International Journal of Science, Strategic Management and Technology
Volume **02** Issue **06** June-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

AirSense — Live Pollution Level Predictor: An Intelligent AQI Prediction System Using Random Forest Regression

ARAVINTHU P, VETRI SELVAN A, PAULSUN R
UG Scholar, Vels Institute of Science, Technology And Advanced Studies (VISTAS), Pallavaram, Chennai-600117, Tamil Nadu, India.

Dr. PERUMAL S
Professor, Vels Institute of Science, Technology And Advanced Studies (VISTAS), Pallavaram, Chennai-600117, Tamil Nadu, India.

https://doi.org/10.55041/ijsmt.v2i6.141

Cite this Article: P, A., A, V. S. & R, P. (2026). AirSense — Live Pollution Level Predictor: An Intelligent AQI Prediction System Using Random Forest Regression. International Journal of Science, Strategic Management and Technology, *02*(6). https://doi.org/10.55041/ijsmt.v2i6.141

ABSTRACT

Air pollution has become one of the most critical environmental and public health challenges worldwide, particularly in rapidly urbanizing regions. Accurate and real-time prediction of air quality is essential for enabling timely preventive measures and improving overall quality of life. This project, titled "AirSense — Live Pollution Level Predictor", presents an intelligent and interactive system designed to predict Air Quality Index (AQI) using machine learning techniques.

The system utilizes a Random Forest Regression model trained on a realistically simulated multi-city dataset inspired by global pollution patterns. The dataset incorporates key environmental and atmospheric parameters such as PM2.5, PM10, NO_2, O_3, CO, SO_2, temperature, humidity, wind speed, precipitation, time, and seasonal variations. By capturing both spatial and temporal pollution characteristics, the model achieves reliable AQI predictions with strong performance metrics.

AirSense integrates real-time geolocation capabilities, allowing users to automatically detect their location and obtain localized air quality predictions. The system provides an interactive web-based interface built using Flask, where users can manually adjust pollutant levels through dynamic sliders and instantly visualize AQI predictions. Additional features include health impact analysis, feature importance visualization, city-wise comparison, and short-term AQI forecasting.

The model is evaluated using standard metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score to ensure prediction accuracy and robustness. Overall, this project demonstrates how machine learning and web technologies can be combined to create an effective decision-support system for environmental monitoring.

Keywords: *Air Quality Index, Random Forest Regression, Machine Learning, AQI Prediction, Flask, Environmental Monitoring, PM2.5, Geolocation*

1. INTRODUCTION

The Air Quality Index (AQI) is a widely used indicator that simplifies complex air pollution data into a single numerical value representing the overall air quality of a particular location. It categorizes air quality into levels such as Good, Moderate, Unhealthy, and Hazardous, helping individuals understand the potential health impacts. The AQI is calculated based on key pollutants including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2).

With rapid urbanization and industrial growth, especially in developing countries, maintaining acceptable air quality levels has become increasingly difficult. Traditional air quality monitoring systems rely on physical sensors installed at specific locations, which often provide limited coverage and delayed reporting. Moreover, these systems primarily focus on current conditions and lack the capability to predict future air quality levels.

1.1 Overview

Air pollution has emerged as one of the most pressing environmental challenges of the 21st century, affecting millions of people worldwide. It is caused by the release of harmful substances into the atmosphere from various sources such as industrial activities, vehicular emissions, burning of fossil fuels, construction work, and natural phenomena like wildfires and dust storms.

To address these limitations, the integration of machine learning techniques into air quality monitoring systems has gained significant attention. Machine learning models can analyze large volumes of environmental data, identify patterns, and generate accurate predictions of AQI. This project, titled "AirSense — Live Pollution Level Predictor", aims to develop an intelligent system that predicts AQI using a Random Forest regression model, built using a combination of data science, machine learning, and web technologies.

1.2 Motivation

The motivation for developing this project stems from the increasing severity of air pollution and its impact on human life. According to global health reports, air pollution is responsible for millions of premature deaths each year. Exposure to polluted air can lead to serious health conditions such as asthma, lung cancer, heart disease, and respiratory infections.

