Journal of Mines, Metals and Fuels, *73(11): 3435-3445; 2025. DOI: 10.18311/jmmf/2025/50123*

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: *www.informaticsjournals.co.in/index.php/jmmf*

An Ensemble based Machine Learning Approach for Air Pollution Detection and Control: A Case Study

Danaveni Pavithra[1], Prashanth Ragam[2*], Singamsetty Ritwika[1], Ashish Nerusu[2], Harathi Chavala[2], Sai Ganesh Reddy Venna[2] and Sai Kumar Addanki[2]

[1]Department of Computer Science Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad - 500090, Telangana, India

[2]School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati - 522241, Andhra Pradesh, India; prashanth.rajam429@gmail.com

Abstract

The rapid advancement of technology and the expansion of industrial activities have significantly compromised air quality, adversely affecting the overall quality of life. Air pollution shows a major environmental challenge with increased consequences negatively affecting human health, ecological balance and economic progress. This leads to a necessity for precise detection and monitoring of air pollution levels. This work uses a dataset of AQI (Air Quality Index) of Hyderabad, the capital city of Telangana, and analyses the level of air pollution in the city. Air Pollution Detection helps in finding the level of air pollution in Hyderabad, where the dataset is split into test and train sets and Machine Learning algorithms are performed. In contrast to traditional estimation algorithms, including statistical and numerical methods, Machine Learning techniques offer enhanced accuracy and efficiency in air quality monitoring systems. In this study, several Machine Learning models, namely Ensemble XGBoost-Random Forest, Gradient Boosting, Random Forest, and XGBoost, were considered for predicting the levels of air pollution. These models use computational power afforded by ML algorithms that could analyse complex datasets and discern patterns and trends that can slip through the sieve of more traditional statistical methods. Using the data covering all environmental factors, examples include particulate matter, ozone, nitrogen dioxide, meteorological variables, and other relevant factors, predictive models are provided to detect the level of air pollution. With careful comparison and evaluation, Ensemble Gradient Boosting comes out as the best model that predicts well the level of air pollution with higher accuracy and reliability. The standalone gradient boosting algorithm gives an MSE of 0.0000998544 and R^2 of 0.99737955298.

Major Findings: *This paper evaluates ensemble ML models (XGBoost-RF, Gradient Boosting, etc.) for Hyderabad's AQI prediction, with Gradient Boosting excelling (MSE: 0.0001, R^2: 0.997). The hybrid XGBoost-RF model proves robust, blending boosting and bagging advantages. ML methods surpass conventional statistical approaches in accuracy for dynamic pollution data. The results offer scalable solutions for real-time urban air quality management.*

Keywords: *Air Pollution Detection, Ensemble XGBoost-RF, Gradient Boosting, Machine Learning, Random Forest, XGBoost*

1.0 Introduction

During recent times, Urban development and its effects have been more pronounced. Air pollution caused by urbanisation has led to many environmental changes. The effects of air pollution on human health and life are widely seen across the world[1]. Industrial plants, burning of straw, coal, and kerosene, mainly cause air pollution. Emission of Hazardous contaminants, including carbon monoxide, carbon dioxide, and particulate matter,

**Author for correspondence*

among others, from these activities are released into our atmosphere[2]. These emissions cause many respiratory and other diseases. According to WHO, 99 per cent of people breathe air which is more polluted than the limits of the WHO and contains a high level of pollutants[3]. According to S. Apte[4], and Dandan Xu[5], inhaling particulate matter causes health effects that cause premature deaths. The life expectancy in the world and the nation has decreased by 1.2-1.9 years due to Air Pollution, which is mostly attributed to the PM 2.5-sized particulate matter in certain countries in Asia and Africa. Lately, this air pollution has caused more effects in urban areas due to industrialisation. Many cities in the world have to come up with different automatic air pollution detection machines for regular maintenance of air quality[6].

The present data set is derived from Kaggle of Hyderabad, one of the major current industrialising cities of India. Over the past 20 years, Hyderabad, the capital of Telangana state in India, has seen an increase in urbanisation of almost 1 %[7]. All the pollutants from vehicles, industries, and even houses of Hyderabad had a bad effect on the weather. Hyderabad, India, frequently experiences air pollution levels that breach the country's regulatory thresholds, with Particulate Matter (PM) posing a persistent challenge. In 2010, *PM10* concentrations recorded averages of 82.2 ±24.6 μg/m³ at commercial zones, 96.2 ±12.1 μg/m³ in industrial areas, and 64.3 ±21.2 μg/m³ across residential neighbourhoods. These measurements notably surpassed India's National Ambient Air Quality Standard (NAAQS) of 60 μg/m³, highlighting widespread non-compliance. The industrial sector exhibited the highest deviation from the mandated limit, followed by commercial and residential regions, underscoring the urgent need for targeted mitigation strategies to address particulate pollution in the city[8]. Hyderabad takes the cake among all the other cities for poor air quality because of high levels of Particulate Matter pollution, surpassing the other cities in southern and western India. According to data from the Central Pollution Control Board, Hyderabad has recorded the maximum number of red days, meaning high levels of PM 2.5 in the air. PM 2.5 particles, which have their diameter of less than 2.5 micrometres, are harmful to health as they can pierce through lung tissues[9]. While Hyderabad's PM 2.5 levels are alarming, its annual average is within the tolerable limit, according to data presented by PCB. Telangana PCB has established 19 manual air quality stations around Hyderabad to check the pollution level.

