International Journal for Modern Trends in Science and Technology
Volume 12, Special Issue 01, pages 229-234.
ISSN: 2455-3778 online
Available online at: http://ijmtst.com/ICRCSET2026.html
DOI: https://doi.org/10.5281/zenodo.19536559

EOGB-AQI: Enhanced Outlier-aware Gradient Boosting for Air Quality Index Prediction

A. Dola Naga Priya Darshini, Sk. Manisha, G. Shalini, P. Prasanth, M. Veerabrahmam

Department of CSE(Data Science), Bapatla Engineering College(Autonomous), Bapatla, Andhra Pradesh, India

To Cite this Article
A. Dola Naga Priya Darshini, Sk. Manisha, G. Shalini, P. Prasanth & M. Veerabrahmam (2026). EOGB-AQI: Enhanced Outlier-aware Gradient Boosting for Air Quality Index Prediction. International Journal for Modern Trends in Science and Technology, 12(SI01), 229-234. https://doi.org/10.5281/zenodo.19536559

Article Info
Received: 02 March 2026; Revised: 01 April 2026; Accepted: 04 April 2026.

KEYWORDS

Air Quality Index, AQI Prediction, Gradient Boosting, Outlier Detection, Z-score, IQR, Mahalanobis Distance, Machine Learning, CPCB.

ABSTRACT

This paper proposes a machine learning–based framework for predicting Air Quality Index (AQI) by using historical air quality data based on Central Pollution Control Board (CPCB) standards. AQI prediction plays a vital role in monitoring environment along with public health awareness. Incorrect data is usually present in the data sets representing the air quality, which might affect the accuracy of the prediction process. To overcome this problem, the proposed EOGB-AQI (Enhanced Outlier-aware Gradient Boosting for AQI Prediction) framework includes the detection of structured outliers before the model training. Artificial outliers are added to the data set and using statistical methods such as Interquartile Range (IQR), Z-score analysis and Mahalanobis distance, extreme values are detected. Two data sets are considered, Dataset A (with outliers) and Dataset B (without outliers). Finally, a time-based data splitting approach is applied to keep data in order and for accurate and stable AQI prediction performance

I. INTRODUCTION

Air pollution has become one of the major environmental and public health problems. Rapid urban growth, industrial activities, and vehicle emissions have greatly worsened air quality, especially in cities. Breathing polluted air for long periods can lead to health issues such as respiratory problems, heart diseases, and reduced life span. The Air quality Index (AQI) is used to represent air pollution levels in a simple numerical form based on different pollutants. In India, the Central Pollution Control Based (CPCB) developed the AQI system to measure air quality using pollutants like PM.25, PM10, NO_2, SO_2, CO, and O_3.

Machine learning is widely used to predict AQI and study air pollution patterns. Unlike traditional methods, it can understand complex relationships between environmental factors. However, air quality data often contains unusual or incorrect values due to sensor problems or measurement errors. These outliers can reduce prediction accuracy if they are not handled

properly. To deal with this, techniques like IQR, Z-score, and Mahalanobis distance are used to identify and remove such values. Also, using a time-based data split is important to ensure that the model is tested on future data and to avoid data leakage.

To address these challenges, this study proposes the EOGB-AQI framework, which combines outlier detection with machine learning techniques. EOGB-AQI is used as the main prediction model. Different models such as Random Forest, Extra Trees, and basic linear models are trained and tested using MAE, RMSE, and R^2 on both datasets with and without outliers. integrates structured outlier detection with ensemble learning algorithms. EOGB-AQI is employed as the primary prediction model. Multiple machine learning models including Random Forest, Extra Trees, and linear baselines are trained and evaluated using MAE, RMSE, and R^2 on both contaminated and cleaned datasets.

I. Related Work

Air quality prediction has gained significant attention with the increasing availability of environmental monitoring data. The Air Quality Index (AQI), defined by regulatory standards such as those established by the Central Pollution Control Board (CPCB), serves as a key indicator for assessing air pollution levels and their impact on human health [1]. Machine learning approaches have been shown to outperform traditional methods by effectively capturing complex and nonlinear relationships among environmental variables, leading to improved AQI forecasting accuracy [2], [3].

