Chemosphere 338 (2023) 139518

Contents lists available at ScienceDirect

Chemosphere

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

Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam

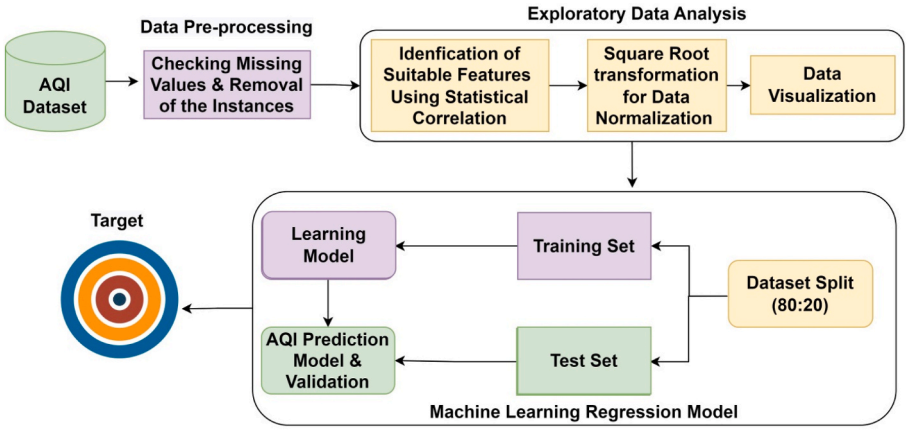
Gokulan Ravindiran [a,b,**], Gasim Hayder [a,c], Karthick Kanagarathinam [d], Avinash Alagumalai [e], Christian Sonne [f,g,*]

[a] Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), Selangor Darul Ehsan, Kajang, 43000, Malaysia
[b] Department of Civil Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, 500090, Telangana, India
[c] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Selangor Darul Ehsan, Kajang, 43000, Malaysia
[d] Department of Electrical and Electronics Engineering, GMR Institute of Technology, Rajam, 532 127, Andhra Pradesh, India
[e] Department of Chemical and Petroleum Engineering, University of Calgary, Calgary, Canada
[f] Aarhus University, Faculty of Technical Sciences, Department of Ecoscience, DK-4000, Roskilde, Denmark
[g] Cluster, School of Engineering, University of Petroleum & Energy Studies, Dehradun, Uttarakhand, 248007, India

HIGHLIGHTS

- We used machine learning models to predict Air Quality Index (AQI).
- Particulate matter, gaseous pollutants and metrological factors were used.
- Meteorological factors contribution in AQI prediction is found negligible.
- Catboost model yielded high prediction accuracy (0.9998) and low RMSE (0.76).
- Using historical data and advanced machine learning assist predictions on air quality.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Volker Matthias

Keywords:
Air quality index
Particulate matter
Gaseous pollutants

ABSTRACT

Clean air is critical component for health and survival of human and wildlife, as atmospheric pollution is associated with a number of significant diseases including cancer. However, due to rapid industrialization and population growth, activities such as transportation, household, agricultural, and industrial processes contribute to air pollution. As a result, air pollution has become a significant problem in many cities, especially in emerging countries like India. To maintain ambient air quality, regular monitoring and forecasting of air pollution is necessary. For that purpose, machine learning has emerged as a promising technique for predicting the Air

* Corresponding author. Aarhus University, Faculty of Technical Sciences, Department of Ecoscience, DK-4000, Roskilde, Denmark.
** Corresponding author. Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), 43000, Kajang, Selangor Darul Ehsan, Malaysia.
E-mail addresses: gokulravi4455@gmail.com (G. Ravindiran), gasim@uniten.edu.my (G. Hayder), karthick.k@gmrit.edu.in (K. Kanagarathinam), avinashandromeda@gmail.com (A. Alagumalai), cs@ecos.au.dk (C. Sonne).

https://doi.org/10.1016/j.chemosphere.2023.139518
Received 10 May 2023; Received in revised form 5 July 2023; Accepted 14 July 2023
Available online 14 July 2023

Meteorological parameters
Climate action

Quality Index (AQI) compared to conventional methods. Here we apply the AQI to the city of Visakhapatnam, Andhra Pradesh, India, focusing on 12 contaminants and 10 meteorological parameters from July 2017 to September 2022. For this purpose, we employed several machine learning models, including LightGBM, Random Forest, Catboost, Adaboost, and XGBoost. The results show that the Catboost model outperformed other models with an R^2 correlation coefficient of 0.9998, a mean absolute error (MAE) of 0.60, a mean square error (MSE) of 0.58, and a root mean square error (RMSE) of 0.76. The Adaboost model had the least effective prediction with an R^2 correlation coefficient of 0.9753. In summary, machine learning is a promising technique for predicting AQI with Catboost being the best-performing model for AQI prediction. Moreover, by leveraging historical data and machine learning algorithms enables accurate predictions of future urban air quality levels on a global scale.

1. Introduction

Population density, industrial activity, agricultural practices, thermal power plants, energy sectors, automotive industries, and transportation all affect air pollution differently (Ravindra, 2019; Ravindra et al., 2020). Aside from harming ecosystems, air pollution also has negative effects on human health, including premature death, skin rashes, lung infections, respiratory tract infections, pneumonia, lung cancer, and heart failure (Manisalidis et al., 2020). Several important factors determine air pollution in a given area including particulate matter, gaseous pollutants and metrological factors, which require AQI estimation by a large number of government and non-governmental organizations worldwide (Bao and Zhang, 2020; Li et al., 2020). Among other, particulate Matter 2.5 ($PM_{2.5}$), Particulate Matter 10 (PM_{10}), Carbon dioxide (CO_2), Carbon monoxide (CO), Sulfur Oxides (SO_X), Nitrogen Oxides (NO_X), Ozone (O_3), and Ammonia (NH_3) are key contributors to AQI. An increased AQI due to these substances have a negative impact on the environment in numerous ways, including global warming, acid rain, the development of smog and aerosols, decreased visibility, and climate change (Balakrishnan et al., 2019).

Greenhouse gases (GHG) are primarily responsible for global warming, and these GHGs have an impact on plant-soil interactions in addition to climate change that have significant negative impacts on agriculture, the environment and also economy (Malhi et al., 2021). Based on information from 2010 to 2019, the World Health Organization (WHO) published a report on worldwide air quality in 2022. This study examined a wide range of air pollutants mentioned above and discovered that $PM_{2.5}$ was rising globally based on studies of 6743 cities in 117 different nations, causing 1.7 million yearly mortalities in India alone. The top 20 cities with the highest air pollution were 18 Indian cities, illustrating the seriousness of India's air pollution and adverse health effects in the years to come (Gurjar et al., 2016; Guttikunda et al., 2014).