Machine Learning (ML) techniques are critical in addressing air pollution challenges, as they enable advanced analysis of intricate environmental data and precise predictive modelling. Unlike conventional statistical approaches, ML methods autonomously identify hidden patterns, dynamic trends, and non-linear interactions within vast datasets, proving exceptionally valuable for predicting air quality metrics and pinpointing contamination sources. This research employed diverse ML frameworks, such as random forest, gradient boosting, XG Boost, and a hybrid Ensemble XG Boost-Random Forest architecture. Random Forest excels at managing noisy data through aggregated decision trees, while Gradient Boosting iteratively refines predictions by addressing residual errors. XG Boost enhances gradient boosting with computational efficiency and scalability, optimising performance for structured datasets. The hybrid ensemble combines the robustness of XGBoost and the versatility of Random Forest to maximise predictive accuracy. Collectively, these approaches underscore ML's transformative potential in creating adaptive, analytics-driven frameworks for real-time pollution surveillance and risk mitigation. Such systems are pivotal for reducing public health hazards and ecological damage caused by deteriorating air quality, empowering stakeholders with actionable insights for timely intervention[10].

2.0 Previous works

One such transformation in air quality predictive models has been from the conventional statistical and numerical models, which included, among others, transfer of chemicals and their dispersion in the atmosphere, to the modern machine learning methods[11,12]. In effect, the role of pollutant concentration prediction in urban areas has gained greater popularity through increased use of such machine learning techniques in this area[13].

2.1 Numerical models

Numerical methods typically utilise mathematical equations to replicate atmospheric processes for the purpose of forecasting air quality. Two notable dispersion models, HIWAY2[14] (developed by the US EPA) and CALINE4[15] (created by the California Department of Transportation),

Table 1. Summary of previous investigations on AQI[24-38]

Author(s)	Technique(s)	Input(s)	R^2	MSE
Reshma Unnikrishnan and G. Madhu	ANN	SO_2, NO_x, NH_3, CO, PM_{10}, $PM_{2.5}$	0.8979	0.0115
Irfan Yaqoob *et al.*	RF, GBM	$PM_{2.5}$	0.88	3.39 (µg/m3)
V. U. Rani, S. Thirisaa	LR, LOR, DT, NB, KNN, RF	PM_{10}, NO_2	0.998	0.004
Theresia Herlina Rochadiani	RF, GBR, AR, BR	$PM_{2.5}$, PM_{10}, O_3, CO, SO_2, NO_2	0.6271	0.6054
Huixiang Liu *et al.*	RF, SVR	NO_x,	0.9776	7.666
Xiaowen Zhang *et al.*	SSA, BiLSTM, LightGBM	CO, SO_2, NO_2	0.9995	0.6897
K. R. Rajagopal, Kumar Narayanan	CNN, BiLSTM	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	0.961	11.92
Emina Džaferović, Kanita Karađuzović-Hadžiabdić	SVR, RF, GB, MLR, MP	PM_{10}, O_3	0.99	2.03
Subashini *et al.*	DT, RF, GBR, AR, ETR, XGBR	$PM_{2.5}$, O_3, NO_2	0.978	21.91
Mihaela Tinca Udristioui *et al.*	DT	PM_1, $PM_{2.5}$, PM_{10},	0.95	0.65
T. Dao *et al.*	RFR, XGBR	NO_2, SO_2, CO, NO	0.942	32.6
Gaurav Narkhede *et al.*	ETR, RFR, LightGBM, KNN	O_3, PM_{10}, $PM_{2.5}$	0.915	254.59
Dat Q. Duong *et al.*	SVM, RF, XB, LightGBM, CB	SO_2, NO_2	0.89	5.77
Nancy F. Chen, Shuang Yang	RF, XGBR, RNN, LSTM	NO_2, SO_2, $PM_{2.5}$, PM_{10}	0.915	13.048
Benakappa *et al.*	ETR, SVR, LightGBM, CB	O_3, $PM_{2.5}$, PM_{10}	0.8660	56.98

Note: Artificial Neural Network (ANN); Random Forest (RF); Gradient Boosting Machines (GBM); Random Forest Regressor (RFR); Gradient Boosting Regressor (GBR); support vector regression (SVR): Singular Spectrum Analysis (SSA); Bidirectional Long Short-Term Memory (BiLSTM); CatBoost (CB); Multilayer Perceptron (MP); Light Gradient Boosting Machine (LightGBM); Convolutional Neural Network (CNN); Extreme Gradient Boosting (GB); Gradient Boosting Regressor (GBR); AdaBoost Regressor (AR); XGBoost Regressor (XGBR); Decision Tree Regressor(DT); K-Nearest Neighbor (KNN); Multiple Linear Regression (MLR); Extra Tree Regressor (ETR); Lasso Regression (LR); Logistic Regression (LOR); Naive Bayes Regressor (NB); Bagging Regressor (BR); Sulfur dioxide (SO_2); Nitrogen oxides (NO_x); Ammonia (NH_3); Carbon monoxide (CO); Particulate matter 10 (PM_{10}); Particulate matter 2.5 ($PM_{2.5}$); Ozone (O_3); Nitrogen dioxide (NO_2); Particulate matter 1 (PM_1); Nitric oxide (NO)

are grounded in the Gaussian Plume model and are primarily employed to estimate vehicular emissions[16]. Another classification of numerical models is known as Chemical transfer models, which are designed to simulate the physical and chemical transformations affecting pollutant concentrations in the atmosphere through mathematical formulations. An example of such a model is WRF-CHEM, a weather research and forecasting model integrated with chemical processes, which has been utilised to predict ozone levels in Shanghai, China. Additional studies have also pointed to the application of other chemical transfer models, such as the Community Multiscale Air Quality model (CMAQ), for estimating pollutant concentrations[17].

2.2 Statistical Models

Statistical models take advantage of historical data to learn the pattern and predict future behaviour with much accuracy. Factor analysis and Cluster analysis are said to be multivariate techniques to make complex data more understandable[18]. In addition to the Cluster analysis, the Louvain algorithm helps in detecting communities or clusters with similar air pollution index[19]. (Jia-bao) Statistical models are more favourable in real-world, occupied environments as they can be trained on large observed datasets[20]. But because of their high complexity and lower efficiency, they fail to estimate the exposed levels of dynamic behaviour precisely.

