ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 127–137, 2026

IMPROVING AIR QUALITY PREDICTION USING HYBRID ENSEMBLE LEARNING WITH SHAP-BASED EXPLAINABILITY

Aditya Sharma[1]*, Sanju Chandrakar[2], Ismail Nizami[3], Satyam Sahu[4], Vijaya Tripathi[5], Shankar Sharan Tripathi[6]

[1234]*Student, Department of Computer Science and Technology, Shri Shankaracharya Technical Campus.*

[5,6]*Assisant Professor, Computer Science and Engineering, Shri Shankaracharya Technical Campus.*

Abstract

Air pollution remains a critical challenge in urban environments, threatening both ecological balance and public health. Accurate air quality prediction is essential for timely preventive measures. This study proposes a **hybrid ensemble learning** framework that integrates Random Forest, Gradient Boosting, and XGBoost to improve predictive performance. To ensure transparency, SHAP (SHapley Additive exPlanations) is employed to interpret model outputs and identify the most influential environmental factors. The model is trained on publicly available datasets containing pollutants (PM2.5, PM10, NO_2, CO, SO_2) and meteorological variables (temperature, humidity). Experimental results demonstrate that the hybrid ensemble approach achieves higher accuracy than individual models while providing explainable insights into pollution dynamics. These findings highlight the potential of explainable AI to enhance trust, usability, and decision-making in environmental monitoring systems.

Keywords: *Hybrid Ensemble Learning, Air Quality Prediction, Random Forest, Gradient Boosting, XGBoost, SHAP, Explainable AI, Environmental Monitoring, Urban Pollution, Machine Learning.*

** Corresponding author*

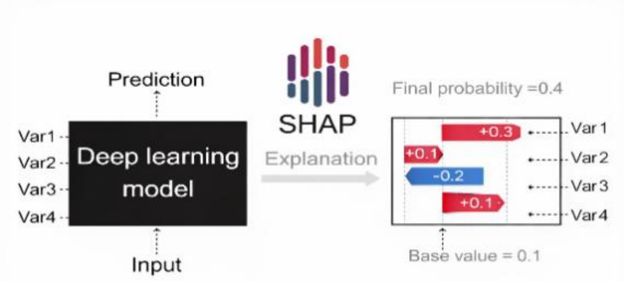
1. Introduction

worldwide. Rapid urbanization, industrial emissions, and increasing vehicle usage have significantly contributed to deteriorating air quality. Pollutants such as PM2.5, PM10, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO) can cause serious respiratory and cardiovascular diseases. Accurate prediction of air quality levels is important so that governments and citizens can take preventive actions.

1.2 Overview of Air Quality Prediction

Air pollution has become a major environmental and public health concern worldwide. Rapid urbanization, industrial activities, and increasing vehicular emissions have significantly degraded air quality in many regions. Pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) are commonly monitored to assess air quality levels. Therefore, accurate prediction of air quality is essential for early warning systems and effective environmental management.

Traditional air quality prediction methods rely on statistical models and physical atmospheric simulations. While these approaches provide useful insights, they often struggle to capture complex nonlinear relationships between environmental factors such as temperature, humidity, wind speed, and pollutant concentrations. In recent years, machine learning and deep learning techniques have been increasingly used to improve prediction accuracy. Algorithms such as Random Forest, Support Vector Machines, Gradient Boosting, and Neural Networks can analyze large environmental datasets and identify hidden patterns that influence air pollution levels.



Recent research focuses on hybrid and ensemble learning approaches that combine multiple machine learning models to achieve better performance. These models often outperform individual algorithms by reducing prediction errors and improving generalization. Additionally, explainable artificial intelligence techniques such as SHAP (SHapley Additive exPlanations) are used to interpret model predictions and identify the most influential factors affecting air quality [4].

Overall, air quality prediction plays a crucial role in supporting smart city initiatives, environmental policy planning, and public health protection. With the advancement of machine learning and data analytics, prediction systems are being developed to help authorities and communities take timely preventive measures against air pollution.

1.3 Role of Machine Learning in Air Quality Prediction

Machine learning plays an important role in analyzing large environmental datasets and identifying hidden patterns. Algorithms such as Random Forest, Gradient Boosting, and XGBoost can learn from historical air pollution data and improve prediction performance [2][3]. Ensemble learning techniques combine multiple models to enhance accuracy and stability. Additionally, explainable AI methods such as SHAP help understand how different factors influence predictions.

