INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Optimized and Explainable Air Quality Index Classification System

Neethu Roy, Jeeson Justin

Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600278

Received: 10 July 2026; Accepted: 15 July 2026; Published: 03 August 2026

ABSTRACT

Air pollution has become a major environmental and public health concern due to rapid urbanization, industrial growth, and increasing vehicular emissions. High concentrations of pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 can significantly impact human health and environmental sustainability. Accurate monitoring and prediction of air quality are therefore essential for effective environmental management and public safety. This paper presents AirAware, a machine learning–based system designed to predict and monitor Air Quality Index (AQI) levels using historical air pollution data and real-time environmental information. The system utilizes the XGBoost algorithm to analyze pollutant parameters and generate accurate AQI predictions and classifications. Data preprocessing techniques such as cleaning, normalization, and SMOTE-based class balancing are applied to improve model performance and ensure reliable predictions across different AQI categories. In addition, the system integrates real-time air pollution data through the OpenWeather API, enabling continuous monitoring of current environmental conditions. The predicted AQI values and pollution trends are displayed through a web-based dashboard, allowing users to visualize air quality patterns and compare real-time data with machine learning predictions. By combining machine learning techniques with real-time data integration, the proposed system provides an effective solution for air quality prediction, monitoring, and environmental awareness.

Keywords: Air Quality Index, XGBoost, Explainable AI, SHAP, SMOTE, Machine Learning.

INTRODUCTION

Air quality prediction has gained significant attention in recent years due to the increasing impact of air pollution on human health and environmental sustainability. Several studies have explored the use of machine learning techniques for predicting air pollution levels. Machine learning models can analyze large environmental datasets and identify complex relationships between pollutant concentrations and meteorological conditions such as temperature, humidity, wind speed, and atmospheric pressure

Among various machine learning approaches, ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost have demonstrated superior performance in environmental prediction tasks. These models improve prediction accuracy by combining multiple decision trees and minimizing prediction errors through iterative optimization techniques. As a result, they are widely used in air quality forecasting and environmental monitoring systems.

Motivated by these advancements, the AirAware system is designed to integrate machine learning algorithms with real-time environmental data to develop an intelligent AQI prediction platform. The system not only predicts air quality levels but also provides visualization tools to help users understand pollution trends and environmental conditions.

Air pollution monitoring systems provide real-time pollution data but often lack accurate prediction capabilities. Without prediction, people cannot prepare for upcoming pollution events or take preventive actions to reduce health risks. Additionally, traditional statistical approaches struggle to model the complex nonlinear relationships between pollutant concentrations and meteorological conditions. Recent studies show that machine learning models can significantly improve air quality prediction accuracy by analyzing historical pollution data and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

environmental factors.

Furthermore, many existing systems focus only on monitoring pollution levels rather than predicting future AQI trends. This creates a gap between environmental monitoring and decision-making. Therefore, there is a need for an intelligent system that can analyze historical air pollution data, integrate real-time pollution information through APIs, and predict AQI levels using advanced machine learning techniques. The proposed system addresses this problem by implementing a machine learning–based AQI prediction model with real-time data visualization and analysis.

Related Work

Air quality prediction shows an increasing use of machine learning techniques to model the complex relationships between air pollutants and the Air Quality Index (AQI). Earlier approaches mainly used statistical and rule-based methods, which often struggled with nonlinear pollutant interactions and large datasets. Recent studies focus on machine learning and ensemble models to improve prediction accuracy and reliability. These studies also highlights the importance of data preprocessing, handling class imbalance, and feature selection for building effective AQI prediction systems.

Long-Term AQI Prediction Using Ensemble Machine Learning Models

A long-term study conducted in eastern Türkiye (2016–2024) employed three machine learning models — XGBoost, LightGBM, and SVM — to predict daily Air Quality Index (AQI) using a dataset comprising four major pollutants (PM_{10}, SO_2, NO_2, O_3) and five meteorological variables including temperature, humidity, precipitation, and wind conditions, evaluated through R^2, RMSE, and MAE metrics. XGBoost emerged as the strongest performer, achieving near-perfect prediction accuracy with $R^2 = 0.999$, RMSE = 0.234, and MAE = 0.158, clearly demonstrating its superior ability to model complex AQI fluctuations driven by environmental factors. The findings confirm that ensemble-based machine learning approaches, particularly XGBoost, hold tremendous practical value for developing reliable air quality forecasting systems, early warning mechanisms, and data-informed environmental health policies at the regional level.

Weighted Voting Ensemble Model with SHAP for Interpretable AQI Prediction

A study introduced a weighted voting ensemble model strategically combining Gradient Boosting, CatBoost, XGBoost, and LightGBM — optimized through GridSearchCV and Optuna with 5-fold cross-validation — trained on the Taiwan Air Quality Dataset (2016–2024) encompassing 4.6 million hourly records from 74 stations, covering six major pollutants and key meteorological parameters, achieving a remarkable validation R^2 of 0.9969 that outperformed 15 baselines including LSTM. SHAP analysis further enhanced model transparency by revealing each pollutant's individual contribution to AQI predictions, reinforcing the system's potential as a reliable, interpretable, and scalable solution for real-time urban air quality management and community health interventions.

Urban Air Quality Prediction Using CatBoost and Ensemble Learning

A study conducted in Visakhapatnam, Andhra Pradesh, India applied multiple machine learning models — LightGBM, Random Forest, CatBoost, AdaBoost, and XGBoost — on a dataset spanning July 2017 to September 2022, covering 12 air contaminants and 10 meteorological parameters to predict AQI in one of India's rapidly industrializing urban centers. CatBoost emerged as the best-performing model with an outstanding R^2 of 0.9998, MAE of 0.60, and RMSE of 0.76, firmly establishing machine learning — and CatBoost in particular — as a highly reliable and scalable approach for accurate urban air quality forecasting on both local and global scales.

Integration of Air Pollution and Clinical Data for Cardiovascular Mortality Prediction

A study in Malaysia merged clinical records from the National Cardiovascular Disease Database with daily air quality data (NOx, SO_2, O_3, PM10) from 2006–2017, applying machine learning models — Logistic Regression, Random Forest, XGBoost, and Ensemble Learning — to predict in-hospital ACS mortality, where the Random Forest model achieved the highest AUC of 0.843, significantly surpassing conventional TIMI risk scores. SHAP analysis further revealed that environmental pollutants NOx and O_3, alongside clinical indicators like Killip class and fasting blood glucose, were the strongest mortality predictors — demonstrating that integrating environmental and clinical data within a machine learning framework offers a far more comprehensive and accurate approach to cardiovascular risk assessment.