2.3 Machine Learning Models

Since AI algorithms can be used for air quality forecasting, they are becoming increasingly more accepted with improvements in technology, but not statistical models, where AI analyses several parameters for a prediction. Data mining techniques are one of the most interesting

approaches in Air Quality prediction. Machine learning models like SVR, RFR and CatBoosting Regression create a highly powerful, efficient and scalable approach in forecasting and managing air pollution in various cities[21]. A hybrid model combining Grey Wolf Optimisation (GWO) and Decision Tree (DT) algorithm gives accurate predictions in various cities[22]. Machine learning algorithms evaluate the air quality index through their computations. Modern-day scientific investigation of this area fails to produce precise results. The implementation of Machine Learning algorithms consists of Linear Regression, Decision Tree, Random Forest, Artificial Neural Network and Support Vector Machine[23].

3.0 Materials and Methods

3.1 Introduction to Air Pollution

Air is said to be polluted if it has harmful substances that negatively affect human health, ecosystems, and the environment is significant. Identifying air pollution is vital for averting health threats, especially those related to respiratory and cardiovascular issues. Hence, understanding the effects and sources of each pollutant is necessary. The existing methods[39-42] detect air pollution of a respective place by analysing the Air Quality Index (AQI), a standardised metric for assessing atmospheric conditions. Four advanced machine learning algorithms—Random Forest, Gradient Boosting, XGBoost, and a hybrid XGBoost-Random Forest (XG Boost-RF) ensemble—were applied to predict and classify pollution severity. The research systematically benchmarks these models' accuracy, computational efficiency, and robustness in interpreting AQI data, aiming to identify the most reliable framework for real-time air quality forecasting. By comparing predictive outcomes across algorithms, the work highlights critical insights into optimising data-driven strategies for environmental monitoring and public health interventions in urban regions facing air quality challenges.

3.2 Data Collection

Actual measurement helps in assessing potential damage and allows for measures of protection of the environment by observing pollution levels, analysing trends, and applying targeted reduction measures. The use of data coming from various sources available on Kaggle has been for this paper[43]. The analysis incorporated a comprehensive range of air pollutants Particulate Matter (PM), Ozone (O_3), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), ammonia (NH_3), benzene (C_6H_6), and carbon monoxide (CO) as well as meteorological parameters like ambient temperature, wind velocity, relative humidity, and atmospheric pressure. The dataset's pollutant concentrations are statistically summarised in the table below, which outlines their minimum, maximum, and standard deviation values to quantify variability and extremes.

3.2.1 Performance Analysis of Machine Learning Algorithms Using Evaluation Metrics

This analysis used different types of machine learning algorithms, finding the best machine learning model based on the performance of each model. Evaluation metrics used in the analysis are Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2 score.

3.2.1 Performance Analysis of Machine Learning Algorithms Using Evaluation Metrics

This analysis used different types of machine learning algorithms, finding the best machine learning model based on the performance of each model. Evaluation metrics used in the analysis are Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2 score.

3.2.2 Machine Learning Techniques for Estimation of Air Pollution

Ensemble machine learning techniques are those specialised portions in the area of artificial intelligence

Table 2. Data collection

Parameter	Minimum	Maximum	Standard Deviation
PM10 (æg/m^3)	76	178	24.180005
SO2 (æg/m^3)	4	50.9	0.207134
NOx (æg/m^3)	18.4	40.3	0.186814
PM2.5 (æg/m^3)	12	87	20.224251
Ammonia - NH_3 (æg/m^3)	11	37	0.224017
O3 (æg/m^3)	2.4	107.6	0.214695
CO (mg/m^3)	0.2	1.52	0.180882
Benzene (æg/m^3)	0	0.9	0.193070
AQI	58	166	0.203208

Table 3. Performance evaluation metrics

Evaluation Metrics	Formula	Ideal Values
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$	Close to 0. MAE= 0, shows no error
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Close to 0 MSE= 0, shows no error
Coefficient of Determination (R2 Score)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum i = 1^{n}(y_i - \bar{y})^2}$	Close to 1. R2=1 shows perfect prediction

that allow computer systems to learn and develop based on experience rather than being explicitly programmed. Increasingly, industries such as healthcare, finance, and manufacturing apply machine learning to make better decisions, build efficiency, and achieve high precision. There exist several machine learning techniques applied in air pollution detection, each with its unique advantages and applications. Such techniques would extract rich patterns and nuances imperative for predictive modelling, considering the enormous datasets of atmospheric variables, meteorological conditions, and pollutant sources. Their ensemble learning paradigms help them adapt and refine predictions iteratively to ensure robust performance even in an evolving regime of environmental dynamics. Synergising the statistical methodology and algorithmic sophistication, such techniques are going to pave the way toward actionable insights into AQI management and public health interventions. With increased pressure from all quarters of this globe to mitigate air pollution, machine learning techniques have begun to arrive as an exciting frontier for the advancement of our understanding and efforts at mitigation against this pervasive environmental challenge. In this current work, the following methodology has been adopted for the prediction of AQI.

