Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 03 Issue: 10 | Oct-2025
ISSN (O) 3107-6696

Air Quality Index Prediction using Machine Learning Techniques

T. Amalraj Victoire[1], M. Vasuki[2], J.Kumaresh[3*]

[1,2]Associate Professor, Department of Master Computer Application, Sri Manakula Vinayagar Engineering College, Pondicherry, India

[4]Student, Department of Master Computer Application, Sri Manakula Vinayagar Engineering College, Pondicherry, India

---***---

Abstract: Air pollution is one of the most important health and environmental issues today. The Air Quality Index (AQI) is gaining popularity as a recognized indicator of the safety of the ambient air that is easily understood by the public and the decision-makers. This study focuses on the use of machine learning models to predict AQI values using five years of pollutant data collected from some Indian cities. The pollutants investigated are sulfur dioxide (SO_2), nitrogen dioxide (NO_2), respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM). Several machine learning models were employed and studied in the analysis, including Linear Regression, Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN), to determine their potential for predicting AQI categories or predicting it as a continuous variable. The analysis showed that Random Forest was the most reliable and accurate approach, as it achieved the best balance between accuracy, interpretability, and generalization. This study raises awareness of the merit of developing predictive models that can and potentially lead to early warning systems and decision support systems for sustainable urban management and public health improvement.

Keywords: Air Quality Index, Machine Learning, Random Forest, Environmental Monitoring, Prediction

1. Introduction

The steady deterioration of ambient air quality has become a critical challenge for densely populated regions, affecting respiratory health, long-term morbidity, and urban livability. Multiple emission sources — including vehicular traffic, industrial outputs, construction dust, and seasonal biomass burning — interact with weather patterns to produce complex pollution dynamics that vary widely across time and location. To communicate air-risk succinctly, authorities use the Air Quality Index (AQI), which aggregates pollutant measurements into a single interpretive value for public guidanceDespite widespread monitoring networks, several limitations constrain current AQI systems: measurement stations are often concentrated in metropolitan zones leaving large areas under-sampled; sensors suffer from intermittent failures and calibration drift; and conventional statistical forecasting methods struggle to capture non-linear pollutant interactions and sudden spikes caused by episodic events. Machine learning offers tools to model such complexity by learning patterns from large datasets, handling noisy observations, and adapting to temporal trends. Motivated by these strengths, this study builds multiple supervised learning models on five years of pollutant data to produce reliable short-term AQI forecasts and categorical risk predictions suitable for operational use.

2. Literature Review

In recent years, numerous studies have explored the use of machine learning (ML) techniques in air quality forecasting. Goyal and Chan (2009) examined regression-based models for predicting air quality in Hong Kong and India but noted limitations in addressing non-linear pollutant interactions. Kumar and Goyal (2011) later proposed hybrid statistical–machine learning models for Delhi, demonstrating improved short-term accuracy.

In Europe, Chaloulakou et al. (2003) implemented artificial neural networks (ANNs) for PM10 prediction in Athens, Greece, successfully capturing non-linear dependencies though interpretability remained a concern due to the “black-box” nature of ANNs. Similarly, Yeganeh et al. (2012) combined support vector machines (SVMs) with genetic algorithms for CO forecasting, achieving high accuracy at the cost of increased computational demand.

Wang et al. (2014) analyzed pollutant trends across 31 Chinese cities, finding seasonal meteorological influences to be a dominant factor affecting AQI variations. Shaban et al. (2016) integrated Internet of Things (IoT) data with predictive models to enable real-time air quality monitoring and alert systems. More recently, Zhang et al. (2018) applied deep learning approaches, such as long short-term memory (LSTM) networks, to capture temporal dependencies in AQI forecasting, resulting in high accuracy. However, deep learning remains resource-intensive and challenging to implement effectively in developing regions like India

