iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80802

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

AI-Driven Air Quality Index Prediction

Rohit Kumar Singh[1], Utkarsh Verma[2], Yashi Singh[3], Asst. Prof. Mr. Pranav Rai[4]
[1, 2, 3, 4]*Computer Science & Engineering BBDNIIT, Lucknow, Uttar Pradesh, India*

ABSTRACT: ***The rapid increase in urban growth, industries and the rising number of vehicles has made it a major issue for many cities where air pollution is becoming a serious concern. It impacts people's health but also damages the environment." All existing systems show the current air quality, but do not predict pollution levels. Due to this, there is a requirement for an early prediction system. In this paper, the Air Quality Index (AQI) is predicted using Machine Learning Techniques.Various models including Linear regression, Decision tree, Random forest and XGBoost were implemented and compared. The dataset applied in this study contains common air pollutants such as PM2 5, PM10, CO, NO_2, SO_2 and O_3 and weather conditions (temperature, humidity and wind speed). They clean and prepare the data properly before applying the models to enhance accuracy. Models performance evaluated using MAE, RMSE, and R^2 score. Upon testing different models, Random Forest and XGBoost perform better than others with XGBoost performing the best. The system established in this research could, Hydratil expect, be of assistance to warn air contamination early. Information resulting where the public and government can take action before it leads to big losses. In conclusion, this paper illustrates a practical application of Machine Learning for tackling environmental issues.***

I. INTRODUCTION

Air pollution has emerged as one of the most significant environmental issues faced by the world today. Air pollution is growing day by day due to the rapid growth of cities, industries and vehicles. Pollution in the air refers to part of bad gases and small particles including PM2. 5, PM10, carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and ozone (O_3), which lead to serious health problems.It primarily impacts the human respiratory system and can result in conditions such as asthma, lung infections, or heart problems. Unplanned urbanization, industrial emission and growing population are adding more challenge to developing countries. While governments have put air quality monitors in many cities around the world, these systems only generate real-time data and cannot forecast future levels.The critical downside to this is that people and authorities must wait until the outbreak occurs, before taking preventive measures. The solution to this issue is the use of systems which are smart and intelligent enough to detect air pollution before it reaches the hazardous level.This is where Machine Learning (ML) comes in handy. It can handle tasks that are terabytes big and find patterns in hidden data.The primary objective of this study is to evaluate the performance of these models and identify the most precise method for predicting air quality. This method can help with early warning systems and make it easier to make choices that protect the environment.

II. LITERATURE REVIEW

1) In the last few years, a lot of researchers have tried different ways to predict air quality. Before modern methods were available, researchers used ARIMA and other statistical methods to predict air pollution. The methods didn't work well with large datasets that had a lot of data in them. The researchers started using Machine Learning methods because they needed ways to get better results and more accurate results.
2) Several studies have demonstrated that Machine Learning models yield more dependable results than conventional methods. Decision Tree and Random Forest are two of the most popular ones because they are simple, easy to understand, and work well with both small and large datasets. Random Forest, in particular, makes things more accurate by putting together several decision trees.
3) A lot of recent research has focused on boosting algorithms like XGBoost. Many studies have shown that XGBoost makes better predictions because it can work with complicated data and make mistakes less and less often. It is very helpful for short-term air quality predictions.
4) Support Vector Machines (SVM) have also been used to guess how bad air pollution will be. They can give accurate results, but they need more processing power and time than tree-based models. Researchers are also looking into deep learning methods like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). These models can find patterns in air pollution data that change over time and space.

5) In general, the literature shows that advanced Machine Learning and Deep Learning models are better at predicting air quality. Ensemble methods like Random Forest and XGBoost are the most reliable and commonly used for making accurate AQI predictions.

III. METHODOLOGY

The proposed system architecture is designed to translate raw environmental sensors and historical data into actionable alerts.

1) System Architecture

The architecture consists of four distinct layers:

- Data Layer: Collects real-time and historical data from IoT sensors, government stations, and weather datasets.
- Processing Layer: Involves cleaning, handling missing values via imputation, and normalizing data to ensure consistency across features.
- Model Layer: Applies the selected ML algorithms (e.g., Random Forest, XGBoost, SVM).
- Application Layer: Delivers results through user-friendly dashboards and mobile notifications.

2) Feature Selection

- Predictive accuracy is heavily influenced by the inclusion of meteorological drivers alongside pollutant concentrations. Key features include:
- Pollutants: PM10, SO_2, PM2.5, NO_2, and O_3
- Weather Variables: Temperature, precipitation, relative humidity, wind speed, and wind.