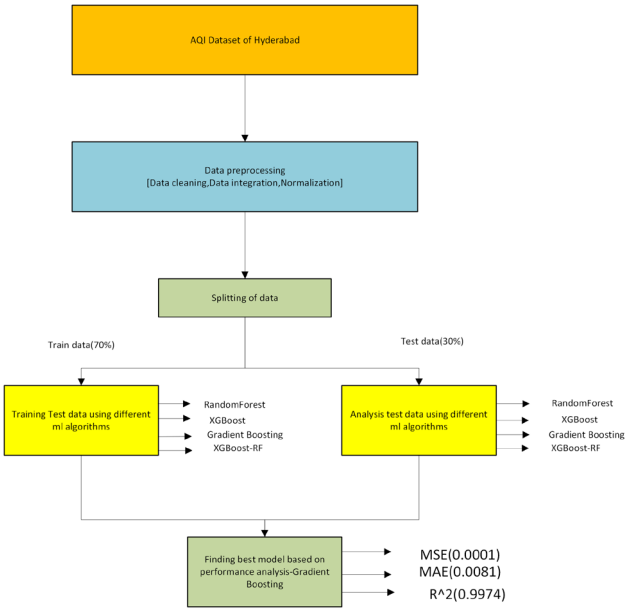


Figure 1. Proposed methodology.

3.2.3 Random Forest Regressor

A very powerful method of doing ensemble learning is the random forest regressor. It works by building thousands of decision trees iteratively when training. Each feature and data point would then be randomly chosen to train each tree. This way, the average of the predictions of the forests will then be used to determine the prediction. One can say that Random Forest Regressor is the best for processing high-dimensional data and finding complex interactions between predictors and the target. They are great in regression problems, including forecasting of pollutant concentrations in the processes of air quality monitoring. It has an attribute of not getting overfit and noise in the data. Finally, Random Forest Regressors are rather simple and scalable models which work efficiently for small-sized as well as large-sized datasets. Providing insights on the features important in the system, Random Forest Regressor aids in the interpretation of a

Table 4. Hyperparameters and metrics for the random forest regressor model

Parameter	Random Forest Regressor Model Value
Max Depth	2
Random State	0
Number of Estimators (n estimators)	100
Min_samples_split	2
Min_samples_leaf	1
bootstrap	True

model, thereby explaining its wide usage across different domains. Prediction for classification:

$$\hat{y} = \text{majority vote}\left(f_1(x), f_2(x), \ldots, f_T(x)\right) \quad (1)$$

Regression prediction:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (2)$$

The Random Forest Regressor was configured with specific parameters to balance model performance and generalisation. By setting max_depth=2, we limited each decision tree's depth, preventing overly complex splits that could lead to overfitting. The ensemble size (n_estimators=100) ensured robust predictions by aggregating outputs from multiple trees, thereby reducing prediction variance. For reproducibility, random_state=0 was fixed, guaranteeing consistent results across repeated experiments. Beyond these key settings, default parameters also played an important role. The min_samples_split=2 threshold ensured nodes only split with sufficient data, maintaining tree stability. Meanwhile, min_samples_leaf=1 allowed leaf nodes to form with minimal samples, capturing fine-grained patterns, though this could increase sensitivity to noise. The default bootstrap=True introduced randomness by training each tree on a bootstrapped data subset, enhancing ensemble diversity and improving generalisation through variance reduction.

3.3 Gradient Boosting

Gradient Boosting is one of the more sophisticated ensemble learning techniques that is widely applied for building predictive models in machine learning. The method operates by utilising a series of weak learners, typically decision trees, which are applied to the residuals generated by the preceding model in a manner akin to descent methodology. The result of the whole process will be at the minimum value of the loss function, giving way to building a strong predictive model that can capture intricate interactions and patterns within the dataset. Unlike the case of some other algorithms, Gradient Boosting is less likely to suffer from overfitting, particularly for high-dimensional data. It finds more applications in a wide range of domains in both tasks of classification and regression. While implementation is achievable, optimal results require meticulous calibration of hyperparameters such as the learning rate and maximum tree depth. Among machine learning methodologies, Gradient Boosting is recognised as a robust approach. This technique iteratively trains multiple decision trees, where each subsequent model targets residual errors from prior iterations. The final aggregate prediction is derived from a weighted combination of all tree outputs. Model performance was evaluated using three metrics: Mean Squared Error (MSE), representing the average of squared differences between observed (btest) and predicted (bpred) values; Mean Absolute Error (MAE), which measures average absolute deviations; and the R^2 coefficient, quantifying the proportion of variance explained by the model.

Prediction update:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \quad (3)$$

Loss function minimization

$$L\left(y, F_{m(x)}\right) = L\left(y, F_{\{m-1\}(x)}\right) + \left\{\partial L\right\} / \left\{\partial F_{\{m-1\}(x)}\right\} \cdot h_{m(x)} \quad (4)$$

Table 5. Hyperparameters and metrics for the gradient boosting regressor model

Parameter	Gradient Boosting Model Value
Max Depth	2
Learning rate	0.1
Random State	0
Number of Estimators (n estimators)	100
subsamples	1.0
min_samples_split	2
min_samples_leaf	1