A high AQI number denotes the most hazardous environment for people, and life is in danger. Consequently, AQI monitoring and forecasting have turned into a crucial tool for international sustainable development (Rybarczyk and Zalakeviciute, 2021). On the basis of statistical, deterministic, physics, machine learning, and deep learning, numerous researchers have created models for AQI prediction. The rigidity of statistical and decision-making models makes them unsuitable for dealing with complex issues. Recent sensor developments have made it simple to identify different air pollution levels, and AQI is automatically calculated. Using the readily available data sets, the AQI forecast is straightforward (Bekkar et al., 2021). The machine learning approach is quite precise and consistently predicts the AQI under all environmental circumstances. Machine learning enables us to produce forecasts of the AQI that are more precise because to the growing amount of historical data that is accessible for research. They are becoming more prevalent in an effort to establish themselves as a viable alternative to established statistical models for time-series forecasting. It is extremely difficult to create a statistical model that can forecast such events due to the highly nonlinear processes involving pollutant concentrations and their incompletely understood dynamics. Machine learning models are an illustration of nonparametric and nonlinear models that use just historical data to determine the correlation between the independent variables, allowing us to create a prediction model that is more accurate.

Visakhapatnam, the capital of Andhra Pradesh and a city on India's east coast also known as Vizag. It is renowned as a port city and houses the administrative centre of the Indian Eastern Navy Command. Weather variables, including wind speed, help the city reduce air pollution naturally during the summer. However, due to the city's low height and seasonal temperature inversion, air pollution becomes a significant issue over the winter (Sumiya et al., 2023). For the state of Andhra Pradesh, the city is regarded as the industrial hub for various industries, including fisheries, shipbuilding, textiles, pharmaceuticals, medical technology, aluminium, and ferrous metals. There is also a 1000 MW thermal power plant located nearby. Two of the most significant naturally existing minerals that contributed to the growth of the heavy manufacturing sectors are manganese and bauxite reserves. Based on historical AQI values and the city's economic growth, Visakhapatnam is identified as one of the emerging polluted cities in Andhra Pradesh, India. In this study, we analyse open-source data from the Central Pollution Control Board (CPCB), covering the period from January 2017 to December 2022. To predict the AQI, we employed five machine-learning algorithms including LightGBM, Random Forest, Catboost, Adaboost, and XGboost. These models demonstrate strong performance across diverse datasets making the selection based on factors such as dataset characteristics, problem requirements, and desired outcomes. LightGBM effectively handles categorical features, offering fast training speed and competitive performance while Random Forest handles both numerical and categorical features and provides feature importance rankings. CatBoost is renowned for its strong performance and generalization ability and AdaBoost adjusts instance weights iteratively to focus on difficult cases, enhancing overall performance, while XGBoost offers regularization techniques for controlling overfitting and allows for flexible model customization. Considering these advantages, we used these models for air quality prediction.

2. Materials and methods

2.1. Study area

The current study examined the air quality index for the Indian state of Andhra Pradesh's Visakhapatnam city. Greater Visakhapatnam Municipal Corporation (GVMC) is site to an air quality monitoring station for the CPCB of India. North Eastern Andhra Pradesh's coastal district of Visakhapatnam is located between 170° 41′ and 170° 59° North latitude and 830° 12′ and 830° 27′ East longitude. With a 550 km^2 size, it is the biggest city in Andhra Pradesh and mostly an industrial one. India's Eastern Naval Command also calls it home. One of the 100 cities with the fastest growth rates worldwide and one of India's top ten richest cities is Visakhapatnam (Fig. 1).

2.2. Air quality and meteorological datasets

The data was gathered from the Central Pollution Board - Central Control Room for Air (CPCBCRR). To compute the AQI as recommended by the CPCB, India, air pollutants including $PM_{2.5}$, PM_{10}, NO_X, NH_3, SO_X,

CO, O_3, Benzene, Toluene, and Xylene were measured. Measurements of any one particulate matter and any three gaseous pollutants are used to construct the AQI (Maximum value of any three Pollutants). AQI is influenced by local weather conditions in addition to air pollutants in a given area. Temperature, Relative Humidity (RH), Wind Speed (WS), Wind Direction (WD), Solar Radiation (SR), Air Pressure (BP), Ambient Temperature (AT), Rainfall (RF) and Total Rainfall (TOT-RF) were among the meteorological parameters that were monitored. A total of 1920 observations covering the period from 01– 07–2017 to 30–09–2022 were discovered in the data sets. In order to conduct the study, a raw dataset of 1920 occurrences containing 23 components (12 Air Pollutants, 10 Metrological Factors, and 1 AQI) was employed. The variable that is being targeted is the AQI. The AQI data sets were presented in Table S1.

2.3. Machine learning methods to predict AQI

To forecast the AQI of the proposed city, machine learning models including LightGBM (Light Gradient Boosting Machine), Random Forest , Catboost, Adaboost, and XGboost were utilised. The structure of the datasets utilised to estimate the AQI was shown in Fig. 2.

2.3.1. LightGBM

A trustworthy tool for implementing gradient boosting in decision trees is LightGBM (Zhou et al., 2022). LightGBM uses tree-based learning strategies, making it a suitable choice for gradient boosting. It provides quicker training and higher production thanks to its decentralised and effective architecture. A histogram-based method called LightGBM conducts variable bucketing, which improves training speed and accuracy while using less memory. When being trained, it works more quickly and can handle large and complex datasets. Support for both parallel and GPU-based learning. The method enables one to infer details about a target Y given only X as input when used in supervised learning scenarios. The LightGBM technique uses a supervised training set (X) and a loss function L (y,f(x)) whose predicted value is to be reduced to accomplish this $\widehat{f}$ (x).

$$\widehat{f} = \arg\ min_f\ E_{y,X}\ L\ (y, f(x)) \tag{1}$$

2.3.2. Random forest

Random forest regression is one form of machine learning-based regression technique (Ganesh et al., 2021). The foundation of this tactic is a combination of bagging and random subspace approaches. Following the bagging of numerous learning trees, an ensemble method is used to merge these trees into a single forecast. From the entire collection of N examples used for training (D), n random instances are chosen to create a bootstrap sample (D_b). It is acceptable to substitute examples when creating bootstrap samples. The input vector x is used to create K distinct regression trees for the bootstrap samples. When performing regression tasks, the random forest prediction is made by averaging K predictions from regression trees h_k (x).

$$Random\ foreset\ prediction = \frac{1}{K}\sum\nolimits_{K=1}^{K} h_K(x) \tag{2}$$

2.3.3. CatBoost

A development of the gradient boosting and decision tree frameworks is CatBoost (Zhang et al., 2020). Boosting is predicated on the idea that numerous, relatively weak models can be joined to create a single, fiercely competitive prediction model that outperforms chance by a slight margin. Gradient boosting decreases errors by fitting a series of decision trees, each of which gains knowledge from the mistakes of the preceding iteration. This process of introducing new functions into the mix is repeated until the chosen loss function is no longer minimised.

