International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 23-34 March 2026-Sept 2026 www.csjournals.com

Machine Learning and Rule-based Expert System for AQI forecasting and Health Risk Mapping

Ritu Khatkar[1] , Dr. Vishal Verma[2] and Mr. Sandeep Rathi[3]

[1]MCA Student, [2]Professor, [3]Asst. Prof.
Deptt. of Comp. Sc. and Application, Chaudhary Ranbir Singh University, Jind, Haryana, India

Abstract: Air Pollution is a major environmental and public health concern due to industrialization and urbanization. Maintaining air pollution is a very essential part of human life. It impacts human life negatively like breathing problems, headache, asthma, weak immune system etc. The dataset is utilized from CPCB website for this study. This study outlines the prediction of AQI which is an important indicator to measure air pollution. It analyze the relationship between AQI and Air Pollution. Increasing AQI levels are directly associated with deteriorating air quality. Higher AQI values indicate increased concerns of harmful pollutants such as PM2.5, PM10, CO, SO_2, and NO_2. A relevant model is developed to analyze and predict AQI based on these pollutants. The model predict AQI value and classify air quality into categories like good, satisfactory, poor and moderate. After classification of AQI categories health diseases are associated with AQI using Rule-based programming. Machine learning techniques are used to predict AQI. Random forest, XGBoost and Linear regression are used to predict AQI. Among all these models XGBoost Regressor display highest accuracy with R2 value 0.983, MSE 11.25 and RMSE 3.35. Random Forest model accuracy with R2 value is 0.982, MSE is 11.66 and RMSE is 3.41. Linear Regression model R2 value is 0.92 ,MSE is 53.60 and RMSE is 7.32.

Keywords: *Air Quality Index(AQI); Ensemble ML; XGBoost; Random Forest; Rule-Based Health Alert*

1. Introduction

In recent years, Air Pollution is a very critical environmental challenges. Poor air quality has a negative impact on public health. There are many factors which increase the air pollution in our daily life. The pollutants which contribute to air pollution are called air pollutants. The presence of these pollutants effects living and non-living things. These pollutants may originate from natural sources like smoke and dust stemming from forest fires or volcanic eruptions. Sometimes, pollutants are also increased due to human activities. It include factories, vehicle smoke, Agriculture & land use activities, transportation, burning fuels like coal and petrol, cutting down trees, crop burning and burning waste etc. Vehicles produce large amounts of carbon monoxide, carbon dioxide, nitrogen oxides, and smoke. Carbon monoxide is produced from the incomplete combustion of fuels like petrol and diesel. When harmful gases in the air are increased it also contributes to air pollution. For example carbon dioxide (CO_2), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NO_2).

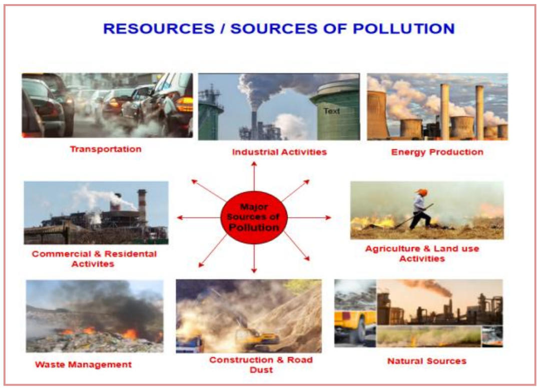


Figure1: Sources of Pollution

Major pollutants which expand air pollution in environment are PM2.5[8], PM10, CO, NO_2, SO_2, O_3.PM2.5, PM10 are the major contributors to increase air pollution. Because the size of these particles is very small. Particulate matter 2.5 have adversarial effect on human life[10]. PM2.5 particles can infilitrate the lower respiratory tract and even enter the bloodstream. Associated with **significant public health risks**: asthma, cardiovascular disease, cancer[9]. It is also said that this particle is increased in winter season due to fog and aerosols[10]. These particles are emerge from vehicle emissions, industrial activities, construction dust, and natural sources like dust storms and soil erosion .

Therefore we need to determine the presence of air pollutants in the atmosphere. We measure pollutants with the help of AQI at a particular place. AQI reveals that how much air is polluted or clean. AQI value also determines public awareness and government for decision-making. Government authorities use AQI data to control pollution sources. Government formulate policies to control pollutions like **BS-VI emission standards** for vehicles.

