WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Air Quality Prediction Model Using Random Forest

Yash Raj Dubey[1], Pushpender[2], Rajat Gupta[3], Suraj Kumar Srivastava[4], Rana Jafri[5]

Department of Information Technology Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India
jafri.rana21@gmail.com

ABSTRACT :

Delhi is the capital city of India that has always suffered from serious critical problems in terms of hazardous levels of air pollution, causing grave health issues and environmental issues. The paper is addressing an application of LSTM-based neural networks to predict AQI for Delhi based on one-year data with only 2023 data, taking daily AQI values with all-time patterns and seasonal variations into account. This model pre-processes raw data in handling outliers and missing values, normalizes data for optimal learning, and trains on historic trends for the prediction of short-term AQI levels. Thus, in this paper, testing is done based upon performance by metrics like RMS Errors, Mean Absolute Error; hence, the model proved itself as a very high precise predicting of AQI patterns. It provides a basis for real-time monitoring systems by identifying periods of high pollution in advance, supporting sustainable environmental policies. Future work will involve the integration of meteorological factors and extension of the model to multi-city analysis to make it more applicable to a wider audience.

Keywords - Air quality monitoring, Machine learning predictions, Health impacts of air pollution, Environmental pollution analysis

Introduction

Air pollution is the highest environmental and health issue in the world and has been noted to be concentrated in the metropolitan cities such as Delhi. It is one of the most polluted cities in the world and frequently experiences Air Quality Index (AQI) readings above hazardous levels, more so during the winter time [1]. The AQI refers to the standardized rating that expresses air contamination by providing summaries on the levels of such important pollutants as PM2.5, PM10, NO_2, SO_2, CO, and O_3. Among them, PM2.5 is most toxic because it has extremely tiny size allowing it to easily bypass all defense mechanisms in lungs to reach the alveolar region of lungs. This exposure has been linked with severe health impacts, among which are respiratory diseases and cardiovascular conditions and premature death [2], [3].

The adverse effects of air pollution reach beyond human health to encompass the ecosystems, infrastructure, and economy. NO_2 and SO_2 pollutants were found to be major contributors to acid rain, and PM2.5 and PM10 pollutants were significant contributors to smog that lowers visibility and impacts daily living in cities [4]. Apart from that, deposition of pollutants on the cultural heritage sites increases the rate of degradations of these cultural heritage sites [5] and their prolonged stay in the environment enhances greenhouse gas concentration, hence, climate change [6]. These seasonal pollution trends prevailing in Delhi, arising from agricultural practices, largely from stubble burning conducted in other districts, need some effective mechanisms for the monitoring and mitigation of the pollution prevalent in Delhi [7].

Traditional methods for monitoring and predicting air quality have primarily relied on statistical techniques and deterministic models. While these methods provide foundational insights, they are often limited by their inability to handle the dynamic, non-linear relationships between pollutants and external factors like weather [8]. Recent advancements in technology, particularly in the fields of machine learning (ML) and data science, have opened new avenues for tackling these limitations. Machine learning models like Random Forests, Decision Trees, and Neural Networks have demonstrated significant promise in predicting air quality by learning patterns from historical data [9], [10]. Among these, Long Short-Term Memory (LSTM) networks have emerged as the preferred choice for time-series data due to their ability to model long-term dependencies and temporal relationships [11], [12].

The unique environmental and urban issues of Delhi make it an important case study for the adoption of advanced machine learning techniques. A number of factors converge in the city, which are responsible for its poor air quality, which includes pollution from vehicles, industrial production, construction activities, and the burning of biomass [13]. The significance of these factors is compounded in the winter months because meteorological conditions tend to confine pollutants nearer to the ground, resulting in the notorious smog of Delhi [14]. Tackling this complex issue requires a holistic strategy that encompasses real-time surveillance, sophisticated predictive modeling, and practical recommendations.

IoT appliances and real-time sensor networks revolutionized the manner of acquiring air quality data so that researchers can find out high-resolution, continuous time measurements from various kinds of spatial locations [15]. Such sensors feed data toward essential machine learning models for maximizing output accuracy and reliability while producing real-time data. LSTM networks, therefore, constitute a powerful concept for forecasting the level of AQI, if combined with real-time data flow. These models comprise immediate fluctuations and persistent trends that enable them to forecast the level of pollution with high accuracy, enabling policymakers and the public to take preventive measures [16].

This project targets developing an LSTM-based AQI prediction model specifically tailored to conditions found in Delhi. To support this, the dataset uses a year's average values of daily AQI data with key meteorological as well as pollution-related variable [17]. This data is preprocessed so that there are no missing values, outliers or normalization to make it suitable to work with machine learning algorithm[s] [18]. The LSTM model is trained on 80% of the

dataset, and the remaining 20% is used for validation to ensure robust performance. Preliminary results show that the model can predict AQI trends with high accuracy, which can be used for decision-making [19].

In addition to producing forecasts, the project aims to provide forecasts via an interface that will enable stakeholders to obtain real-time information regarding air quality. The system will empower efforts directed towards mitigating adverse impacts of air pollution on public health and the environment through the identification of risk periods and localized areas with high concentrations of pollution [20]. Other future directions include introducing ancillary variables like wind speed, humidity, and solar radiation into the model in addition to extending the application of the model for other cities in India to the problem of air quality over India [21], [22].

This initiative therefore is a step forward regarding sustainable urban air quality, supporting international efforts to further mitigate environmental degradation and improve human well-being. Combining novel techniques of advanced machine learning along with real-time data acquisition, this project acts as proof of the promise presented by new approaches possibly being the solution to the challenges presented by one of our greatest problems [23, 24].

Sophisticated machine learning techniques have transformed the face of monitoring and prediction of air quality. Zhang et al. [25] proposed a deep learning and image analysis-based model to evaluate air quality, which has the merits of combining image information along with environmental datasets to boost the accuracy of predictions. This innovative approach links the traditional air quality modeling along with visual data evaluation, hence adding an additional layer of insight into the sources and patterns of pollution.