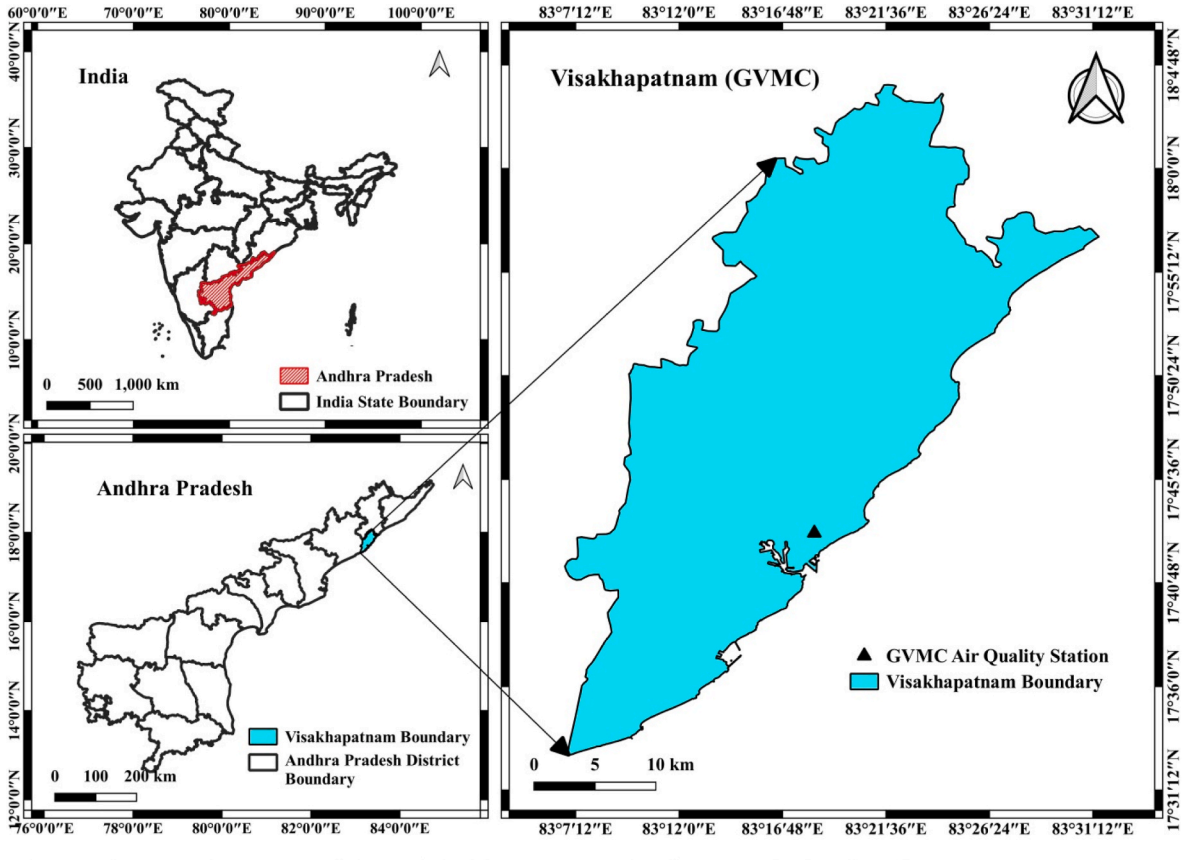


Fig. 1. Location map of the Visakhapatnam, Andhra Pradesh of India.

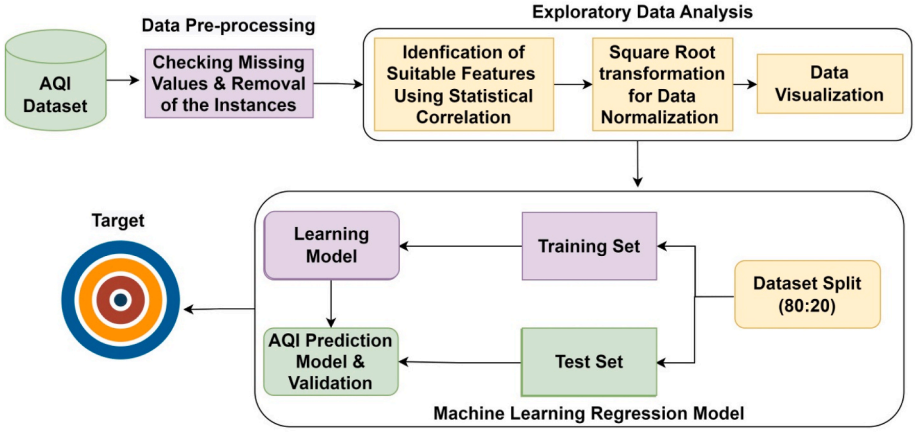


Fig. 2. Machine learning flowchart for the prediction of AQI.

CatBoost's method for creating decision trees differs from traditional gradient boosting models. Instead, CatBoost creates "oblivious trees," which are made possible by requiring that all nodes at the same level test the same predictor under the same conditions. This allows for the computation of a leaf's index using just bitwise operations.

By arbitrary ordering the components of D using a random permutation, σ, CatBoost chooses the data to be utilised for fitting h^{t+1}. $D_k = \{x_1, x_2, \ldots, x_{k-1}\}$ where $x_1, x_2, \ldots, x_{k-1}$ are the elements of D arranged by the random permutation and (k) is the kth element of D under permutation,σ. Instead of strictly adhering to Eq. (3), CatBoost uses a variation of it in its analysis to define the encoded value, $\widehat{x}_k^i$ for the ith categorical value during Decision Tree h^{t+1} fitting.

$$\widehat{x}_k^i = \frac{\sum x_j \in D_k 1\ x_j^i = x_k^i.\ y_j + ap}{\sum x_j \in D_k 1\ x_j^i = x_k^i + a} \tag{3}$$

Here $1\ x_j^i = x_k^i$ is the indicator function.

2.3.4. Adaptive boosting (AdaBoost) regressor

AdaBoost (Mishra et al., 2020) one of the first boosting algorithms employed, was used to address difficulties. AdaBoost modifies the data of each training sample (x_i, y_i) by applying a weight $w_1, w_2, \ldots, w_N$. The fundamental learner gives each observation the same amount of thought in the initial stage. Once weights have been assigned to each observation, the weak learner may be utilised for prediction. In this method, the predictions made by the base learner after the weak learner are more likely to be accurate. This process will be carried out again until the tth iteration, after which the T_i base learning algorithm's limit will be reached. The outputs of weak learners can be combined to generate more robust learners, which enhance the ability to predict outcomes.

2.3.5. Extreme gradient boosting (XGBoost)

Boosting is a technique for combining numerous weak classifiers into a single effective one. Gradient Boosting served as the basis for the development of the technique known as XGBoost (Mahesh et al., 2022). Gradient Boosting's XGBoost variation outperforms the original in terms of computing efficiency, scalability, and generalisation performance. Data organization is of utmost importance while utilising XGBoost. All category data will be transformed into their numeric equivalents because XGBoost only takes numeric vectors as input. This encoding change can be made with a single hot encoding. The feature engineering and data purification stages come next. We can obtain the estimated model by using the universal function, as indicated by the following formula:

$$\widehat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

where,

$\widehat{y}_i^t$ = forecasts at the stage t
$f_t(x_i)$ = a learner at stage t
x_i = the input variable
$\widehat{y}_i^{(t-1)}$ = forecasts at the stage t-1

2.4. Data pre-processing

A total of 1920 instances were recorded from the open source data collected from January 2017 to December 2022. The characteristics' lacking values were initially erased. After the missing values are taken out, there are 1675 total instances with 25 characteristics. On AQI, a type conversion from object to float data type has been performed. The statistical data from the datasets for air pollutants and meteorological parameters are summarised in Tables S3 and S4. The performance evaluation of the various models was conducted by splitting the available dataset into a training set (representing 80% of the data) and a testing set (representing 20% of the data). The models' ability to forecast AQI assessed using evaluation metrics such as root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), and R^2 as widely used in regression models when evaluating model accuracy and performance from datasets.