Different types of pollutants are chlorofluorocarbons (CFCs), which are used in refrigerators, air conditioners, and aerosol sprays. CFCs damage the ozone layer of the atmosphere. Fortunately, less harmful chemicals are used in place of CFCs. In addition to these gases, very small particles are also produced by the combustion of diesel and petrol in automated vehicles, which remain suspended in the air for a long time. In transportation CNG , electric vehicles and public transport system are mainly used to reduce pollution.

It is well known that Ozone layer act as a protective barrier from harmful ultraviolet rays originates from the sun. O_3 also increase air pollution but not directly. It is formed with the help of sunlight reaction and other pollutants. CFCs decrease the thickness of Ozone layer. It increase the risk of skin cancer, harms plants and marine life.

Air pollution contribute to many diseases that affect human health. Diseases which arise due to air pollution are respiratory diseases such as asthma, eye irritation, and lung cancer etc. These diseases affect different organs of our body. The impact is more severe on vulnerable groups such as children and elder people. Long term exposure of air pollution also reduce life expectancy. Particulate matter 2.5 also contributes to lung cancer [10]. The results showed that the estimated avoidable mortality, linked to ischemic heart disease and lung cancer, was 2,773 for every 100,000 people, which accounts for 2,024,290 preventable deaths of the total population[10]. It causes difficulties in breathing and wheezing in children. It affects people more severely who already suffer from breathing problems. PM2.5 also linked to respiratory diseases, asthma, bronchitis.

When inhaled, they enter the body and cause diseases. These particles are also emitted by industrial processes such as steel manufacturing and mining. Particulate matters produced from power plants also responsible for atmospheric pollution. Fog during winter season also contributes to air pollution but not directly. When it interacts with smoke from vehicles, industries, or burning crop residue, it forms **smog**. Smog is extremely harmful than normal fog or pollution alone. Fog reduces visibility, but when it converts into smog it degrades the quality of air. It makes smogy mornings and which causes feel extremely unhealthy.

Objectives of the Study:

To address the identified gaps in existing environmental forecasting systems, the primary objecti core research immperatives are formulated as follows:

To analyze and preprocess localized environmental data: To systematically analyze historical multi-dimensional air quality and meteorological data (sourced from CPCB, Jind) to identify complex, non-linear pollution patterns.

To develop a comparative machine learning pipeline: To design and evaluate a multi-step forecasting architecture by benchmarking a baseline linear model (Linear Regression) against advanced ensemble models (Random Forest and XGBoost) for accurate, continuous AQI prediction.

To design a deterministic rule-based health mapping engine: To program an automated expert system using conditional logic (if-then constructs) that seamlessly maps the forecasted numerical AQI values to established WHO and CPCB epidemiological guidelines.

To generate actionable respiratory health warnings: To bridge the gap between environmental data and public health by automatically translating the numerical predictions into proactive, disease-specific vulnerability alerts (such as Asthma and COPD risks).

2. Literature Review

Table1: Literature Review of Existing Machine Learning Models

S.No.	Author & Year	Dataset Used	Techniques / Models	Key Features	Results	Limitations
1	Tırınk, 2025	2016-24 (Turkey) AQI +	XGBoost, LightGBM, SVM	Long-term pollutant + meteorologica	XGBoost: R^2 = 0.999, RMSE =	Region-specific; limited

		weather data		l integration	0.234	generalization
2	Yenkikar et al., 2025	India (Kaggle multi-year AQI + pollutants)	Hybrid RF + ARIMA, SHAP	Nonlinear AQI modeling + residual correction; feature importance (PM2.5, NO2, SO2)	R^2 = 0.94, MSE = 508.46	Limited external validation; dataset bias
3	M. Karmoude (2025)	Multiple AQI datasets	Non-neural, DL, Hybrid	Comparative analysis, AQI prediction	~85–95%	No real-time system
4	M.Rajesh (2025)	Sensor, satellite, demographic data	RF, XGBoost	Real-time prediction, health mapping	~90–96%	Complex integration
5	M. Abdelmalek (2025)	Pollutant datasets	GPR, SVM, DT,Ensemble	Handle complex datasets	~80–90%	High computation
6	Manuel Méndez(2023)	Review of 150+ papers	Various ML techniques	Trend analysis of existing research	No exact accuracy	No implementation
7	Kumar & Pande (2022)	Indian AQI data	Machine Learning models	Focus on anthropogenic impact	~80–90%	Region-specific
8	Delhi PM2.5 Study (~2018)	CPCB (2016–2017), 8 stations	GIS (IDW), AQI EPA method	Spatial PM2.5 mapping + weather correlation	PM2.5 up to 278 µg/m³; AQI Unhealthy Hazardous	Short duration; sparse monitoring
9	Zoran et al., 2020	PM2.5, PM10, AQI, meteorological data (Jan–Apr 2020, Milan)	Pearson correlation	Air pollution + climate vs COVID-19 cases	PM2.5/PM10 linked with daily new cases; dry air aids spread; warm season doesn't stop transmission	Short-term, observational; limited to Milan
10	Hassan et al., 2021	PM2.5 data from 312 sites in Pakistan (2017)	BenMAP-CE, rollback scenario, VSL method	PM2.5 reduction analysis; mortality assessment	≈2.02 million preventable deaths; economic burden ≈ US	Used US-based health functions; only PM2.5 and mortality