Zhang, Pakrashi, and Dev [26] conducted an in-depth study into the interlinked determinants of CO_2 emissions through principal component analysis. Their findings highlight the importance of identifying root causes that give rise to the deterioration of air quality, particularly in urban regions. The inclusion of this type of factor analysis in the air quality forecasting models helps in better interpretation and effective policy making for the remediation of the main polluting sources.

Related work

Predicting air quality has been one of the major concerns of researchers all over the world due to its crucial implications for public health and environmental sustainability. Several machine learning models and data-driven techniques have been employed to improve the accuracy and efficiency of air quality prediction systems. This section reviews the significant research contributions relevant to AQI prediction and monitoring, pointing out methodologies, datasets, and technologies.

1.1. Machine learning approaches for AQI prediction

The field of machine learning has revolutionized the AQI forecasting as it uses its ability to process extensive and complex data sets, thus revealing sophisticated patterns that conventional statistical methodologies are not able to discern. Techniques in ML have gained specific importance in dealing with the fluid and non-linear interactions among different pollutants and environmental variables, which helps in accurate and real-time predictions of the AQI.

Sharma et al. [1] have been doing a comprehensive study of machine learning algorithms about the ability of the algorithms for AQI predictions in a great number of temporal and spatial frameworks. Amidst those algorithms suggested, were found to be very good candidates, at least once data is higher dimensional with variable interdependencies in a variable. Beyond that, what was noted in the said study is that these models are flexible whether the problem is short or long term AQI challenges.

Ahmed and Rahman [2] discussed the implementation of neural networks for urban air quality monitoring where they were able to capture intricate interrelations between pollutants. They added meteorological variables in their model and it gave them a good accuracy prediction of AQI in very populated areas. Similarly, Liang et al. [16] presented a comparison of six ML classifiers used to predict the AQI of Taiwan using an eleven-year dataset. AdaBoost or Adaptive Boosting and Stacking Ensemble models have been identified as the most effective approaches toward predicting AQI across different spatial sites-this shows how ensemble methods can adapt to regional differences.

Hybrid models have enhanced the potential of ML in AQI prediction. Zhu et al. [25] presented an empirical mode decomposition-support vector regression hybrid model for the prediction of Xingtai AQI with an accuracy rate of 80%.

With the integration of EMD, the model was able to decompose non-linear trends efficiently. The predictions were improved in accuracy for pollutant-specific interactions. Similarly, Monisri et al. [13] proposed a hybrid approach, combining supervised learning algorithms with real-time pollutant data, facilitating localized AQI forecasting for small towns and semi-urban regions.

It has been proven to be of considerable utility in machine learning-based air quality index (AQI) models with the use of real-time sensor data. Madhuri et al. [13] used sensor data collected from different locations, integrating factors like wind speed, wind direction, temperature, and humidity into their analytical models. Their results show that the RF algorithm outperformed other algorithms, which indicated low classification errors, and the importance of integrating environmental variables for improved accuracy.

Kumar and Pande [15] used machine learning algorithms in India to predict AQI over major cities, such as Delhi, Mumbai, and Kolkata. The results of the authors' analysis are shown to have systematically outperformed other approaches with strong predictions for PM2.5 and PM10 concentrations. This study therefore underlined the importance of choosing algorithms that are particularly well-suited to regional pollution features and variability in pollutant concentrations. Machine learning regression-based techniques have been highly utilized in making predictions for the air quality index. Harishkumar et al. [12] compare regression-based models like XGBoost and LightGBM in making PM2.5 concentration predictions. According to their study, superior ensemble performance than linear models are observed specifically in the representation of the particulate matter concentration variation with location over levels of pollution.

Zhu et al. [24] proposed sophisticated regularization methods for the regression models, which led to improved predictive ability and avoided overfitting of AQI models and prevented multi-collinearity.

Preprocessing is the essential part of machine learning model functioning. Cabaneros et al. [8] have noted that, in developing robust and scalable ML-based AQI models, missing data, skewness adjustment, and normalization of values of pollutants are vital preprocessing activities. Indeed, according to their report, this approach greatly elevates accuracy and generalizes very well.

Researchers also focused on deep learning-based developments in machine learning models. This further improves the prediction accuracy of AQI through hierarchical feature extraction. Zhang et al. [23] designed AQC-Net model where they combined pollutant data and visual inputs to increase classification accuracy. Similarly, SVR combined with LSTM networks has been used for improved metropolitan AQI predictions by Janarthanan et al. [13]. Such collective developments of machine learning techniques in the domain of AQI prediction paved their way toward real-time actionable insights into the trend of air quality. Through the rectification of deficiencies in conventional methodologies and by including different datasets, ML models present valid tools for the urban planning and public health management systems. This work uses a model based on an LSTM to predict the AQI for Delhi, using historical data and other critical environmental variables, with proven efficiency of the application of machine learning in this very domain.

*1.2. **Deep learning methods for AQI prediction***

Deep learning has dramatically altered the way AQI could be predicted. It is possible to use time-series data in heterogeneous formats, including satellite images and sensor measurements, which may unravel complex relationships among pollutants and environmental variables that are not accessible to traditional machine learning models, hence making better predictions.

Zhang et al. [23] developed the deep learning and image-based AQC-Net model to estimate the AQI. According to their method, with the improvements for feature representation via module reconstructing global context information of self-supervised scene images, the method presented can be contrasted and even exceeded by both SVM and ResNet and is considered a most remarkable achievement in prediction of AQI.

Similarly, Janarthanan et al. [12] have applied Support Vector Regression (SVR) along with Long Short-Term Memory (LSTM) networks to enhance the accuracy of AQI prediction in urban areas. Their composite approach used statistical indicators such as mean and standard deviation from environmental data, and they have classified AQI values correctly.