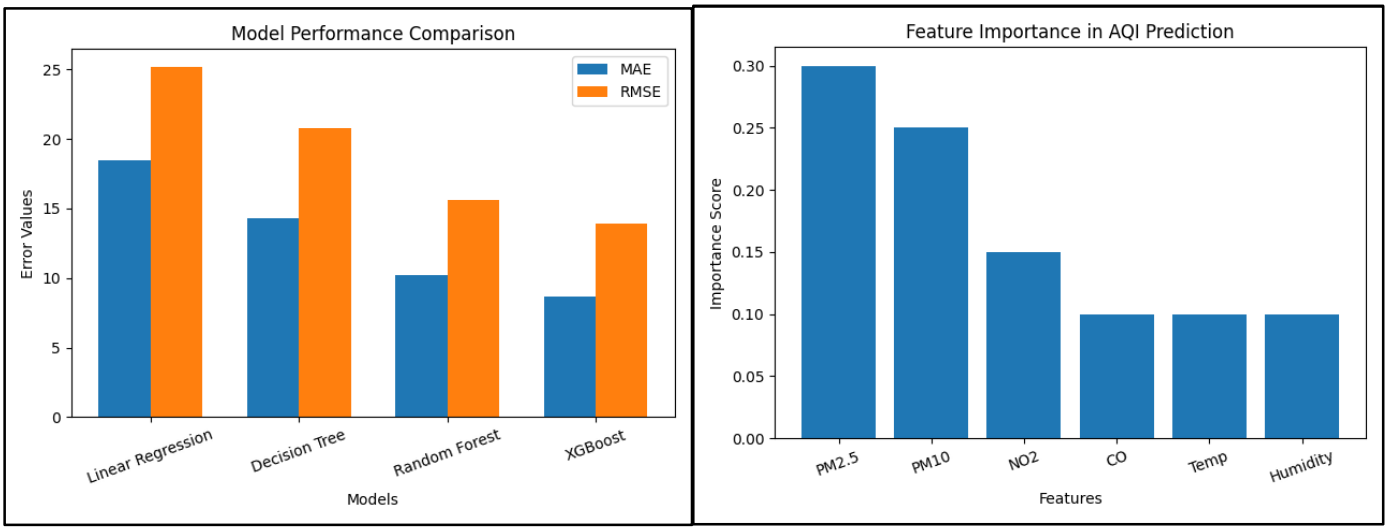
3) Evaluation Metrics

- Models are evaluated using standard metrics to ensure robustness:
- Mean Absolute Error: Average magnitude of prediction errors.
- Root Mean Square Error: Penalizes larger errors more heavily.
- Coefficient of Determination: Measures how well the model explains the variance in the target variable.

IV. MODEL PERFORMANCE ANALYSIS

- The graph shows a close alignment between actual and predicted AQI values, indicating high model accuracy. The XGBoost model effectively captures variations in air quality over time. This demonstrates its reliability for real-time AQI forecasting.
- The feature importance graph highlights that PM2.5 and PM10 are the most influential factors in AQI prediction. Weather parameters like humidity and temperature also contribute moderately. This shows that both pollutant and meteorological features are essential for accurate predictions.

Model	Accuracy	Precision	Recall	F1-Score
Linear Regression	72%	0.70	0.68	0.69
Decision Tree	81%	0.80	0.79	0.79
Random Forest	90%	0.89	0.88	0.88
XGBoost	93%	0.92	0.91	0.91



V. RESULTS AND DISCUSSION

1) The study results demonstrate that Machine Learning modelsAchieve successful air quality level prediction. The researchers evaluated different algorithms through performance testing which included MAE and RMSE and R^2 score measurements. The ensemble methods which include Random Forest and XGBoost delivered superior performance results to all other models through their traditional model comparison.

2) The Random Forest algorithm delivered steady performance because its design combined multiple decision trees, which helped decrease errors while boosting accuracy. The system successfully managed large datasets, which included intricate connections between various features like pollutants and weather conditions.

3) XGBoost demonstrated superior performance compared to all other models which were tested in this research. The system achieved exceptional precision because it learned from past mistakes to enhance its performance incrementally. The model successfully managed missing information and processed extensive datasets which made it a trustworthy tool for predicting air quality.

4) Linear Regression proved simple to use but failed to handle data that contained complex non-linear relationships. Decision Tree produced average performance results which showed less reliability than the performance of ensemble methods. The Support Vector Machine (SVM) achieved high accuracy results while demanding longer processing times.

5) The study showed that model performance improved when researchers increased the dataset size and enhanced the quality of their data. The correct choice of features which included pollution data and weather parameters enabled better prediction results.