2.5. Exploratory data analysis

Exploratory data analysis will be used to uncover hidden patterns that are typically present in the datasets. It is necessary to perform exploratory data analysis before using machine learning methods. Exploratory data analysis is performed to establish the connection between different air contaminants that lead to the highest AQI (Langer and Meisen, 2021). Additionally, it is utilised to examine the status and trends of several air contaminants from 2017 to 2022. Additionally, the pollutants will be categorised according to the greatest effect any one air pollutant can have on the AQI. The heatmap for the above mentioned parameters (dataset) is shown in Figure S1. The correlation matrix method, which establishes the size of the relationship between several variables, was used to evaluate the heat map. Scores for the Heat Map will range from +1 to −1 (Li et al., 2016). The directions are indicated by positive and negative indicators. While negative values represent the ideal inversion coefficient, positive ones represent the most ideal positive correlation. Figure S1 demonstrates clearly that metrological considerations have little impact on AQI prediction. The report showed that for some criteria, the association was positive and for others, it was negative. Less will be affected by the negative association when predicting AQI. As a result, the prediction employed the normal dataset with a correlation threshold of 0.5 When the value was greater than 0.5,

the connection was quite positive. The correlation coefficient between AQI and $PM_{2.5}$, AQI and PM_{10} was obtained as 0.94. Similarly 0.58 for NO_2, 0.57 for CO, and 0.56 for NOx. The heatmap demonstrates that particulate matter, followed by NO_2, CO, and NOx, is the most important variable in AQI prediction, with other pollutants contributing less than 0.5 and a negative relationship with metrological factors. The effect of pollutants on AQI forecasts NO_2, CO, and NOx accounting for 8% of the total, with $PM_{2.5}$ and PM_{10} accounting for 13% (Figure S3). These findings correspond to the heatmap, and the contribution of metrological factors to AQI forecasting was extremely minimal, with the exception of BP, which accounts for 4% of AQI prediction.

2.6. Data transformation

Table 1 summarises the skewness and kurtosis of the major air pollutants both before and after data processing. Skewness and kurtosis are required for all types of data sets. Skewness is a metric for measuring loss of symmetry or, to be more precise, asymmetry. The datasets are referred to as symmetrical if the left and right data sets are equally distanced from the centre. Kurtosis is a statistic used to assess how strongly or weakly tailed data are in comparison to a normal distribution (Lord et al., 2021). The datasets must be transformed into a normal distribution because they are not normally distributed. Additional methods for transforming datasets include square root transformation, log transformation, and Box Cox transformation. Current datasets contain a small amount of skew.

The data was changed to make it more normal by applying the square root transformation. According to Table 1, the values for skewness and kurtosis were quite high for NO, NH_3, Xylene, and TOT-RF. Kurtosis levels between −10 and +10 and skewness values between −3 and +3 are both acceptable ranges. As a result, data transformation has been utilised to make the data more normal. The values of skewness and kurtosis were reduced and confirmed to be normal after square root treatment (Schneider and Wheeler-Kingshott, 2014). Figure S3 displays the skewness and kurtosis ranges of various air pollutants. Figure S4 displays NO, NH_3, Xylene, and TOT-RF data both before and after the square root change. The datasets were left skewed or extremely positive skewed and following transformation, normal distribution was obtained.

3. Result and discussion

3.1. AQI data summary

3.1.1. Particulate matters ($PM_{2.5}$ and PM_{10})

Fig. 3 displays the average $PM_{2.5}$ and PM_{10} concentrations on a monthly and annual basis. The results showed that PM_{10} concentrations were continuously greater than 100 μg/m^3 between January and March and between October and December. In the same way, PM_{10} and $PM_{2.5}$ concentrations in 2020 were found to be fewer than 100 and 40 μg/m^3, respectively, due to Covid-19 lockdown for five months (Singh and Chauhan, 2020). The $PM_{2.5}$ and PM_{10} concentrations were below 40 μg/m^3 and 100 μg/m^3, respectively, between April and September. These findings made it clear that $PM_{2.5}$ and PM_{10} levels were above the permitted range for the first six months and then fell below the permitted range. Figs. 3 and 4 demonstrated that the AQI values and PM_{10} values were very similar. This implies that the AQI prediction is significantly impacted by the Particulate Matter. Additionally, the mean $PM_{2.5}$ and PM_{10} concentrations for the five-year period from July 2017 to September 2022 were 43 and 106 μg/m^3, respectively. These readings were higher than the CPCB of India's annual PM_{10} National Ambient Air Quality Standard (NAAQS), which is 60 μg/m^3 . The $PM_{2.5}$ and PM_{10} had maximum concentrations of 202 and 312 μg/m^3, respectively. The permitted limits for $PM_{2.5}$ and PM_{10} had been exceeded by 7.5 and 6%, respectively, over the past five years.

The city's high levels of particulate matter may be caused by the following activities. Coal combustion was recognised as one of the main and principal sources because a coal-fired thermal power plant was close to the city. There will be increased levels of arsenic, zinc, and chromium as a result of these thermal power plants. Because Visakhapatnam is a beachfront city, the sea salt spray is a major source of chloride and sodium. Potassium was discovered in particle matter, which was likely caused by the home and industrial burning of biomass. Road traffic and the metal and petroleum sectors are further major source (Altlkulaç et al., 2022).

IQAir published their most recent analysis on air quality in 2021. It consists of 7323 cities dispersed among 131 nations, union territories, and geographic areas. The annual concentration of $PM_{2.5}$ was 58.1 g/m^3, according to the IQAir study, while the WHO permitted limit is less than 5 g/m^3 and the allowable level was 40 g/m^3 according to the recommendations Indian Air quality regulations. The average $PM_{2.5}$ level (53.3 g/m^3) in India was 11 times higher than recommended by the WHO. Delhi had an annual $PM_{2.5}$ concentration of 85 g/m^3, which is seventeen times the WHO recommendation. India is one of the most polluted nations in South East Asia, according to the IQAir 2021 report, with 12 of its cities ranking among the top 15.

Additionally, this fluctuation was caused by atmospheric, metrological, or temperature inversion effects throughout the year (Khillare and Sarkar, 2012). During summer, the dense and warm air moves faster as compared to the winter period, making winter air pollution thereby trapped and persist for longer periods. These conditions contribute to an increased concentration of particulate matter in the atmosphere making smog as a significant issue during recent winters (Javed et al., 2021). Long winters are common in countries like India, which prolongs the formation of smog. These formations affect the quality of the surrounding air and raise the AQI (Garg and Gupta, 2020). This AQI level is closely tied to the overall amount of rainfall. India's seasonal variance was classified using its four distinct seasons: winter (December to February), summer (March to June), monsoon (July to September), and post-monsoon (October to November) (Bose and Roy Chowdhury, 2023).

AQI and TOT-RF were compared in Fig. 4. Even though a number of other factors are crucial to the AQI Prediction, rainfall is greatly impacted by AQI. More rainfall will result in a lower AQI, and vice versa. It is seen that from July to October, the city receives significantly more rain than it does during the other seven months. Rainfall and AQI are closely associated environmental variables. More precipitation causes surface water pollution because it mixes with the particulate matter and other pollutants in the air when it falls to the ground. As a result, there will be less particulate matter in the air. The statistics clearly show that the AQI is lower and vice versa whenever there is more precipitation. The results of the data analysis showed that there is a direct relationship between AQI and both rainfall quantity and climate (Chandrappa and Kulshrestha, 2016). India experiences its wettest weather from July to

Table 1
Skew & Kurtosis Values of selected Features before and after transformation.