Comparison of Machine Learning and Deep Learning for Particulate Matter Forecasting

A study evaluating particulate matter forecasting models using air quality data from Maharashtra (2019–2023) compared traditional machine learning approaches — Linear Regression, Decision Tree, Random Forest, and XGBoost — against a deep learning-based Long Short-Term Memory (LSTM) model, assessed through R^2 Score and RMSE metrics. LSTM decisively operformed all conventional models, achieving superior R^2 scores between 0.99 and 0.998 across five cities — compared to R^2 scores as low as 0.15 for traditional models — firmly establishing deep learning as a more powerful and reliable solution for capturing complex temporal dependencies in PM2.5 and PM10 concentration forecasting.

Short-Term Air Quality Forecasting Using Machine Learning and SHAP-Based Feature Selection

A study developed short-term air quality forecasting models for Macau using diurnal measurements collected from 2016 to 2021, applying five machine learning algorithms — ANN, Random Forest, XGBoost, SVM, and Multiple Linear Regression — to predict PM2.5, PM10, and CO concentrations across both 24-hour and 48-hour forecasting horizons, with 12 SHAP analysis further employed to refine feature selection and improve model performance. Among all tested models, Random Forest and SVM consistently delivered the strongest predictive accuracy for both PM2.5 and PM10 across the two forecasting windows, confirming that well-tuned machine learning models can serve as reliable and practical tools for early public warning systems in pollution-prone urban environments like Macau.

SMOTE-Based Data Balancing for Air Pollution Classification

A study addressing the challenge of data imbalance in air pollution classification implemented the SMOTE technique alongside XGBoost and Random Forest algorithms, evaluating model performance across an 80:20 train-test split to determine the most effective approach for accurate pollution-related health risk classification. Random Forest with SMOTE emerged as the best-performing combination, achieving an impressive accuracy of 92.4%, AUC of 0.98, and log loss of 0.2366 — demonstrating that applying SMOTE significantly enhances model performance in identifying minority classes, with both XGBoost and Random Forest surpassing 90% accuracy post-resampling.

Explainable Artificial Intelligence for Transparent Air Quality Prediction

A study explored the integration of Explainable Artificial Intelligence (XAI) with deep learning to enhance the transparency and reliability of air quality prediction, employing XGBoost and KNN for data classification while addressing the well-known limitation of traditional methodologies — their lack of interpretability and stakeholder trust. SHAP and LIME explainability techniques were subsequently applied to identify the most influential variables driving air quality predictions, demonstrating that combining powerful classification models with explainability frameworks significantly improves decision-making clarity and supports the development of more effective, evidence-based air quality management strategies.

Hybrid Explainable AI Framework for Real-Time Air Quality Prediction

A study developed a hybrid AI framework for real-time air quality prediction in Afghanistan — combining ensemble models (Random Forest and XGBoost) with deep learning architectures (LSTM, CNN, and TSMixer) — trained on historical pollution data from NEPA alongside real-time meteorological data from OpenWeather API, with geospatial clustering, SHAP, and LIME employed to enhance regional accuracy and model interpretability, all deployed through a Django-based API and interactive dashboards. TSMixer achieved the

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

highest regression performance (R^2 = 0.9861), while Random Forest led classification tasks 13 with an outstanding accuracy of 99.96%, with NO_2 and PM_{10} identified as the most influential pollution indicators — collectively demonstrating that interpretable, hybrid AI systems can serve as powerful and practical tools for real-time air quality governance in under-resourced regions.

Hybrid Machine Learning Framework for Meteorological Feature-Based AQI Prediction

A study conducted in Jinan, China (July 2020–July 2021) introduced an advanced air quality prediction methodology that combined rigorous correlation analysis, univariate and multivariate significance testing, and Random Forest-based feature ranking to identify ten key meteorological factors influencing six major pollutants, while also incorporating seasonal characteristic analysis to capture the distinct impact of temperature, humidity, and atmospheric pressure on pollution levels. Among all evaluated models, LightGBM emerged as the strongest classifier with 97.5% accuracy and an F1 score of 93.3%, while LSTM proved superior for AQI regression tasks — achieving a goodness-of-fit of 91.37% for AQI and 90.46% for O_3 predictions — collectively affirming that hybrid machine learning frameworks leveraging both measured and forecasted meteorological data offer a highly reliable approach to air quality forecasting.

From the above studies, it is evident that machine learning models such as XGBoost, Random Forest, LightGBM, and LSTM have significantly improved AQI prediction accuracy across various environmental datasets, with techniques like SMOTE, SHAP, and real-time API integration further strengthening their practical reliability and interpretability. However, most existing works address prediction or monitoring in isolation, leaving a clear gap for a unified system that combines both capabilities. Therefore, the proposed system bridges this gap by integrating SMOTE-based data balancing, Random Forest and XGBoost model comparison, and real-time AQI visualization through an interactive Flask-based web dashboard — delivering a comprehensive and accessible platform for informed environmental decision making.

Proposed System

The proposed Air Quality Index (AQI) Prediction System is built to deliver a reliable and intelligent solution for forecasting air pollution levels by harnessing the power of machine learning, a pressing necessity given the escalating impact of air pollution on both public health and environmental sustainability. At its core, the system leverages two powerful algorithms — Random Forest and XGBoost — trained on historical air quality data to improve prediction accuracy and capture complex relationships between pollutant parameters. The Synthetic Minority Over-Sampling Technique (SMOTE) is strategically applied to tackle class imbalance, ensuring the model learns fairly and accurately across all AQI categories.

The trained model is seamlessly deployed within a Flask-based web application, offering users an intuitive and interactive dashboard that visualizes predicted AQI values alongside live pollution data streamed through an external real-time API. This enables direct comparison between forecasted and actual air quality conditions and supports continuous monitoring of environmental data. Altogether, the system bridges the worlds of predictive analytics and real-time environmental monitoring, empowering individuals, researchers, and policymakers with actionable insights to make informed, data-driven decisions toward a cleaner and healthier environment.