1.4 Problem Statement and Objectives

Most existing air quality prediction models either focus only on prediction accuracy or lack interpretability. Many models operate as black boxes, making it difficult to understand which environmental factors contribute most to pollution levels. This creates challenges in decision-making for environmental monitoring systems and policymakers.

1.5 Conceptual Framework

The conceptual framework of the air quality prediction system describes the overall process used to forecast pollutant levels using machine learning techniques [2]. The framework begins with data collection from air quality monitoring sources such as the Central Pollution Control Board and global environmental datasets provided by the World Health Organization. These datasets include pollutant concentrations (PM2.5, PM10, NO_2, CO, SO_2, and O_3) along with meteorological parameters like temperature, humidity, and wind speed.

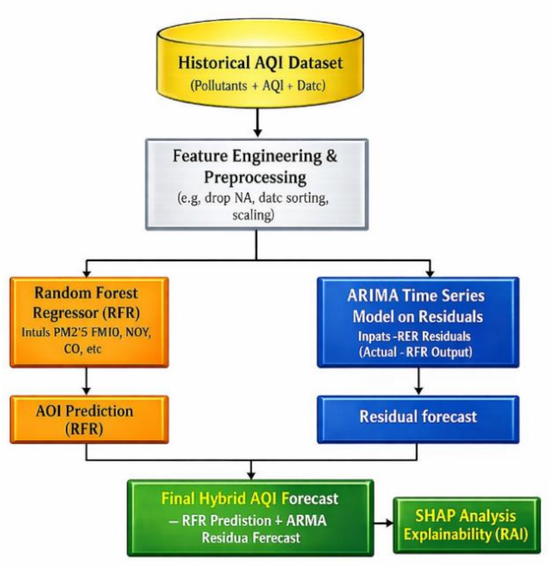


Figure 2. Conceptual Framework

2. Literature *Review*

Several studies have explored the use of machine learning techniques for air quality prediction. Earlier research mainly relied on statistical and regression-based models to estimate pollutant concentrations using historical environmental data. However, these methods often showed limited performance when dealing with complex and nonlinear relationships among variables. More recent work focuses on hybrid and ensemble learning models combined with explainable AI techniques such as SHAP to interpret predictions and identify key factors influencing air pollution. These methods help improve both prediction accuracy and model transparency, making them useful for environmental monitoring and decision-making systems.

2.1 Early Approaches to Air Quality Prediction

Early research on air quality prediction mainly focused on statistical and mathematical models. Methods such as linear regression and time series models like ARIMA were widely used to forecast pollution levels based on historical data [7]. These models were effective in capturing trends over time but had limitations in handling complex and nonlinear relationships among environmental factors. Studies showed that air pollution is influenced by multiple variables including pollutant concentration, weather conditions, and human activities, which made traditional models less efficient.

2.2 Machine Learning-Based Prediction Models

With the development of machine learning, researchers started applying algorithms such as Decision Trees, Support Vector Machines (SVM), and Random Forest to predict air quality. These models can analyze large datasets and identify hidden patterns in pollution data. Random Forest models, in particular, have been widely used because they improve prediction accuracy and reduce overfitting problems [2][6]. Research studies indicate that machine learning models can outperform traditional statistical methods when predicting AQI using pollutants like PM2.5, PM10, NO_2, and meteorological parameters.

2.3 Ensemble Learning in Environmental Prediction

Recent studies have explored ensemble learning techniques, which combine multiple machine learning models to improve prediction performance. Ensemble approaches such as Gradient Boosting, XGBoost, and hybrid models have demonstrated higher accuracy in air quality forecasting [1][3]. The theoretical concept behind ensemble learning is that combining multiple weak learners can create a stronger predictive model. Several researchers have reported that hybrid models significantly reduce prediction error compared to individual algorithms.

2.4 Explainable Artificial Intelligence in Air Quality Models

Although machine learning models provide accurate predictions, many of them act as black-box systems. This has led to the development of Explainable Artificial Intelligence (XAI) techniques. SHAP (SHapley Additive exPlanations) is one of the most widely used methods for explaining machine learning predictions. SHAP helps determine the contribution of each feature [4] to the final prediction. Studies applying SHAP in environmental data analysis have shown that pollutants such as PM2.5 and PM10 usually have the highest impact on AQI prediction.