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 23-34 March 2026-Sept 2026 www.csjournals.com

					$1 billion	considered; limited to 2017 data

3. Research Methodology:

3.1 Steps to be followed to create a model of Air Quality and Public Health Impact using Machine Learning:-

- Data Collection
- Data Preprocessing
- Feature Selection & Outlier Removal
- Model Training & Selection
- Model Evaluation & Analysis

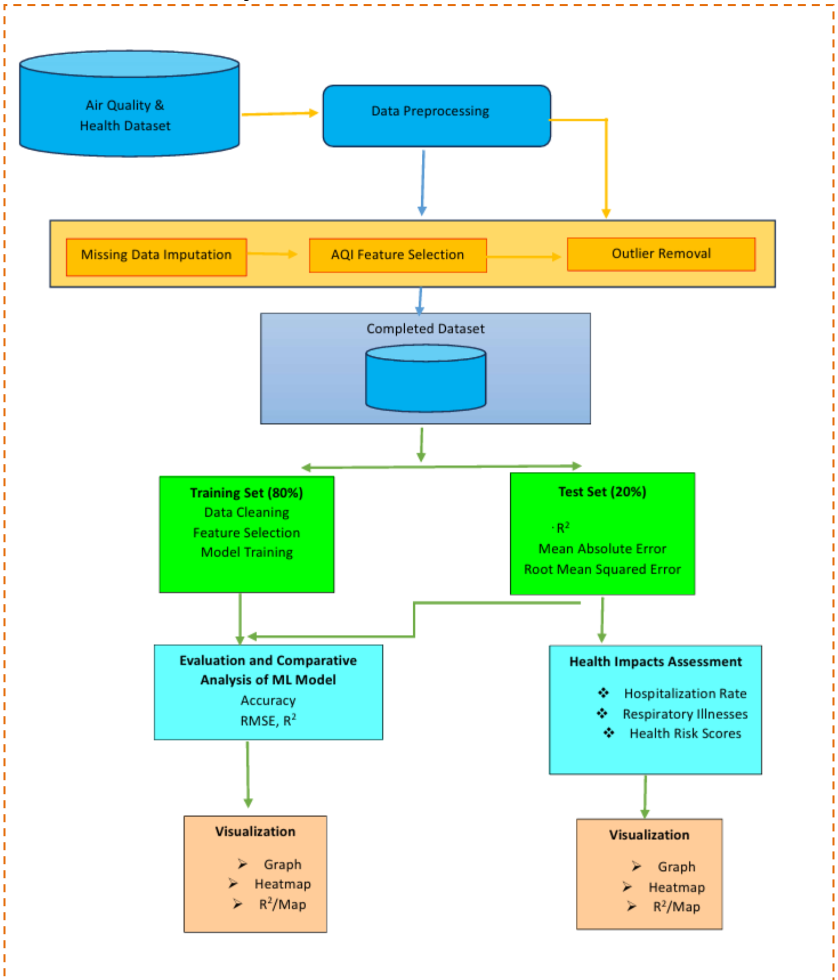


Figure 2: Workflow of the overall framework

3.1 Data Collection: For this study we utilized pre-existing historical to predict AQI. It is obtained from the Central Pollution Control Board (CPCB)[8], Government of India. The dataset includes year wise records of major pollutants like PM2.5, PM10, NO_2, and SO_2 ,O3 etc. The data were retrieved from the official CPCB website for multiple years(e.g., 2019–2024). Each year's dataset was downloaded separately and then preprocessed to ensure consistency in format and variables.