Modular Description

Air pollution has emerged as one of the most pressing environmental and public health crises facing urban communities today, as steadily rising concentrations of harmful pollutants continue to threaten human health and disrupt ecological balance in ways that demand urgent attention. Accurately forecasting the Air Quality Index (AQI) in a timely manner is no longer optional, it is a fundamental requirement for effective environmental monitoring, informed policy-making, and meaningful public awareness. Yet, conventional AQI prediction approaches often fall short, struggling to untangle the complex and nonlinear interactions between pollutants, cope with noisy real-world environmental data, and handle the persistent challenge of class imbalance across AQI categories, all of which can reduce 15 prediction accuracy and reliability. This reality underscores the urgent need for an intelligent, data-driven AQI prediction system capable of addressing these challenges with greater precision and consistency. Such a system must be grounded in robust data preprocessing, equipped to handle

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

class imbalance effectively, and powered by well-chosen machine learning models capable of capturing the complex relationships between pollutant parameters and environmental conditions. By combining advanced machine learning techniques with seamless real-time pollution data integration, the proposed system aims to enhance environmental monitoring, strengthen data-driven decision making, and empower communities with timely and accurate air quality insights.

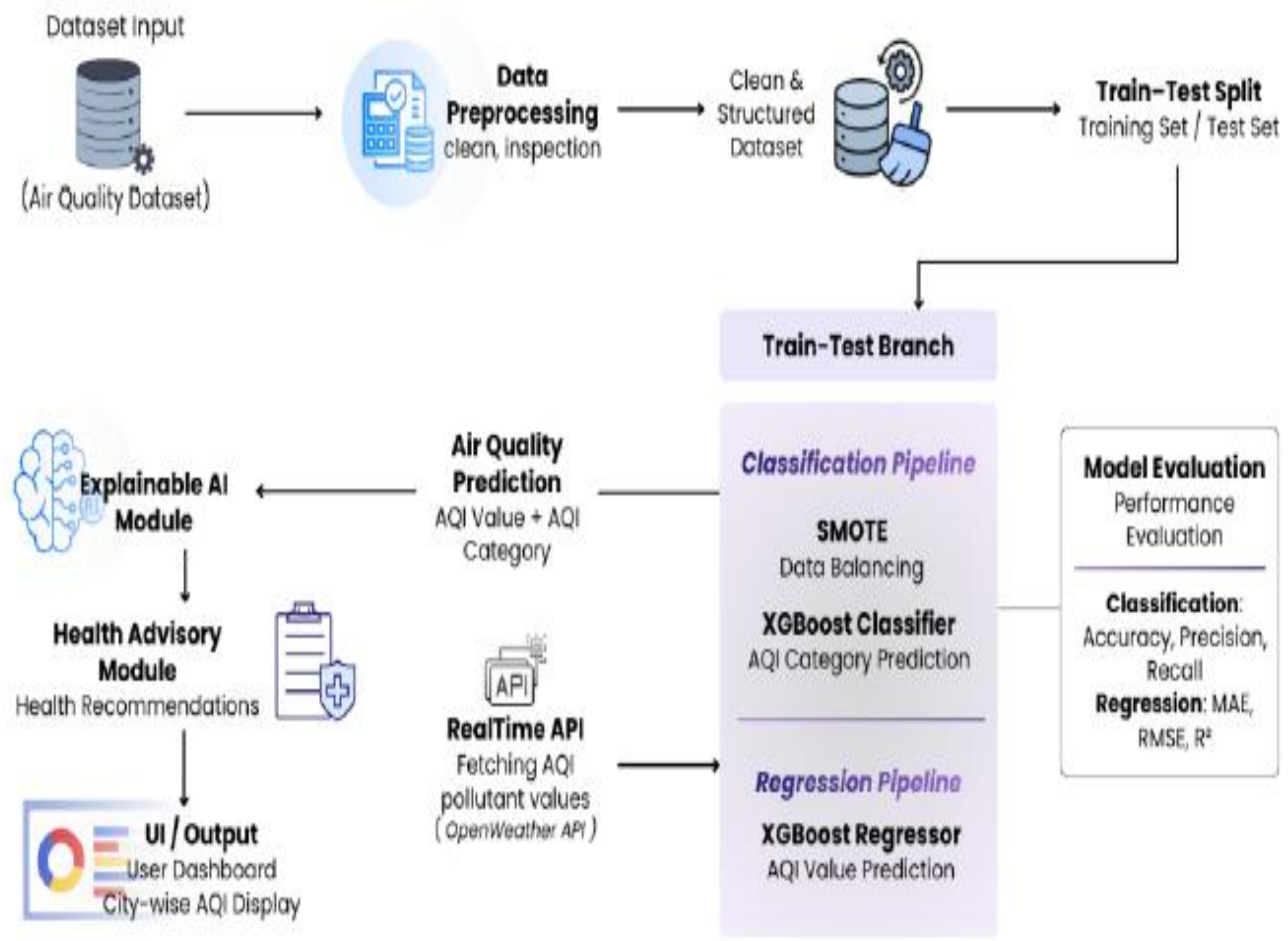
Software Description

The proposed AQI Prediction System is built on a carefully selected stack of programming languages, libraries, and development tools — each chosen to ensure seamless data processing, efficient model training, and meaningful result visualization. Python serves as the backbone of the entire system due to its extensive ecosystem for machine learning and data analysis. Raw pollution data is cleaned, transformed, and structured using NumPy and Pandas, while Matplotlib and Seaborn are used to visualize AQI trends, pollutant distributions, and model performance metrics in an intuitive and informative manner. For predictive modeling, Random Forest and XGBoost are implemented through Scikit-learn and the dedicated XGBoost library, enabling the system to capture complex nonlinear relationships between pollutants and AQI values with high accuracy. To address the persistent challenge of class imbalance across AQI categories, the SMOTE technique from the Imbalanced-learn library is applied during training, ensuring the models learn fairly and generalize effectively across all pollution levels. Model development and experimentation are carried out using Jupyter Notebook and Google Colab, which provide flexible and interactive environments for iterative training and performance evaluation. Once trained, the optimized model is deployed within a Flask-based web application that powers an interactive dashboard for real-time AQI visualization and prediction display. To further enhance the system's capabilities, live air pollution data is continuously fetched through an external API, enabling users to compare real-time environmental conditions with the model's forecasted AQI values. Together, these integrated components form a robust and scalable AQI prediction and monitoring platform designed to support environmental 16 awareness and data-driven decision-making.