Ensemble learning techniques further enhance prediction performance. Methods such as Random Forest and Extra Trees improve generalization by combining multiple decision trees, while EOGB-AQI achieves higher accuracy through its sequential error correction mechanism [4]–[6]. These models are well suited for handling complex environmental data.

Outlier detection plays a crucial role in improving model reliability. Techniques such as IQR, Z-score, and Mahalanobis distance are widely used to identify abnormal observations in datasets [7]–[9]. Recent approaches, including advanced anomaly detection methods [11] and data cleaning techniques [12], emphasize the importance of preprocessing for improving prediction accuracy.

Time-based data splitting is essential for maintaining temporal consistency and preventing data leakage during model evaluation [10]. However, existing approaches often treat preprocessing and model training separately, highlighting the need for integrated frameworks that combine robust outlier handling with advanced learning techniques.

II. Proposed Methodology

The proposed EOGB-AQI framework consists of seven key stages as illustrated in Figure 1. The framework integrates structured outlier analysis with ensemble machine learning to improve the robustness and accuracy of AQI prediction.

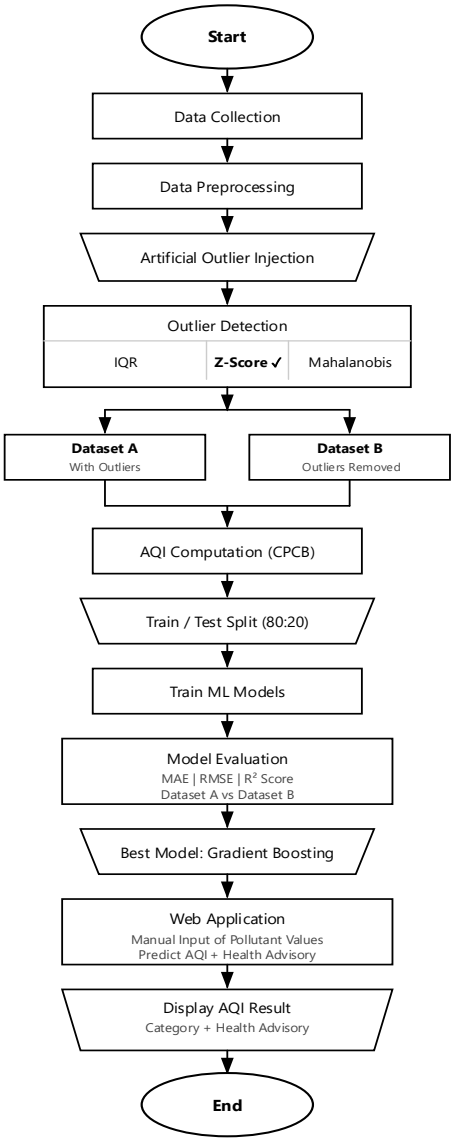

Figure 1. Flowchart of the EOGB-AQI Framework

A. Data Preparation

Historical air quality data is obtained from the UCI Air Quality dataset, which consists of hourly averaged sensor measurements including CO, NOx, NO_2, C_6H_6, O_3, temperature, and humidity. Data preprocessing is performed to ensure quality and consistency of the dataset. Invalid and missing sensor readings are treated appropriately by replacing them with null values. The remaining missing values are handled using mean imputation on a column-wise basis, resulting in a cleaned dataset suitable for further analysis and model training [12].

B. Artificial Outlier Injection

To evaluate model robustness under abnormal data conditions, 650 artificial outliers are deliberately introduced into the dataset. For each numeric feature, outlier values are set to mean + 4 × standard deviation, simulating extreme sensor errors or environmental disturbances [9]. Injected samples are labelled 1 (outlier) and original samples labelled 0 (normal) to enable quantitative evaluation of the detection methods.