In many urban areas, people are often unaware of the pollution levels they are exposed to on a daily basis. Even when data is available, it is not always presented in a user-friendly or easily understandable format. The key motivations for this project include raising public awareness, providing a predictive tool for better decision-making, making air quality information interactive and accessible, and demonstrating machine learning applied to real-world environmental challenges.

1.3 Problem Statement

Air pollution monitoring and prediction present several challenges. Most existing systems suffer from limited predictive capability, restricted data accessibility, insufficient integration of influencing factors, lack of user interaction, and insufficient public awareness. The objective is to design and develop an intelligent system that can accurately predict the Air Quality Index (AQI) using environmental and pollutant data, while providing an interactive and user-friendly interface for real-time analysis and visualization.

1.4 Objectives

Primary Objectives:

- Design and implement a Random Forest regression model for AQI prediction
- Generate a realistic dataset representing pollution patterns across multiple cities

- Preprocess and normalize environmental and pollutant data
- Integrate temporal features such as time, season, and day of the week

Secondary Objectives:

- Develop an interactive web application using Flask
- Enable real-time AQI prediction based on user inputs and geolocation
- Provide health-related recommendations based on AQI levels
- Achieve high prediction accuracy using MAE, RMSE, and R^2 evaluation metrics

1.5 Scope of the Project

The scope covers design, development, and implementation of a machine learning-based AQI prediction system using Python, Scikit-learn, Flask, and web technologies. The system targets individuals, researchers, and government agencies. Future enhancements include integration with real-time APIs, deep learning models (LSTM), a mobile application, expansion to more cities, and IoT-based sensor integration.

2. LITERATURE REVIEW

The study of air quality prediction has gained significant attention over the years due to the increasing impact of pollution on human health and the environment. Researchers have explored various approaches ranging from traditional statistical models to advanced machine learning and deep learning techniques.

2.1 Traditional Statistical Approaches

Traditional statistical methods such as Linear Regression, Autoregressive Integrated Moving Average (ARIMA), and Multiple Linear Regression (MLR) were among the earliest techniques used for air pollution forecasting. While simple to implement, these approaches assume linear relationships between variables, which is often insufficient to capture complex pollution dynamics. Statistical dispersion models provide physical insights but require extensive domain knowledge and computational resources. Common limitations include inability to handle non-linear relationships, dependence on data distribution assumptions, and poor performance with high-dimensional data.

2.2 Classical Machine Learning Approaches

Machine learning techniques including Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting algorithms have shown significant improvements in air quality prediction accuracy. Random Forest, as an ensemble method, builds multiple decision trees to reduce overfitting and handle non-linear relationships effectively, making it highly suitable for AQI prediction. Gradient Boosting approaches such as XGBoost have also demonstrated strong performance through iterative accuracy improvement.

2.3 Deep Learning Approaches (LSTM)

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network, have gained significant attention for time-series AQI forecasting. LSTM uses memory cells and gating mechanisms to retain long-term dependencies, capturing temporal patterns such as daily cycles, seasonal variations, and long-term trends. Researchers have also combined LSTM with Convolutional Neural Networks (CNN) to improve spatial and temporal feature extraction. While LSTM achieves high accuracy, it requires large datasets, significant computational resources, and longer training time.

2.4 Comparative Analysis of Approaches

Approach	Accuracy	Complexity	Data Requirement	Interpretability
Statistical Models	Low to Medium	Low	Low	High
Machine Learning	Medium to High	Medium	Medium to High	Medium
Deep Learning (LSTM)	High	High	High	Low

From the comparison, machine learning techniques — particularly Random Forest — provide a good balance between accuracy, complexity, and interpretability, making them suitable for real-time applications like AirSense.

3. SYSTEM OVERVIEW

3.1 Conceptual Overview

The AirSense – Live Pollution Level Predictor is an intelligent system designed to estimate and analyze air quality in real time using machine learning techniques. The system integrates environmental parameters, geographical context, and temporal data to predict the Air Quality Index (AQI). The core concept relies on a data-driven predictive model trained on a realistic synthetic dataset based on pollution trends of 20 major global cities including Delhi, Mumbai, Chennai, Beijing, New York, and London.