S.No	Attributes	Before Transformation		After Transformation	
		Skewness	kurtosis	Skewness	kurtosis
1	$PM_{2.5}$ (μg/m^3)	1.44	2.83	1.44	2.83
2	PM_{10} (μg/m^3)	1.06	1.40	1.06	1.40
3	NO (μg/m^3)	3.44	20.41	1.02	2.58
4	NO_2 (μg/m^3)	0.90	1.90	0.90	1.90
5	NO_x (ppb)	1.62	5.01	1.62	5.01
6	NH_3 (μg/m^3)	5.17	44.87	1.60	7.71
7	SO_2 (μg/m^3)	1.88	8.69	1.88	8.69
8	CO (μg/m^3)	0.51	1.06	0.51	1.06
9	O_3 (μg/m^3)	1.84	3.88	1.84	3.88
10	Benzene (μg/m^3)	0.96	4.15	0.96	4.15
11	Toluene (μg/m^3)	2.91	15.31	2.91	15.31
12	Xylene (μg/m^3)	8.63	91.96	3.33	20.95
13	TOT-RF (mm)	7.85	81.99	3.61	15.50
14	AQI	1.61	3.23	1.61	3.23

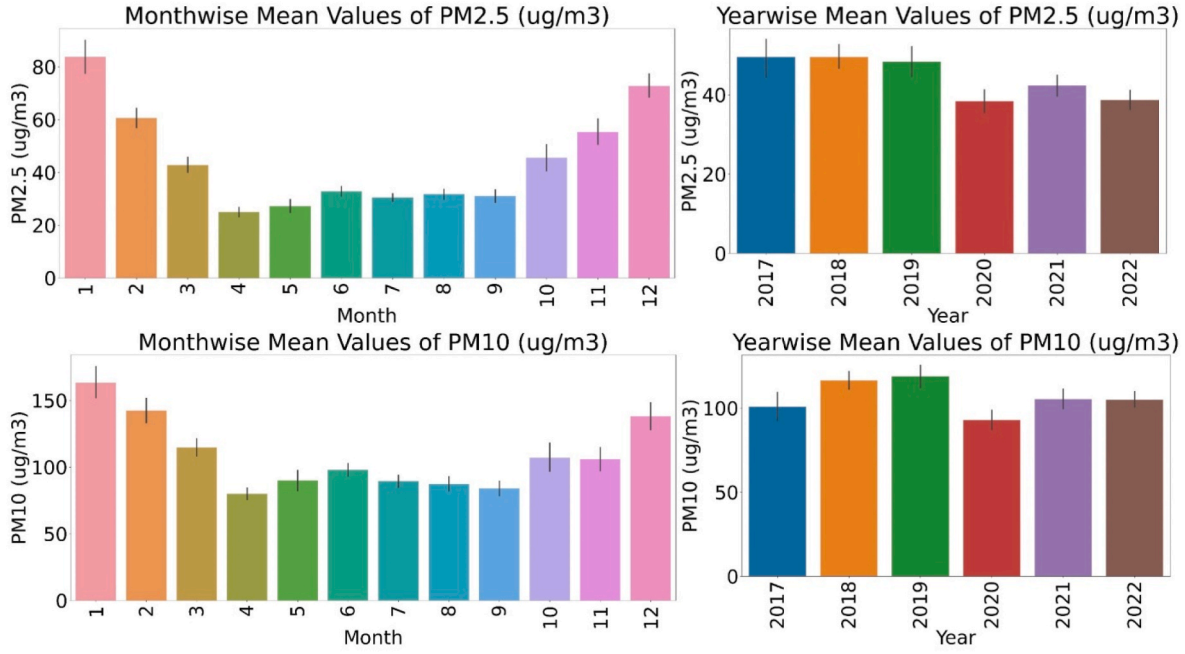


Fig. 3. Mean annual and monthly variation of particulate matters ($PM_{2.5}$ and PM_{10}) of Vishakhapatnam city.

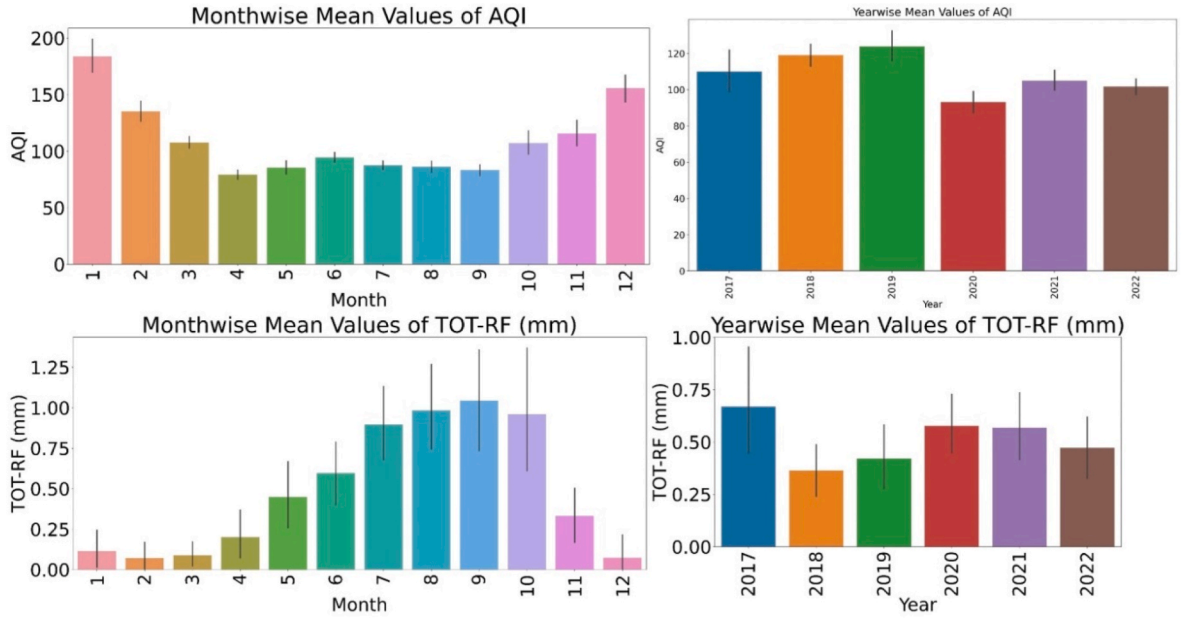


Fig. 4. Mean Annual and monthly variation of AQI and TOT- RF of Visakhapatnam City.

October, and the winter months of November to February. This might be the reason for the seasonal change in the AQI.