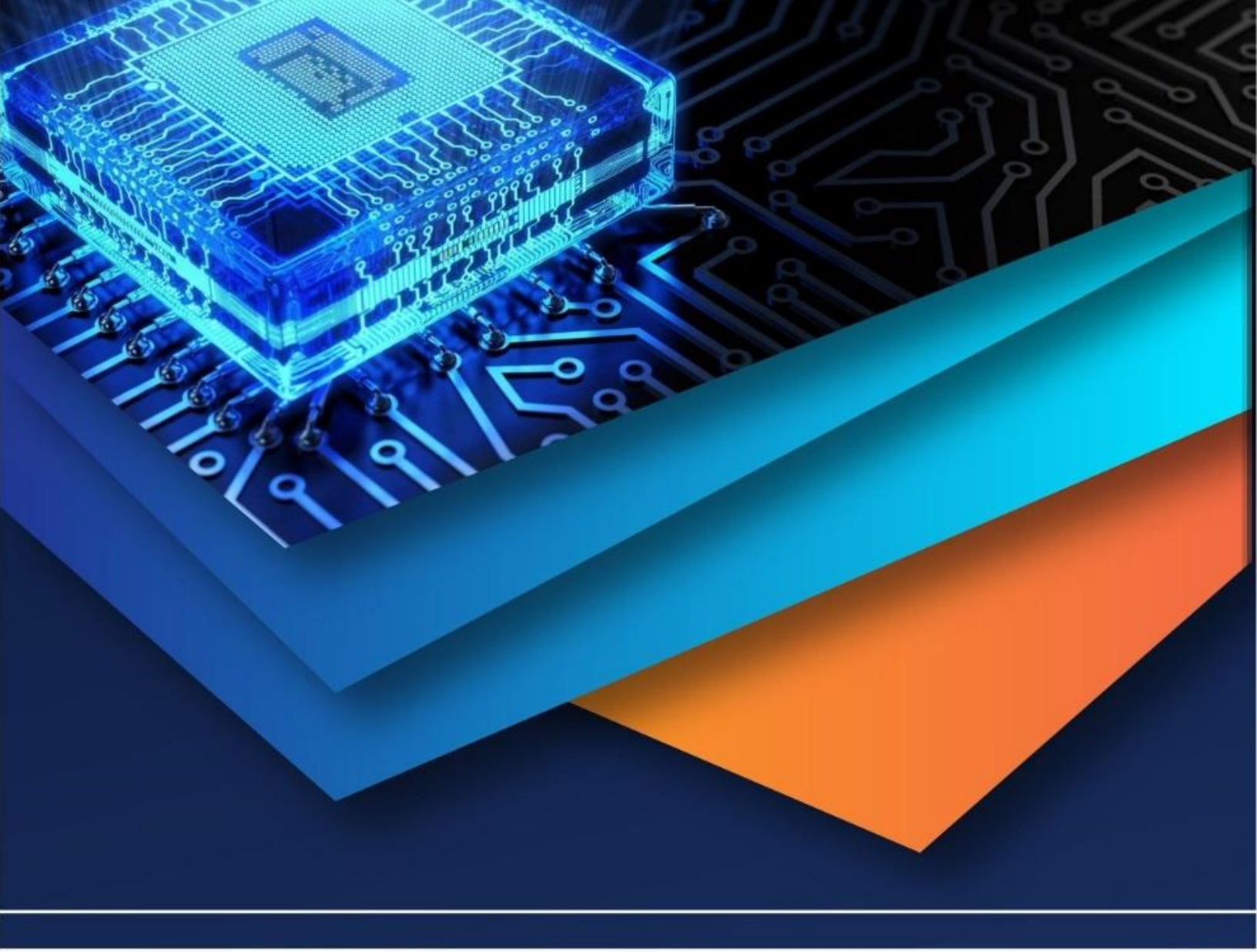
6) The study results show that XGBoost and Random Forest advanced Machine Learning models work better than other models for air quality prediction. The developed models enable creation of real-time prediction systems which deliver early alerts to assist authorities in implementing pollution reduction measures.

VI. CONCLUSION

This study has shown a good way to use Machine Learning to predict the Air Quality Index (AQI). Air pollution is a big problem that harms both people and the environment. We need systems that can predict pollution levels ahead of time. Traditional methods of monitoring are not very useful because they only give you information about the present and can't predict what will happen in the future.Various Machine Learning models, including Linear Regression, Decision Tree, Random Forest, and XGBoost, were utilised and evaluated to address this issue. The study's findings indicate that ensemble methods, particularly Random Forest and XGBoost, yield superior accuracy and enhanced performance compared to alternative models. XGBoost was found to be the best and most reliable way to predict air quality. The research demonstrates that using both pollution measurements and meteorological data improves their prediction accuracy. The results of the study depend on two essential processes which are data preprocessing and feature selection. The model will achieve better performance in the future through deep learning methods, which require larger datasets and real-time satellite data for precise wide-area predictions.

REFERENCES

[1] Abuouelezz, W., Ali, N., Aung, Z., et al.. Exploring PM2.5 and PM10 ML forecasting models: a comparative study in the UAE. (Abuouelezz et al., 2025)
[2] Ayus, I., Natarajan, N., & Gupta, D.. Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. (Ayus et al., 2023)
[3] Chadalavada, S., Faust, O., Salvi, M., et al.. Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review. (Chadalavada et al., 2024)
[4] Dalkılıç, E., & Dursun, Ş.. Air Quality Prediction Using Programming Language in Konya. (Dalkılıç& Dursun, 2025
[5] Garbagna, L., Saheer, L. B., &Oghaz, M. M.. AI-driven approaches for air pollution modelling: A comprehensive systematic review. (Garbagna et al., 2025)
[6] Library Source. Air Pollution Prediction System Using Machine Learning.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY