, all available CPU cores were utilized by setting the n_jobs parameter to −1. The XGB Regressor model was then fitted to the training data using the fit() method of the GridSearchCV object. Additional arguments, such as early_stopping_rounds and eval_set, were

passed to the fit() method to prevent overfitting and monitor the model's performance on a validation set. To suppress the output, the verbose parameter was set to 'False'. After training, the best hyperparameters were determined to be 'learning_rate': 0.1, 'max_depth': 10, and 'n_estimators': 100. These optimized settings allowed the XGBoost model to achieve its highest performance when predicting the AQI.

2.7.2. Random forest

Random forest is a supervised learning method that involves constructing an ensemble of decision trees and aggregating their outputs to generate a final prediction(Wang et al., 2018). For random forest regression, individual decision trees are created using a random subset of features and training instances. To implement the random forest regression, the necessary classes from the scikit-learn library are imported, including the RandomForestRegressor class for creating the random forest model and the GridSearchCV and k-fold classes for hyperparameter tuning and cross-validation, respectively. A RandomForestRegressor object is then created with a specified random_state. The values for the defined hyperparameters are 'max_depth': 15, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, and 'n_estimators': 100.

Next, a GridSearchCV object is created with the random forest model, hyperparameter dictionary, 5-fold cross-validation, and the R-squared metric for scoring. The n_jobs parameter is set to −1 to utilize all available CPU cores for faster computation. The model is fitted to the training data using the fit() method of the GridSearchCV object. The fit method performs a grid search over the specified hyperparameters and returns the best hyperparameters found. The best hyperparameters are found to be 'max_depth': 15, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, and 'n_estimators': 100. Finally, the model is used to predict the target variable on the test set.

2.7.3. Bagging Regressor

Bootstrap aggregating, or Bagging for short, is a method in machine learning that trains multiple models on various subsets of the training data to improve the overall model's performance(Pérez-Rodríguez et al., 2023). In the case of bagging regression, the instances are decision trees, and the final prediction is obtained by averaging the predictions of all decision trees. To develop the bagging regression model, we import the necessary classes from the scikit-learn library, including the BaggingRegressor class for creating the model, and the GridSearchCV and KFold classes for performing hyperparameter tuning and cross-validation, respectively. Next, we create a BaggingRegressor object with a specified random_state and define a dictionary of hyperparameters to tune. The hyperparameters include the number of decision trees and the proportion of training samples to use for each decision tree. The specified values are 'n_estimators': [100, 200, 300] and 'max_samples': [0.5, 0.7, 0.9].

A GridSearchCV object is created with the bagging regression model, hyperparameter dictionary, 5-fold cross-validation, and the R-squared metric for scoring. The n_jobs parameter is set to −1 to use all available CPU cores for faster computation. The model is fit on the training data using the fit() method of the GridSearchCV object, which performs a grid search over the specified hyperparameters and returns the best hyperparameters found. The best hyperparameters are found to be 'n_estimators': 300 and 'max_samples': 0.9.Finally, the bagging regression model is used to predict the target variable on the test set, and its efficacy is assessed on both the training and test sets.

2.7.4. LGBM Regressor

LightGBM is a fast and scalable gradient-boosting framework that's widely used in machine learning(Chen et al., 2023). LightGBM is scalable and can handle large datasets efficiently. It has built-in support for categorical features without requiring one-hot encoding. It can handle a wide range of data types, including numerical, categorical, and even sparse data. LightGBM supports L1 and L2 regularization, allowing for better control of model complexity and preventing overfitting. To start, we import the necessary classes from both scikit-learn and LightGBM libraries, such as LGBMRegressor for creating a LightGBM regressor, GridSearchCV for performing hyperparameter tuning, KFold for cross-validation, and mean_squared_error for evaluating model performance. Then, we define a dictionary of hyperparameters that we want to tune, including the number of estimators, maximum depth of tree, learning rate, and number of leaves. The initial values of these parameters are 'n_estimators': [50, 100, 200], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.1, 1], and 'num_leaves': [31, 61, 121].

Next, we create a GridSearchCV object with the LightGBM regressor, hyperparameter dictionary, 5-fold cross-validation, and the negative mean squared error metric for scoring. We also set the n_jobs parameter to −1 to use all available CPU cores for faster computation.

We fit the model on the training data using the fit() method of the GridSearchCV object, which performs a grid search over the specified hyperparameters and returns the best hyperparameters found. The best hyperparameters are printed on the console for reference. After that, we use the model to predict the target variable on the test set and evaluate its performance on the same set. Finally, we use the best hyperparameters found by the grid search (i.e., 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'num_leaves': 61) to create the final model.

3. Result and discussion

3.1. Air quality index

Fig. 4 depicts the monthly and annual variations in the AQI. According to the results, the AQI has crossed 50 and it is in the range of 50–120, indicating that the city is moderately polluted. The results also show that AQI levels were extremely high in December and January. Seasonal variation may be the reason for the highest level of AQI in these months. These months fall under the winter season when temperatures are lower and more mist formation is observed. Because of the presence of moist air in the atmosphere, this could result in the formation of a temperature inversion and the pollutants emitted from the source get retained in the atmosphere (Chandrappa and Kulshrestha, 2016a). The air is denser, and there is no space for pollutants to escape into the atmosphere, increasing the concentration of air pollutants (Khillare and Sarkar, 2012). These will cause global warming, which will lead to climate change.

It is also clear from the results that the AQI range is less than 50 (Good) from April to August. In India, the rainy season generally lasts from June to September, with maximum rainfall occurring with high intensity (Chandrappa and Kulshrestha, 2016b). Rainfall causes $PM_{2.5}$ and other gaseous pollutants to settle, lowering the concentration of air pollutants in the atmosphere. Furthermore, the annual AQI ranges were extremely high in 2017 and 2018, with AQI levels of 100 and 119, respectively. It is observed that the AQI value in 2020 is very low, the lowest in the last five years. This could be due to the nationwide lockdown caused by Covid 19 from March to July(Singh and Chauhan, 2020). During this time, the city's electricity consumption dropped to 17.5% because most industries were closed, resulting in less air pollution. However, the AQI has been observed to rise to 63 and 85 for the years 2021 and 2022, respectively.

3.2. Particulate matter

Particulate matter is categorized into three sizes: PM_{10}, $PM_{2.5}$, and $PM_{0.5}$. These particulate matter are released due to construction activities, transportation, and coal-fired thermal power plants (Izzotti et al., 2022). These particulate matter are harmful to human health and green vegetation since they are respirable and reach the lungs, while very fine particles reach the bloodstream and it is also considered as a possible cause for cancer. As a result, particulate matter is considered a very

Fig. 4. Seasonal and annual variation of AQI.

important pollutant in calculating AQI (Manisalidis et al., 2020b). Fig. 5 depicts seasonal and annual variations of $PM_{2.5}$. It is concluded from Figs. 4 and 5 that the AQI and $PM_{2.5}$ patterns are similar, indicating that $PM_{2.5}$ has a significant impact on AQI calculation. $PM_{2.5}$ trend is similar to the AQI trend and the highest concentration is observed from November to February of each year. This is because, during the winter, $PM_{2.5}$ accumulates in the atmosphere and results in increased concentration in the atmosphere (Xia et al., 2022).

3.3. Gaseous pollutants

A gaseous pollutant is regarded as the primary source of long-term impact in the atmosphere, resulting in climate change. Secondary air pollutants are created when gaseous pollutants are released into the atmosphere (Doble and Kumar, 2005). Fig. 6 depicts seasonal and monthly variations in various gaseous pollutants in the atmosphere. Fig. 6 shows that NO, NO_2, and NOx seasonal and monthly variations followed the trend of AQI and $PM_{2.5}$. Because of the natural phenomena discussed above, the concentration was maximum in the winter and minimum in the summer. The trend of other pollutants is inversely proportional to the concentration of ozone in the atmosphere. Fig. 6 shows that ozone concentrations are very high during the summer and very low during the winter(Lu et al., 2021). When compared to solar radiation, the maximum solar radiation of the sun is observed in summer and less in winter (Fig. 7). It is clear that solar radiation is high, formation of ground-level ozone is also high (Khillare and Sarkar, 2012). When NO_2 and hydrocarbons react with sunlight, ozone gas is formed in the atmosphere. Transportation, electricity generation, and industrial activities all emit NO_2, whereas petroleum and power plants (coal-fired) are the primary sources of hydrocarbon emissions. The primary source of photochemical smog is the formation of ozone gas (Manisalidis et al., 2020b).

3.4. Meteorological factors on AQI

Meteorological parameters of a region or city play a major role in the transport of air pollutants from one location to another. Wind speed and direction vary significantly in the study area of Chennai city, which is located very close to the Bay of Bengal. Wind speed allows air pollutants to disperse more easily into the atmosphere and be carried to another location (Krishna et al., 2018; Shelton et al., 2022). The seasonal variations of the metrological parameters are depicted in Fig. 7. The results show that the wind speed was highest in the summer and lowest in the winter. A maximum wind speed of 0.26 m/s was recorded in August,

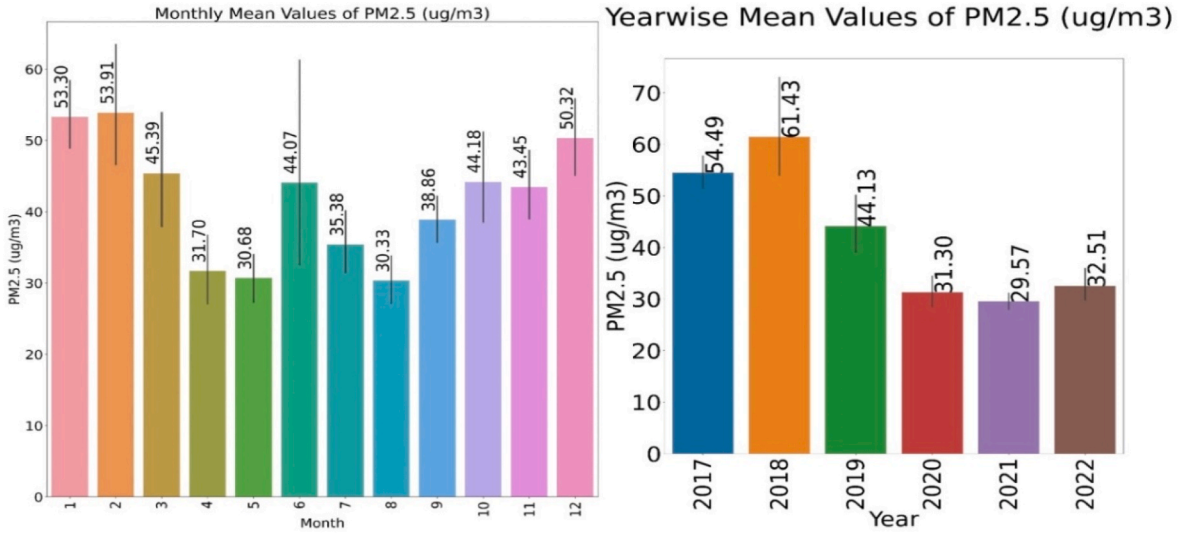

Fig. 5. Seasonal and annual variation of $PM_{2.5}$.

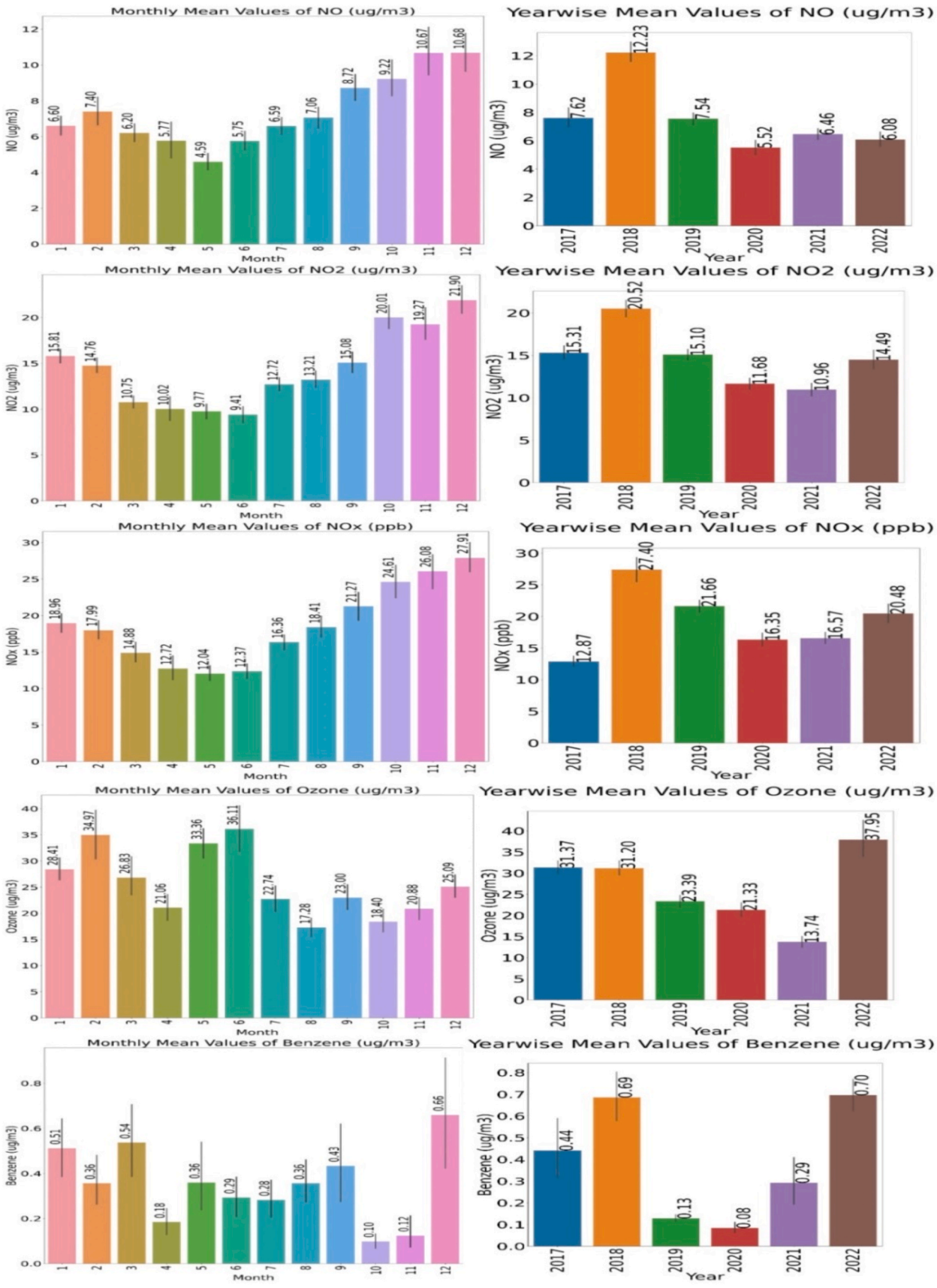

Fig. 6. Monthly and Annual variation of Gaseous Pollutants.