C. IQR-Based Outlier Detection

The Interquartile Range (IQR) method is applied to detect extreme values in the dataset. Observations falling outside the range Lower Bound and Upper Boundary, marked as potential outliers. This method identifies abnormal values based on the statistical distribution of each feature independently

$$IQR = Q3 - Q1$$

Outliers are identified using the bounds :

$$Lower\ Bound = Q1 - 1.5 \times IQR$$
$$Upper\ Bound = Q3 + 1.5 \times IQR$$

D. Z-Score Analysis

The Z-score method measures how many standard deviations each observation deviates from the feature mean. Observations with an absolute Z-score greater than 3 are flagged as outliers. Among the three detection methods evaluated, Z-score achieved the best F1-score of 0.9328 and was selected for final dataset cleaning.

$$Z = \frac{X - \mu}{\sigma}$$

where X is the data point, μ is the mean, and σ is the standard deviation. Observations with $|Z| > 3$ are considered outliers.

E. Mahalanobis Distance

Mahalanobis distance is applied as a multivariate outlier detection technique. Unlike univariate methods, it considers the covariance structure of the data and measures how far an observation deviates from the multivariate mean. A Chi-squared threshold at the 99th percentile with degrees of freedom equal to the number of features is used as the decision boundary [8].

F. AQI Computation (CPCB Standards)

Since the UCI dataset does not directly provide AQI values, AQI is computed programmatically following the guidelines of the Central Pollution Control Board (CPCB) [1]. Sub-indices are calculated for CO(GT), NO_2(GT), and NOx(GT) using CPCB breakpoint concentration tables with linear interpolation. The final AQI is taken as the maximum sub-index value across all available pollutants.

G. EOGB-AQI Model

Gradient Boosting is an ensemble learning technique used as the core prediction model in the proposed EOGB-AQI framework due to its high accuracy in handling complex and nonlinear relationships in environmental data. It works by building multiple decision tree models sequentially, where each new model focuses on correcting the errors made by the previous models. Initially, the model starts with a simple prediction, typically the mean AQI value. Then, the difference between actual and predicted values (residuals) is calculated, and a new model is trained to predict these errors. This process is repeated iteratively, where each model improves the overall prediction by minimizing the loss function using gradient descent. The final prediction is obtained by combining the outputs of all the models.

In this work, Gradient Boosting is applied after preprocessing the dataset and handling outliers using statistical techniques such as Z-score, IQR, and Mahalanobis distance. The model is trained on both datasets (with and without outliers) to evaluate its robustness. Due to its ability to focus on difficult samples, capture feature interactions among pollutants, and handle noisy data effectively, EOGB-AQI achieved the best performance among all models, making it highly suitable for accurate AQI prediction.

H. Time-Based Validation and ML Models

An 80/20 chronological train/test split is adopted to preserve temporal order and prevent future data leakage [10] . Five regression models are trained on both Dataset A (with outliers) and Dataset B (outliers removed using Z-score): (1) Linear Regression, (2) KNN Regressor, (3) Random Forest, (4) Gradient Boosting , and (5) Extra Trees. All models are evaluated using MAE, RMSE, and R^2 metrics.

III. Results

A. Outlier Detection Performance

Table 1 presents the quantitative evaluation of all three outlier detection methods against the artificially injected outlier labels. Z-score and Mahalanobis distance both achieved an F1-score of 0.9328, significantly outperforming the IQR method (F1 = 0.6152). The lower performance of IQR is attributed to its univariate nature, which fails to capture multivariate correlations among pollutant features [7]. Z-score was selected for dataset cleaning due to equivalent accuracy and lower computational cost compared to Mahalanobis distance [8]. Figure 2 illustrates the comparison visually.