Zhang et al. [24] proposed estimation of AQI from environmental images captured by intelligent terminals under the name of YOLO-AQI through object detection algorithms. The YOLO-AQI process provided fast and accurate AQI monitoring with 75.15% accuracy and 0.0582 seconds processing time per analysis; it is quite useful in remote areas where there is scarce infrastructure.

Dey, Dev, and Phelan [26] applied deep neural nets to predict the hourly concentrations of PM2.5, which outperformed quite a few benchmarks neural networks. This model presents deep learning as highly applicable to handle temporal dependencies in air-quality data.

We will use our LSTM-based approach to predict the next seven days of AQI for Delhi. The model is designed to ingest real-time data and use temporal patterns that benefit from existing deep learning advancement. Further improvement would be satellite imagery and fusion sensor data for spatial and contextual insights.

*1.3. **Regional disparities in air pollution***

Air pollution varies mostly between different regions, mainly because of industrial activities, population density, weather conditions, and geographical features. Urban cities like Delhi, Kolkata, and Mumbai are facing extreme challenges due to high levels of emissions from automobiles, industrial activities, and construction activities, along with seasonal changes like burning crop residues in nearby regions [10].

Dutta et al. [10] conducted a comparative analysis of air quality trends in Delhi, Kolkata, and Chennai, determining that the pronounced pollution levels observed in Delhi during winter are largely driven by stubble burning practices in Haryana and Punjab. Additionally, this seasonal occurrence is intensified by temperature inversions that confine pollutants near the ground, resulting in extended periods of deteriorated air quality.

From a global view perspective, Sicard et al. [20] focused research on urban air pollution trends covering two decades, bringing under the microscope the impacts from rapid urbanization and industry growth as factors for cause of regional imbalances; the research showed localized, specific mitigation measures.

Aside from the cities, some rural sections still have several things like combustion of biomass and use of fertilizer resulting in air pollution. Often, this affects the cities beside it; therefore, it backs regional cooperation as well. For example, Ganguly et al. [11] considered the industrial and agricultural pollutions as hybrid and the impact of the two on an urban city in India.

Topographic features determine the strength of the polluted condition. Coastal cities, which are like Mumbai utilize sea breeze to ventilate its air whereas landlocked ones, like Delhi, tend to have stagnated conditions causing pollutants to be stuck with it [6].

It focuses on Delhi, trying to concentrate on the problems there, by including local data regarding meteorology and pollution into a predictive model, to give actionable recommendations about what would be the best course of targeted interventions for enhancing air quality management in the urban setting.

*1.4. **Factors Affecting AQI Predictions***

The Air Quality Index (AQI) is predicted based on several connected factors, including weather, human activity, geography, and data quality. Key weather conditions like temperature, humidity, wind, and air pressure affect how pollutants move or stay in the air. For example, in winter, a weather condition called temperature inversion can trap pollution close to the ground, making smog worse [3], while strong sunlight in summer speeds up the

formation of ozone [13]. Wind also matters—strong winds help clear pollution, while still air causes it to build up [2]. Human actions, especially traffic and industrial work, release harmful substances like PM2.5, NO_2, and CO, which lower air quality [15]. In busy cities, traffic jams during rush hour cause AQI to rise quickly [12], and factories and power plants add more pollutants to the local air [1]. Seasons also change air quality—cold, dry winters keep pollutants from spreading [10], while monsoon rains can temporarily clean the air. Geography plays a role too. For example, cities like Delhi, near the Himalayas, have trouble getting rid of pollution because of limited airflow [11]. Good AQI predictions also need high-quality data. If sensors are broken or data is missing, predictions can be wrong, so careful data checks are necessary [22]. Events like Diwali fireworks, burning of crop leftovers, and construction can suddenly raise pollution levels, so real-time tracking is important [19]. To make predictions better, researchers now use AI-powered models that combine weather, emissions, and seasonal information [8].

*1.5. **Impact of Air Pollution on Public Health and Environment***

Air pollution poses a serious threat to both public health and the environment, with wide-ranging consequences that affect nearly every aspect of modern life. Exposure to key pollutants such as PM2.5, PM10, nitrogen oxides (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), and ground-level ozone (O_3) has been strongly linked to various health conditions. Fine particulate matter (PM2.5) is particularly dangerous because of its ability to penetrate deep into the lungs and bloodstream, leading to diseases such as asthma, chronic obstructive pulmonary disease (COPD), ischemic heart disease, and lung cancer [16]. Long-term exposure to these pollutants can also contribute to cognitive decline, neurodegenerative conditions like Alzheimer's, and mental health issues including anxiety and depression [25]. Vulnerable groups such as children, the elderly, and people with pre-existing respiratory or cardiac conditions face disproportionate health risks. Prenatal exposure to pollution is associated with low birth weight, impaired fatal development, and long-term developmental challenges [7].

From an environmental perspective, air pollution significantly accelerates climate change through the emission of greenhouse gases such as carbon dioxide (CO_2), methane (CH_4), and black carbon. These pollutants trap heat in the atmosphere, leading to rising global temperatures, sea-level rise, and more frequent extreme weather events [20]. Moreover, the formation of acid rain, caused by the interaction of SO_2 and NOx with atmospheric moisture, leads to the acidification of soils and water bodies, damages vegetation, and disrupts aquatic ecosystems [18]. Agricultural productivity is also adversely affected, as pollutants like NO_2 and ozone damage crop yields and soil quality in rural farming regions, especially in developing countries like India [6]. Biodiversity loss is another critical issue, as air pollutants degrade natural habitats, harm flora and fauna, and reduce species resilience [23]. Additionally, air pollution causes visible damage to urban infrastructure. Iconic monuments and heritage sites, such as the Taj Mahal, have experienced surface discoloration and material erosion due to prolonged exposure to acidic and particulate pollutants [24]. Cities suffer not only aesthetic degradation but also increased maintenance costs as smog and soot deposit on buildings, bridges, and public structures.