3. Problem Statement

Accurate aqi prediction in india remains a major challenge due to multiple limitations existing monitoring networks are concentrated in urban centers leaving semi-urban and rural areas under-observed moreover government sensor data often contain missing readings and calibration errors reducing reliability short-term spikes caused by events such as stubble burning dust storms and fireworks further complicate prediction traditional aqi systems primarily provide real-time updates rather than forecasts offering limited scope for preventive measures the highly dynamic and non-linear relationship between various pollutants emission sources and weather conditions requires advanced models capable of adaptive learning and generalization hence there is an urgent need for robust aqi prediction models powered by machine learning such systems should be capable of learning from large diverse datasets accounting for spatial and seasonal variations and generating reliable forecasts that can aid governments policymakers and citizens in making informed proactive decisions**.**

4. Existing System

Current AQI monitoring and prediction systems in India and other countries mainly rely on statistical methods and conventional sensor-based frameworks. The Central Pollution Control Board (CPCB) manages fixed monitoring stations across major cities that record pollutant levels and provide hourly AQI readings. However, these systems suffer from several drawbacks, including limited spatial coverage, since rural and semi-urban regions often lack real-time monitoring stations, leading to data gaps.

Traditional short-term forecasting models depend largely on linear regression or time-series techniques. While computationally efficient, these models fail to represent the complex, non-linear relationships between multiple pollutants and meteorological conditions. Consequently, they perform poorly during sudden pollution surges caused by events like stubble burning, Diwali fireworks, or industrial accidents.

Additionally, existing AQI systems rely heavily on the accuracy and uptime of sensors. In practice, issues such as calibration errors, sensor downtime, and missing data reduce the reliability and precision of reported AQI values. As a result, the current frameworks fall short of delivering dependable forecasts or comprehensive spatial coverage.

5. Proposed System

The framework proposed in this paper is a novel AI-based system for Air Quality Index (AQI) prediction, which improve upon the shortcomings of traditional models. The methods existing today are usually limited due to purely linear approaches. This proposal has better suitability to AI by introducing methods based on supervised machine learning, such as Random Forest, Decision Trees, Logistic Regression and K-Nearest Neighbors (KNN), which appropriately specialize in recognizing nonlinear relationships among variables, such as pollutants and environmental factors, providing improved predictions in complex and highly variable environments. The proposed framework focuses on using five years of pollutant data, which includes SO_2, NO_2, RSPM and SPM. After data collection, we will smooth the data, then use the smoothing processes to address missing data, normalization and feature selection. In this way, the model ensures all data input is clean and current - suited for predictive analysis. The main strength of the proposal is the kind of outcomes forecasted, continuous for AQI values (fit for research and in-depth analysis), as well as categorical AQI levels (indicating a quick communication to the public), providing versatility in the proposed models. Along with better AI-based predictive model to achieve AQI predictions and historic data holdings, we will also value scalability, general interpretation as well as ease of use to the model. The options of ensemble learning methods, specifically Random Forest, permits higher accuracy than traditional statistical models and appropriate consideration of overfitting of also.

6. Methodology

This study follows a systematic process to generate accurate and reliable AQI predictions. The workflow involves data collection, preprocessing, feature engineering, AQI calculation, model training, and evaluation using multiple algorithms to determine the best-performing approach.

Data Collection: Five years of pollutant data (2016–2021) were collected from official air quality monitoring stations across India. The dataset includes SO_2, NO_2, RSPM, and SPM values, representing diverse seasonal conditions such as winter smog, monsoon variations, and summer dust storms, ensuring sufficient diversity for model generalization.

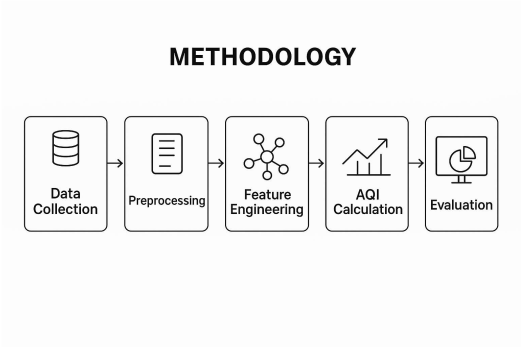
Data Cleaning: Raw environmental data often contain inconsistencies such as duplicate entries, irregular timestamps, or sensor calibration errors. The cleaning stage involved removing duplicates, normalizing timestamps, and correcting obvious anomalies, ensuring consistency for further analysis.

