IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13328

AIR POLLUTION PREDICTIONS USING MACHINE LEARNING

Sampoorna S[1], Dr. J. Savitha[2]

Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[1]

Professor, Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[2]

Abstract: Air pollution has become one of the most critical environmental challenges affecting human health, climate stability, and sustainable development worldwide. Accurate prediction of air pollutant levels is essential for early warning systems, policy planning, and effective environmental management. This study proposes a machine learning-based framework for air pollution prediction using historical air quality and meteorological data. The system integrates data collection from air quality monitoring stations, weather parameters such as temperature, humidity, wind speed, and rainfall, and location-specific information. Data preprocessing techniques including data cleaning, handling missing values, feature selection, and normalization are applied to enhance model performance. The proposed model utilizes advanced machine learning algorithms, particularly ensemble learning techniques such as XGBoost, to predict major pollutant concentrations including PM2.5, PM10, NO_2, SO_2, and CO. The predicted pollutant values are further used to compute the Air Quality Index (AQI) and classify air quality into categories such as Good, Moderate, Poor, Very Poor, and Severe. Experimental evaluation demonstrates that the proposed approach improves prediction accuracy and reduces error compared to traditional statistical models. The developed system also supports real-time alerts, hotspot identification, and decision-making support for environmental authorities. This research contributes to sustainable urban planning and public health protection through intelligent air quality forecasting.

Keywords: Air Pollution Prediction, Machine Learning, XG Boost, Air Quality Index (AQI), PM2.5, PM10, Environmental Monitoring, Data Preprocessing, Ensemble Learning, Real-Time Air Quality Forecasting, Smart Cities, Environmental Sustainability.

I. INTRODUCTION

Air pollution is one of the most significant environmental and public health challenges faced by modern society. Rapid industrialization, urbanization, vehicular emissions, and increased energy consumption have contributed to the continuous deterioration of air quality across major cities worldwide. Exposure to harmful pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), and carbon monoxide (CO) can lead to severe respiratory and cardiovascular diseases, reduced life expectancy, and environmental degradation. Therefore, continuous monitoring and accurate prediction of air pollution levels have become essential for sustainable development and public health protection. Traditional air quality prediction methods rely on statistical and deterministic models, which often struggle to capture the complex and nonlinear relationships between meteorological parameters and pollutant concentrations. With the advancement of computational intelligence, machine learning (ML) techniques have emerged as powerful tools for analysing large volumes of environmental data and generating accurate predictions. Machine learning models can learn hidden patterns from historical air quality data and meteorological variables such as temperature, humidity, wind speed, and rainfall, enabling more reliable forecasting of pollutant levels. In recent years, ensemble learning algorithms such as Extreme Gradient Boosting (XGBoost) have gained popularity due to their high predictive accuracy, efficiency, and ability to handle missing or noisy data. These models are particularly effective in modeling nonlinear dependencies and interactions among environmental variables. By leveraging machine learning techniques, air pollution prediction systems can provide early warnings, identify pollution hotspots, and support data-driven decision-making for government authorities and environmental agencies. This study proposes a machine learning-based air pollution prediction framework that integrates data preprocessing, feature selection, model training, and Air Quality Index (AQI) computation. The system aims to predict major pollutant concentrations and classify air quality levels to assist in real-time monitoring and preventive measures. The proposed approach contributes to smart city initiatives, environmental sustainability, and improved public awareness through accurate and timely air quality forecasting.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13328

II. LITERATURE REVIEW

Air pollution prediction has attracted significant research attention due to its impact on public health and environmental sustainability. Over the past decade, various statistical and machine learning approaches have been proposed to forecast air pollutant concentrations and Air Quality Index (AQI) levels. Early studies primarily relied on traditional statistical models such as Linear Regression, Autoregressive Integrated Moving Average (ARIMA), and Multiple Linear Regression (MLR) to predict pollutant concentrations. Although these models provided baseline forecasting performance, they were limited in capturing nonlinear relationships between meteorological parameters and air pollutants. As air pollution is influenced by complex interactions among multiple environmental factors, purely statistical methods often resulted in reduced accuracy. With advancements in artificial intelligence, machine learning techniques have increasingly been adopted for air quality forecasting. Algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forest, and Artificial Neural Networks (ANN) have demonstrated improved predictive performance compared to traditional models. Random Forest, in particular, has been widely used due to its ability to handle high-dimensional datasets and reduce overfitting. Similarly, Artificial Neural Networks have shown strong capability in modeling nonlinear patterns within environmental datasets. Recent studies have explored deep learning approaches such as Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNN) for time-series air quality prediction. These models are effective in capturing temporal dependencies in sequential air pollution data. However, deep learning models often require large datasets, high computational resources, and extensive training time. Among ensemble learning techniques, Extreme Gradient Boosting (XGBoost) has gained considerable attention due to its high efficiency, scalability, and superior performance in regression tasks. XGBoost improves prediction accuracy through gradient boosting optimization and regularization techniques, making it suitable for complex environmental datasets. Several comparative studies have reported that XGBoost outperforms traditional machine learning models in terms of accuracy, error reduction, and computational efficiency for AQI prediction. Despite significant progress, challenges remain in handling missing data, real-time forecasting, spatial variability, and interpretability of predictions. Therefore, integrating efficient preprocessing techniques, robust feature selection methods, and optimized ensemble learning models is essential to enhance prediction reliability. Based on the reviewed literature, machine learning—particularly ensemble-based approaches such as XGBoost-provides a promising solution for accurate air pollution forecasting. This study builds upon existing research by implementing an optimized machine learning framework that integrates data preprocessing, feature engineering, and AQI classification for improved air quality prediction.

III. PROBLEM STATEMENT

Air pollution has become a critical environmental and public health issue worldwide due to rapid industrialization, urban growth, increased vehicular emissions, and energy consumption. Elevated levels of pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) pose severe health risks, including respiratory disorders, cardiovascular diseases, and premature mortality. Effective monitoring and early prediction of air pollution levels are therefore essential to reduce health hazards and support environmental policy decisions. Conventional air quality forecasting methods rely on statistical models and rule-based approaches, which often fail to capture the complex, nonlinear, and dynamic relationships between meteorological variables and pollutant concentrations. These traditional models typically exhibit limited prediction accuracy, especially when dealing with large-scale, high-dimensional, and time-series environmental datasets. Furthermore, air pollution data often contain missing values, noise, and inconsistencies, which further reduce forecasting reliability. Real-time prediction and accurate Air Quality Index (AQI) classification remain challenging due to the variability of climatic conditions and pollutant interactions. Existing systems may also struggle with scalability, computational efficiency, and adaptability to different geographical regions.

Therefore, there is a need for a robust and efficient machine learning-based framework that can:

- Accurately predict air pollutant concentrations using historical and meteorological data.
- Handle nonlinear relationships and high-dimensional datasets effectively.
- Reduce prediction errors and improve forecasting reliability.
- Provide early warnings through accurate AQI classification.

This study addresses these challenges by proposing a machine learning-driven air pollution prediction system that integrates data preprocessing, feature selection, and ensemble learning techniques to enhance prediction performance and support informed decision-making for environmental management.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13328

IV. METHODOLOGY

The proposed air pollution prediction system follows a structured machine learning framework consisting of data collection, preprocessing, model development, prediction, and evaluation. The overall methodology is designed to improve prediction accuracy and ensure reliable Air Quality Index (AQI) forecasting.

1. Data Collection

Air quality and meteorological datasets are collected from monitoring stations and publicly available environmental databases. The dataset includes major pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, and CO. Meteorological parameters including temperature, humidity, wind speed, and rainfall are also incorporated, as these variables significantly influence pollutant dispersion and accumulation. Historical AQI records are used to train and validate the prediction model.

2. Data Preprocessing

Raw environmental data often contain missing values, noise, and inconsistencies. Therefore, preprocessing is performed to enhance data quality. The steps include:

- **Data Cleaning:** Removal of duplicate records and correction of erroneous values.
- **Handling Missing Values:** Application of imputation techniques such as mean, median, or interpolation methods.
- **Feature Selection:** Identification of significant pollutant and meteorological features using correlation analysis.
- **Data Normalization:** Scaling features to ensure uniformity and improve model convergence.
- **Data Splitting:** Division of the dataset into training and testing sets to evaluate model performance.