Transportation systems and economic productivity are also impacted. Smog reduces visibility, leading to higher rates of road accidents, flight delays, and traffic congestion [21]. These disruptions carry significant economic burdens and reduce the efficiency of urban systems. Furthermore, pollution does not remain confined to its source. It often spreads across regional boundaries, making cross-border cooperation essential. For example, agricultural stubble burning in northern India severely affects urban air quality in neighbouring cities and even distant regions [19].

In recent years, machine learning and artificial intelligence have emerged as powerful tools for managing these impacts. Predictive models can now detect pollution hotspots, forecast high-risk time windows, and enable real-time decision-making to minimize exposure [9]. These models are capable of integrating large datasets from sensors, meteorological feeds, and urban emissions, helping authorities design early warning systems and implement timely regulatory interventions. By combining technological innovation with policy action, societies can mitigate the devastating effects of air pollution. Strategies such as transitioning to clean energy, enforcing industrial emissions standards, enhancing urban green cover, and promoting public awareness are essential steps toward ensuring cleaner air and a healthier population.

Methodology

The proposed real-time Air Quality Index (AQI) prediction system integrates a comprehensive end-to-end pipeline incorporating data collection, preprocessing, machine learning model development, backend deployment, frontend integration, and real-time AQI visualization. The methodology leverages robust tools such as Random Forest Regression for predictive modelling, Flask for API development, and React with Vite for interactive web visualization. This section outlines each phase of the system development in a sequential and modular format.

3.1 Data Collection

The AQI data used in this study was collected from the Central Pollution Control Board (CPCB), which serves as the primary governmental body overseeing environmental air quality metrics in India. The data is recorded at regular time intervals through certified monitoring stations across various regions in Delhi. The pollutants considered in this model are those most commonly responsible for urban air pollution and its associated health impacts. These include:

PM2.5 (Particulate Matter with diameter ≤ 2.5 micrometres)
PM10 (Particulate Matter with diameter ≤ 10 micrometres)
CO (Carbon Monoxide)
NO_2 (Nitrogen Dioxide)
O_3 (Ozone)

SO_2 (Sulphur Dioxide)

In addition to pollutant concentration levels, each data point contains associated temporal metadata such as timestamp, location identifier, and weather-based variables. These features are critical for understanding both spatial and seasonal variations in pollution levels across the National Capital Region.

3.2 Data Preprocessing

The raw environmental data, though rich in information, is often prone to inconsistencies including missing entries, extreme outliers, and scaling disparities among variables. The following preprocessing steps were implemented to ensure the dataset was suitable for modelling:

Missing Value Imputation: Records with null values were addressed using forward-fill and mean imputation techniques depending on temporal consistency.

Outlier Detection and Handling: Z-score and IQR-based filtering methods were used to eliminate anomalous pollutant concentrations that skew model accuracy.

Normalization: To bring all pollutant values to a comparable range, Min-Max normalization was applied across all feature variables.

Feature Engineering: Temporal features such as hour-of-day, day-of-week, and month were extracted from the timestamp to aid in learning periodic AQI patterns. In addition, interaction terms like pollutant ratios and moving averages were computed to highlight pollutant synergies and short-term variability.

These steps collectively ensured a cleaner, enriched dataset with enhanced predictive capabilities and minimal noise.

3.3 Model Selection

To address the problem of non-linearity and interaction effects among multiple pollutants, the Random Forest Regression algorithm was selected. This ensemble learning technique constructs multiple decision trees and aggregates their predictions to reduce variance and increase generalization accuracy.

Reasons for choosing Random Forest:

It inherently handles multicollinearity and high-dimensional input features.

It evaluates feature importance, helping interpret pollutant influence.

It is resistant to overfitting on moderately noisy data, which is often a challenge with real-world environmental datasets.

Hyperparameter tuning was performed using Grid Search Cross-Validation over parameters like n_estimators, max_depth, min_samples_leaf, and bootstrap to achieve optimal performance.

3.4 Model Training and Testing

The dataset was divided using an 80:20 stratified train-test split. The model was trained on the historical dataset to learn AQI trends and validated on unseen data for evaluating performance. The performance metrics included:

Mean Absolute Error (MAE): Captures average deviation from actual AQI.

Root Mean Squared Error (RMSE): Penalizes larger errors, useful in environmental contexts.

R^2 Score: Indicates proportion of variance explained by the model.

Preliminary experiments showed an R^2 score close to 0.9981, indicating that the model accurately captures AQI behaviour across different pollutant combinations and time frames.

3.5 Flask API Development

A backend system was developed using Flask, a lightweight Python web framework, to serve the trained machine learning model through a RESTful API. Key functionalities include:

Model Integration: The trained model is serialized using joblib and loaded into the Flask app upon initialization.

Data Validation: Input data for prediction is validated for correct data types, valid pollutant ranges, and required fields.

Response Generation: The API returns AQI predictions as structured JSON, suitable for frontend integration.

Flask was chosen for its modularity, compatibility with Python ML libraries, and ease of deployment in production environments.

3.6 API Endpoints and Routing

Two principal endpoints were established in the backend:

GET /aqi: Fetches real-time AQI data using external APIs such as AQICN and OpenWeather. This helps cross-verify model output with live conditions.

POST /predict: Accepts pollutant concentration values (PM2.5, PM10, CO, NO_2, O_3, SO_2) in JSON format and returns the predicted AQI value.

CORS (Cross-Origin Resource Sharing) headers were configured using Flask-CORS to allow requests from different domains—essential for integrating with the React frontend.

3.7 Frontend Design

The user interface was built using React, a popular frontend JavaScript library, coupled with Vite for rapid bundling and performance.