Handling Missing Values: Approximately 8–10% of readings were missing due to sensor downtime or technical issues. To address this, imputation methods such as interpolation, mean substitution, and forward filling were applied, maintaining dataset integrity and preventing imbalance between training and test samples.

Normalization: Since pollutant concentrations vary in scale (e.g., SO_2 and NO_2 in parts per million (PPM), RSPM and SPM in µg/m³), Min-Max normalization was employed to standardize all features into a common range, improving model training efficiency and comparison.

AQI Calculation: AQI values were computed based on Central Pollution Control Board (CPCB) standards, deriving sub-indices for each pollutant and selecting the maximum sub-index as the final AQI. This ensured that results aligned with official air quality categorization and public communication standards.

Feature Selection: Feature importance analysis identified the pollutants most influential in determining AQI. Correlation results showed that NO_2 and RSPM together contributed over 65% of AQI variance. Redundant or less significant attributes were removed to avoid model overfitting and enhance predictive efficiency.



7. Working Process

The working process of the proposed AQI prediction system follows a linear pipeline that converts raw data to decisions. The process starts by acquiring pollutant data from official monitoring stations, with continuous records of 5 years for the prescribed pollutants, SO_2, NO_2, RSPM, and SPM. Once acquired, the dataset is preprocessed, which includes cleaning, normalizing, and ensuring that missing values are accounted for prior to assessment of AQI. Next, following CPCB standards, AQI were calculated, and through correlation and calculations of feature importance, appropriate meaningful features were derived. The data then undergoes a splitting phase where the training and testing sets are defined, and models are trained utilizing multiple machine learning methods. During the training phase, tuning of hyperparameters may take place, and the testing phase includes evaluating model performance on unseen data. After training and testing, results are evaluated using appropriate and specified metrics, and displayed in charts/graphs which allows policymakers, researchers, and the public to interpret the predicted air quality trends easier. This method shows a thoughtful ordered workflow to develop and initiate the use of the AQI prediction system while understanding the accountability it takes to create something worthwhile with quantitative findings.

8. Implementation

Frontend Design: A user-friendly web interface is created to display AQI predictions in real time. The interface features interactive charts, graphs, and color-coded AQI categories, enabling users to quickly interpret air quality results without requiring technical knowledge.

Backend Processing: The backend is developed using Python within the Flask framework, ensuring efficient data flow and execution. It handles data preprocessing, AQI computation, and machine learning model operations seamlessly, resulting in rapid response times.

Dataset Management: Pollutant data records are stored in structured CSV or SQL databases. Python libraries such as Pandas and NumPy are used to clean, organize, and manipulate the data efficiently during training and prediction stages.

Model Integration: The trained machine learning models are serialized using the joblib library and integrated into the backend. This setup enables real-time AQI predictions whenever new pollutant data is received.

Training and Testing: Models are trained on 70% of the dataset, with the remaining 30% used for testing. Hyperparameter tuning ensures optimal performance across varying pollution levels and conditions.

Deployment and User Interface: The system is deployed as a Flask-based web application. Users can input pollutant concentrations and instantly view predicted AQI values along with visualizations such as trend graphs and health category indicators.

9. Tools and Technology

The AQI prediction model utilizes a combination of modern technologies and data science tools to achieve scalability, accuracy, and efficiency. Python serves as the primary programming language due to its strong ecosystem for machine learning and data analysis.

For data preprocessing and visualization, **Pandas** and **NumPy** are used for efficient dataset handling, while **Matplotlib** and

Seaborn generate informative visual outputs. Model development employs **Scikit-learn**, which provides a comprehensive suite of algorithms such as Random Forest, Decision Tree, Logistic Regression, and KNN, along with utilities for tuning and evaluation.

The **Flask** framework powers the backend, allowing smooth integration between trained models and the web application. Datasets are stored as CSV files but can be scaled to SQL databases for larger deployments. The system can operate on a standard workstation with 8GB RAM and a multi-core processor or be scaled to cloud platforms like AWS or Google Cloud for larger implementations.