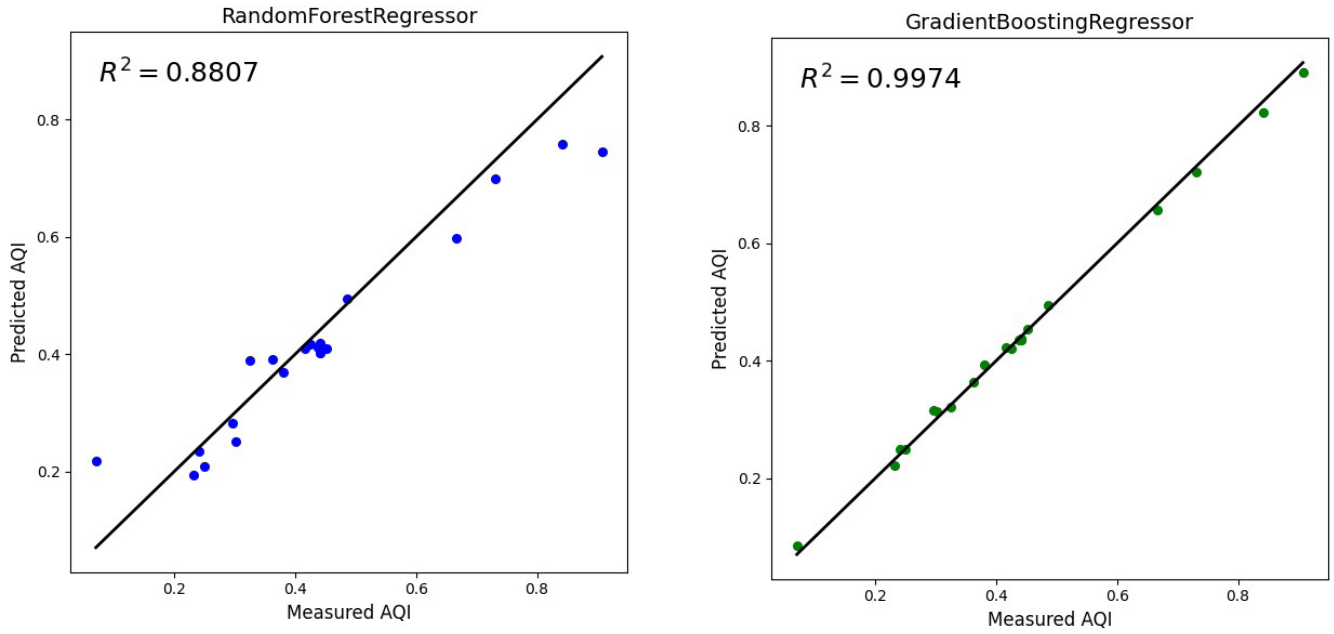


Figure 2. Random forest regressor and gradient boosting regressor predicted model.

The Gradient Boosting Regressor was implemented with carefully selected hyperparameters to optimise predictive performance while maintaining generalisation capability. Tree complexity was constrained by setting max_depth=2, effectively limiting the depth of individual weak learners to prevent overfitting and ensure model interpretability. A conservative learning rate (η=0.1) was employed to facilitate stable optimisation through gradual, additive model updates, achieving an optimal balance between training efficiency and predictive accuracy. The ensemble comprised n_estimators=100 sequential regression trees, providing sufficient model capacity while avoiding excessive computational overhead. Experimental reproducibility was ensured through random_state=0 initialisation, enabling exact replication of results.

Default parameters were retained where appropriate, though their influence on model behaviour warrants discussion. The subsample ratio maintained its default value of 1.0, utilising all available training samples for each boosting iteration. While this approach maximises data utilisation, future work could explore stochastic gradient boosting (subsample<1.0) to potentially enhance generalisation through variance reduction. Node splitting criteria required min_samples_split=2 instances to partition internal nodes, while terminal nodes were permitted to form with min_samples_leaf=1 observation. This configuration allows for detailed pattern capture while potentially increasing sensitivity to training set noise, suggesting a trade-off between model fidelity and robustness that may require domain-specific tuning.

3.3.1 XGBOOST

XGBoost is a refined distributed gradient boosting library that focuses on achieving high levels of efficiency, flexibility, and portability. It includes a variety of machine learning algorithms that are based on the Gradient Boosting framework. The name XGBoost is derived from Extreme Gradient Boosting. The developer is Tianqi Chen, an engineering Ph.D. student at the University of Washington. One of the reasons it is gaining popularity is the ability to handle high cardinality data and speed, which can handle very large datasets significantly faster than other gradient boosting libraries. It has a mechanism to handle missing values automatically. Missing value issues are common for real-world datasets, and this is just one of the reasons XGBoost is becoming popular. It supports both regression and classification types of problems. It also involves the term of regularisation to control the complexity of the model and prevent overfitting. Hence, XGBoost can be regarded as one summit of the gradient boosting algorithms in the context of accuracy and efficiency provided by it in almost all the applications involved in machine learning.

Table 6. Hyperparameters and metrics for the XGBoost Model

Parameter	XGBoost Model Value
Max Depth	2
Learning rate	0.1
Random State	0
Number of Estimators (n estimators)	100
booster	'gbtree'
subsample	1.0

Objective function:

$$Obj(\Theta) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t)}}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad (5)$$

Where, $\Omega\left(f_k\right) = \gamma T + 12\lambda \sum_{j=1}^{T} w_j^2$

$$Leaf\ \ weight\ \ update: w_j = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \ddot{e}}$$

3.3.2 XG BOOST-RF

XGBoost-RF is a new ensemble-based hybrid of compound because it benefits from the outcomes of both Random Forest and Extreme Gradient Boosting techniques. By incorporating feature randomness as found in Random Forests with the boosting mechanism of XGBoost, the hybrid model is supposed to boost the strength and robustness of prediction. An ensemble of decision trees is built by XGBoost-RF at training according to the diversity of the Random Forest, which captures complex patterns in the data. Then, it uses the boosting technique by XGBoost to iteratively improve the performance of the model by minimising the loss function. A highly effective model for a variety of tasks, from regression to classification, this method combines different methodologies. Nevertheless, effective parameter tuning and model selection are critical to fully exploit the capabilities of XG Boost -RF, XG Boost.