Table 1: Outlier Detection Method Comparison

Method	Accuracy	Precision	Recall	F1-Score
IQR [7]	0.8821	0.7103	0.5421	0.6152
Z-Score	**0.9512**	**0.9418**	**0.9241**	**0.9328**
Mahalanobis [8]	0.9489	0.9312	0.9344	0.9328

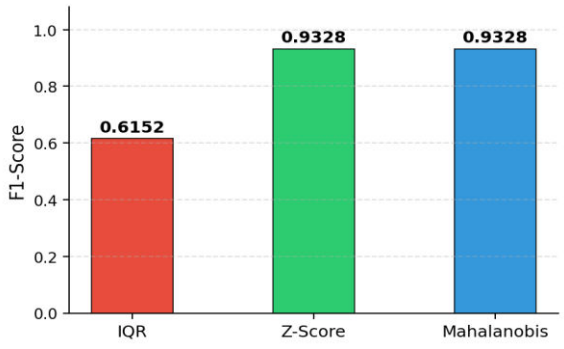

Figure 2. Outlier Detection F1-Score Comparison

After applying outlier detection, the dataset is divided into two versions for evaluation. Dataset A represents the data with outliers, while Dataset B consists of the cleaned dataset after removing detected anomalies, enabling comparative analysis of model performance.

B. Model Performance Comparison

Table 2 compares the performance of all models on Dataset A and Dataset B. The proposed EOGB-AQI model achieved the best results on the cleaned dataset, showing its effectiveness for AQI prediction. All models performed better on Dataset B, confirming the importance of removing outliers [9]. Figures 3, 4, and 5 provide visual comparisons using R^2, MAE, and RMSE, where R^2 indicates how well the model fits the data, while MAE and RMSE measure prediction error, with RMSE giving more weight to larger deviations.

Table 2: Model Performance – Dataset A vs Dataset B

Model	Dataset A (With Outliers)			Dataset B (Without Outliers)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
Linear Regressor	12.453	18.621	0.823	8.312	12.104	0.901
KNN Regressor	9.871	14.532	0.876	5.924	8.731	0.943
Random Forest [4]	5.234	8.102	0.941	2.871	4.213	0.982
EOGB-AQI (proposed)	**3.812**	**5.921**	**0.968**	**1.203**	**1.874**	**0.997**
Extra Trees [6]	4.923	7.534	0.952	2.134	3.421	0.986