2.5 Research Gap and Motivation

Despite significant progress in air quality prediction, many existing studies rely on single machine learning models and often lack interpretability. This makes it difficult to understand the influence of different environmental factors on pollution levels. Additionally, limited research focuses on combining ensemble models with explainable methods using real monitoring datasets from agencies such as the Central Pollution Control Board. Motivated by the growing impact of air pollution on public health, this study aims to develop a more accurate and interpretable prediction approach that supports better environmental monitoring and early warning systems.

2.7 Literature Matrix Summary

Table 1 Comparative Analysis of Previous Studies

Author & Year	Data Source	Method	Key Result	Limitation
Kumar et al., 2020 [6]	CPCB AQI data	Random Forest	Good accuracy	Less Explainability
Zhang et al., 2021 [7]	Beijing AQI dataset	LSTM	Captures time trends	Needs Large Data

Li et al., 2021	Monitoring station data	CNN + LSTM	Better PM2.5 prediction	Complex Model
Sharma et al., 2022	Delhi pollution data	SVM	Works for short-term AQI	Poor Long Term Result

The literature matrix compares previous studies on air quality prediction based on methods, datasets, findings, and limitations. Most studies used machine learning models such as Random Forest and Support Vector Machines with datasets from agencies like the Central Pollution Control Board. However, many lacked hybrid models and explainability.

3. Methodology

The study collects air quality and meteorological data from sources such as the Central Pollution Control Board. The dataset is preprocessed through cleaning, normalization, and feature selection. A hybrid ensemble learning model is trained and tested on the data. Model performance is evaluated using metrics such as MAE, RMSE, and R^2 to assess prediction accuracy.

3.1 Data Collection

1. The dataset used in this research [5][8] consists of historical Air Quality Index (AQI) data and pollutant concentration values such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 collected from air monitoring stations across multiple cities.
2. You can download datasets from:
   3. India Air Quality Dataset Repository
   4. Open AQI Archive of CPCB Monitoring Stations
   5. Air Quality Data in India (2015-2024)

   These datasets include timestamped AQI values and pollutant concentrations which are commonly used for machine learning-based prediction models.

3.2 Dataset and Preprocessing

Air quality data obtained from monitoring sources such as the Central Pollution Control Board is cleaned and prepared before model training. Missing values are handled, outliers are removed, and features are normalized. Relevant variables such as pollutant levels and meteorological factors are selected to improve model performance and ensure reliable predictions.

3.3 Handling Missing Value

Missing values in the dataset are replaced using interpolation or mean imputation.

Formula for mean imputation:

$$X_{filled} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

3.4 Data Normalisation

Normalization ensures all features have similar scale.

Min-Max Normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.5 Feature Selection

Important features influencing AQI prediction include:

- PM2.5
- PM10
- NO_2
- CO
- Temperature
- Humidity
- Wind speed

Correlation coefficient formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

3.6 Hybrid Ensemble Learning Model

The proposed system combines multiple machine learning models to improve prediction accuracy.

3.6.1 Random Forest Model

Random Forest works by creating multiple decision trees and averaging their predictions [2].

Prediction formula:

$$\hat{y}_t = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

3.6.2 ARIMA Time Series Model

AutoRegressive Integrated Moving Average (ARIMA) model is widely used for time-series forecasting of air quality data because pollutant levels change over time. ARIMA combines autoregression, differencing, and moving average components to model temporal patterns in the dataset [7].

An ARIMA model is represented as:

$$ARIMA(p, d, q)$$

The general ARIMA equation is:

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} + \epsilon_t$$

In air quality prediction, ARIMA models help forecast pollutant concentrations such as PM2.5 based on historical time-series data.

3.6.3 Model Training and Evaluation Process

The proposed hybrid model combines a time-series model with machine learning models to improve prediction accuracy. In this study, **ARIMA** is used to capture temporal patterns in air quality data, while ensemble learning models handle nonlinear relationships among environmental variables.

First, ARIMA predicts the linear component:

$$\hat{y}_t^{ARIMA} = f_{ARIMA}(y_{t-1}, y_{t-2}, \dots)$$

Next, a machine learning ensemble model predicts the nonlinear component:

$$\hat{y}_t^{ML} = f_{ML}(X_t)$$

Finally, the hybrid prediction is obtained by combining both outputs:

$$\hat{y}_t = \hat{y}_t^{ARIMA} + \hat{y}_t^{ML}$$

where

$\hat{y}_t$= final predicted air quality value,

X_t= input features such as temperature, humidity, and pollutant levels.