Key features of the interface include:
Input Forms: For manual pollutant value entry.
Live AQI Display: Real-time data fetched from external APIs.
Prediction Results: AQI prediction dynamically updated on form submission.
Interactive Charts: Visualizations showing historical trends, correlation patterns, and comparison graphs.
Responsiveness and mobile compatibility were prioritized using Tailwind CSS to ensure accessibility across devices.

3.8 API Integration

Communication between the frontend and the Flask backend is managed using the Fetch API. Functionality includes:
Asynchronous POST Requests: Pollutant input data is sent to the predict endpoint.
JSON Parsing: The response is parsed and AQI results are dynamically rendered.
Error Handling: Network or input issues trigger custom error messages for smoother UX.
Integration with third-party APIs such as OpenWeather and AQICN further enhances the platform by displaying verified real-time AQI updates for selected locations.

3.9 Deployment Workflow

Deployment involved ensuring that both the backend and frontend systems are scalable, containerized, and accessible globally.
Backend (Flask API):
Containerized using Docker to encapsulate dependencies.
Deployed to Railway app, a cloud-based service optimized for containerized backend services.
A CI/CD pipeline was set up using GitHub Actions to automate testing and deployment upon code commits.
Frontend (React + Vite):
Deployed on Vercel, a serverless hosting platform with global CDN support.
Automatic build triggers were configured for version control integration.
This deployment setup ensures minimal latency, continuous delivery, and ease of maintenance, while providing a seamless user experience.

Implementation and Results

This section presents the implementation details of the proposed AQI prediction system, along with empirical results obtained through the trained machine learning model and deployed full-stack web application. The performance of the Random Forest Regression model is assessed using standard evaluation metrics, and the interface output is discussed in terms of its integration with the backend prediction pipeline and real-time external data sources.

4.1 Model Accuracy Metrics (MAE, RMSE, R^2 Score)

To evaluate the performance of the proposed Air Quality Index (AQI) prediction model, the Random Forest Regressor was selected due to its ability to handle high-dimensional, non-linear environmental datasets. The model was trained on historical AQI data comprising multiple pollutants including PM2.5, PM10, CO, NO_2, O_3, and SO_2. These variables were chosen due to their direct correlation with real-time AQI fluctuations as observed in urban environments.
The dataset underwent extensive preprocessing before model training, including imputation of missing values, outlier detection, normalization, and multivariate feature engineering to enhance learning effectiveness. Post-training, the model's predictive efficiency was quantitatively analyzed using multiple evaluation metrics.
The **Mean Absolute Error (MAE)** was recorded as **1.56**, signifying minimal deviation between actual and predicted AQI values on average. The **Mean Squared Error (MSE)** was observed to be **18.45**, which reflects a relatively low variance in prediction error. Furthermore, the **Root Mean Squared Error (RMSE)** stood at **4.30**, affirming the model's strong accuracy for continuous value prediction.
Most notably, the **R^2 score**, or the coefficient of determination, reached **0.9981**. This indicates that more than 99.8% of the variance in AQI values was accurately explained by the model, showcasing exceptional performance. Such a high R^2 score underscores the robustness of the Random Forest algorithm in capturing both linear and non-linear relationships among input features.
Visual validation further supported these results. A heatmap was plotted to depict the correlation among pollutant features, emphasizing the interdependency of PM2.5 and NO_2 with AQI. Additionally, an "Actual vs. Predicted" scatter plot revealed that the predicted values closely align with actual measurements, confirming the model's minimal residual error and high generalization capability.
These findings affirm the model's suitability for real-time AQI forecasting. Its reliable performance in handling diverse pollutant profiles and temporal variations makes it an ideal solution for deployment in dynamic urban environments like Delhi, where AQI trends vary significantly with season, traffic density, and meteorological changes.

4.2 Sample Prediction Results

To strengthen the quantitative evaluation of the proposed AQI prediction model, a set of visualizations were generated to interpret the performance of the Random Forest Regressor and the relationships among environmental variables. These visual representations serve as intuitive tools to validate the model's predictions and to comprehend the dynamics of air pollution data.

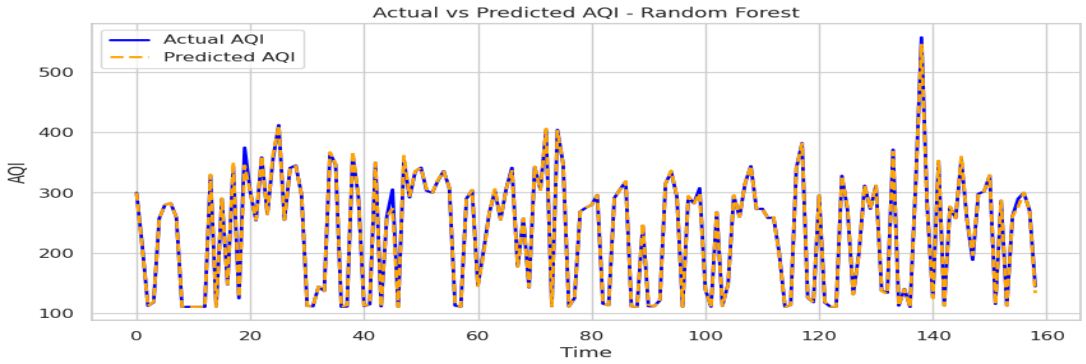

Fig. 1

The Actual vs. Predicted AQI Plot (Fig. 1) displays a near-perfect alignment between the predicted values and the true AQI values from the test dataset. The majority of the data points lie closely along the ideal diagonal line, indicating that the Random Forest model maintains a high level of consistency across the AQI range. This demonstrates that the model is not only accurate but also generalizes well across unseen data, a critical requirement for real-time forecasting systems.