3.1.2. *Gaseous pollutants*

Important pollutants, including CO, NO, NH_3, NO_2, NO_X, and SO_X, were compared in Fig. 5. Particulate matter and gaseous pollutants play an important role in the calculation of AQI. It is observed that the trend of $PM_{2.5}$ is similar to that of the AQI. $PM_{2.5}$ and AQI follow the same trend throughout the year, indicating that $PM_{2.5}$ plays important role in AQI calculation. Similarly, CO follows the same trend with respect to AQI, indicating that CO also has a significant impact on AQI. Fig. 5 and Table S4 reflect the impact of gaseous pollutants that determine the AQI category. It is observed that CO, NH_3, and SO_2 have lower concentrations and fall into the AQI Category "Good (0–50)". However, during winter, it is observed that NO_2 falls into the AQI category "Satisfactory (51–100)"; reflecting NO_2 has a significant impact on the overall AQI, and CO follows the trend of AQI throughout the year.

Benzene, ozone, xylene, and toluene have over recent year's shown to emerge indicating that the monthly average ozone value was equivalent to that of $PM_{2.5}$ and PM_{10} (Fig. 6). The ozone's monthly mean value peaked from November to February, which also happens to be the winter season in India. Ground level ozone is produced when sunlight and volatile organic compounds (VOCs) interact; it is not immediately released into the atmosphere. These pollutants are released into the atmosphere by vehicles, the automotive sector, thermal power plants, bio refineries, chemical facilities, and a number of other sectors. The

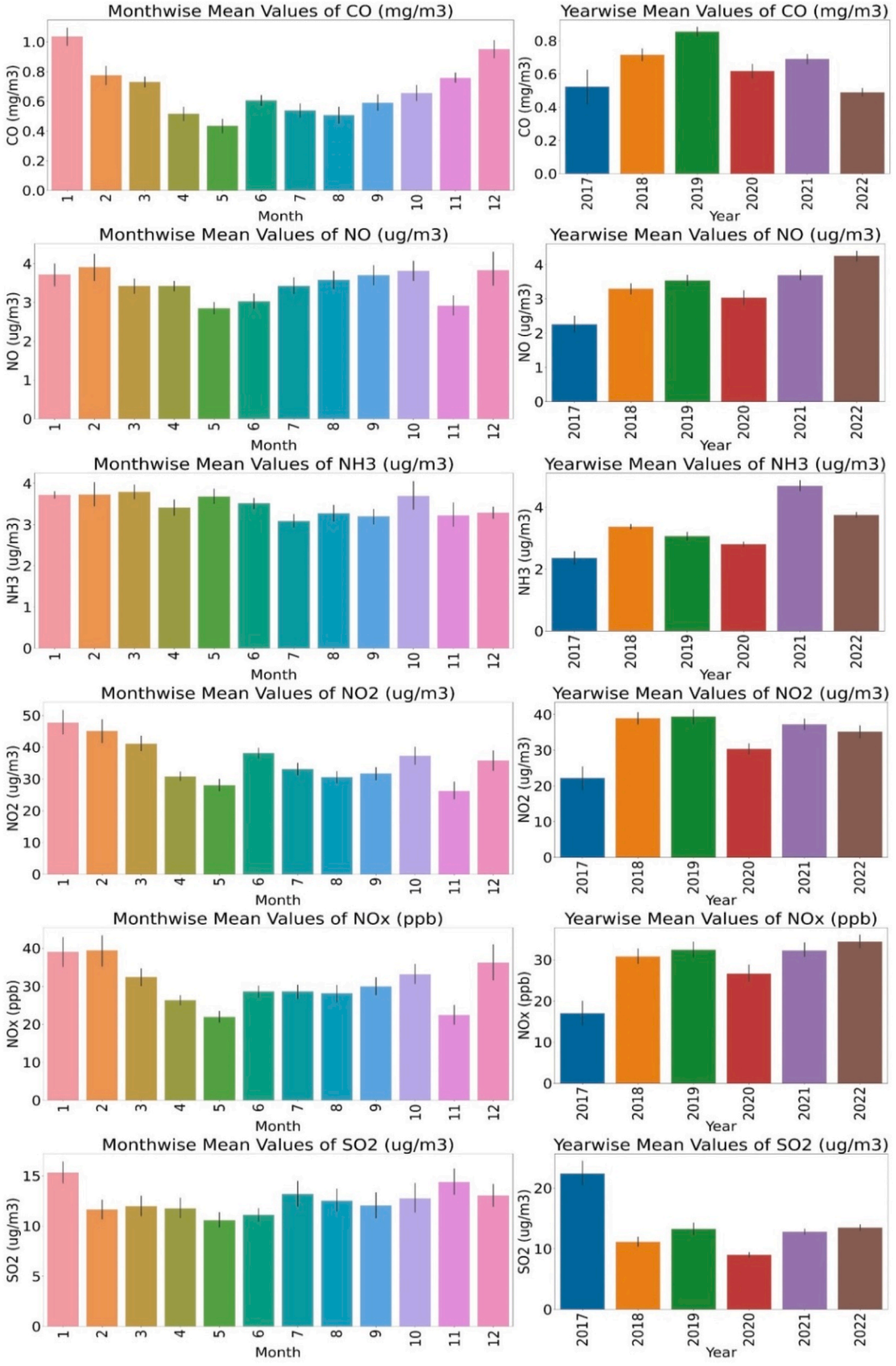


Fig. 5. Mean annual and monthly variation of CO, NO, NH_3, NO_2, NO_X and SO_X of Visakhapatnam City.

city's heavy industries are well-known for producing large volumes of VOCs, which interact photochemical with nitrous oxide to create ozone (Manisalidis et al., 2020). Because of the wintertime temperature inversion, these VOCs will be surrounded by cold, thick air and won't be able to escape into the atmosphere (Khillare and Sarkar, 2012). When the sun reflects off the surface of the earth, VOC and nitrogen oxides will react, creating an excessive amount of ozone. Due to the high temperatures in the summer, which heat the air and allow VOC and nitrous