3.2 Data Preprocessing: In this process we clean our dataset so that model can easily understand it. Missing values are filled with the help of mean, mode and median. Mean and median is used to fill numerical values. Mode is used to fill categorical missing values. Data transformation is also a part of data preprocessing. In this we convert data in a proper format. Data is scaled between 0 and 1 by applying normalization and standardization.

3.3 Exploratory Data Analytics: Exploratory Data Analysis (EDA) is used to understand the dataset structure and underlying patterns. Statistical summary and visualization are used to analyze data distribution. Histograms, bar charts and heatmap are used for data analysis. It detects missing values and outliers from our dataset.

3.4 Feature Selection & Feature Engineering: It is used to selecting, creating, and transforming raw data into meaningful features that improve a machine learning model's performance. In feature selection features are selected to train model like reduce dimensionality, it helps in minimize noise and reduce overfitting.

3.5 Model Training & Selection: In this step model is trained to predict AQI on the bases of other particles existence in the environment. For model training dataset is spilt / divided into training and testing sets. Dataset is spilt in the ratio of 80:20 or 70:30. After model training models are selected to apply on the dataset for predictions. In this phase multiple machine learning techniques[7] are used like Random Forest[2][4], XGBoost[1][4], Decision Tree[5] and Light GBM[1].

❖ Proposed Models:

3.5.1 Linear Regression

It is a simple mathematical model which find out the relationship between input and output or target variable. In this model we predict AQI value on the basis of other pollutants exists in the environment.

3.5.2 Random Forest

It is an ensemble model it means that it create multiple decision trees at one time. After creating multiple decision tree and aggregates their outputs to improve prediction accuracy. It reduce overfitting, and provides robust predictions.

3.5.3 XGBoost Regressor

XGBoost Regressor is a gradient boosting-based composite predictive system. It develops sequential structure where each models reduce residual errors. It provides high predictive accuracy, handles noisy environmental data effectively, and incorporates regularization techniques to reduce overfitting[1][4][7].

❖ Proposed Methodology Framework:-

This framework defines the system workflow from data collection to AQI prediction. The framework ingest air quality data from CPCB website. A supervised learning ML model use data to predict AQI values. Then health risk and health diseases are defined on the basis of AQI value predicted by the model with the help of rule-based system.