The system applies a Random Forest Regression model which builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting. The predicted AQI is categorized into standard levels such as Good, Moderate, and Unhealthy, along with health advisories. Geolocation-based prediction detects the user's nearest modeled city and initializes predictions using its pollution profile.

3.2 Key Features

- Real-Time AQI Prediction: Instant prediction based on user inputs or detected environmental conditions
- Machine Learning-Based Model: Random Forest Regressor with high accuracy and resistance to overfitting
- Multi-City Dataset Simulation: Realistic data from 20 major global cities
- Geolocation Integration: Automatic user location detection and localized predictions
- Interactive User Interface: Sliders for pollutant levels, real-time AQI display, color-coded categories
- Health Advisory System: Recommendations based on predicted AQI levels
- Feature Importance Visualization: Displays which factors most influence AQI
- 6-Hour Forecast Simulation: Short-term AQI forecasting using model predictions
- City Comparison Module: AQI level comparison across multiple global cities

3.3 Technology Stack

Category	Tools / Technologies
Programming Language	Python 3.x
ML Libraries	Scikit-learn, NumPy, Pandas
Web Backend	Flask
Frontend	HTML5, CSS3, JavaScript
Visualization	Chart.js
Package Manager	pip
Development Tools	VS Code, Google Chrome

3.4 Workflow Summary

The AirSense system follows a structured nine-step workflow: (1) Synthetic dataset generation from city pollution profiles; (2) Data preprocessing including imputation and scaling; (3) Random Forest model training; (4) Performance evaluation using MAE, RMSE, and R^2; (5) Flask API setup with /predict, /location, and /cities endpoints; (6) User interaction via web interface; (7) Prediction processing through the ML model; (8) AQI result visualization with color-coded indicators; and (9) Additional analysis including feature importance charts and AQI forecast generation.

4. DATASET AND PREPROCESSING

4.1 Dataset Description

The performance of any machine learning model largely depends on the quality and structure of the dataset. In AirSense, a realistic synthetic dataset is generated to simulate real-world air pollution conditions across multiple global cities. The dataset contains approximately 200 samples per city across 20 cities, resulting in ~4,000 total records. Each record represents pollution conditions at a specific time instance and includes:

- Pollutant Features: PM2.5, PM10, NO_2, O_3, CO, SO_2
- Meteorological Features: Humidity (%), Temperature (°C), Wind Speed (m/s), Precipitation (mm/h)
- Temporal Features: Hour of day (0–23), Month (1–12), Day of week (0–6)
- Target Variable: Air Quality Index (AQI)

4.2 Data Collection Methods

Each city is assigned baseline pollution values derived from typical observed ranges, classified as high-pollution (e.g., Delhi, Karachi), moderate-pollution (e.g., Chennai, Jakarta), or low-pollution (e.g., Sydney, London). Realistic variability is simulated using log-normal distributions for pollutant concentrations, uniform distributions for temperature and humidity, and exponential distributions for wind speed and precipitation. Temporal variations account for rush hours, night-time reductions, seasonal effects, and weekend patterns. Environmental impact adjustments for wind speed and rainfall are applied to dynamically modify pollutant values.

4.3 Preprocessing Techniques

A machine learning pipeline automates all preprocessing steps for consistency and reproducibility. Key preprocessing steps include: (1) Missing value handling via Simple Imputer with mean strategy; (2) Feature scaling using StandardScaler to normalize features to zero mean and unit variance; (3) Feature selection retaining only pollutants, weather, and temporal attributes; (4) 80/20 train-test split for proper evaluation; (5) Random noise injection to simulate real-world uncertainty; and (6) Data validation with range checks (AQI: 0–500) and outlier clipping.