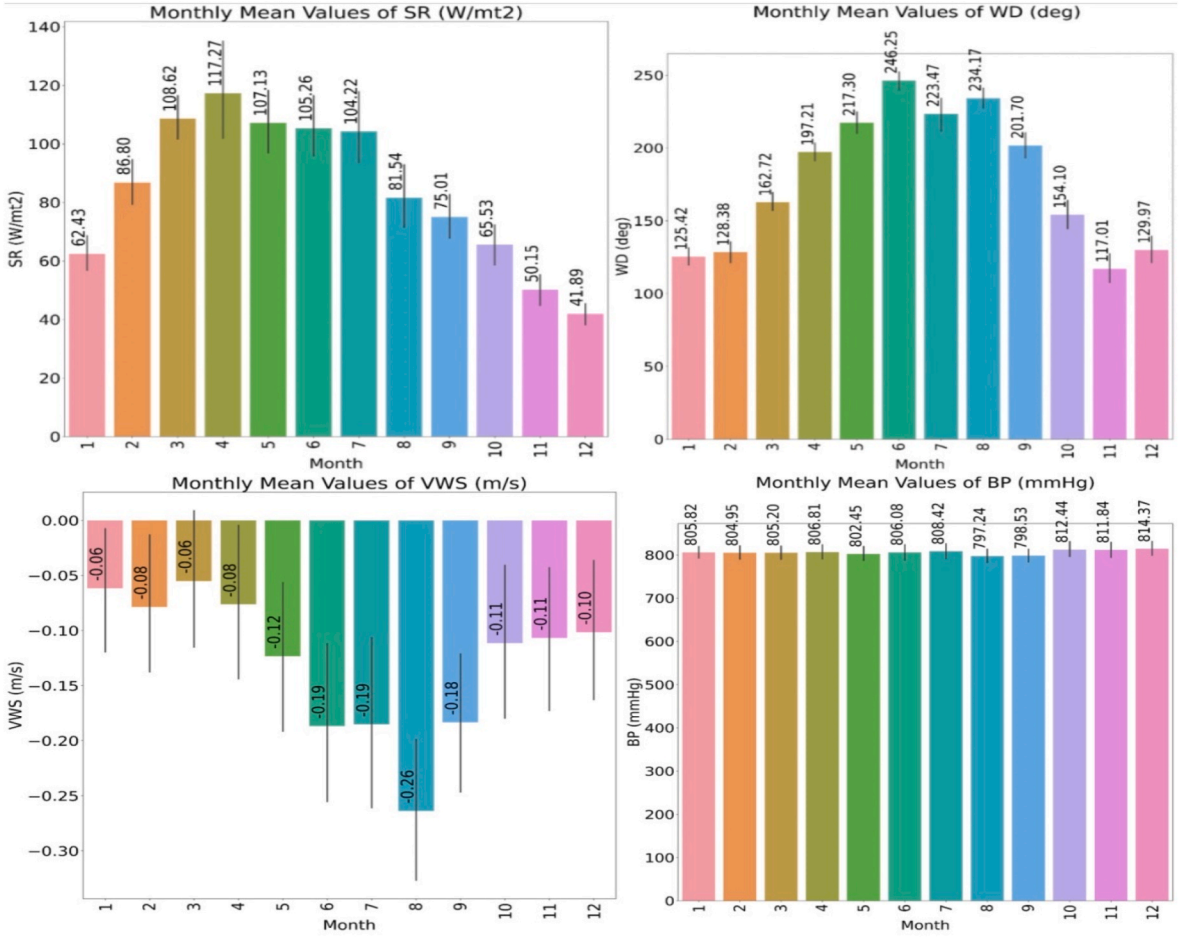

Fig. 7. Monthly and annual variation of meteorological factors.

with a very low wind speed of 0.06 m/s recorded in January and March. The negative values of the wind speed indicate that the wind is coming from south to north. When these wind speed results are compared to the AQI results (Fig. 4), it is clear that the AQI is highest in January with an AQI value of 104.66 and lowest in August with an AQI value of 63.87. This interpretation leads to the conclusion that wind speed plays a major role in the transport of pollutants and affects the AQI of a locality/city.

Similarly, wind direction has a significant impact on a location's AQI. Fig. 8 illustrates the wind rose diagram for the study area based on the datasets available from 2017 to 2022. A wind rose diagram has been developed based on wind direction, and there are 16 different directions based on the degree of directions (Kilabanur et al., 2022). The result shows that from November to February, the wind direction ranges from 116 to 129°, indicating that the wind is from the southeast (SE). March and October experienced wind from the South-South East (S/SE), April and September experienced wind from the South-South West (S/SW), and May to August experienced wind from the South West (SW). According to the results, whenever the city experienced a wind direction from the southeast, the AQI ranges were very high, whereas when the city experienced a wind direction from the southwest, the pollution was very low. South-west winds have high speeds ranging from 0.12 to 0.16 m/s, which can result in the dispersion of air pollutants. South-west winds blow from the sea (Bay of Bengal) to the land, bringing with them a lot of moist air and resulting in heavy rainfall. Almost 35% of the yearly rainfall will be received during the southwest monsoon period, resulting in very low levels of air pollutants in the atmosphere (Bose and Roy Chowdhury, 2023).

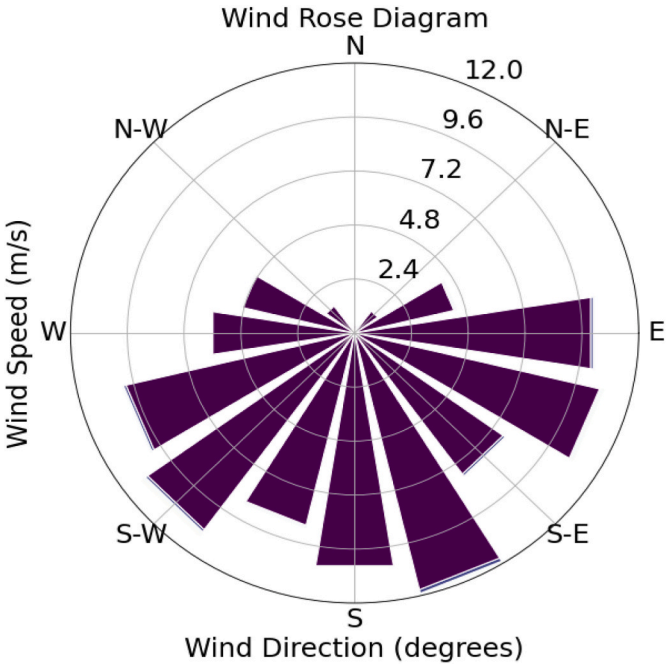

Fig. 8. Wind Rose diagram of Chennai city based on the historical data (2017–2022).

3.5. AQI predictions using machine learning models

Cross-validation is employed during the optimization process to obtain the optimum parameters of the model across multiple subsets of the data. By training and evaluating the model on different folds of the data, a more robust estimate of its generalization ability is obtained. The model consistently demonstrates strong performance across all folds, which indicates that it is not overfitting specific portions of the data (Sahner and Spellmeyer, 2020). Figs. 9–11 depict the normal distribution, residual error, and residual histograms of the various models developed using the available datasets. Table 3 summarizes the performance of the various models in AQI prediction, which was based on 80% of the training datasets and 20% of the testing datasets. MAE, MSE, RMSE, and R^2 are important metrics for evaluating the performance and accuracy of the developed models (Bao and Zhang, 2020; Wu et al., 2019). R^2 is commonly used to compare the best fit and however, it is always necessary to compare different model errors as well as R^2 to determine the best-fit model (Oswalt Manoj et al., 2022).

According to Table 5, XGBoost has a high R^2 value of 0.9985 with very low errors of 0.0262 (MAE), 0.0017 (MSE), and 0.0420 (RMSE) for training data sets. Similarly, R^2 is 0.9252, with errors of 0.057 (MAE), 0.0225 (MSE), and 0.1501 (RMSE) for testing datasets, and it outperforms all other models. In terms of specific features in the data, XGBoost has been able to capture the relationships between the predictors (such as particulate matter, gaseous pollutants, and meteorological parameters) and the target variable (AQI) more effectively. It could have identified important interaction effects or non-linear dependencies that other models might have missed. The models' performance and predictions will apply to scenarios where the factors influencing air pollution and the composition of pollutants are similar to those observed in Chennai during the study period. Different air pollution scenarios, such as unique emission sources or distinct pollutant profiles, may require specific adjustments or retraining of the models to ensure their validity.

3.6. Impact of air pollutants on climate change and human health

Air pollution not only causes a serious threat to human health but also to climate change. The most common air pollutants that harm human health and climate change are CO_2 and methane and these two air pollutants result in global warming. Due to global warming, the earth is facing several adverse impacts namely a rise in the earth's average temperature, the disappearance of glaciers, rising sea levels, increased flooding and natural disasters. The average temperature across global surfaces in 2022 was 1.55 °F (0.86 °C) higher than the 20th-century average of 57.0 °F (13.9 °C) (U.S. Global Change Research Program, 2018). Some mitigation measures need to be adopted by every sector to reduce the carbon emissions or footprint since CO_2 is considered the major GHG. Carbon footprint is considered one of the tools that can make society understand carbon emissions and develop a carbon capture technology. If the carbon footprint is reduced, definitely society can combat climate change. From the study, it is clear that the AQI of Chennai City has on increasing for the last three years and as per the data, in 2022 the AQI was around 85 and now it is in the range of

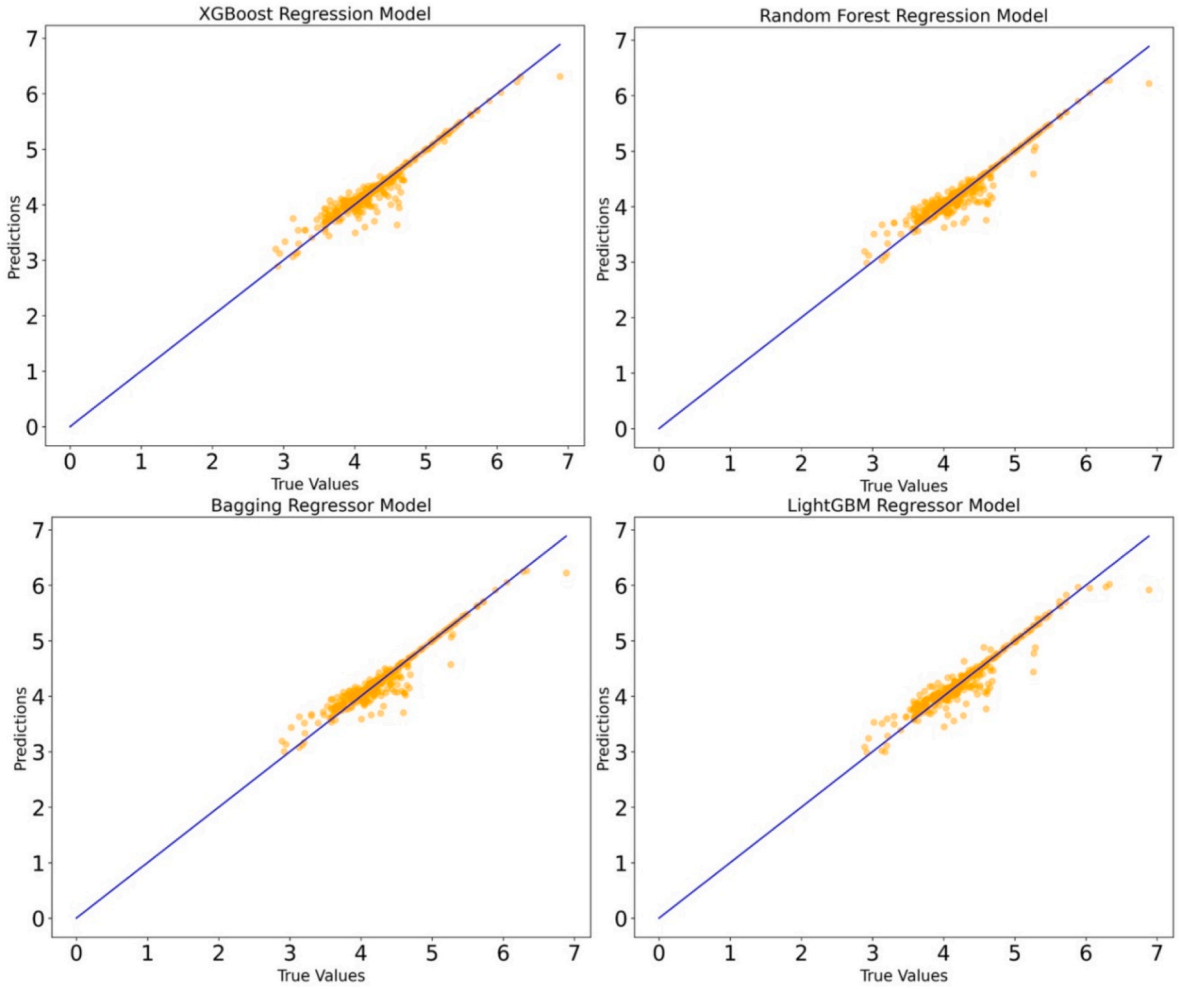

Fig. 9. Normal distribution of datasets of different predictive models.

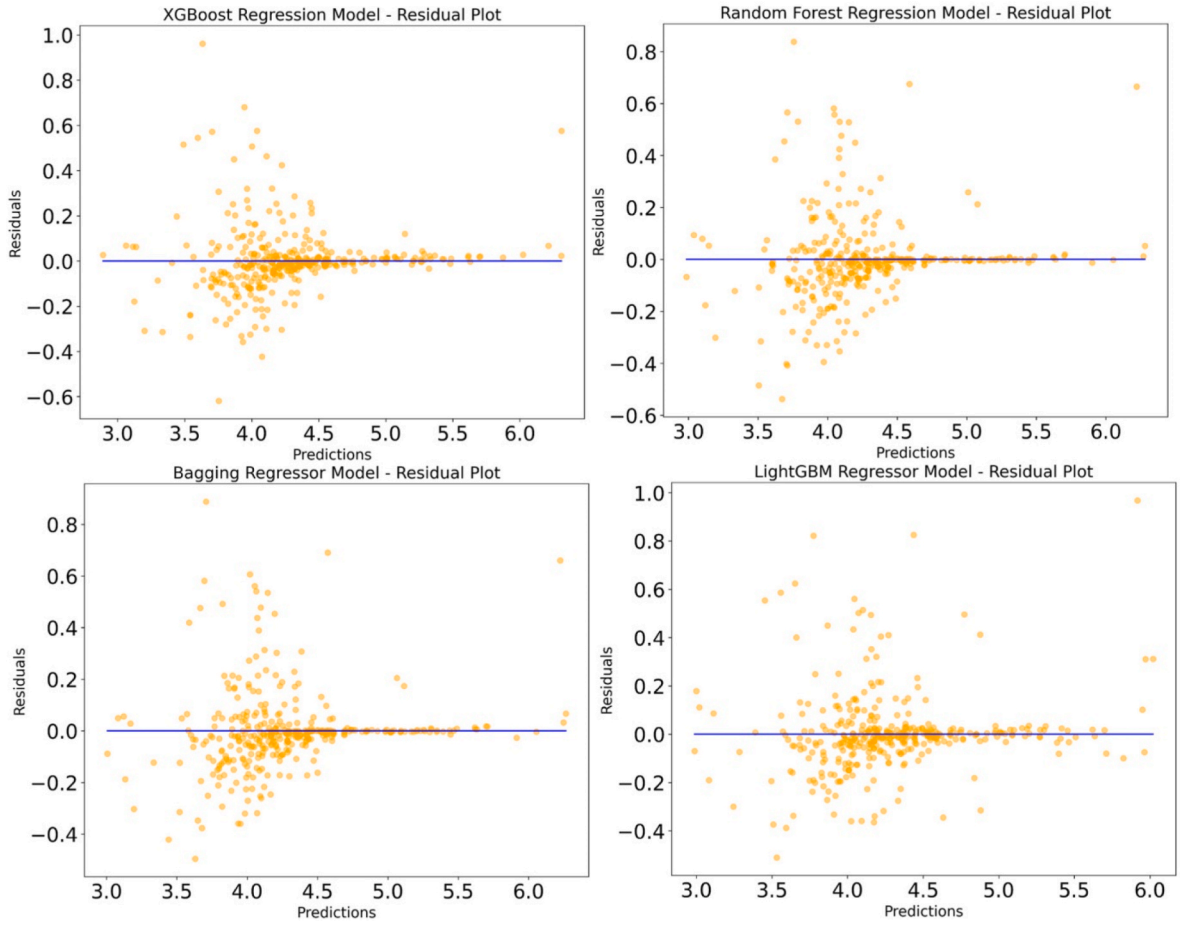

Fig. 10. Residual Plots of different predictive models.

satisfactory. In the near future, if the AQI values are crossing 100, it will have a moderate health impact on humans and breathing discomfort will be caused. Globally air pollution is the third major reason for the maximum death next to high blood pressure and smoking. In 2019, around 6.71 million people died due to asthma, lung infection and respiratory tract infection and all these diseases occur due to poor air. In India, air pollution ranks first among all other diseases and around 1.67 million people died in the year 2019. So, India is in an emerging situation to control the air pollutants and this can be achieved by predicting the AQI and adapting mitigation measures to control air pollution.

3.7. Global initiatives towards air pollution

Climate change mitigation techniques include dramatically reducing GHG emissions, reducing deforestation, and increasing forest cover. GHG reduction can be accomplished by reducing the use of coal and fossil fuels while increasing the use of green energy such as solar, wind and hydropower, as well as biogas from biomass, biofuels, natural gas and hydrogen gas. (Perera, 2018; Wang et al., 2021). GHGs can also be lowered by altering existing equipment and implementing decarbonization processes. Climate change and Air Pollution is regarded as global concern, and mitigation measures implemented by a single country or small region will have little positive impact. Global mitigation measures are currently at the diplomatic level, with international, national, government, and some non-governmental organizations (NGOs) active (Chen et al., 2022). According to the 6th IPCC climate change study's emergency report, the United States and numerous European countries have vowed to reduce net zero carbon emissions by 2050. China agreed to cut by the year 2060, and India agreed to reduce by the year 2070. Many countries throughout the world have promised to reduce methane emissions by 30% by 2030 compared to 2020.

4. Conclusion

The current study examined and predicted the AQI for Chennai based on historical data from 2017 to 2022. The year with the highest AQI was 2018, followed by 2017. AQI values were reduced to less than 60 in 2020 due to the nationwide lockdown caused by COVID-19. However, after 2020, the AQI began to rise again, reaching 63 in the year 2021 (partial lockdown due to COVID-19) and then rising to 85 in the year 2022. According to the results, $PM_{2.5}$ played a significant role in AQI prediction, with a correlation of 0.91, while other parameters had a very low correlation of less than 0.5. All other parameters have a negligible impact on AQI predictions. The AQI was accurately predicted using ML models, with XGBoost and Random Forest showing maximum correlations of 0.9935 and 0.9865, respectively, for training datasets. As a result, machine learning algorithms can accurately forecast AQI levels. Future research could concentrate on developing machine learning models for AQI prediction across the country, to create a unique model that predicts AQI for any city/region. In this study, the missing values were initially removed, which could potentially lead to a reduction in the size of the dataset and the loss of valuable information. To overcome these limitations, future research can consider addressing the missing data issue by employing techniques such as imputation or advanced data cleaning methods. The data limitations include the fact that air quality can be influenced by changes in emission regulations and policies, potentially impacting air quality trends. However, the model development process did not account for the impact of these regulatory changes. The integration of these models into pre-existing air quality monitoring systems or policy frameworks may indeed face certain practical

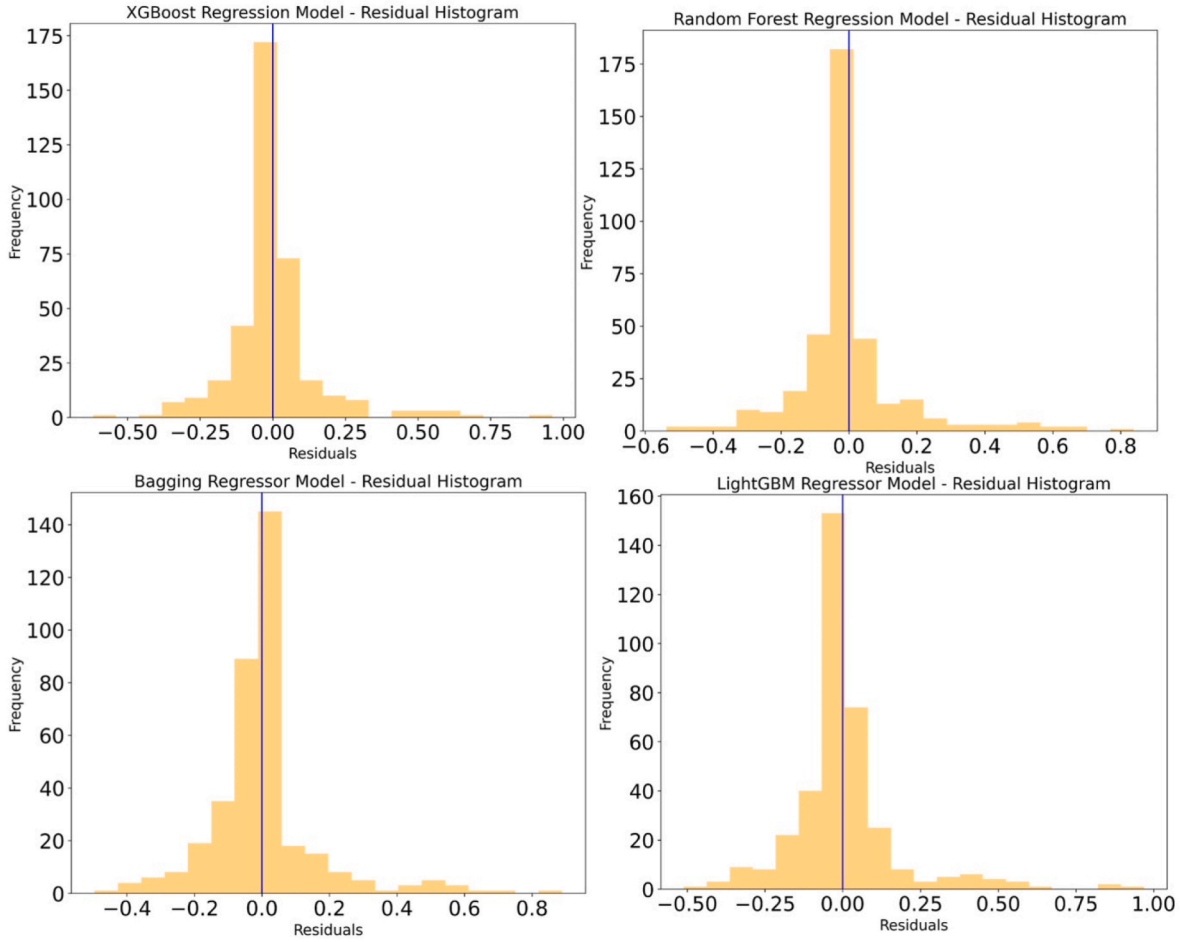

Fig. 11. Residual Histogram of different predictive models.

Table 5
Performance of different ML models in AQI Prediction of Chennai City.

S.No	Model Name	Training				Validation/Testing			
		MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
1	XGBoost	0.0262	0.0017	0.0420	0.9935	0.0857	0.0225	0.1501	0.9252
2	Random Forest (RF)	0.0339	0.0035	0.0598	0.9869	0.0915	0.0256	0.1601	0.9149
3	Bagging Regressor	0.0362	0.0040	0.0637	0.9852	0.09032	0.0251	0.1584	0.9167
4	LightGBM	0.0513	0.0069	0.0830	0.9748	0.0971	0.0285	0.1688	0.9054

considerations. It's crucial to assess the compatibility of existing data with the model's input requirements. In a real-world scenario, if the data is insufficient or of low quality, steps should be taken to improve data collection methods.

CRediT authorship contribution statement

Gokulan Ravindiran: Writing – original draft, Writing – review & editing. **Sivarethinamohan Rajamanickam:** Conceptualization, Writing – review & editing. **Karthick Kanagarathinam:** Writing – original draft, Writing – review & editing. **Gasim Hayder:** Writing – review & editing. **Gorti Janardhan:** Conceptualization, Writing – review & editing. **Arun Kumar Priya:** Conceptualization, Writing – review & editing. **Sivakumar Arunachalam:** Conceptualization, Writing – review & editing. **Abeer A. AlObaid:** Writing – review & editing. **Ismail Warad:** Writing – review & editing. **Senthil Kumar Muniasamy:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by Tenaga Nasional Berhad (TNB) and Universiti Tenaga Nasional (UNITEN) through the BOLD Refresh Postdoctoral Fellowships under the project code of J510050002-IC-6 BOLDREFRESH2025-Centre of Excellence. The authors extend their appreciation to the Researchers Supporting Project number (RSP2023R381), King Saud University, Riyadh, Saudi Arabia.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2023.117354.

References

Abirami, G., Girija, R., Das, A., Sreenivasan, N., 2022. Predicting air quality index with machine learning models. Mach. Learn. Deep Learn. Effic. Improv. Healthc. Syst. 353–371. https://doi.org/10.1201/9781003189053-16.

Ambient (outdoor) air pollution [WWW Document], n.d. URL https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health?gclid=CjwKCAjwgsqoBhBNEiwAwe5w09xdoVlHzRKOjbeMHsO_fMNgXzplczt5fltufPX4Im2TsUjUT9OXLxoCA7IQAvD_BwE (accessed 9.26.23).

Ayus, I., Natarajan, N., Gupta, D., 2023. Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. Asian J. Atmos. Environ. 17, 1–22. https://doi.org/10.1007/S44273-023-00005-W/FIGURES/14.

Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R.S., Brauer, M., Cohen, A.J., Stanaway, J. D., Beig, G., Joshi, T.K., Aggarwal, A.N., Sabde, Y., Sadhu, H., Frostad, J., Causey, K., Godwin, W., Shukla, D.K., Kumar, G.A., Varghese, C.M., Muraleedharan, P., Agrawal, A., Anjana, R.M., Bhansali, A., Bhardwaj, D., Burkart, K., Cercy, K., Chakma, J.K., Chowdhury, S., Christopher, D.J., Dutta, E., Furtado, M., Ghosh, S., Ghoshal, A.G., Glenn, S.D., Guleria, R., Gupta, R., Jeemon, P., Kant, R., Kant, S., Kaur, T., Koul, P.A., Krish, V., Krishna, B., Larson, S.L., Madhipatla, K., Mahesh, P.A., Mohan, V., Mukhopadhyay, S., Mutreja, P., Naik, N., Nair, S., Nguyen, G., Odell, C. M., Pandian, J.D., Prabhakaran, D., Prabhakaran, P., Roy, A., Salvi, S., Sambandam, S., Saraf, D., Sharma, M., Shrivastava, A., Singh, V., Tandon, N., Thomas, N.J., Torre, A., Xavier, D., Yadav, G., Singh, S., Shekhar, C., Vos, T., Dandona, R., Reddy, K.S., Lim, S.S., Murray, C.J.L., Venkatesh, S., Dandona, L., 2019. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. Lancet Planet. Health 3, e26–e39. https://doi.org/10.1016/S2542-5196(18)30261-4.

Bao, R., Zhang, A., 2020. Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Sci. Total Environ. 731, 139052 https://doi.org/10.1016/J.SCITOTENV.2020.139052.

Bekkar, A., Hssina, B., Douzi, S., Douzi, K., 2021. Air-pollution prediction in smart city, deep learning approach. J. Big Data 8, 1–21. https://doi.org/10.1186/S40537-021-00548-1/FIGURES/17.

Bhalgat, P., 2019. Air quality prediction using machine learning algorithms. Int. J. Comput. Appl. Technol. Res. 8, 367–370.

Bodor, Z., Bodor, K., Keresztesi, Á., Szép, R., 2020. Major air pollutants seasonal variation analysis and long-range transport of PM10 in an urban environment with specific climate condition in Transylvania (Romania). Environ. Sci. Pollut. Res. 27, 38181–38199. https://doi.org/10.1007/S11356-020-09838-2/FIGURES/13.

Bose, A., Roy Chowdhury, I., 2023. Investigating the association between air pollutants' concentration and meteorological parameters in a rapidly growing urban center of West Bengal, India: a statistical modeling-based approach. Model. Earth Syst. Environ. 1, 1–16. https://doi.org/10.1007/S40808-022-01670-6/FIGURES/9.

Chandrappa, R., Kulshrestha, U.C., 2016a. Major issues of air pollution. Sustain. Air Pollut. Manag. 143, 1. https://doi.org/10.1007/978-3-319-21596-9_1.

Chandrappa, R., Kulshrestha, U.C., 2016b. Air pollution and disasters. Sustain. Air Pollut. Manag. 143, 325. https://doi.org/10.1007/978-3-319-21596-9_8.

Chen, H., Li, X., Feng, Z., Wang, L., Qin, Y., Skibniewski, M.J., Chen, Z.S., Liu, Y., 2023. Shield attitude prediction based on Bayesian-LGBM machine learning. Inf. Sci. 632, 105–129. https://doi.org/10.1016/J.INS.2023.03.004.

Chen, L., Msigwa, G., Yang, M., Osman, A.I., Fawzy, S., Rooney, D.W., Yap, P.S., 2022. Strategies to achieve a carbon neutral society: a review. Environ. Chem. Lett. 20, 2277–2310. https://doi.org/10.1007/S10311-022-01435-8/METRICS.

Doble, M., Kumar, A., 2005. Gaseous pollutants and volatile organics. Biotreat. Ind. Effluents 301–312. https://doi.org/10.1016/B978-075067838-4/50031-2.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and its implications for data analysis. Shanghai Arch. Psychiatry 26, 105. https://doi.org/10.3969/J.ISSN.1002-0829.2014.02.009.

Feng, C., Wang, H., Lu, N., Tu, X.M., 2013. Log transformation: application and interpretation in biomedical research. Stat. Med. 32, 230–239. https://doi.org/10.1002/SIM.5486.

Gupta, N.S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., Arulkumaran, G., 2023. Prediction of air quality index using machine learning techniques: a comparative analysis. J. Environ. Public Health 2023, 1–26. https://doi.org/10.1155/2023/4916267.

IQAir. First in Air Quality [WWW Document], n.d. URL. https://www.iqair.com/world-air-quality-report, 7.14.23.

Izzotti, A., Spatera, P., Khalid, Z., Pulliero, A., 2022. Importance of punctual monitoring to evaluate the health effects of airborne particulate matter. Int. J. Environ. Res. Publ. Health 19, 10587. https://doi.org/10.3390/IJERPH191710587.

Joseph, V.R., Vakayil, A., 2022. SPlit: an optimal method for data splitting. Technometrics 64, 166–176. https://doi.org/10.1080/00401706.2021.1921037/SUPPL_FILE/UTCH_A_1921037_SM8231.PDF.

Khillare, P.S., Sarkar, S., 2012. Airborne inhalable metals in residential areas of Delhi, India: distribution, source apportionment and health risks. Atmos. Pollut. Res. 3, 46–54. https://doi.org/10.5094/APR.2012.004.

Kilabanur, P., Dharek, M.S., Sunagar, P., Ya, al, Tyukalov, Y., Gopi, R., Saravanakumar, R., Elango, K.S., Chandrasekar, A., Navaneethan, K.S., Gopal, N., 2022. Construction Emission Management Using Wind Rose Plot and AERMOD Application You May Also like Enhancing Index and Strength Properties of Black Cotton Soil Using Combination of Geopolymer and Flyash Calculation of the Circular Plates' Stability in Stresses Construction Emission Management Using Wind Rose Plot and AERMOD Application. https://doi.org/10.1088/1757-899X/1145/1/012225.

Krishna, K.R., Beig, G., Krishna, K.R., Beig, G., 2018. Influence of meteorology on particulate matter (PM) and vice-versa over two Indian metropolitan cities. Open J. Air Pollut. 7, 244–262. https://doi.org/10.4236/OJAP.2018.73012.

Kumar, K., Pande, B.P., 2022. Air pollution prediction with machine learning: a case study of Indian cities. Int. J. Environ. Sci. Technol. 1–16. https://doi.org/10.1007/S13762-022-04241-5/TABLES/7.

Liang, Y.C., Maimury, Y., Chen, A.H.L., Juarez, J.R.C., 2020. Machine learning-based prediction of air quality. Appl. Sci. 2020 10. https://doi.org/10.3390/APP10249151, 9151 10, 9151.

Lu, X., Zhang, L., Shen, L., 2021. Tropospheric ozone interacts with weather and climate. Air Pollution, Clim. Heal. An Integr. Perspect. Their Interact. 15–46. https://doi.org/10.1016/B978-0-12-820123-7.00006-1.

Madan, T., Sagar, S., Virmani, D., 2020. Air quality prediction using machine learning algorithms-A review. In: Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN. https://doi.org/10.1109/ICACCCN51052.2020.9362912, 2020 140–145.

Mahesh, T.R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H.K., Swapna, B., Guluwadi, S., 2022. Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. J. Sensor 2022. https://doi.org/10.1155/2022/4649510.

Maltare, N.N., Vahora, S., 2023. Air Quality Index prediction using machine learning for Ahmedabad city. Digit. Chem. Eng. 7, 100093 https://doi.org/10.1016/J.DCHE.2023.100093.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020a. Environmental and health impacts of air pollution: a review. Front. Public Health 8, 14. https://doi.org/10.3389/FPUBH.2020.00014.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E., 2020b. Environmental and health impacts of air pollution: a review. Front. Public Health 8, 14. https://doi.org/10.3389/FPUBH.2020.00014.

Mehta, Y., Rajan, J., 2017. Peer-review under responsibility of the organizing committee of the 13th global congress on manufacturing and management. NC-ND license Procedia Eng. 174, 90–104. https://doi.org/10.1016/j.proeng.2017.01.173. http://creativecommons.org/licenses/by-nc-nd/4.0/. http://creativecommons.org/licenses/by-nc-nd/4.0/. Peer-review under responsibility of the organizing committee of the 13th Global Congress on Manufacturing and Management ScienceDirect 2016 Global Congress on Manufacturing and Management Manufacturing Sectors in India: Outlook and Challenges-NC-ND license.

Oswalt Manoj, S., Ananth, J.P., Rohini, M., Dhanka, B., Pooranam, N., Ram Arumugam, S., 2022. FWS-DL: forecasting wind speed based on deep learning algorithms. Artif. Intell. Renew. Energy Syst. 353–374. https://doi.org/10.1016/B978-0-323-90396-7.00007-9.

Perera, F., 2018. Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: solutions exist. Int. J. Environ. Res. Publ. Health 15. https://doi.org/10.3390/IJERPH15010016.

Pérez-Rodríguez, J., Fernández-Navarro, F., Ashley, T., 2023. Estimating ensemble weights for bagging regressors based on the mean–variance portfolio framework. Expert Syst. Appl. 229, 120462 https://doi.org/10.1016/J.ESWA.2023.120462.

Ramírez, A.S., Ramondt, S., Van Bogart, K., Perez-Zuniga, R., 2019. Public awareness of air pollution and health threats: challenges and opportunities for communication strategies to improve environmental health literacy. J. Health Commun. 24, 75. https://doi.org/10.1080/10810730.2019.1574320.

Ravindra, K., 2019. Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. Environ. Int. 122, 201–212. https://doi.org/10.1016/J.ENVINT.2018.11.008.

Ravindra, K., Singh, T., Pandey, V., Mor, S., 2020. Air pollution trend in Chandigarh city situated in Indo-Gangetic Plains: understanding seasonality and impact of mitigation strategies. Sci. Total Environ. 729, 138717 https://doi.org/10.1016/J.SCITOTENV.2020.138717.

Rybarczyk, Y., Zalakeviciute, R., Rybarczyk, Y., Zalakeviciute, R., 2017. Regression models to predict air pollution from affordable data collections. Mach. Learn. - Adv. Tech. Emerg. Appl. https://doi.org/10.5772/INTECHOPEN.71848.

Sahner, D., Spellmeyer, D.C., 2020. Artificial intelligence: emerging applications in biotechnology and pharma. Biotechnol. Entrep. Leading, Manag. Commer. Innov. Technol. 399–417. https://doi.org/10.1016/B978-0-12-815585-1.00028-0.

Saravanan, S.P., Desmet, M., Kanniperumal, A.N.P., Ramasamy, S., Shumskikh, N., Grosbois, C., 2019. Geochemical footprint of megacities on river sediments: a case study of the fourth most populous area in India, Chennai. Miner. 2019 9, 688. https://doi.org/10.3390/MIN9110688. Page 688 9.

Sekeroglu, B., Kirsal Ever, Y., Dimililer, K., Al-Turjman, F., 2022. Comparative evaluation and comprehensive analysis of machine learning models for regression problems under a creative commons attribution 4.0 international (CC BY 4.0) license. Comparative evaluation and comprehensive analysis of machine learning models for regression problems. Probl. Data Intell. 4, 620–652. https://doi.org/10.1162/dint_a_00155.

Shelton, S., Liyanage, G., Jayasekara, S., Pushpawela, B., Rathnayake, U., Jayasundara, A., Jayasooriya, L.D., 2022. Seasonal variability of air pollutants and their relationships to meteorological parameters in an urban environment. Adv. Meteorol. 2022, 1–18. https://doi.org/10.1155/2022/5628911.

Singh, R.P., Chauhan, A., 2020. Impact of lockdown on air quality in India during COVID-19 pandemic. Air Qual. Atmos. Heal. 13, 921–928. https://doi.org/10.1007/S11869-020-00863-1/FIGURES/5.

Singh, V., Singh, S., Biswal, A., 2021a. Exceedances and trends of particulate matter (PM2.5) in five Indian megacities. Sci. Total Environ. 750, 141461 https://doi.org/10.1016/J.SCITOTENV.2020.141461.

Singh, V., Singh, S., Biswal, A., 2021b. Exceedances and trends of particulate matter (PM2.5) in five Indian megacities. Sci. Total Environ. 750, 141461 https://doi.org/10.1016/J.SCITOTENV.2020.141461.

U.S. Global Change Research Program, 2018. Climate science special report: fourth national climate assessment, volume I. U.S. Glob. Chang. Res. Progr. 1, 470. https://doi.org/10.7930/J0J964J6.

Villanueva, F., Notario, A., Tapia, A., Albaladejo, J., Cabañas, B., Martínez, E., 2016. Ambient levels of volatile organic compounds and criteria pollutants in the most industrialized area of central Iberian Peninsula: intercomparison with an urban site. Environ. Technol. 37, 983–996. https://doi.org/10.1080/09593330.2015.1096309.

Wang, Fang, Harindintwali, J.D., Yuan, Zhizhang, Wang, M., Wang, Faming, Li, S., Yin, Z., Huang, L., Fu, Y., Li, L., Chang, S.X., Zhang, L., Rinklebe, J., Yuan, Zuoqiang, Zhu, Q., Xiang, L., Tsang, D.C.W., Xu, L., Jiang, X., Liu, J., Wei, N., Kästner, M., Zou, Y., Ok, Y.S., Shen, J., Peng, D., Zhang, W., Barceló, D., Zhou, Y., Bai, Z., Li, B., Zhang, B., Wei, K., Cao, H., Tan, Z., Zhao, L. bin, He, X., Zheng, J., Bolan, N., Liu, X., Huang, C., Dietmann, S., Luo, M., Sun, N., Gong, J., Gong, Y., Brahushi, F., Zhang, T., Xiao, C., Li, X., Chen, W., Jiao, N., Lehmann, J., Zhu, Y.G., Jin, H., Schäffer, A., Tiedje, J.M., Chen, J.M., 2021. Technologies and perspectives for achieving carbon neutrality. Innov 2, 100180. https://doi.org/10.1016/J.XINN.2021.100180.

Wang, J., Li, X., Jin, L., Li, J., Sun, Q., Wang, H., 2022. An air quality index prediction model based on CNN-ILSTM. Sci. Rep. 12 https://doi.org/10.1038/S41598-022-12355-6.

Wang, S., Aggarwal, C., Liu, H., 2018. Random-forest-inspired neural networks. ACM Trans. Intell. Syst. Technol. 9 https://doi.org/10.1145/3232230.

Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H., 2019. Hyperparameter optimization for machine learning models based on bayesian optimization. J. Electron. Sci. Technol. 17, 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120.

Xia, X., Zhang, K., Yang, R., Zhang, Y., Xu, D., Bai, K., Guo, J., 2022. Impact of near-surface turbulence on PM2.5 concentration in Chengdu during the COVID-19 pandemic. Atmos. Environ. 268, 118848 https://doi.org/10.1016/J.ATMOSENV.2021.118848, 1994.