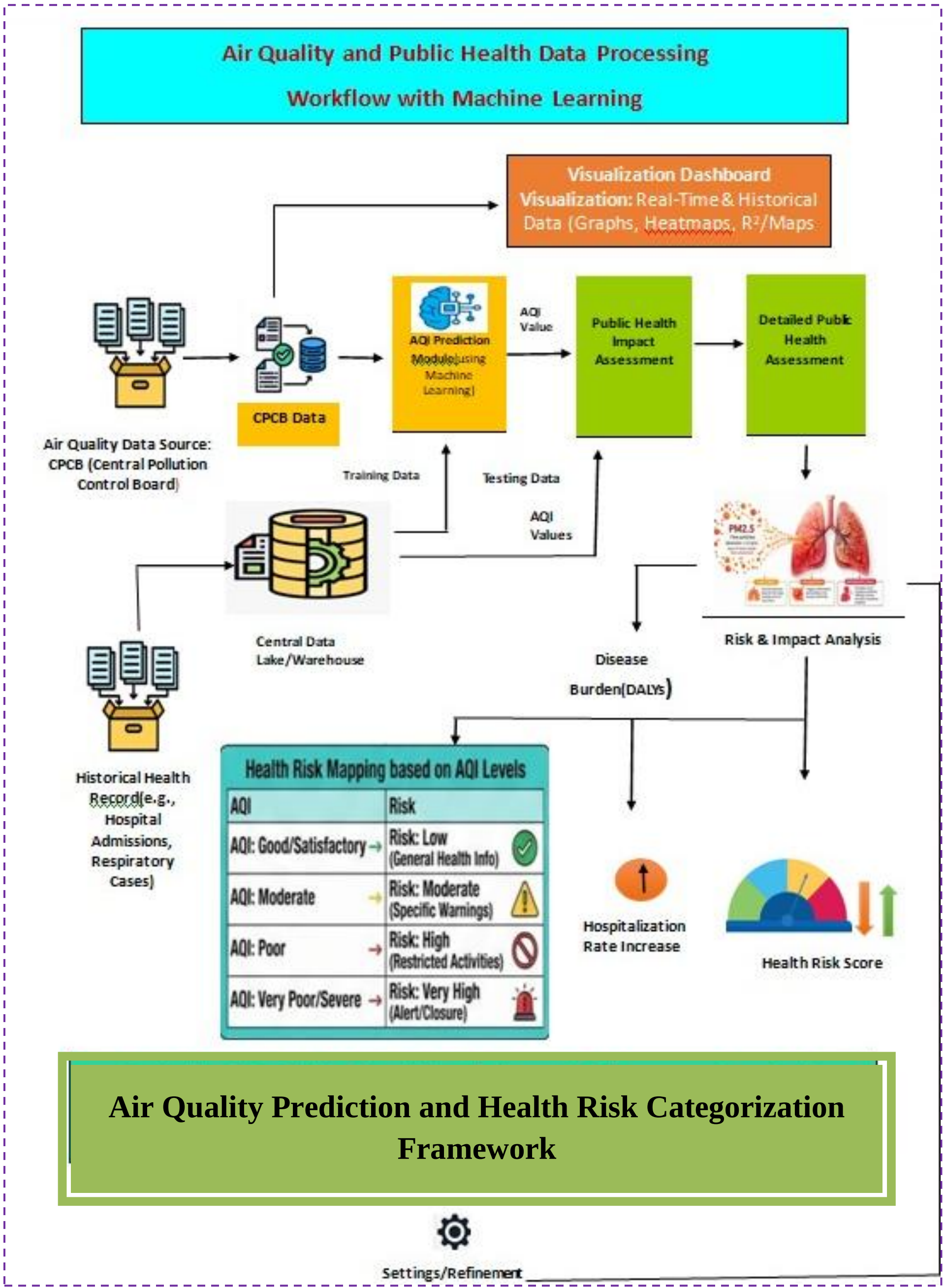


Figure3: System Architecture of the Framework

3.6 Model Validation: After model training and selection we can access the predictive efficeny.of our model. In model evaluation we can use appropriate metrics. Metrics used for classification problems are Predictive Efficency, Precision, Recal/Detection Rate, F1-score Mean. For regression problems Mean Absolute Error (MAE), MSE , RMSE, R^2 score metrics are evaluated. Classification report is also used for model evaluation.

- **Health Risk Mapping:** After predict AQI value for next seven days we mapped health risk like in existing papers [8]. The AQI values were categorized into six air quality classes using predefined threshold ranges. AQI values between 0–50 were

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 23-34 March 2026-Sept 2026 www.csjournals.com

labeled as **Good**, 51–100 as **Satisfactory**, 101–200 as **Moderate**, 201–300 as **Poor**, 301–400 as **Very Poor**, and values above 400 were classified as **Severe**.