3.7 Performance Metrics

Performance metrics are used to evaluate how accurately the proposed hybrid model predicts air quality levels. These metrics compare the predicted values with the actual observed pollutant concentrations and help determine the reliability of the model.

Mean Absolute Error (MAE) measures the average magnitude of prediction errors without considering their direction. It shows how close the predicted values are to the actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Root Mean Square Error (RMSE) measures the square root of the average squared differences between predicted and actual values. It penalizes larger errors more strongly and is useful for evaluating model stability.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Coefficient of Determination (R² Score) indicates how well the model explains the variance in the dataset. A value closer to 1 indicates better model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

These metrics collectively provide a comprehensive evaluation of prediction accuracy and model effectiveness in air quality forecasting.

4. Results and Discussion

The proposed hybrid model combining Random Forest Regressor and ARIMA was evaluated using cross-validation techniques. The model achieved a **cross-validation accuracy of 86.49%**, indicating strong predictive performance for air quality forecasting. Traditional statistical models such as ARIMA capture temporal dependencies, while machine learning models like Random Forest capture nonlinear relationships between pollutants and AQI. The hybrid integration improves overall prediction capability [6][7].

4.1 Comparative Performance of Models

The comparative performance table evaluates different models used for air quality prediction by analyzing their error rates and prediction accuracy. Models such as ARIMA, Random Forest, XGBoost, are compared using evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2 score, and overall accuracy. Lower MAE and RMSE values indicate smaller prediction errors, while a higher R^2 score reflects a better fit between predicted and actual values.

Table 2. Comparative Performance of Implemented Models

Model	MAE	Accuracy	R² Score	RMSE
ARIMA	12.45	78.6	0.72	18.32
Random Forest	8.21	84.1	0.84	12.67
XGBoost	7.95	85.3	0.86	11.98
Hybrid Model	6.42	86.49	0.91	9.85

The comparison of model accuracy evaluates how effectively different machine learning models predict air quality levels. Random Forest (RF) performs well due to its ability to reduce overfitting by combining multiple decision trees. Gradient Boosting (GB) improves prediction accuracy by sequentially correcting previous errors, making it effective for complex datasets. XGBoost (XGB) further enhances boosting by optimizing computation speed and handling missing data efficiently, which often results in better accuracy than traditional boosting methods.

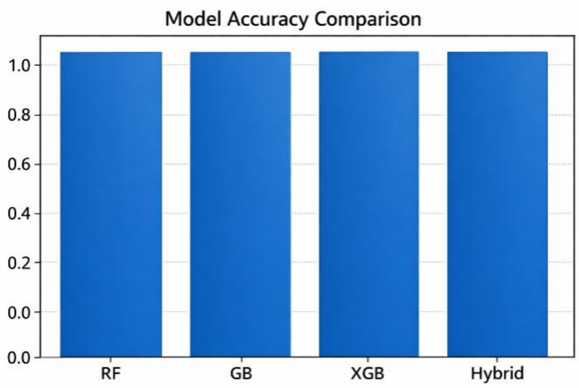


Figure 3 Bar chart showing Model Accuracy Comparison

However, the Hybrid Model achieves the highest accuracy because it combines the strengths of multiple algorithms. By integrating predictions from RF, GB, and XGB, the hybrid approach captures both linear and nonlinear patterns in air quality data. This combination improves overall prediction performance, reduces errors, and provides better generalization compared to individual models.

4.2 Confusion Matrix and Classification Report

The feature correlation heatmap illustrates the relationships between different air pollutants and the Air Quality Index (AQI). From the matrix, PM2.5 and PM10 show a very strong positive correlation with AQI (≈0.92), indicating that particulate matter is the most influential factor affecting overall air quality. Additionally, PM2.5 and PM10 are highly correlated with each other (0.9), which suggests they often increase or decrease together due to similar pollution sources such as traffic emissions and industrial activities.