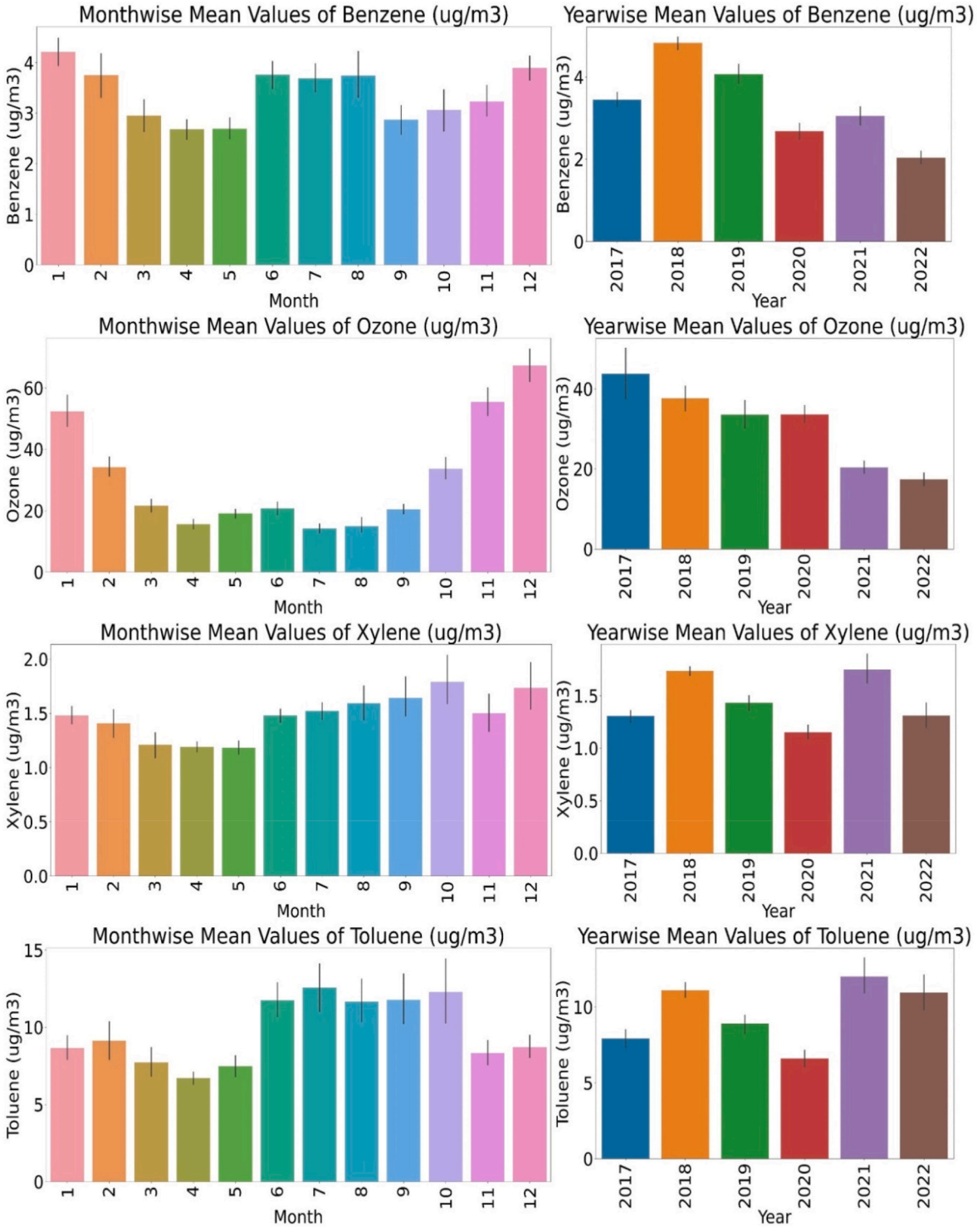


Fig. 6. Mean Annual and monthly variation of benzene, ozone, xylene and toluene of Visakhapatnam City.

oxide to escape, this behaviour is reversed. As a result, summertime ground-level ozone formation is lower. When it comes to carbon monoxide, the similar pattern is seen. According to the guidelines of the Central Pollution Control Board (CPCB) of India, Table S4 detail the yearly average allowed limit of air pollutants that contributes to AQI. Six major categories can be used to classify AQI generally.

The temporal variations of all pollutants from 2017 to 2022 were shown in Figure S5. Month after month and year after year, air pollution rises. Due to temperature inversion, Particulate Matter, CO, and Ozone levels are still at their maximum in the winter and somewhat lower in the summer (Khillare and Sarkar, 2012). It is seen that in 2021, all pollution levels increased, whereas they were lower in 2020. This might be the case because heavy industries carried on operating continuously after the Covid-19 Lockdown, perhaps causing air pollution levels to rise. Fig. 6 made it abundantly evident that the concentrations of benzene, xylene, and toluene in the environment are not constant and do not follow any pattern. This suggests that the city's industrial operations were to blame for the contaminants' accumulation.

A report on the major health impact on people in India was published in 2018 by the Health Effects Institute (HEI). According to HEI, household burning, coal combustion followed by agricultural burning, anthropogenic emission, and transport, diesel, and brick kilns are significant contributors to the discharge of main air pollutants. According to estimates from 2015, air pollution was responsible for 10% of all deaths in India. According to the study's summary, 0.169 million deaths were attributed to coal combustion, 0.268 million to residual biomass burning, 0.1 million to dust, 0.06 million to agricultural burning, and 0.065 million to transportation, kilns, and diesel. Additionally, it is predicted that by the year 2050, 3.6 million deaths would have occurred due to air pollution, an increase of 84% over the number of deaths that occurred in 2015. Significant air pollution was recorded in Visakhapatnam in 2018, making it one of the most polluted cities in south India. There are annual emissions of 47,800 tonnes of $PM_{2.5}$, 65,000 tonnes of PM_{10}, 3250 tonnes of O_3, 182,100 tonnes of NO_X, 188,550 tonnes of CO, 39,500 tonnes of VOC, and 41,250 tonnes of SO_2. The industrial emissions from small, medium, and large businesses were responsible for almost 70% of these pollutants (Police et al., 2016).

3.2. Machine learning models to predict AQI

Figures S6 – S10 depict the expected AQI, which has been generated using machine learning. The hyperparameters of the proposed machine learning models have been tailored to the dataset to achieve the best performance. The Grid-search technique has been utilised to identify the best hyperparameters that produce the most accurate predictions. Grid search involves an exhaustive search of a manually specified subset of the hyperparameter space of the proposed machine learning method (Wu et al., 2019). Table 2 provides a summary of the performance evaluation of various models on the training set (which represents 80% of the data) and the testing set (which represents 20% of the data). Table 2 presents an overview of the models' ability to forecast AQI based on both the training and testing datasets. The MAE is calculated using the absolute differences between the dataset's values and the predicted values. The RMSE is the square root of the MSE, which is the difference between the actual and anticipated values generated from the squares of the average difference between datasets. Because the MSE values are typically greater than other types of values, RMSE is frequently used to reduce the difference between the two types of error by taking the square root of MSE (Oswalt Manoj et al., 2022). To determine which model performs the best, it is not appropriate to compare MSE, MAE, and RMSE. Instead, the performance of different models must be compared based on the MAE of all other models, and the same applies to MSE and RMSE. Furthermore, the MAE, MSE, and RMSE values for the testing and training datasets should always be comparable to each other. If there is a significant difference between the training set and the testing set, the presence of additional outliers in the datasets is suggested. As a result, in addition to the R^2 value, it is necessary to assess the model's performance using several dataset error analyses. According to Table 2, both Random Forest and Catboost ML models had the highest correlation for training datasets, with 0.9936 for Random Forest and 0.9998 for Catboost ML. The high prediction accuracy of Random Forest and Catboost are owing to the distinct features of these algorithms. For instance, Random Forest models work together to precisely represent feature importance. On the other hand, CatBoost was specifically developed to construct high-performing models at an incredible speed for massive datasets.

3.3. Implications and perspectives

The study's results demonstrated that the Random Forest and CatBoost algorithms outperformed other machine learning models like LightGBM, Adaboost, and XGboost in predicting AQI accurately. Boosting algorithms, while effective, are prone to overfitting due to their reliance on tree-based methods. Additionally, parallelizing the training process with tree algorithms can be challenging. However, CatBoost incorporates parameters that help mitigate overfitting in datasets . These findings suggest that Random Forest and CatBoost algorithms could be explored in other urban areas or regions with different air quality characteristics to assess their effectiveness across various scenarios.

Moreover, considering the potential of these algorithms, it may be valuable to develop real-time AQI prediction systems that leverage these models. Such systems could provide policymakers with up-to-date and accurate information, enabling them to implement timely and effective measures to improve air quality. The scalability and application of these algorithms in other areas and regions warrant further investigation and exploration.