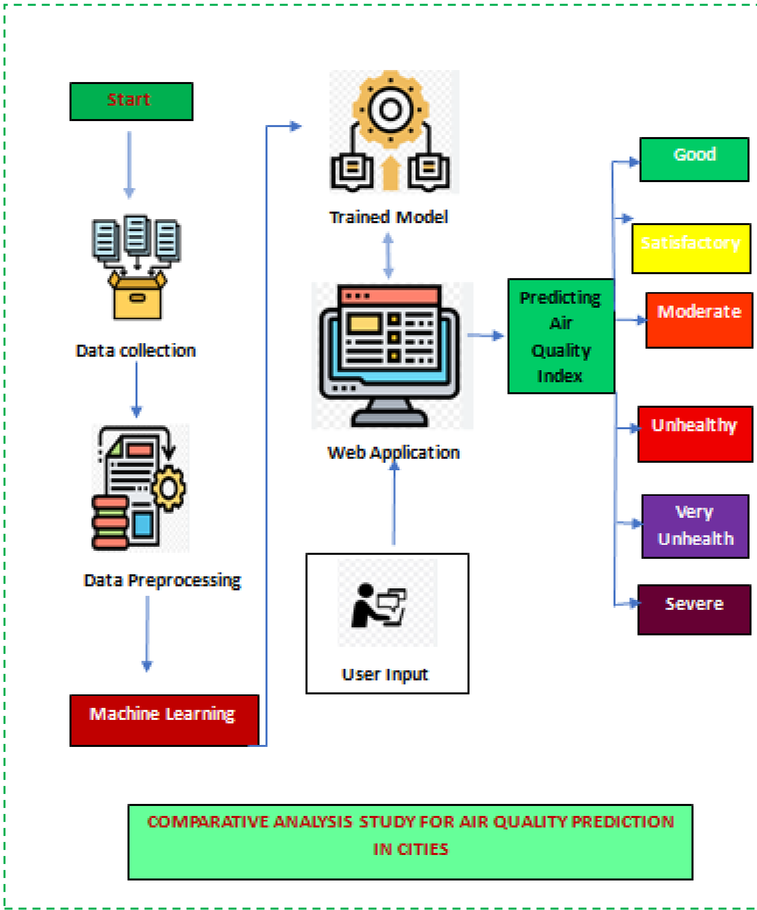


Figure4: Health Risk & Diseases Mapping

- **Health Diseases Prediction**: After categorized AQI into classes, health diseases prediction model is integrated with it with the help of Rule-based programming. Because machine learning models cannot predict health diseases on the basis of AQI. Many diseases are occurred due to poor air pollution or AQI and PM2.5[8][10] like lung diseases, respiratory problems, ischemic heart disease and asthma.

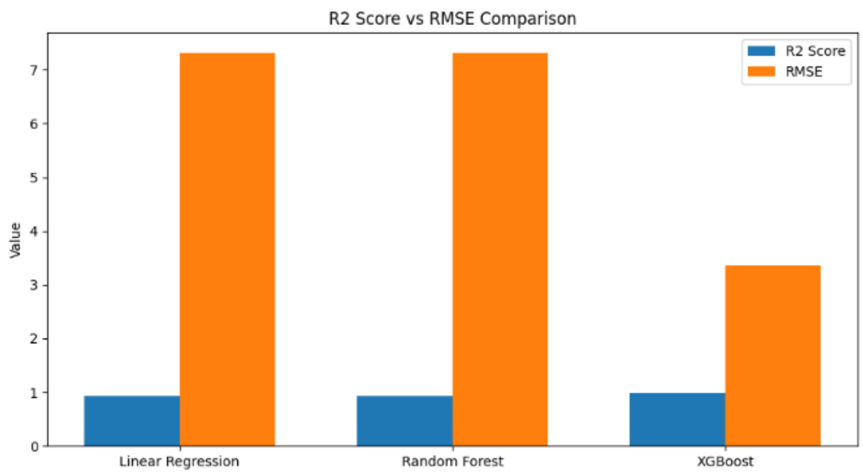
3.7 Results: In this section, we provide the outcomes and evaluation metrics of the proposed machine learning models implemented in this segment. Three machine learning models were used to predict AQI by using CPCB dataset. From evaluation metrics we find that XGBoost Regression demonstrates architecture superiority to the other models with R2 score of 0.983, RMSE of 3.35 and MAE of 11.25. Followed by Random Forest R2 score of 0.982, RMSE of 3.41 and MAE of 0.28. The evaluation metrics of Linear Regression model are R2 score of 0.86, RMSE of 7.32 and MAE of 4.73.

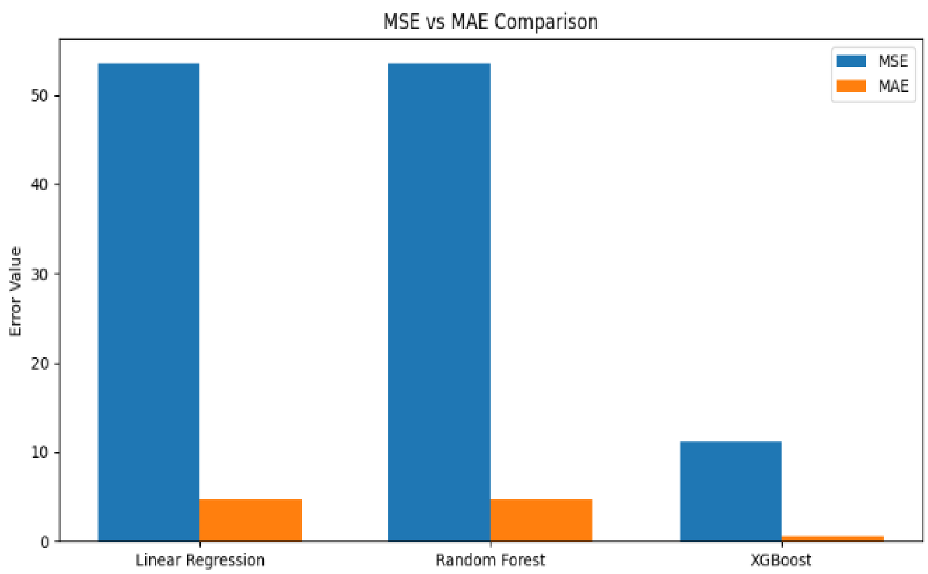
International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 23-34 March 2026-Sept 2026 www.csjournals.com