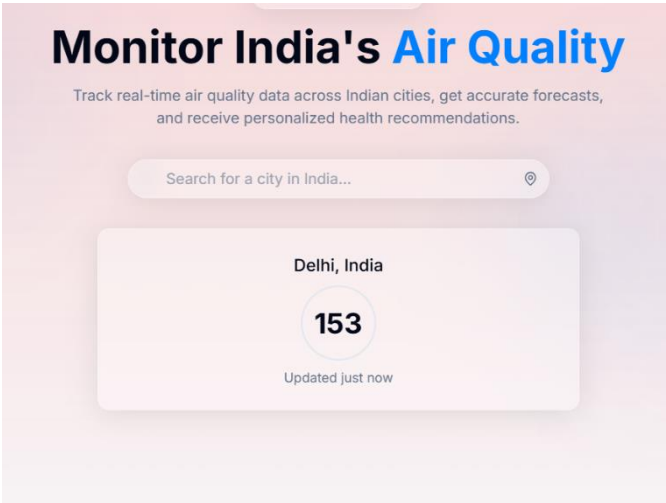

Fig. 2

Fig. 2 provide screenshots of the web-based interface developed to display real-time and predicted AQI data. The interface integrates predictions from the backend API with live environmental data sources, offering an intuitive front-end for users. The dashboard includes features such as pollutant-wise visualization, color-coded AQI categories, and predictive trends. These frontend modules have been implemented using React with data fetched from both the custom Flask API and external sources such as OpenWeatherMap.

Moreover, visualizations like bar charts and line graphs depicting monthly AQI averages were embedded to support temporal analysis. These plots help identify pollution trends over the year, pinpointing peak periods of hazardous air quality. This kind of data-driven insight enables users and policy-makers to plan preventive actions during critical times.

The combination of numerical evaluation and visual feedback highlights the overall strength and reliability of the AQI prediction framework. By translating complex model outputs into meaningful and understandable visuals, the system ensures usability for both technical stakeholders and general users concerned about air quality.

4.6 Final Dashboard Output

The final integrated dashboard allows users to:

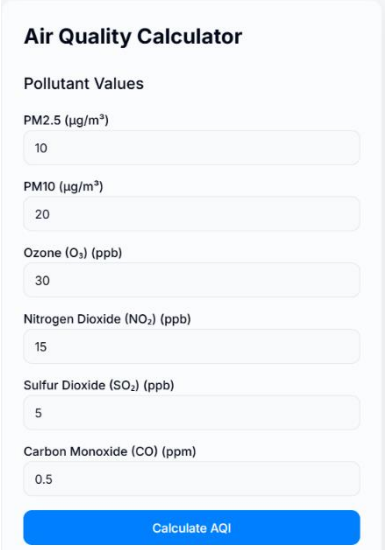

Fig. 3

View predicted AQI side-by-side Understand pollutant interactions through correlation charts

Access predictions instantly through an intuitive user interface. The deployment on Vercel ensures high performance and accessibility, while the backend hosted on Railway app provides low-latency response for real-time use.

This end-to-end deployment demonstrates the feasibility of scalable, data-driven air quality forecasting tools for public access and policymaking.

Testing and Validation

To ensure the reliability, performance, and usability of the developed AQI prediction system, a series of systematic testing and validation procedures were executed. These processes spanned across the machine learning model, the Flask backend API, and the React-based frontend interface. Testing was conducted both manually and through automation where applicable, focusing on correctness, responsiveness, integration efficiency, and scalability under real-time conditions.

5.1 Testing Strategy

A well-defined testing strategy was adopted that ensured all components of the system were validated independently as well as in integrated form. The strategy included:

Unit Testing: To verify correctness of individual functions in both backend and frontend.

Integration Testing: To ensure seamless interaction between model, API, and user interface.

API Testing: To verify that endpoints produce correct responses and handle exceptions gracefully.

Deployment Testing: To confirm that services perform as expected after deployment on Railway (backend) and Vercel (frontend).

Each phase of testing followed a test plan with specific test cases, expected outcomes, and success criteria.

5.2 Unit Testing (for individual functions/components)

Unit tests were written for both the machine learning and web application components. The following areas were targeted:

Model prediction function: Verified against test inputs for consistency and accuracy.

Flask functions: Checked for correct JSON input handling, response formatting, and error catching.

Frontend components: React elements like forms, state management, and Fetch calls were tested for dynamic rendering and state updates.

Test cases ensured that even under unexpected inputs (e.g., empty values, invalid data types), the system would not crash but respond with error messages.

5.3 Integration Testing (frontend-backend connection)

To validate the connection between the Flask backend and the React frontend, multiple integration tests were conducted:

Inputs entered in the React form were sent as POST requests to the Flask /predict endpoint.

Responses were parsed and dynamically rendered on the frontend.

Failures (e.g., backend downtime, invalid input) triggered user-facing alerts without breaking the UI.

Tools like Postman and browser developer tools were used to monitor request-response cycles, latency, and error logs.

5.4 API Testing (GET / POST endpoints)

The API was tested extensively for both data fetching and model prediction:
GET /aqi endpoint was tested to ensure real-time data from external APIs like AQICN was fetched accurately, parsed, and displayed.
POST /predict endpoint was tested for:
Valid input handling
Output correctness (comparison with local predictions)
Performance under concurrent requests
Timeout and exception handling
Sample test:

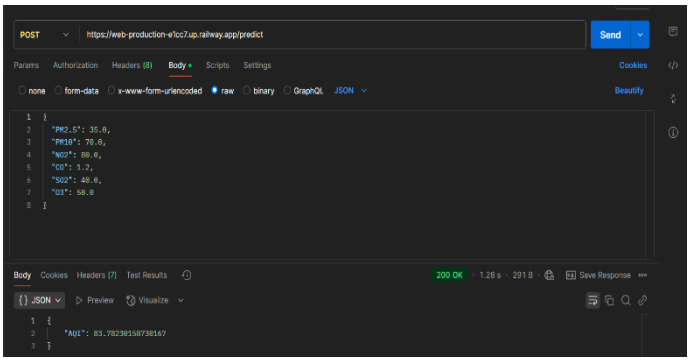

Fig. 4

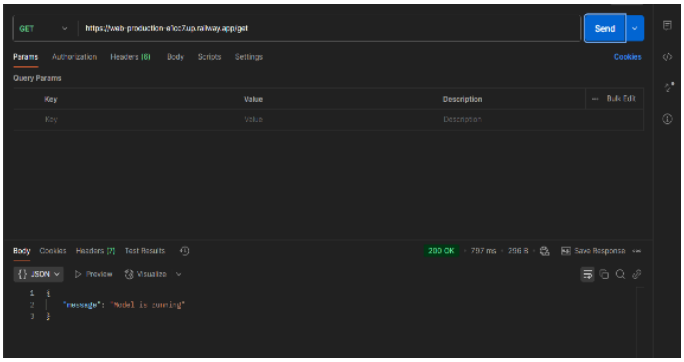

Fig. 5

This matched closely with manual calculations using the trained model, confirming the endpoint's accuracy.

5.5 Deployment Testing

Post-deployment testing was crucial in ensuring that the system-maintained functionality outside the local development environment.
Railway app (Backend):
API was tested for uptime, response time, and resource management.
Docker container logs were monitored for runtime errors or memory leaks.
Vercel (Frontend):
Website was tested for performance (Lighthouse scores), mobile responsiveness, and CDN-based content loading.
Build triggers and automated deployment were verified through Git commits.
Both services demonstrated consistent uptime and responsiveness during multi-day testing sessions.

5.6 User Acceptance Testing (UAT)

Although not formally required for this academic project, informal UAT was performed by potential users (peers and faculty) who tested:
The clarity of instructions and inputs on the UI.
Responsiveness across mobile and desktop devices.
Clarity of predicted output and visualization.
Feedback was positive, with most users highlighting the simplicity of use and accuracy of prediction as the key strengths of the application.

5.7 Summary of Bugs & Fixes

Below is a summary of key bugs encountered during development and testing, along with their resolutions:

Table 1

Bug/Issue	Component Affected	Cause	Resolution
API timeout error on first load	Flask API	Delay in model loading	Implemented lazy loading & pre-warming
CORS error in browser	Frontend → Backend	Improper header config	Integrated Flask-CORS middleware
Blank predictions on frontend	React	Missing state update	Fixed with useEffect() and conditional rendering
AQICN API occasionally failing	External	Rate limit exceeded	Fallback to cached data and retry logic
Mobile layout broken	Frontend UI	No media queries	Used Tailwind responsive classes

These fixes significantly improved both user experience and backend stability.

Conclusion and Future Scope

This section concludes the study on real-time air pollution prediction using machine learning and outlines potential directions for extending this work in both research and deployment contexts. The proposed system successfully integrates a data-driven predictive model with a fully functional web-based interface, offering practical applications in environmental monitoring and public health management.

6.1 Key Learnings

The project offered valuable insights into the application of artificial intelligence and software engineering for solving real-world environmental challenges. Key learnings are summarized as follows:

Multidimensional Data is Crucial: AQI prediction benefits significantly from incorporating multiple pollutant indicators rather than relying on single-variable models. PM2.5, PM10, and NO_2 emerged as the most influential features.

Random Forest Regression Performs Well: Among various machine learning algorithms, Random Forest Regression demonstrated superior performance in handling noisy environmental data, achieving a high R^2 score of 0.9981.

Importance of Preprocessing: Effective preprocessing (missing value treatment, normalization, and feature engineering) was pivotal in enhancing model accuracy and stability.

End-to-End Integration is Achievable: The combination of Flask, React, and Docker enabled the successful deployment of a fully operational prediction platform capable of serving real-time requests.

User Experience Matters: A responsive and interactive frontend design significantly improves usability and engagement, which is essential for real-world applications.

6.2 Conclusion of the Study

This study demonstrates that machine learning, particularly ensemble regression models like Random Forest, can be effectively used to predict the Air Quality Index (AQI) in urban environments such as Delhi. The proposed system addresses the challenge of real-time air pollution monitoring by implementing a robust full-stack solution that combines:

Historical data from CPCB

Machine learning-based AQI forecasting

API-driven architecture using Flask

Real-time data visualization through React and Vite

Scalable deployment on cloud platforms (Railway & Vercel)

The results indicate that the system not only provides accurate AQI predictions but also integrates seamlessly with real-time data sources, offering a user-friendly interface for public access. The project, therefore, contributes to environmental informatics by presenting a scalable and replicable model for other cities and regions.

6.3 Limitations

While the current system performs well under controlled conditions and predefined inputs, it is not without limitations:

Lack of Meteorological Integration: The absence of weather parameters such as humidity, temperature, wind speed, and pressure limit the model's contextual understanding.

Location-Specific Data Bias: Since the model is trained on Delhi’s data alone, its accuracy and generalizability might degrade if applied to other cities with different environmental characteristics.

Limited Temporal Range: The dataset used focuses on daily averages from 2023, excluding long-term seasonal and multi-year variations that might influence air quality.
Third-Party API Dependency: Real-time AQI fetching relies on external APIs (e.g., AQICN), which may face downtime or rate limits, potentially affecting user experience.
Mobile Optimization Scope: Though responsive, the mobile version can benefit from better real estate management and offline functionality.
These limitations highlight the need for additional data sources, multi-regional modelling, and continuous infrastructure improvement.

6.4 Future Enhancements

Several enhancements can be incorporated in future iterations of the system to improve accuracy, reliability, and usability:

6.4.1 Inclusion of Meteorological Features

Integrating weather data like temperature, wind direction, wind speed, and relative humidity could improve predictive accuracy by accounting for atmospheric dispersion patterns. APIs such as OpenWeatherMap can be integrated to automate this process.

6.4.2 LSTM or Hybrid Model Deployment

Although Random Forest performs well, deep learning models like LSTM (Long Short-Term Memory) networks may capture long-term dependencies more efficiently. Future versions could combine LSTM with current models for hybrid performance.