3. Model Development

An ensemble machine learning algorithm, Extreme Gradient Boosting (XGBoost), is employed for pollutant concentration prediction. XGBoost is chosen due to its high efficiency, scalability, and ability to handle nonlinear relationships. The model is trained using historical data to learn complex interactions between environmental variables. Hyperparameter tuning techniques are applied to optimize model performance and reduce overfitting.

4. Prediction and AQI Calculation

The trained model predicts future pollutant concentrations. Based on the predicted pollutant values, the Air Quality Index (AQI) is computed using standard AQI calculation formulas. The AQI is then categorized into predefined classes such as Good, Moderate, Poor, Very Poor, and Severe for easier interpretation.

5. Model Evaluation

- The performance of the proposed model is evaluated using statistical metrics such as:
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- R-squared (R^2) score

These evaluation metrics help measure prediction accuracy and compare the proposed model with baseline approaches.

6. Deployment and Real-Time Monitoring

The final model can be integrated into a web or mobile-based dashboard to provide real-time air quality forecasts, pollution hotspot identification, and early warning alerts. This enables environmental authorities and the public to take preventive actions.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13328

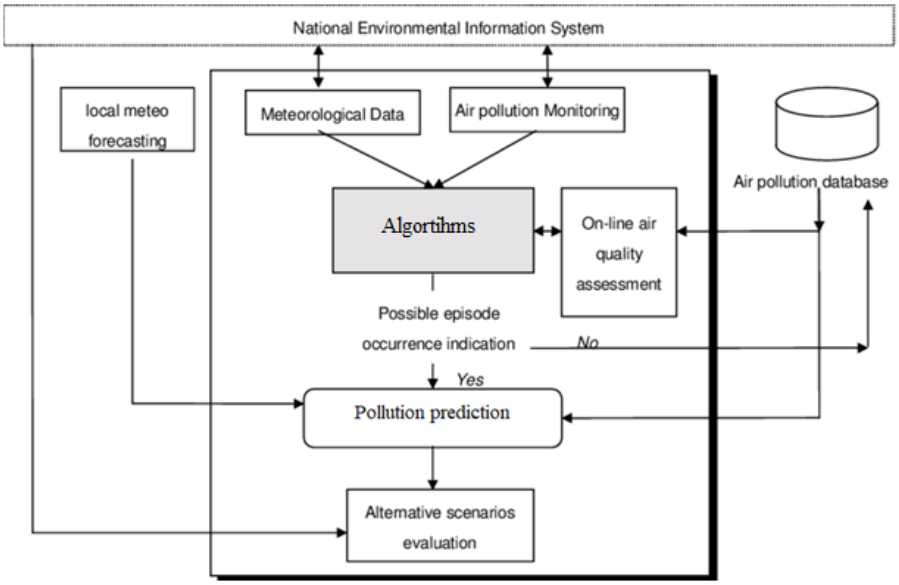

Figure1:System Architecture

V. RESULTS

The proposed machine learning-based air pollution prediction model was evaluated using historical air quality and meteorological datasets. The dataset was divided into training and testing sets to assess model generalization performance. The XGBoost model demonstrated strong predictive capability in estimating pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, and CO. The performance of the model was measured using standard evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) score.

The experimental results indicate:

- Low MAE and RMSE values, showing minimal prediction error.
- High R^2 score, indicating strong correlation between predicted and actual pollutant values.
- Improved accuracy compared to traditional regression models.

The model effectively captured nonlinear relationships between meteorological parameters and pollutant concentrations. Additionally, AQI classification based on predicted pollutant levels showed reliable categorization into Good, Moderate, Poor, Very Poor, and Severe levels.The system also successfully identified pollution hotspots and supported real-time forecasting. These findings confirm that ensemble learning techniques such as XGBoost provide higher efficiency and accuracy for air pollution prediction compared to conventional statistical approaches.