The combination of these tools creates a robust, modular, and easily deployable system suitable for real-world AQI forecasting applications

10. Challenges

Incomplete and Missing Data: Monitoring stations often experience equipment malfunctions or outages, resulting in data gaps that may reduce the accuracy of predictive outcomes.

Data Quality and Calibration: Errors in sensor calibration or inconsistent data recording can introduce noise, which misleads model interpretation and decreases reliability.

Seasonal and Climatic Variations: India's seasonal diversity — including monsoon, summer dust storms, and winter smog — leads to fluctuating pollutant behavior, demanding adaptable prediction mechanisms.

Computational Limitations: Large-scale datasets and advanced ML algorithms can require high processing power and storage, slowing training or deployment in low-resource environments.

Overfitting and Model Generalization: Some algorithms may perform well on training data but fail on unseen samples. Proper validation and tuning are essential to ensure model generality.

Lack of Interpretability: Although models like Random Forest achieve high accuracy, their internal processes are complex. For effective policy adoption, simplified explanations of predictions are required.

Scalability and Infrastructure: Extending AQI prediction systems to nationwide coverage demands robust cloud frameworks and IoT connectivity, which remain difficult to implement consistently.

11. Future Work

Deep Learning Models:Future research could leverage more advanced deep learning models, such as LSTM and CNN, to capture long-term temporal patterns in order to increase the accuracy of AQI forecasting across different seasons.
Integrating IoT Sensors:Low-cost IoT sensors may have the potential to enhance coverage by being used in rural and semi-urban areas. Real-time IoT data together with ML could provide more accurate predictions in a shorter time frame and more localized AQI predictions.
Satellite and Meteorological Data:Combining meteorological data (e.g., temperature, humidity, wind speed) with satellite imagery may improve spatial forecasting capabilities and recognize possible external environmental factors that may affect AQI.
Smart City Applications:The system could be used or adapted for smart cities. AQI prediction can inform traffic regulations, industrial regulations, and healthcare advisories in order to improve urban living standards.

12. Conclusion

This research has shown that machine-learning frameworks can generate accurate short-term reaches of the Air Quality Index based on multi-year pollutant records. Using, at a minimum, SO_2, NO_2, RSPM and SPM records as input features, we trained and compared several algorithms. In process, we noted that Random Forest consistently strikes the best balance of predictive accuracy and tolerance to imperfect data. The system has applications for both continuous estimation of AQI, as well as use case in establishing health-level categories. Our framework can be utilized by either researcher or practitioner. The ability to identify deteriorating air conditions even before their onset will help provide city managers and policymakers with more planning time to undertake targeted interventions and issue public advisories as serious conditions emerge. Future efforts — for example, by incorporating metering of richer meteorological inputs or increased sensor density — exist to add additional operational value to the framework and advance applied use

13. Reference

[1] Goyal, P., & Chan, A. T. (2009). Air quality modeling and forecasting. *Environmental Monitoring and Assessment*, Springer.

[2] Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using hybrid model. *Atmospheric Pollution Research*.

[3] Breiman, L. (2001). Random Forests. *Machine Learning Journal*, 45(1), 5–32.

[4] Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). Neural network and multiple regression models for PM10 prediction in Athens. *Atmospheric Environment*, 37(8), 977–985.

[5] Wang, Y., Ying, Q., Hu, J., & Zhang, H. (2014). Spatial and temporal variations of six criteria air pollutants in 31 capital cities in China. *Environment International*, 73, 413–422.

[6] Shaban, K. B., Kadri, A., & Rezk, E. (2016). Urban air pollution monitoring system with forecasting models. *IEEE Sensors Journal*, 16(8), 2598–2606.

[7] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting: Part I. History, techniques and current status. *Atmospheric Environment*, 60, 632–655.

[8] Jain, S., & Khare, M. (2010). Urban air quality prediction using artificial neural networks. *Environmental Modelling & Software*, 25(4), 494–502