Workflow of Model

The workflow of the proposed AQI Prediction and Monitoring System follows a well-structured, end-to-end pipeline designed to deliver accurate air quality forecasts and meaningful real-time insights. The process begins with collecting historical air pollution data from publicly available sources such as Kaggle, encompassing key pollutant parameters including PM2.5, PM10, NO_2, SO_2, CO, and O_3, which are commonly used indicators for calculating the Air Quality Index. This raw data is then passed through a comprehensive preprocessing stage where missing values are handled, irrelevant attributes are removed, and the most influential pollutant features are selected, transforming the dataset into a clean and model-ready format. With a refined dataset prepared, the class distribution across AQI categories is analyzed, and the SMOTE technique is applied to generate synthetic samples for underrepresented minority classes, ensuring that the models are trained on a balanced and representative dataset. The processed data is then divided into training and testing sets, upon which two machine learning algorithms — Random Forest and XGBoost — are trained to capture the complex nonlinear relationships between pollutant concentrations and AQI values. Model performance is evaluated using standard regression metrics including R^2 Score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), providing a comprehensive basis for assessing model accuracy and prediction reliability. Among the evaluated models, XGBoost demonstrates superior performance and is therefore selected as the final prediction model due to its high predictive accuracy and ability to handle complex environmental datasets efficiently.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026



Datasets

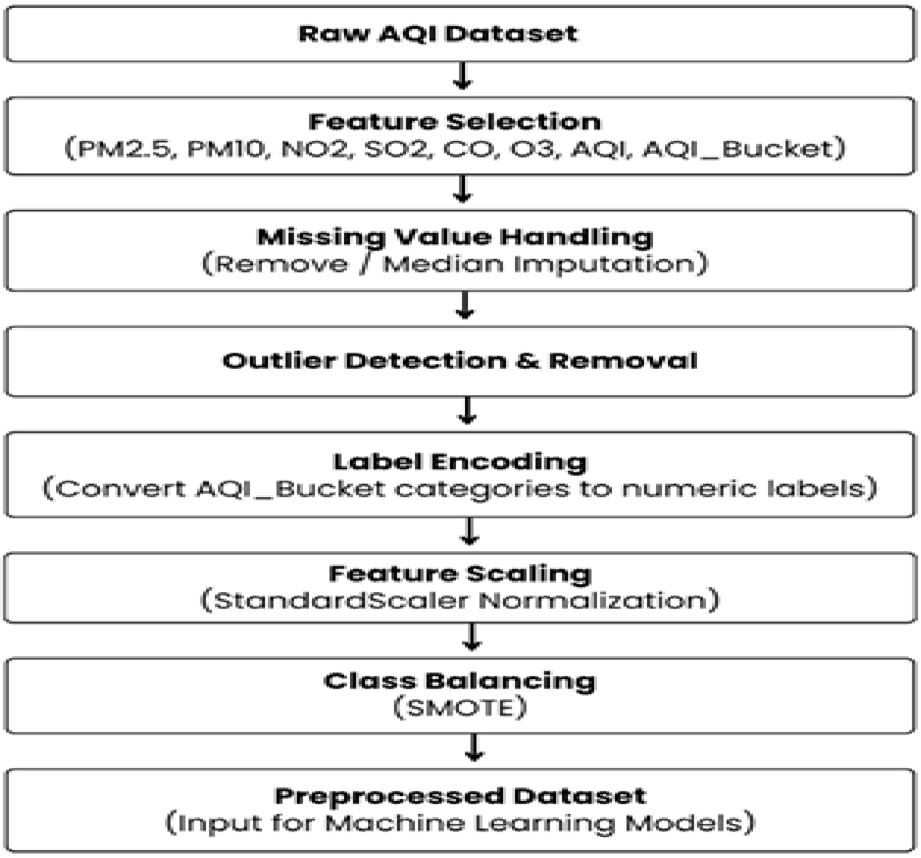
The dataset used in the proposed AQI Classification System comprises comprehensive historical air quality records collected from multiple Indian cities and obtained from the Air Quality Data in India dataset available on the Kaggle platform, containing daily measurements of major air pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with other relevant environmental factors where available. In addition to pollutant concentrations, the dataset includes calculated AQI values and their corresponding categorized AQI levels—Good, Satisfactory, Moderate, Poor, Very Poor, and Severe—based on standard air quality guidelines. Since real-world environmental data often contains missing entries, noise, and imbalanced class distribution across AQI categories, appropriate preprocessing techniques are applied to clean, normalize, and balance the data. This well-structured and processed dataset forms the core foundation for training and testing the machine learning model, enabling accurate, reliable, and city-specific AQI classification by capturing pollution patterns and temporal trends across different regions.

Data Preprocessing

Data preprocessing forms the backbone of the proposed AQI Prediction System, ensuring that the raw air quality dataset is thoroughly cleaned, structured, and prepared before machine learning models are applied [4]. The process begins by filtering the dataset to retain only the most relevant pollutant parameters — PM2.5, PM10, NO_2, SO_2, CO, and O_3 — which are widely used indicators for evaluating air quality and calculating AQI values. Missing values are addressed through median-based imputation, inconsistent records are removed, and IQR-based outlier detection is applied to eliminate abnormal pollutant readings that could distort model learning. Next, AQI category labels are converted into numerical form using Label Encoding, while StandardScaler is applied to normalize pollutant values so that each feature contributes equally during model training. The dataset

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

is then divided into training and testing sets to enable reliable and unbiased model evaluation. To address the issue of class imbalance across AQI categories, the SMOTE (Synthetic Minority Over-sampling Technique) is applied to generate synthetic samples for minority classes, ensuring the machine learning models are trained on a balanced dataset. These preprocessing steps significantly improve dataset quality and enable machine learning algorithms such as Random Forest and XGBoost to learn the complex relationships between pollutant concentrations and AQI levels more effectively, resulting in accurate and reliable AQI predictions.