Table 7. Hyper parameters and metrics for XGBOOST-RF model

Parameter	XGBoost-RF Model Value
Max Depth	2
Random State	0
Number of Estimators (n estimators)	100
subsample	0.8
colsample_bynode	0.8
reg_lambda	1
booster	'gbtree'

Objective function:

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t)}}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad (6)$$

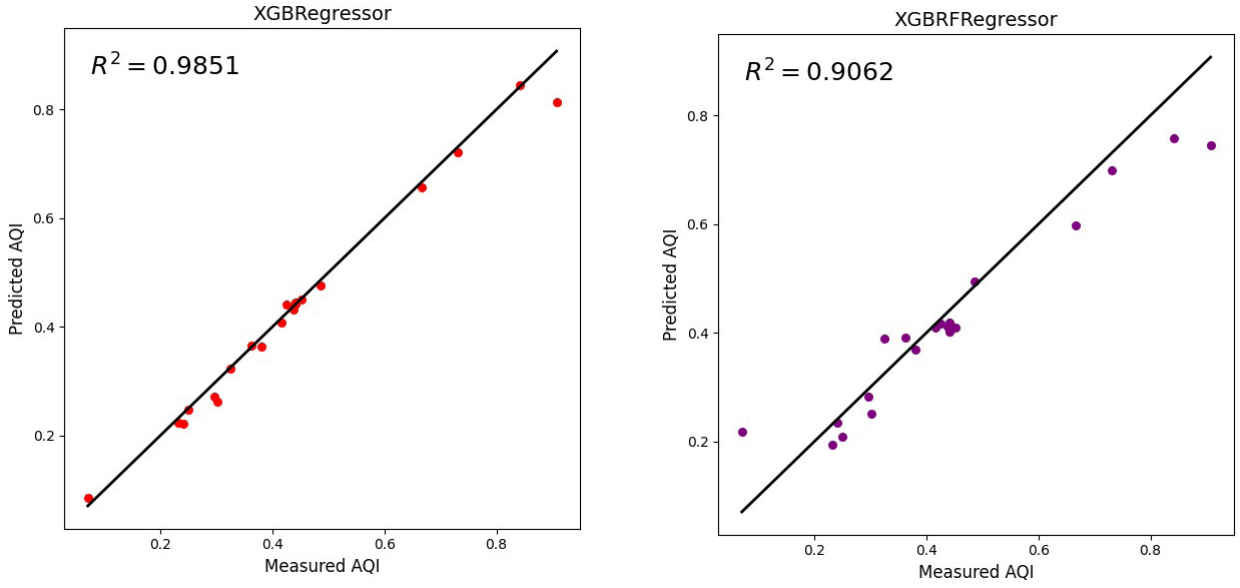


Figure 3. XGB Regressor and XG BOOST-RF regressor model.

4.0 Results and Discussions

This study presents an experimental analysis of the efficiency of four distinct machine learning algorithms utilised for air pollution detection, specifically focusing on Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), and a hybrid model combining XGBoost with Random Forest (XGBRF). These models were trained and tested using a dataset comprising hourly measurements of air quality, meteorological data, and geographical features from various monitoring stations.

Table 8 summarises the R^2, MAE and MSE values of each prediction model. The key findings of the analysis shows that MAE and MSE for Gradient Boosting algorithm were the smallest, indicating high precision in air pollution levels prediction compared to remaining models. With a high R^2 of 0.9974, this model can capture a considerable part of variance in data by the predictions.

Table 8. Comparison of evaluation metrics for all developed ML algorithms

Model	MAE	MSE	R2
Random forest	0.3844	0.0029	0.8807
Gradient boosting	0.0081	0.0001	0.9974
XGBOOST	0.1348	0.0005	0.9851
XGBOOST-RF	0.0439	0.0036	0.9062

Hence, the outstanding performance of Gradient Boosting for the task of predicting air quality proves to be very effective.

5.0 Conclusion

In the current work, the analysed air pollution data is related to the Air Quality Index. The observation of interest for this analysis is:

- The model formulation incorporates various inputs, including the following: Months, PM10 measured in $\mu g/m^3$, SO_2 in $\mu g/m^3$, NO_x in $\mu g/m^3$, *PM2.5* in $\mu g/m^3$, Ammonia (NH_3) in $\mu g/m^3$, O_3 in $\mu g/m^3$, CO in mg/m^3, and Benzene in $\mu g/m^3$, with the output variable being the Air Quality Index (AQI).
- A total of four machine learning algorithms have been developed to address the AQI prediction challenge, specifically XGBoost, Random Forest, and Gradient Boosting, along with their ensemble methods. The efficacy of these algorithms is assessed through the coefficient of determination (R^2) and Mean Squared Error (MSE).
- The results indicate that the Gradient Boosting algorithm achieves the highest accuracy with an R^2 of 0.99737955298 and the lowest error with an MSE of 0.0000998544. Additionally, the standalone XGBoost algorithm provides an R^2 of 0.9850808453 and an MSE of 0.000568507.