VI. CONCLUSION

Air pollution prediction plays a crucial role in protecting public health and supporting environmental sustainability. This study proposed a machine learning-based framework using the XGBoost algorithm to predict air pollutant concentrations and compute Air Quality Index (AQI) levels.The experimental results confirm that ensemble learning techniques significantly improve prediction accuracy compared to conventional statistical methods. The integration of data preprocessing, feature selection, and optimized model training enhanced the reliability of air quality forecasting.The proposed system can assist environmental authorities in early warning generation, pollution control planning, and policy implementation. It also supports smart city initiatives by enabling real-time monitoring and informed decision-making.Overall, the study demonstrates that machine learning provides an efficient, scalable, and accurate solution for air pollution prediction and management.The offered solution provides an alternative to the risks and expenses of the physical RC vehicles and provides a convenient platform to learn about vehicle control and the integration of the embedded systems. It can also be expanded with the use of AI-based automation and autonomous navigation, which allows its usage in future research and education.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13328

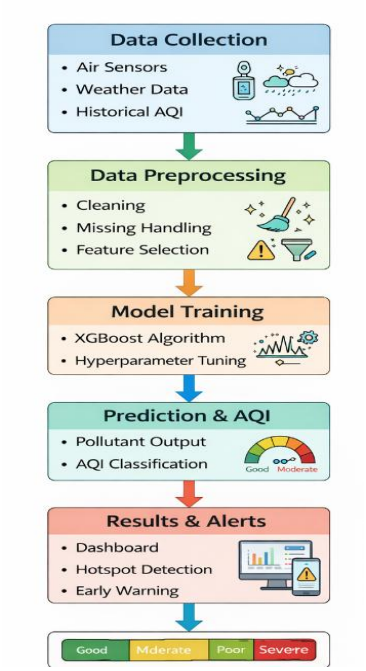

Figure2: Final process of the project

REFERENCES

[1]. Poonam Paul, Ritik Gupta, Sanjana Tiwari, Ashutosh Sharma, "IoT based Air Pollution Monitoring System with Arduino", IJART, May 2005.

[2]. Zishan Khan, Abbas Ali, Moin Moghal, "IoT based Air Pollution using NodeMCU and Thingspeak", IRANS, pp. 11-16, March 2014.

[3]. SaiKumar, M. Reji, P.C. KishoreRaja "AirQuality Index in India", IEEE conference Chennai, August 2014.

[4]. Mohan Joshi,"Research Paper on IoT based Air and Sound Pollution monitoring system", IETS Journal, pp. 11-17, September 2015.

[5]. "Malaya Ranjan, Rai kumar,"Understanding Parts per million in real time air quality index", Journal of Mathematics and advanced sciences, pp. 23-29, September 2009

[6]. D. Bandyopadhyay and J. Sen, "Internet of Things: Applications and Challenges in Technology and Standardization," Wirel. Pers. Commun., vol.58, no. 1, pp. 49–69, May 2011.

[7]. L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey,"Comput. Netw., vol. 54, no. 15, pp. 2787–2805, October 2010.

[8]. H. Kopetz, Real-Time Systems: Design Principles for Distributed Embedded Applications. Boston, MA: Springer US, 2011, ch. Internet of Things, pp. 307– 323.

[9]. A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo,"A Survey on Facilities for Experimental Internet of Things Research," IEEE Communications Magazine, vol. 49, no. 11, pp. 58–67, November 2011.

[10]. J. Kim, J. Lee, J. Kim, and J. Yun, "M2M Service Platforms: Survey, Issues, and Enabling Technologies," IEEE Communications Surveys Tutorials, vol. 16, no. 1, pp. 61–76, January 2014.

[11]. Jen-Hao Liu, Yu-Fan Chen, Tzu-Shiang Lin, And Da-Wei Lai ,Tzai-Hung Wen, Chih-Hong Sun, And Jehn-Yih Juang,Joe-Air Jiangdeveloped Urban Air Quality Monitoring System Based On Wireless Sensor Networks 2011 IEEE.

[12]. Srinivas Devarakonda, Parveen Sevusu, Hongz Hang Liu, Ruilin Liu, Liviu Iftode, Badri Nath Urbcomp" Real-Time Air Quality Monitoring Through Mobile Sensing In Metropolitan Areas"13, August 2013 Acm.