Table 2. Quantitative Model Evaluation Diagnostics

Machine Learning Techniques	R2 Score	RMSE	MSE	MAE
Linear Regression	0.92	7.32	53.60	4.73
Random Forest	0.98	3.42	11.66	0.29
XGBoost Regression	0.98	3.35	11.25	0.50

Model Performance Evaluation Metrics Graph





International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 23-34 March 2026-Sept 2026 www.csjournals.com

Feature importance Graph:

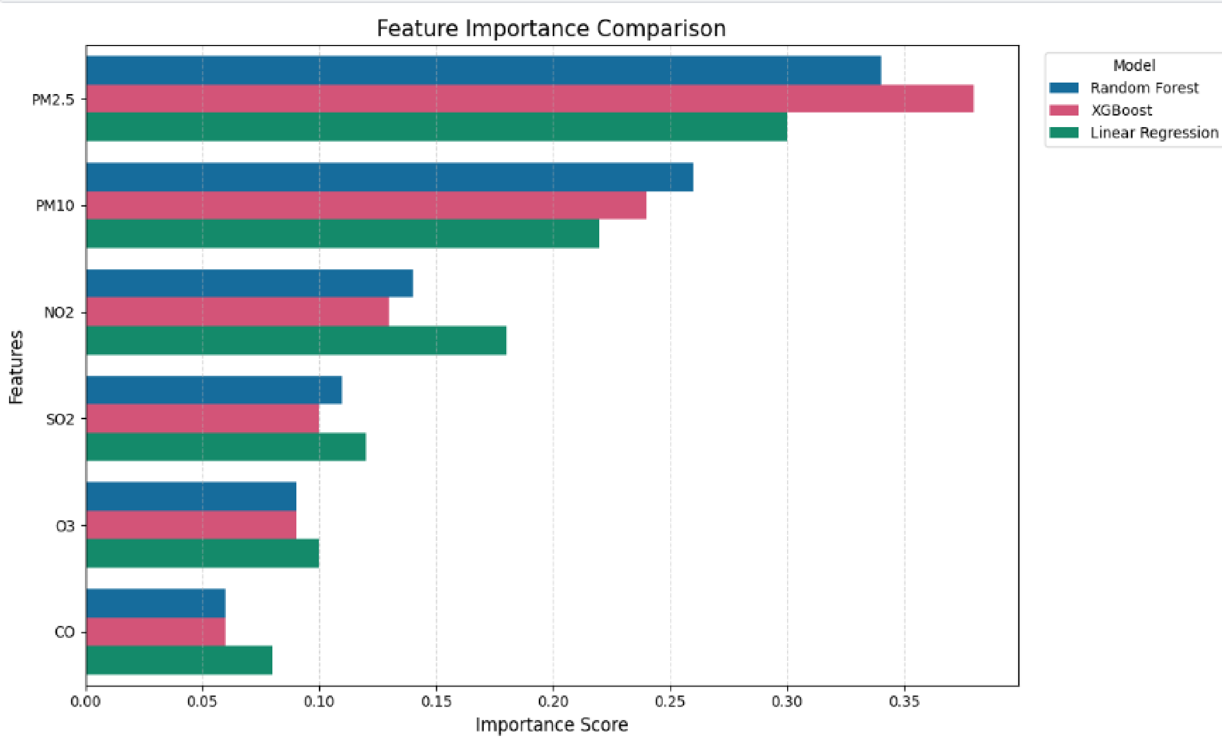
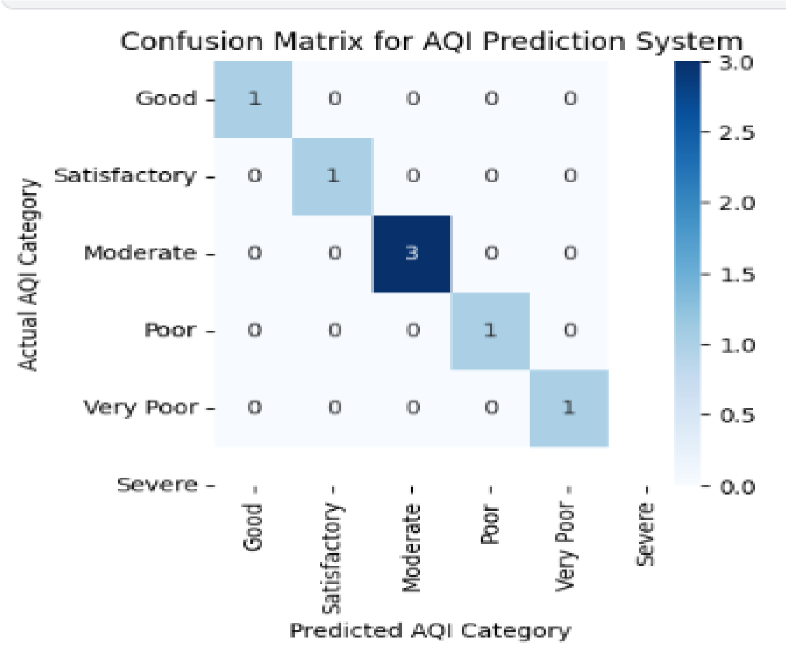