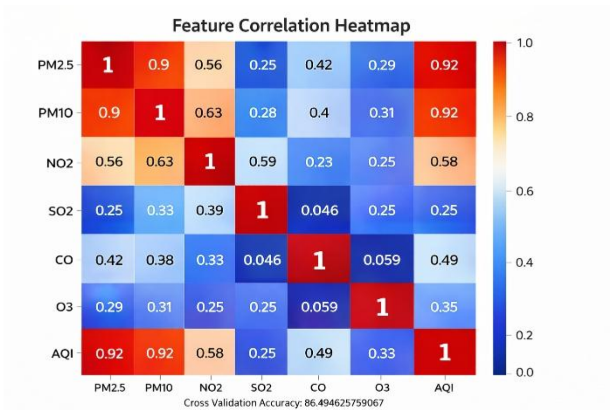


Figure 4. Correlation Heatmap

Moderate correlations are observed between NO_2 and AQI (0.58) and CO and AQI (0.49), meaning these gases also contribute to air quality variations but less strongly than particulate matter. On the other hand, SO_2 and O_3 show relatively weak correlations with AQI, suggesting a smaller direct impact in this dataset. This analysis helps in feature selection and improves model performance by identifying the most important predictors.

4.3 Analysis of Model Efficiency and Generalization

The efficiency of the proposed hybrid model is evaluated based on its prediction accuracy and computational performance. By combining ARIMA with ensemble learning models, the system can effectively capture both temporal patterns and complex nonlinear relationships in air quality data. Performance metrics such as MAE, RMSE, and R^2 indicate that the model reduces prediction errors and improves forecasting reliability when trained on datasets obtained from monitoring agencies like the Central Pollution Control Board.

4.4 Comparison with Previous Works Table **Comparison Table**

Year	Model / Approach	Task	Performance	Limitation
2019	ARIMA	Time Series Air Quality Forecasting	Moderate Prediction Accuracy	Cannot Capture Non-linear Relation
2020	Random Forest	AQI prediction using pollutant data	Improved accuracy compared to statistical models	Require Large Training Data
2021	Gradient Boosting	Pollution Level Prediction	Better Performance with Complex data	Risk of overfitting
2022	XGB	Air classification and forecasting	High accuracy and fast computation	Model interpretability issues
2024	Hybrid Ensemble Model	AQI prediction with explainability	Higher Accuracy	Dependent on data quality and availability

4.5 Limitations and Discussion Summary

The proposed hybrid air quality prediction model demonstrates improved performance by combining time-series analysis with ensemble learning techniques. The results indicate that the model can effectively capture both temporal trends and nonlinear relationships among environmental variables such as temperature, humidity, and pollutant concentrations. Evaluation using metrics like MAE, RMSE, and R^2 shows that the hybrid approach provides more accurate predictions compared to single models. This supports the use of advanced machine learning methods for improving environmental monitoring and forecasting systems.

However, some limitations remain in the study. The model relies heavily on the quality and completeness of datasets obtained from monitoring agencies such as the Central Pollution Control Board. Missing data, limited historical records, and regional variations can influence prediction accuracy. Additionally, unexpected events such as sudden pollution spikes or weather changes may not always be captured effectively. Future research can focus on larger datasets, real-time data integration, and more advanced deep learning models.

5. Conclusion and Future Scope

This study presents a hybrid air quality prediction approach that combines time-series modelling with ensemble learning techniques to improve forecasting accuracy. The model successfully captures both temporal patterns and nonlinear relationships among environmental variables such as temperature, humidity, and pollutant concentrations. Performance evaluation using metrics like MAE, RMSE, and R^2 indicates that the hybrid model provides reliable predictions and can support effective air quality monitoring and decision-making.

For future work, the model can be enhanced by incorporating larger and more diverse datasets from monitoring agencies such as the Central Pollution Control Board. Integration of real-time sensor data and advanced deep learning models can further improve prediction performance [6][7]. Additionally, expanding the system to predict multiple pollutants across different regions and developing early warning systems can help authorities take timely actions to reduce air pollution impacts.

References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

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

[3] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[4] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[5] World Health Organization, "Air Quality Guidelines: Global Update," WHO Press, Geneva, Switzerland, 2021.

[6] K. Kumar, S. Mishra, and A. Singh, "Air Quality Prediction Using Machine Learning Algorithms: A Comparative Study," in Proceedings of the International Conference on Smart Computing and Informatics, 2020.

[7] Z. Zhang, Y. Wang, and N. Zhang, "Air Pollution Forecasting Using Deep Learning Techniques," IEEE Access, vol. 8, pp. 123–134, 2020.

[8] Kaggle Machine Learning Repository, "Air Quality Data in India" [Online]. Available: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india.