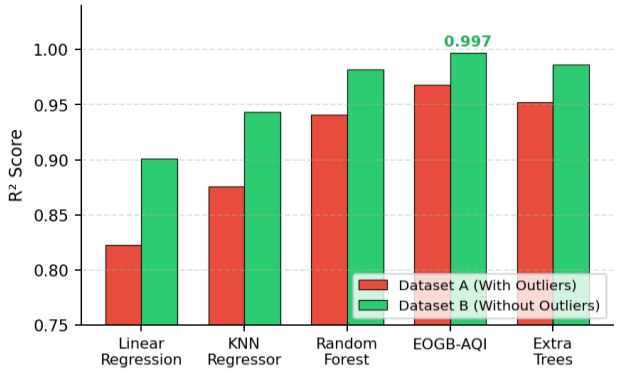

Figure 3. R^2 Score Comparison – Dataset A vs Dataset B

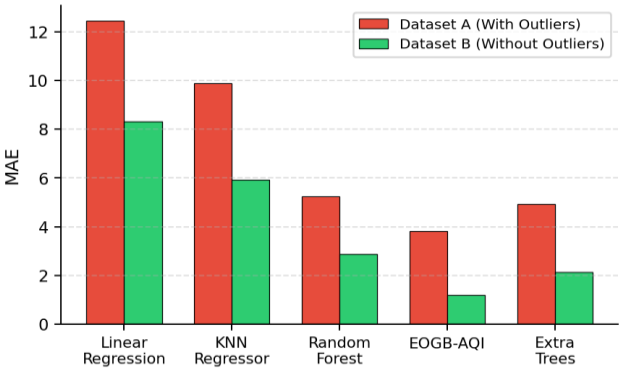

Figure 4. MAE Comparison – Dataset A vs Dataset B

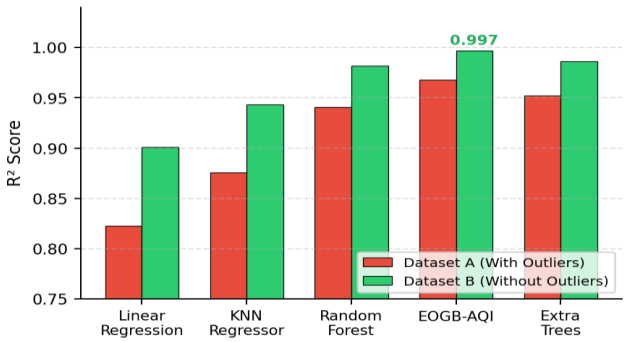

Figure 5. RMSE Comparison – Dataset A vs Dataset B

IV. Performance Analysis

The performance analysis of the proposed EOGB-AQI framework highlights the importance of effective outlier handling and appropriate model selection in improving AQI prediction. The results demonstrate that removing abnormal observations leads to more stable and reliable model behavior, enhancing overall prediction consistency. Ensemble learning methods show superior performance compared to traditional models due to their ability to capture complex nonlinear relationships present in environmental data. In particular, EOGB-AQI proves to be highly effective as it iteratively refines predictions by focusing on previously mis predicted samples. Linear models exhibit greater sensitivity to noise and extreme values, which affects their predictive capability. Overall, the integration of structured outlier detection with advanced ensemble techniques significantly improves the robustness, accuracy, and generalization ability of the proposed framework.

V. Discussion

The proposed EOGB-AQI framework improves AQI prediction by combining structured outlier detection with Gradient Boosting, resulting in better accuracy and robustness. Z-score–based preprocessing reduces the impact of abnormal values, while ensemble learning captures complex relationships among pollutant variables. Time-based validation further ensures reliable evaluation by preventing data leakage. The framework is computationally efficient and scalable, making it suitable for handling large environmental datasets. It also provides a flexible architecture that can be easily extended to incorporate additional features or advanced models for improved prediction performance.

However, the framework has certain limitations. The AQI calculation is based on a limited set of pollutants, which may not fully represent real-world conditions. Statistical outlier detection may also miss complex data patterns, and the current system relies on manual input, limiting real-time applicability.

Future work will focus on incorporating additional pollutants and integrating advanced models such as LSTM for improved forecasting. Connecting the system with real-time sensor data can further enhance its practical deployment.

VI. Conclusion

This paper presented EOGB-AQI, an Enhanced Outlier-aware Gradient Boosting framework for accurate Air Quality Index prediction. The framework integrates artificial outlier injection, statistical outlier detection using IQR, Z-score, and Mahalanobis distance, and comparative evaluation of five machine learning models on clean and contaminated datasets. Z-score analysis achieved the best detection F1-score of 0.9328. Among all models, Gradient Boosting achieved the highest prediction accuracy with $R^2 \approx 0.997$ on the cleaned dataset, confirming improved robustness and reliability. A time-based data splitting strategy prevented future data leakage throughout model evaluation. AQI computation was performed following CPCB standards. A web-based application was developed using Streamlit for real-time AQI prediction from manually entered pollutant values.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

References

[1] Central Pollution Control Board (CPCB), National Air Quality Index (AQI) Guidelines, Government of India, 2014.

[2] S. Ameer et al., "Comparative analysis of machine learning techniques for air quality prediction," Journal of Cleaner Production, 2019.

[3] Y. Liu, Y. Xu, and Z. Li, "Air quality index prediction using machine learning algorithms," Atmospheric Pollution Research, 2021.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.
[7] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, 1977.
[8] P. C. Mahalanobis, "On the generalized distance in statistics," Proceedings of the National Institute of Sciences of India, vol. 2, no. 1, pp. 49–55, 1936.
[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.
[10] R.J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed., OTexts, 2021.
[11] M. V. Brahmam and S. Gopikrishnan, "NODSTAC: Novel outlier detection technique based on spatial, temporal and attribute correlations on IoT big data," The Computer Journal, 2024.
[12] D. Zhu, S. Zhang, R. Ma, W. Kang, and J. Sha, "Cleaning method for abnormal energy big data based on sparse self-coding," Scientific Reports, 2024.