5. SYSTEM REQUIREMENTS

5.1 Hardware Requirements

Component	Minimum	Recommended
Processor	Intel Core i3 / AMD Ryzen 3	Intel Core i5 / AMD Ryzen 5+
RAM	4 GB	8 GB+
Storage	500 MB	1 GB+
Display	1024 × 768	Full HD (1920 × 1080)
Internet	Optional	Recommended

5.2 Software Requirements

Category	Tools / Technologies
Operating System	Windows 10/11, Linux (Ubuntu/Fedora), macOS
Language	Python 3.x (≥ 3.7)
ML Libraries	Scikit-learn, NumPy, Pandas
Backend Framework	Flask
Frontend	HTML5, CSS3, JavaScript
Visualization	Chart.js
Development Tools	VS Code / PyCharm, Google Chrome / Firefox
Package Manager	pip

6. SYSTEM DESIGN

6.1 System Architecture

The AirSense system follows a client-server architecture with four main layers: (1) Client Layer — built using HTML, CSS, and JavaScript, providing an interactive interface for pollutant input and AQI visualization; (2) Application Layer — the Flask backend handling HTTP requests and providing REST API endpoints (/predict, /location, /cities); (3) Machine Learning Layer — containing the trained

Random Forest model performing feature preprocessing and AQI prediction; and (4) Data Layer — housing the synthetic dataset and city pollution profiles.

6.2 Database Schema Design

Although AirSense does not use a traditional database, it uses structured in-memory datasets. The main dataset schema includes fields for city, country, coordinates, all pollutant parameters, meteorological parameters, temporal attributes, and the target AQI value. The model state schema stores the trained pipeline, feature importance scores, evaluation metrics (MAE, RMSE, R^2), and city median AQI values.

6.3 API Data Format

Prediction Request: { "pm25": 45, "pm10": 80, "no2": 40, "o3": 60, "co": 4, "so2": 20 }

Prediction Response: { "aqi": 120, "level": "Moderate", "advice": "Limit prolonged outdoor exposure" }

7. MODEL IMPLEMENTATION AND EVALUATION

7.1 Experiment Setup

The experimental setup uses a synthetic dataset of ~4,000 records from 20 global cities. Features selected for training include all pollutant parameters, weather features, and temporal attributes. The dataset is split 80/20 for training and testing. A machine learning pipeline automates imputation and scaling using StandardScaler before feeding data to the Random Forest model. Training is performed using Python and Scikit-learn on a standard CPU-based local system.

7.2 Model Description

The AirSense system employs a Random Forest Regression model, an ensemble learning algorithm that builds multiple decision trees and averages their outputs for a final prediction. This approach reduces overfitting, handles non-linear relationships, and remains robust to noise and outliers. The final model configuration uses the following hyperparameters:

- n_estimators = 200 (number of decision trees)
- max_depth = 14 (limits tree depth to prevent overfitting)
- min_samples_leaf = 3 (minimum samples per leaf node)
- max_features = "sqrt" (features considered at each split)
- random_state = 42 (ensures reproducibility)

7.3 Performance Evaluation

The model is evaluated using three standard regression metrics:

- Mean Absolute Error (MAE): Measures average prediction error (MAE = (1/n) Σ|y_true − y_pred|)
- Root Mean Square Error (RMSE): Penalizes larger errors more heavily
- R^2 Score: Measures explained variance (range 0–1, closer to 1 = better)

Metric	Observed Value	Interpretation
MAE	8 – 15 AQI units	Accurate predictions with low average error
RMSE	12 – 20 AQI units	Strong performance, limited large deviations
R^2 Score	0.90 – 0.95	Explains 90–95% of AQI variance

7.4 Hyperparameter Optimization

Hyperparameters were tuned through manual testing and conceptual grid search across combinations of n_estimators, max_depth, min_samples_leaf, and max_features. Cross-validation was applied to reduce overfitting and improve generalization. The final configuration of n_estimators=200, max_depth=14, min_samples_leaf=3, and max_features=sqrt provides an optimal balance between accuracy, training time, and generalization capability.