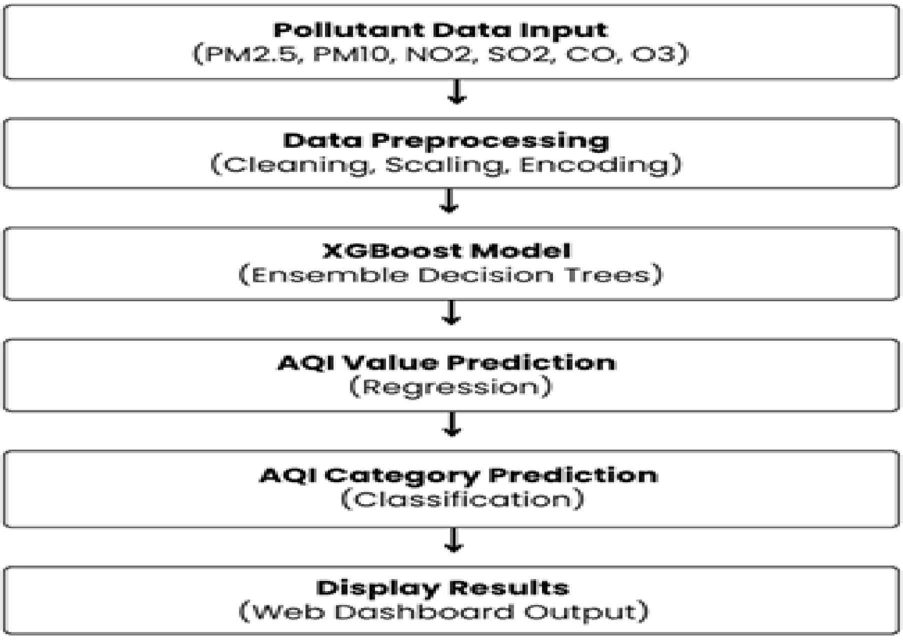


Model and Algorithm Used

The proposed AQI Prediction System employs two powerful machine learning algorithms — Random Forest and XGBoost — to analyze complex pollutant parameters including PM2.5, PM10, NO_2, SO_2, CO, and O_3, and accurately forecast both AQI values and their corresponding pollution categories. Random Forest builds multiple decision trees during training and combines their outputs to produce stable and reliable predictions while reducing the risk of overfitting in high-dimensional environmental dataset. XGBoost further enhances prediction performance by sequentially constructing decision trees where each iteration corrects the errors of the previous model, enabling it to effectively capture the nonlinear relationships between pollutant parameters and AQI levels. In the proposed system, XGBoost performs both AQI value regression and AQI category classification, consistently outperforming Random Forest during model evaluation.

Both models are evaluated using standard performance metrics including Accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score, which provide a comprehensive measure of prediction accuracy and model reliability. Based on the evaluation results, XGBoost demonstrates superior performance and is selected as the final prediction model. The optimized model is then integrated into the Flask-based web application, enabling real-time AQI prediction and visualization through an interactive dashboard for users and environmental monitoring stakeholders.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026



Model Evaluation

Model evaluation is conducted to objectively measure the predictive performance of the two machine learning models — Random Forest and XGBoost — both trained on historical air pollution data containing key pollutant parameters including PM2.5, PM10, NO_2, SO_2, CO, and O_3. Model performance is assessed using a set of standard evaluation metrics, including Accuracy for classification tasks and Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score for regression analysis, providing a comprehensive basis for evaluating prediction accuracy and model reliability.

Experimental results show that XGBoost consistently outperforms Random Forest, achieving higher classification accuracy and lower prediction error. The classification performance is further validated using a confusion matrix and classification report, which analyze the prediction results across AQI categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. For regression evaluation, scatter plots comparing actual versus predicted AQI values demonstrate a strong correlation, confirming the model's capability to accurately capture pollutant–AQI relationships. The optimized XGBoost model is then integrated into a Flask-based web application, where live pollutant data obtained through an external air quality API is used as real-time input to generate AQI predictions. The system successfully predicts AQI values and their corresponding pollution categories for selected cities, demonstrating the effectiveness and practical applicability of machine learning techniques for real-time air quality monitoring and environmental decision-making.