4. Conclusions

The study analysed the forecast for Visakhapatnam's AQI between 2017 and 2022. The AQI levels were observed to rise between 2017 and 2019, followed by a decline in 2020 due to the nationwide lockdown enforced in response to Covid-19. However, the AQI levels continued to rise thereafter. $PM_{2.5}$ and PM_{10} were found to be crucial factors in determining the AQI values, while the metrological characteristics had minimal impact. The machine learning models used in the study accurately predicted the AQI, with the Random Forest and Catboost models showing maximum correlations of 0.9998 and 0.9936, respectively, for training datasets. Therefore, machine learning algorithms can effectively forecast AQI levels. To upscale the models to other regions, it is necessary to validate their performance under diverse air quality conditions. Additionally, assessing the transferability of the models by applying them to other cities or countries and evaluating their predictive capabilities in those contexts is essential.

Author statement

Author statement for "Air quality prediction by machine learning models: A predictive study on the east coast of India". Gokulan Ravindiran, Karthick Kanagarathinam, Avinash Alagumalai: writing original draft, reviewing, editing, data curation. Gasim Hayder, Christian Sonne: reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial

Table 2
Machine learning Models with their performance factors in prediction of AQI.

S. No	Model Name	Training Set (80%)				Validation/Testing Set (20%)			
		MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
1	LightGBM	1.80	27.62	5.25	0.9915	1.80	27.62	5.25	0.9915
2	Random Forest	1.10	20.54	4.53	0.9936	0.41	3.03	1.74	0.9990
3	Catboost	2.01	25.37	5.03	0.9922	0.60	0.58	0.76	0.9998
4	Adaboost	7.69	84.74	9.20	0.9739	7.38	77.79	8.82	0.9753
5	XGboost	2.04	24.17	4.91	0.9925	1.48	5.93	2.43	0.9981

MAE- Mean Absolute error; MSE- Mean squared error: RMSE- Root Mean square Error R^2- Correlation Coefficient.

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 Post-doctoral Fellowships under the project code of J510050002-IC-6 BOLDREFRESH2025-Centre of Excellence. The authors also extend their appreciation to Department of Civil Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, 500090, Telangana, India.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2023.139518.

References

Altlkulaç, A., Turhan, Ş., Kurnaz, A., Gören, E., Duran, C., Hançerlioğullarl, A., Uğur, F. A., 2022. Assessment of the enrichment of heavy metals in coal and its combustion residues. ACS Omega 7, 21239–21245. https://doi.org/10.1021/ACSOMEGA.2C02308/ASSET/IMAGES/LARGE/AO2C02308_0004.

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.

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

Ganesh, N., Jain, P., Choudhury, A., Dutta, P., Kalita, K., Barsocchi, P., 2021. Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes, 2021 Processes 9. https://doi.org/10.3390/PR9112095, 2095 9, 2095.

Garg, A., Gupta, N.C., 2020. The great smog month and spatial and monthly variation in air quality in ambient air in Delhi, India. J. Heal. Pollut. 10 https://doi.org/10.5696/2156-9614-10.27.200910.

Gurjar, B.R., Ravindra, K., Nagpure, A.S., 2016. Air pollution trends over Indian megacities and their local-to-global implications. Atmos. Environ. 142, 475–495. https://doi.org/10.1016/J.ATMOSENV.2016.06.030.

Guttikunda, S.K., Goel, R., Pant, P., 2014. Nature of air pollution, emission sources, and management in the Indian cities. Atmos. Environ. 95, 501–510. https://doi.org/10.1016/J.ATMOSENV.2014.07.006.

Javed, A., Aamir, F., Gohar, U.F., Mukhtar, H., Zia-Ui-haq, M., Alotaibi, M.O., Bin-Jumah, M.N., Marc, R.A., Pop, O.L., 2021. The potential impact of smog spell on humans' health amid COVID-19 rages. Int. J. Environ. Res. Publ. Health 18. https://doi.org/10.3390/IJERPH182111408.

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.

Langer, T., Meisen, T., 2021. System design to utilize domain expertise for visual exploratory data analysis, 2021 OR Inf. 12, 140. https://doi.org/10.3390/INFO12040140. Page 140 12.

Li, H., Fan, H., Mao, F., 2016. A visualization approach to air pollution data exploration—a case study of air quality index (PM2.5) in beijing, China, 2016 Atmósfera 7, 35. https://doi.org/10.3390/ATMOS7030035. Page 35 7.

Li, L., Li, Q., Huang, L., Wang, Q., Zhu, A., Xu, J., Liu, Ziyi, Li, H., Shi, L., Li, R., Azari, M., Wang, Y., Zhang, X., Liu, Zhiqiang, Zhu, Y., Zhang, K., Xue, S., Ooi, M.C.G., Zhang, D., Chan, A., 2020. Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: an insight into the impact of human activity pattern changes on air pollution variation. Sci. Total Environ. 732, 139282 https://doi.org/10.1016/J.SCITOTENV.2020.139282.

Lord, D., Qin, X., Geedipally, S.R., 2021. Exploratory analyses of safety data. Highw. Saf. Anal. Model. 135–177. https://doi.org/10.1016/B978-0-12-816818-9.00015-9.

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. Sens. 2022 https://doi.org/10.1155/2022/4649510.

Malhi, G.S., Kaur, M., Kaushik, P., 2021. Impact of climate change on agriculture and its mitigation strategies: a review, 2021 Sustain. Times 13, 1318. https://doi.org/10.3390/SU13031318. Page 1318 13.

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

Mishra, S., Mishra, D., Santra, G.H., 2020. Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: an empirical assessment. J. King Saud Univ. - Comput. Inf. Sci. 32, 949–964. https://doi.org/10.1016/J.JKSUCI.2017.12.004.

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.

Police, S., Sahu, S.K., Pandit, G.G., 2016. Chemical characterization of atmospheric particulate matter and their source apportionment at an emerging industrial coastal city, Visakhapatnam, India. Atmos. Pollut. Res. 7, 725–733. https://doi.org/10.1016/J.APR.2016.03.007.

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., 2021. Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys. Res. Lett. 48, e2020GL091202 https://doi.org/10.1029/2020GL091202.

Schneider, T., Wheeler-Kingshott, C.A.M., 2014. Q-space imaging: a model-free approach. Quant. MRI Spinal Cord 146–155. https://doi.org/10.1016/B978-0-12-396973-6.00010-1.

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.

Sumiya, E., Dorligjav, S., Purevtseren, M., Gombodorj, G., Byamba-Ochir, M., Dugerjav, O., Sugar, M., Batsuuri, B., Tsegmid, B., 2023. Climate patterns affecting cold season air pollution of ulaanbaatar city, Mongolia. Climate 11, 4. https://doi.org/10.3390/CLI11010004/S1.

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.

Zhang, Y., Zhao, Z., Zheng, J., 2020. CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. J. Hydrol. 588, 125087 https://doi.org/10.1016/J.JHYDROL.2020.125087.

Zhou, Y., Wang, W., Wang, K., Song, J., 2022. Application of LightGBM algorithm in the initial design of a library in the cold area of China based on comprehensive performance, 2022 Build 12, 1309. https://doi.org/10.3390/BUILDINGS12091309. Page 1309 12.