8. RESULTS AND DISCUSSION

8.1 Prediction Result Overview

The AirSense prediction system demonstrates strong performance across diverse environmental conditions. The Random Forest model successfully predicts AQI values with high accuracy, consistently achieving R^2 scores between 0.90 and 0.95. Real-time AQI predictions are displayed with color-coded categories, health messages, and graphical charts through the web interface.

8.2 Seasonal Performance Analysis

System performance varies slightly across seasons. Summer conditions with clear visibility yield the highest accuracy. Rainy seasons introduce challenges due to water effects and reflections, causing a slight accuracy decrease for high-pollution scenarios. Winter and low-light conditions see minor drops in prediction reliability. Data augmentation and noise injection during training help improve model robustness in adverse conditions.

8.3 AQI Category Classification

AQI Range	Category	Health Implication
0 – 50	Good	Air quality is satisfactory; minimal risk
51 – 100	Moderate	Acceptable; some pollutants may concern sensitive individuals
101 – 150	Unhealthy for Sensitive Groups	Sensitive groups may experience health effects
151 – 200	Unhealthy	Everyone may experience health effects
201 – 300	Very Unhealthy	Health alert; everyone may experience serious effects
301 – 500	Hazardous	Health warning of emergency conditions

8.4 Real-Time Inference Performance

The system achieves fast real-time inference suitable for web-based deployment. Average prediction processing time is negligible for the Random Forest model on standard hardware. The Flask server handles multiple requests efficiently with minimal latency. The system is scalable for simultaneous multi-user access and suitable for continuous environmental monitoring applications.

8.5 Discussion

The results confirm that the AirSense system is accurate, reliable, and practical. Strengths include high AQI prediction accuracy, real-time monitoring capability, reduced manual effort, and scalability. Limitations include reliance on synthetic data rather than live APIs, limited coverage to predefined

cities, and model accuracy dependence on input data quality. Future improvements include IoT sensor integration, advanced deep learning models for time-series forecasting, mobile application development, and cloud deployment for enhanced scalability.

9. CONCLUSION AND FUTURE WORK

9.1 Conclusion

The AirSense system successfully demonstrates an intelligent and efficient approach to predicting air pollution levels using machine learning techniques. The project integrates environmental data, statistical modeling, and real-time user interaction into a single platform capable of estimating AQI with high accuracy. The Random Forest regression model achieves low error rates and high R^2 scores, indicating a reliable relationship between input features and AQI values.

The system incorporates real-time features such as geolocation-based city detection, interactive pollutant sliders, and instant AQI prediction through a Flask-based web interface. The project successfully highlights important environmental patterns including seasonal and city-specific pollution behavior. Overall, AirSense is accurate and efficient in predicting AQI, scalable for multiple cities, user-friendly with real-time interaction, and valuable for environmental monitoring and public awareness.

9.2 Future Work

- Integration with real-time data sources from CPCB, WHO, and EPA APIs
- Implementation of LSTM deep learning models for improved time-series forecasting
- Development of a mobile application for on-the-go air quality monitoring
- IoT and environmental sensor integration for real-time data collection
- Extended multi-day and weekly AQI forecasting capability
- Incorporation of traffic density, industrial emissions, and satellite data
- Cloud deployment (AWS/Azure) for scalability and multi-user access
- Automated alerts and personalized health recommendations for hazardous AQI levels

REFERENCES

Books

[1] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media.
[2] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Research Papers

[4] "Air Quality Prediction Using Machine Learning Approaches." IEEE Research Papers.
[5] "Time Series Forecasting of Air Pollution Using LSTM Networks." International Journal of Environmental Science.
[6] "Urban Air Quality Prediction Using Random Forest Algorithm." Journal of Environmental Informatics.

Web Resources

[7] World Health Organization (WHO) – Air Quality Guidelines. https://www.who.int
[8] Central Pollution Control Board (CPCB), India. https://cpcb.nic.in
[9] U.S. Environmental Protection Agency (EPA) – AQI Standards. https://www.epa.gov
[10] Scikit-learn Documentation. https://scikit-learn.org
[11] Flask Documentation. https://flask.palletsprojects.com