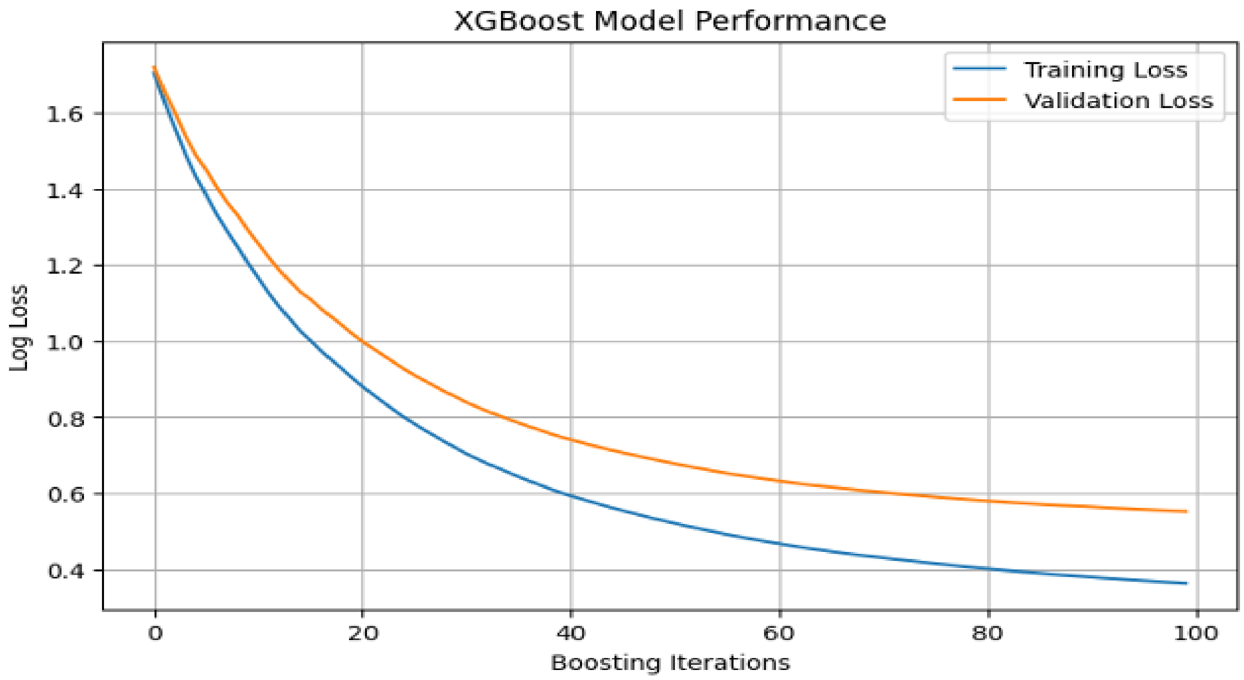
Analysis and interpretation

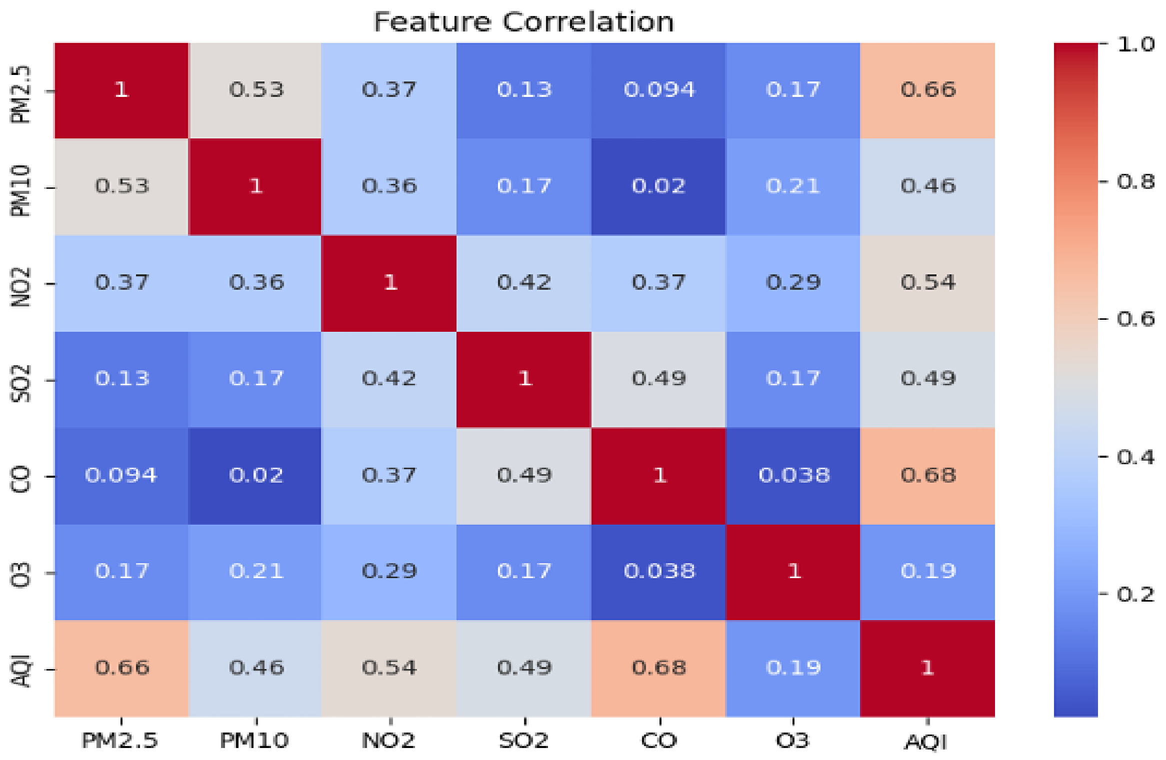
The performance of the proposed AQI prediction system is evaluated using several standard machine learning evaluation metrics to measure the effectiveness and reliability of the trained models. The XGBoost classifier is assessed using metrics such as Accuracy, Precision, Recall, and F1-score, which provide a comprehensive understanding of the model's classification performance. Accuracy measures the overall correctness of the predictions, while Precision and Recall evaluate how effectively the model identifies AQI categories without producing excessive false positives or false negatives. The F1-score provides a balanced evaluation by combining both precision and recall. Additionally, a confusion matrix is used to analyze the distribution of correct and incorrect predictions across different AQI categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe, providing deeper insight into the classification performance of the model. For AQI value prediction, the XGBoost regression model is evaluated using graphical analysis such as the Actual vs Predicted AQI scatter plot, which demonstrates a strong correlation between predicted and actual AQI values. The training

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

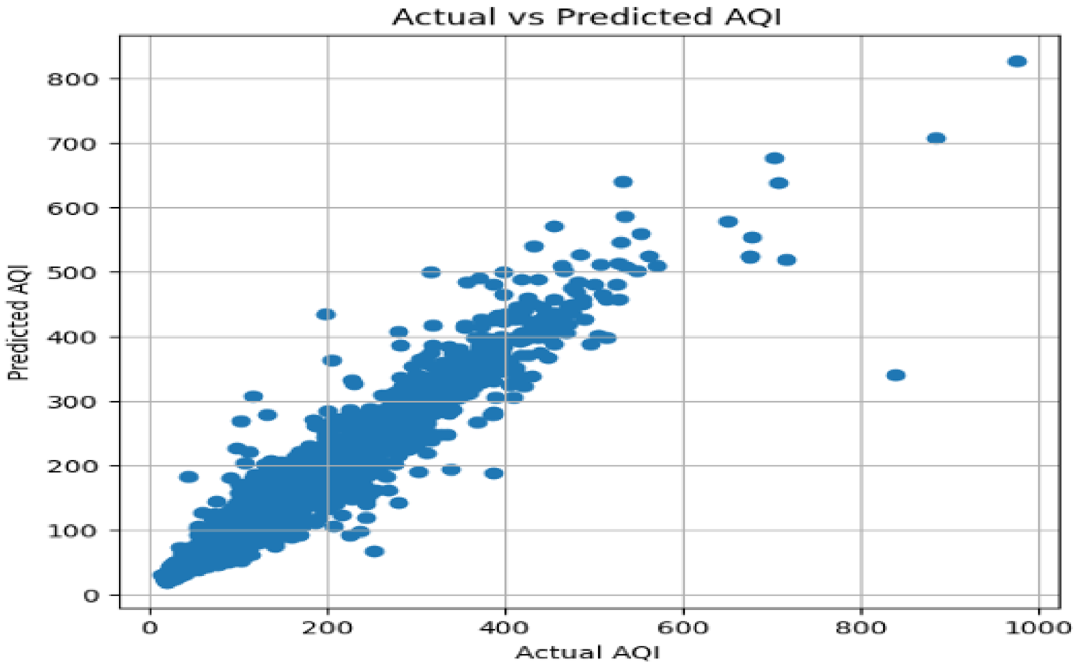
and validation loss curves further illustrate the model's learning behavior during boosting iterations, indicating stable training and good generalization without significant overfitting.