Table3: Feature Importance scores for each ML models

Features	Random Forest	XGBoost	Linear Regression
PM2.5	0.3400	0.3800	0.3000
PM10	0.2600	0.2400	0.2200
NO2	0.1400	0.1300	0.1800
SO2	0.1100	0.1000	0.1200
O3	0.0900	0.0900	0.1000
CO	0.0600	0.0600	0.0800

Confusion Matrix of AQI Predicted and Actual Values:



3.8 Conclusion:

This study presents a hybrid framework that integrates machine learning model (XGBoost) with a rule-based system. The goal is to predict air quality and related health risks. The model learns patterns from past CPCB data and predicts AQI for the next 7 days. The system also converts AQI values into simple health risk categories. These categories show risks for diseases like asthma and COPD. Overall, the system helps in early warning for air pollution and supports better decisions for public health.

3.9 Future Scope: Although the current model predict AQI and associate health risk using ML techniques. There are several opportunities for further enhancement and improvements. By integrating advanced technologies like Deep learning, cloud computing and Blockchain etc. In future you can deploy this model on cloud platform where data is fetched using real-time APIs because currently model runs on a local machine using Python. In future you can store data on Blockchain-based infrastructure which is collected from CPCB website. Using smart contract system ensure that data used for ML techniques is tamper-proof and transparent.

4 References:

[1] Tırınk, S. (2025). *Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM*. PLoS ONE, 20(10), e0334252. DOI: https://doi.org/10.1371/journal.pone.0334252

[2] Yenkikar, A., Mishra, V. P., Bali, M., & Ara, T. (2025). *Explainable forecasting of air quality index using a hybrid random forest and ARIMA model*. MethodsX, 15**,** 103517. DOI: https://doi.org/10.1016/j.mex.2025.103517

[3] Karmoude, M., Munhungewarwa, B., Chiraira, I., et al., Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks, Sci. Total Environ., 1002 (2025) 180593

[4] Rajesh, M., Ganesh Babu, R., Moorthy, U., Easwaramoorthy, S. V., Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping, Sci. Rep. 15 (2025) 28801

[5] Abdelmalek, M. M., Mahmoud, H., & Shokry, H., Prognosis of air quality index and air pollution using machine learning techniques, Sci. Rep. 15 (2025) 25890.

[6] Méndez, M., Merayo, M. G., Núñez, M., Machine learning algorithms to forecast air quality: a survey, Artif. Intell. Rev. (2023)

[7] Kumar, K., & Pande, B. P., Air pollution prediction with machine learning: a case study of Indian cities, Environ. Monit. Assess., 20(5):5333–5348 (2022).

[8] A. K. Gorai, P. B. Tchounwou, S. S. Biswal, and F. Tuluri,“Spatio-temporal variation of concentrations and its health impacts in a mega city, Delhi in India,” *Environmental Health Insights*, vol. 12, pp. 1–13, 2018. doi: 10.1177/1178630218792861

[9] M. A. Zoran, R. S. Savastru, D. M. Savastru, and M. N. Tautan, “Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy,” *Science of the Total Environment*, vol. 738, p. 139825, 2020. doi: 10.1016/j.scitotenv.2020.139825

[10] Hassan, A., Ilyas, S. Z., Agathopoulos, S., Hussain, S. M., Jalil, A., Ahmed, S., & Baqir, Y. (2021). *Evaluation of adverse effects of particulate matter on human life*. Heliyon, **7**(1), e05968. https://doi.org/10.1016/j.heliyon.2021.e05968