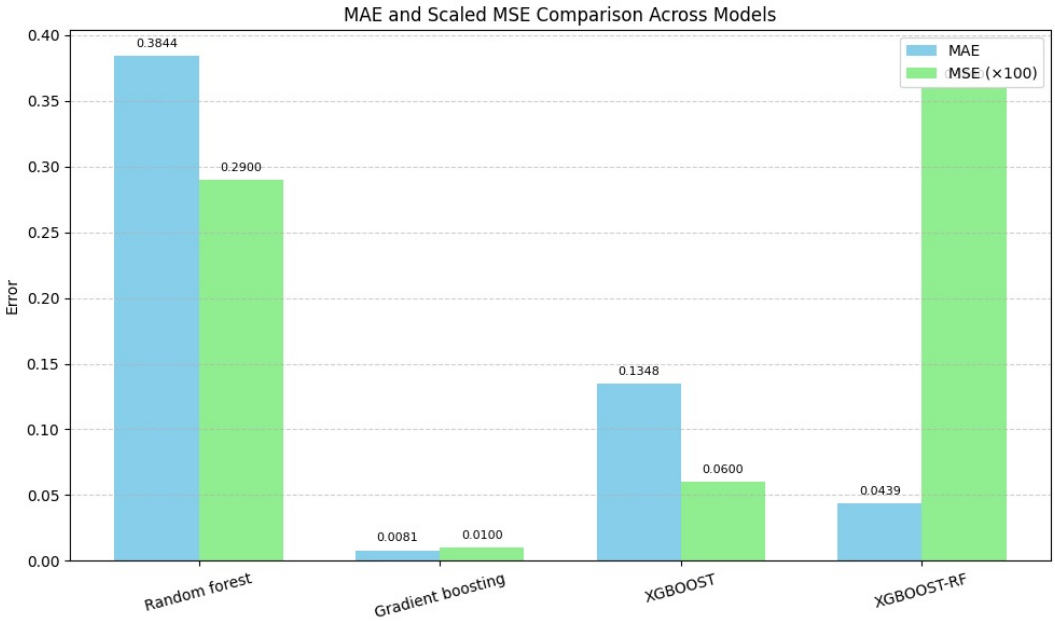


Figure 4. Comparison of scaled MAE and MSE of models.

- The findings presented in Table 8 indicate that subsequent research reveals the ensemble model Gradient Boosting outperforms individual algorithms, including Random Forest, XG Boost and XG Boost-RF. Consequently, these results suggest that gradient boosting serves as an effective model for predicting air quality. Consequently, the research findings indicate that gradient boosting is highly effective in the domain of air quality prediction, significantly surpassing the performance of algorithms employed individually.

6.0 References

1. Shaban KB, Kadri A, Rezk E. Urban air pollution monitoring system with forecasting models. IEEE Sens J. 2016; 16(8):2598-606. https://doi.org/10.1109/JSEN.2016.2514378
2. Kumar K, Pande BP. Air pollution prediction with machine learning: a case study of Indian cities. Int J Environ Sci Technol. 2022; 20. https://doi.org/10.1007/s13762-022-04241-5
3. Pai SJ, Carter TS, Heald CL, Kroll JH. Updated World Health Organisation air quality guidelines highlight the importance of non-anthropogenic PM2.5. Environ Sci Technol Lett. 2022; 9(6):501-6. https://doi.org/10.1021/acs.estlett.2c00203
4. Apte JS, Brauer M, Cohen AJ, Ezzati M, Pope CA 3rd. Ambient PM2.5 reduces global and regional life expectancy. Environ Sci Technol Lett. 2018; 5(9):546-51. https://doi.org/10.1021/acs.estlett.8b00360
5. Xu D, Zhang Y, Zhou L, *et al*. Acute effects of PM2.5 on lung function parameters in schoolchildren in Nanjing, China: a panel study. Environ Sci Pollut Res. 2018; 25:14989-95. https://doi.org/10.1007/s11356-018-1693-z
6. Lourenco Niza I, Bueno AM, Gameiro da Silva M, Broday EE. Air quality and ventilation: exploring solutions for healthy and sustainable urban environments in times of climate change. Results Eng. 2024; 24:103157. https://doi.org/10.1016/j.rineng.2024.103157
7. Jayachandran V, Narayana Rao T. Long-term regional air pollution characteristics in and around Hyderabad, India: effects of natural and anthropogenic sources. Atmos Environ X. 2024; 22:100254. https://doi.org/10.1016/j.aeaoa.2024.100254
8. Rao PV, Rao K, Raveendhar N, Swamy AVVS. Status of air pollution in Hyderabad City, Telangana State. Int J Innov Res Sci Eng Technol. 2016; 5(4):4769-80.
9. Memon DA, Brohi KM, Panhwar MA, Memon AR, Bhutto AA. Effect of air pollution on human health in manufacturing industries at SITE area, Hyderabad. Indian J Sci Technol. 2019; 12(23). https://doi.org/10.17485/ijst/2019/v12i23/144321
10. Vaddiraju SC. Mapping of air pollution using GIS: a case study of Hyderabad. Int J Res Advent Technol. 2019; 7(5). https://doi.org/10.32622/ijrat.752019328
11. Ravindiran G, Hayder G, Kanagarathinam K, Alagumalai A, Sonne C. Air quality prediction by machine learning models: a predictive study on the Indian coastal city of Visakhapatnam. Chemosphere. 2023; 338:139518. https://doi.org/10.1016/j.chemosphere.2023.139518
12. Watson AY, Bates RR, Kennedy D, editors. Air pollution, the automobile, and public health. Washington (DC): National Academies Press; 1988.
13. Appel KW, Gilliland AB, Sarwar G, Gilliam RC. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: sensitivities impacting model performance-part I: ozone. Atmos Environ. 2007; 41(40):9603-15. https://doi.org/10.1016/j.atmosenv.2007.08.044
14. Nath P, Saha P, Middya AI, Roy S. Long-term time-series pollution forecast using statistical and deep learning methods. Neural Comput Appl. 2021; 33(19):12551-70. https://doi.org/10.1007/s00521-021-05901-2
15. Drewil GI, Al-Bahadili RJ. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. Meas Sens. 2022; 24:100546. https://doi.org/10.1016/j.measen.2022.100546
16. Liang M, Chao Y, Tu Y, Xu T. Vehicle pollutant dispersion in the urban atmospheric environment: a review. Atmosphere. 2023; 14(2):279. https://doi.org/10.3390/atmos14020279
17. Hou T, Yu S, Jiang Y, *et al*. Impacts of chemical initial conditions in the WRF-CMAQ model on the ozone forecasts in Eastern China. Aerosol Air Qual Res. 2022; 22:210402. https://doi.org/10.4209/aaqr.210402
18. Wu EM, Kuo SL. A study on the use of a statistical analysis model to monitor air pollution status in an air quality total quantity control district. Atmosphere. 2013; 4(4):349-64. https://doi.org/10.3390/atmos4040349
19. Liu J, Zheng Y, Lee C. Statistical analysis of the regional air quality index of Yangtze River Delta based on complex network theory. Appl Energy. 2024; 357:122529. https://doi.org/10.1016/j.apenergy.2023.122529
20. Wei W, Ramalho O, Malingre L, *et al*. Machine learning and statistical models for predicting indoor air quality. Indoor Air. 2019; 29(5):704-26. https://doi.org/10.1111/ina.12580
21. Gupta S, Mohta Y, Heda K, *et al*. Prediction of air quality index using machine learning techniques: a comparative