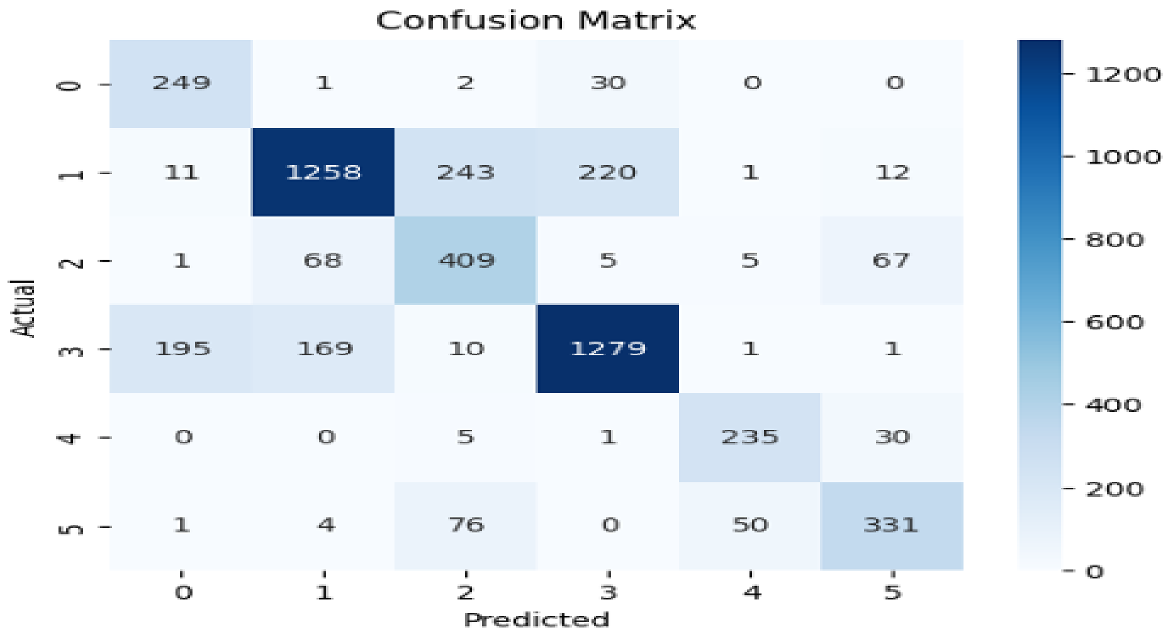
Overall, the evaluation results confirm that the proposed machine learning-based AQI prediction system provides reliable and accurate results for air quality monitoring and environmental analysis.