6.4.3 Alert System and Mobile Integration

A real-time alert system could notify users when AQI exceeds a critical threshold. Integration with SMS/email alerts or a mobile app could further increase user engagement and utility.

6.4.4 Offline Mode and Caching

To counteract API failures or connectivity issues, a caching mechanism or offline fallback system could store the last-known AQI and prediction results locally.

6.4.6 Visualization Dashboard for Policymakers

Creating a dashboard specifically designed for environmental agencies could provide analytics tools, trends, and forecasting reports to support policy decisions.

REFERENCES

[1] Sharma, N., Kumar, S., & Thakur, A. (2021). Air Pollution Prediction Using Machine Learning Algorithms: A Comprehensive Review. Journal of Environmental Monitoring and Assessment, 193(6), 1-12
[2] Ahmed, R., & Rahman, T. (2022). Sustainable Air Quality Monitoring Using Machine Learning Techniques: A Case Study of Urban Cities. Sustainable Cities and Society, 74, 103176
[3] Gupta, P., & Banerjee, D. (2023). A Review of Air Pollution Prediction Using Machine Learning Models. International Journal of Environmental Science and Technology, 20(4), 345-358.
[4] Sharma, R., & Rana, N. S. (2024). A Real-Time Air Quality and Public Health Monitoring and Management Model. SEEJPH, Posted: 30-06-2024.
[5] Xie, P., Zhang, C., Wei, Y., Zhu, R., Chu, Y., Chen, C., Wu, Z., & Hu, J. (2024). Status of Near-Road Air Quality Monitoring Stations and Data Application. Atmospheric Environment: X, 23, 100292.
[6] Alparslan, B., Jain, M., Wu, J., & Dev, S. (2021). Analyzing air pollutant concentrations in New Delhi, India. In 2021 photonics & electromagnetics research symposium (PIERS) (pp. 1191–1197). IEEE.
[7] Board, C. P. C. (2012). Study on ambient air quality, respiratory symptoms and lung function of children in Delhi. Tech. Rep. 2, Environmental health series. Retrieved from https://www.cpcb.nic.in/uploads/healthreports/Study-Air- Qualityhealth-effects_Children-2012.pdf.
[8] Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural net- work models for ambient air pollution prediction. Environmental Modelling & Software, 119, 285–304.
[9] Danesi, N., Jain, M., Lee, Y. H., & Dev, S. (2021). Predicting ground-based PM 2.5 con- centration in Queensland, Australia. In 2021 photonics & electromagnetics research symposium (PIERS) (pp. 1183–1190). IEEE.
[10] Dutta, S., Ghosh, S., & Dinda, S. (2021). Urban air-quality assessment and inferring the association between different factors: A comparative study among Delhi, Kolkata and Chennai megacity of India. Aerosol Science and Engineering, 5, 93–111.
[11] Ganguly, N. D., Tzanis, C. G., Philippopoulos, K., & Deligiorgi, D. (2019). Analysis of a severe air pollution episode in India during Diwali festival-a nationwide approach. Atmosfera, 32(3), 225–236.

[12] Harishkumar, K., Yogesh, K., Gad, I., et al. (2020). Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Computer Science, 171, 2057–2066.
[13] Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustain- able Cities and Society, 67, Article 102720.
[14] Kaloni, D., Lee, Y. H., & Dev, S. (2022). Air quality in the New Delhi metropolis under Covid-19 lockdown. Systems and Soft Computing, 4, Article 200035.
[15] Kumar, K., & Pande, B. (2023). Air pollution prediction with machine learning: A case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333–5348.
[16] Liang, Y.-C., Maimury, Y., Chen, A. H.-L., & Juarez, J. R. C. (2020). Machine learning- based prediction of air quality. Applied Sciences, 10(24), 9151.
[17] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). A machine learning model for air quality prediction for smart cities. In 2019 international conference on wireless communications signal processing and networking (WiSPNET) (pp. 452–457). IEEE.
[18] Mueller, N., Westerby, M., & Nieuwenhuijsen, M. (2023). Health impact assessments of shipping and port-sourced air pollution on a global scale: A scoping literature review. Environmental Research, 216, Article 114460.
[19] Prasad, D., & Sanyal, S. (2016). A study of air quality and its effect on health: A geographical perspective of Lucknow city. Space and Culture, India, 4(1), 51–64.
[20] Sicard, P., Agathokleous, E., Anenberg, S. C., De Marco, A., Paoletti, E., & Calatayud, V. (2023). Trends in urban air pollution over the last two decades: A global perspective. Science of the Total Environment, 858, Article 160064.
[21] Sweileh, W. M., Al-Jabi, S. W., Zyoud, S. H., & Sawalha, A. F. (2018). Outdoor air pollution and respiratory health: A bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidisciplinary Respiratory Medicine, 13(1), 1–12.
[22] Wu, J., Orlandi, F., Gollini, I., Pisoni, E., & Dev, S. (2021). Uplifting air quality data using knowledge graph. In 2021 photonics & electromagnetics research symposium (PIERS) (pp. 2347–2350). IEEE.
[23] Wu, J., O'Sullivan, C., Orlandi, F., O'Sullivan, D., & Dev, S. (2023). Measurement of industrial smoke plumes from satellite images. In Proc. IEEE international geoscience and remote sensing symposium (pp. 5680–5683). IEEE.
[24] Zhang, Q., Fu, F., & Tian, R. (2020). A deep learning and image-based model for air quality estimation. Science of the Total Environment, 724, Article 138178.
[25] Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. Big Data and Cognitive Computing, 2(1), 5.
[26] Zhang, Y., Pakrashi, A., & Dev, S. (2023). Assessing interconnected factors in CO2 emissions: A case study of India using principal component analysis. In Proc. IEEE conference on energy Internet and energy system integration. IEEE