analysis. J Environ Public Health. 2023; 2023:4916267. https://doi.org/10.1155/2023/4916267
22. Natarajan SK, Shanmurthy P, Arockiam D, *et al*. Optimised machine learning model for air quality index prediction in major cities in India. Sci Rep. 2024; 14:6795. https://doi.org/10.1038/s41598-024-54807-1
23. Madan T, Sagar S, Virmani D. Air quality prediction using machine learning algorithms-a review. In: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN); 2020; Greater Noida, India; 2020 .p. 140-5. https://doi.org/10.1109/ICACCCN51052.2020.9362912
24. Unnikrishnan R, Madhu G. Comparative study on the effects of meteorological and pollutant parameters on ANN modelling for prediction of SO2. SN Appl Sci. 2019; 1:1394. https://doi.org/10.1007/s42452-019-1440-1
25. Yaqoob I, Kumar V, Chaudhry SA. Machine learning calibration of low-cost sensor PM2.5 data. In: 2024 IEEE International Symposium on Systems Engineering (ISSE); 2024; 1-8. https://doi.org/10.1109/ISSE63315.2024.10741149
26. Rani VU, Thirisaa S. Analysis of various machine learning regression models for AQI prediction. In: 2023 International Conference on Intelligent Systems for Communication, IoT and Security (CISCT); 2023. p. 1-5. https://doi.org/10.1109/CISCT57197.2023.10351428
27. Rochadiani TH. Prediction of air quality index using ensemble models. J Appl Inform Comput. 2024; 8(2):384-9. https://doi.org/10.30871/jaic.v8i2.8532
28. Liu H, Li Q, Yu D, Gu Y. Air quality index and air pollutant concentration prediction based on machine learning algorithms. Appl Sci. 2019; 9(19):4069. https://doi.org/10.3390/app9194069
29. Zhang X, Jiang X, Li Y. Prediction of air quality index based on the SSA-BiLSTM-LightGBM model. Sci Rep. 2023; 13(1):6627. https://doi.org/10.1038/s41598-023-32775-2
30. Rajagopal KR, Narayanan K. A novel approach for air quality index prognostication using hybrid optimization techniques. Int Res J Multidiscipl Technovation. 2024; 6(2):84-99. https://doi.org/10.54392/irjmt2427
31. Džaferović E, Karađuzović-Hadžiabdić K. Air quality prediction using machine learning methods: a case study of Bjelave neighborhood, Sarajevo, BiH. In: 2020 International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS); 2020; 423-34. https://doi.org/10.1007/978-3-030-54765-3_29
32. Subashini S, Deenumol A, Arjun A, *et al*. Comparative analysis of machine learning algorithms for AQI prediction. SSRN. 2025. https://doi.org/10.2139/ssrn.5089080
33. Udristioiu MT, EL Mghoughi Y, Yildizhan H. Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. J Clean Prod. 2023; 428:138496. https://doi.org/10.1016/j.jclepro.2023.138496
34. Dao T, Trung HQ, Dieu VH, *et al*. Analysis and prediction for air quality using various machine learning models. In: 2022 Federated Conference on Computer Science and Information Systems (FedCSIS); 2022; 89-94. https://doi.org/10.15439/2022R03
35. Narkhede G, Deore A, Kolte B. Machine learning models for daily AQI prediction: an in-depth analysis. In: 2024 International Conference on Intelligent Computing and Control Systems (ICICCS); 2024; 1-6. https://doi.org/10.1109/ICICEC62498.2024.10808250
36. Duong DQ, Le QM, Nguyen-Tai T, *et al*. Multi-source machine learning for AQI estimation. In: 2020 IEEE International Conference on Big Data (Big Data); 2020; 4567-76. https://doi.org/10.1109/BigData50022.2020.9378322
37. Chen NF, Yang S. Research on prediction model of Xiamen air quality index based on machine learning. In: 2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT); 2022.
38. M B. Predictive modelling for air quality: a multi-model regression approach. Int J Sci Technol Eng. 2024; 12(12):1368-78. https://doi.org/10.22214/ijraset.2024.66036
39. Su Y. Prediction of air quality based on the gradient boosting machine method. In: 2020 International Conference on Big Data and Informatisation Education (ICBDIE); 2020; 395-7. https://doi.org/10.1109/ICBDIE50010.2020.00099
40. Ketu S. Spatial air quality index and air pollutant concentration prediction using linear regression based Recursive Feature Elimination With Random Forest Regression (RFERF): a case study in India. Nat Hazards. 2022; 114(2):2109-38. https://doi.org/10.1007/s11069-022-05463-z
41. Ma J, Cheng JCP, Xu Z, *et al*. Identification of the most influential areas for air pollution control using XGBoost and grid importance rank. J Clean Prod. 2020; 274:122835. https://doi.org/10.1016/j.jclepro.2020.122835
42. Lin L, Liang Y, Liu L, *et al*. Estimating PM2.5 concentrations using the machine learning RF-XGBoost model in Guanzhong urban agglomeration, China. Remote Sens. 2022; 14(20):5239. https://doi.org/10.3390/rs14205239
43. Atom. Air pollution analysis: Hyderabad, India. Kaggle; 2024. Available from: https://www.kaggle.com/code/atom1991/air-pollution-analysis-hyderabad-india