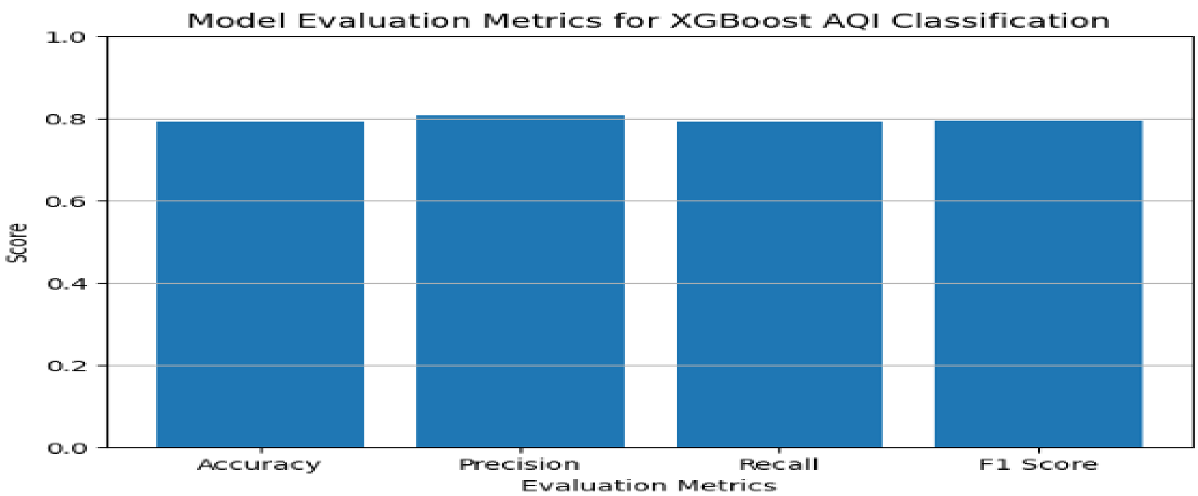




ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026







ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

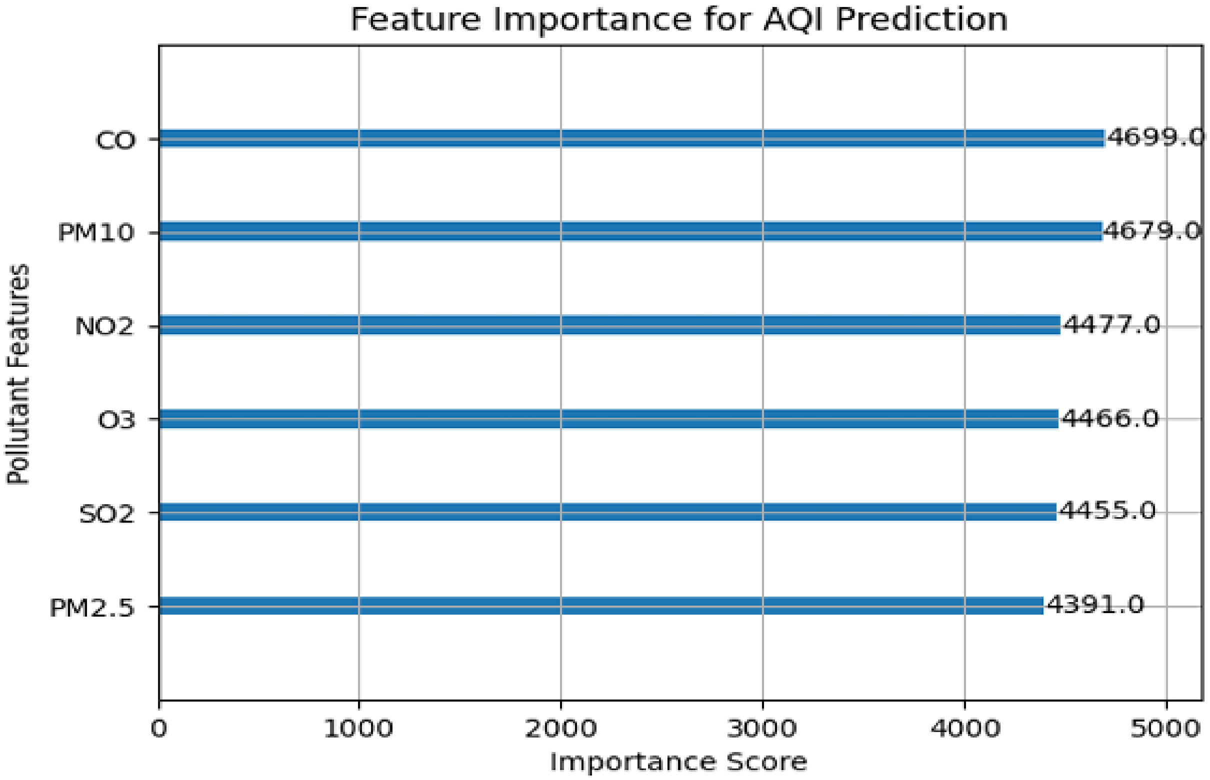


Fig. 1. Title of the figure with 8 pt. size

CONCLUSION

The proposed AQI Prediction System effectively demonstrates the power of machine learning in analyzing complex air pollution data, leveraging key pollutant parameters — PM2.5, PM10, NO_2, SO_2, CO, and O_3 — alongside robust preprocessing techniques including missing value handling, outlier removal, feature scaling, and SMOTE-based class balancing to build a reliable and high-performing prediction pipeline. Following a rigorous comparison between Random Forest and XGBoost, the latter emerged as the superior model and was seamlessly deployed within a Flask-based web application that retrieves live pollutant data via an external API, delivering accurate, real-time AQI predictions and category classifications through an intuitive interactive dashboard that promotes meaningful environmental awareness.

Future Scope

Although the proposed AirAware system achieved promising AQI prediction performance using the XGBoost algorithm, several opportunities remain for further improvement. Future work will focus on incorporating larger and more geographically diverse datasets collected from multiple regions and climatic conditions to enhance the model's robustness and generalizability. A comprehensive comparative analysis involving additional machine learning and deep learning algorithms, including LightGBM, CatBoost, Long Short-Term Memory (LSTM), and Transformer-based architectures, will be conducted to establish the most suitable model for both short-term and long-term AQI forecasting. Model evaluation will be strengthened through k-fold cross-validation and a wider range of performance metrics, including accuracy, precision, recall, F1-score, RMSE, MAE, and R^2 score, to provide a more rigorous assessment of predictive performance. Furthermore, advanced Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME will be integrated to improve model transparency by identifying the contribution of individual pollutants to AQI predictions, thereby increasing user trust and supporting informed environmental decision-making. Future enhancements may also include incorporating

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

meteorological variables, IoT sensor networks, and real-time edge-based deployment to enable scalable, intelligent, and explainable air quality monitoring systems.

REFERENCES

1. S. Tırınk, "Machine Learning-Based Forecasting of Air Quality Index under Long-Term Environmental Patterns: A Comparative Approach with XGBoost, LightGBM, and SVM," PLOS One, vol. 20, no. 10, 2025.
2. S. Singh, M. Kumar, V. Sengar, A. Kumar, K. Abhishek, and B. M. A. Shafeeq, "Ensemble Learning for Air Quality Index Prediction: Integrating Gradient Boosting, XGBoost, and Stacking with SHAP-Based Interpretability," Scientific Reports, 2026.
3. G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air Quality Prediction by Machine Learning Models: A Predictive Study on the Indian Coastal City of Visakhapatnam," Chemosphere, vol. 338, 2023.
4. S. Kasim, S. Malek, S. Cheen, P. N. Fatin, K. X. Ning, H. Hamidi, W. A. Wan Ahmad, K. S. Ibrahim, K. Negishi, M. N. Sulaiman, and A. Fong, "Machine Learning-Based Prediction of Mortality Risk from Air Pollution-Induced Acute Coronary Syndrome in the Western Pacific Region," Scientific Reports, 2026.
5. P. Patel, S. Patel, K. Shah, K. Desai, S. Patel, and M. Shah, "A Systematic Study on PM2.5 and PM10 Concentration Prediction in Air Pollution Using Machine Learning and Deep Learning Model," Environmental Chemistry and Ecotoxicology, vol. 7, 2025.
6. T. M. T. Lei, S. C. W. Ng, and S. W. I. Siu, "Application of ANN, XGBoost, and Other Machine Learning Methods to Forecast Air Quality in Macau," Sustainability, vol. 15, 2023.
7. F. P. Arifianti and A. Salam, "XGBoost and Random Forest Optimization Using SMOTE to Classify Air Quality," Advance Sustainable Science, Engineering and Technology (ASSET), vol. 6, no. 1, 2024.
8. S. Chakraborty, B. Misra, and N. Dey, "Explainable Artificial Intelligence (XAI) for Air Quality Assessment," in Design Studies and Intelligence Engineering, IOS Press, 2024.
9. M. W. Jalali, B. Saidi, H. Farahmand, M. A. R. Panah, and E. N. Saruhan, "Scalable AI Driven Air Quality Forecasting and Classification for Public Health Applications," Discover Atmosphere, vol. 3, 2025.
10. Q. Liu, B. Cui, and Z. Liu, "Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling," Atmosphere, vol. 15, 2024.
11. J. David Sukeerthi Kumar, P. Rizwana, M. Pushpanjali, Y. Suma Mallika, P. Akshaya Sree, and T. Meghana, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," International Journal of Progressive Research in Engineering Management and Science (IJPREMS), vol. 4, no. 4, pp. 1784–1790, 2024.
12. S. Pavani Muthyala, A. Anju Shaik, V. Reddy Peram, and N. Sai Priyanka, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," Journal of Nonlinear Analysis and Optimization, vol. 15, no. 1, 2024.
13. S. Kumar Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized Machine Learning Model for Air Quality Index Prediction in Major Cities in India," Scientific Reports, vol. 14, 2024.
14. M. Karmoude, B. Munhungewarwa, I. Chiraira, R. McKenzie, J. Kong, B. Smith, G. Ayana, N. Njara, T. Mathaha, M. Kumar, and B. Mellado, "Machine Learning for Air Quality Prediction and Data Analysis: Review on Recent Advancements, Challenges, and Outlooks," Science of the Total Environment, vol. 1002, 2025.
15. C. Liu, G. Pan, D. Song, and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," IEEE Access, vol. 11, 2023.
16. C. Liu, G. Pan, D. Song, and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," IEEE Access, 2023.
17. S. Kumar Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized Machine Learning Model for Air Quality Index Prediction in Major Cities in India," Scientific Reports, vol. 14, 2024.