Scalable IoT Solutions for Urban Air Quality Monitoring: ML Approaches for Low-Power WAN Data Analysis

Dr. Ayesha banu[1*], Ennapureddy Madhusri[2], Sankuri Bhavyathrika[2], Mogudumpuram Kalyan[2]

[1]Professor & Head, [2]UG student, [1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: ayeshabanuvce@gmail.com

ABSTRACT

The increasing urbanization and industrialization worldwide have necessitated the deployment of scalable and efficient air quality monitoring systems to ensure public health and environmental safety. This project proposes a robust IoT-based solution for urban air quality monitoring by leveraging Machine Learning (ML) approaches to analyze data collected from Low-Power Wide Area Network (LPWAN) sensors. The primary objective of this research is to enhance the accuracy of air quality prediction models while maintaining scalability and computational efficiency for real-time monitoring applications. The existing system employs Linear Regression (LR) as a baseline predictive model due to its simplicity and ease of implementation. While Linear Regression provides a fundamental understanding of the relationships between variables, it exhibits limitations in handling non-linear interactions and complex patterns within the data, resulting in suboptimal prediction performance. To address these challenges, the proposed system introduces Random Forest Regression (RFR), an ensemble learning technique that constructs multiple decision trees during training and outputs the average prediction of the individual trees. This approach enhances prediction accuracy by effectively managing non-linearity, reducing overfitting, and providing better generalization capabilities compared to traditional linear models. Comprehensive Exploratory Data Analysis (EDA) is performed to visualize and understand the underlying patterns, distributions, trends, and correlations within the dataset. Visualization techniques such as scatter plots, histograms, heatmaps, and box plots are employed to identify significant features, detect anomalies, and evaluate data quality. EDA also assists in feature selection, which further contributes to the improvement of model performance. Experimental results demonstrate that the Random Forest Regression model significantly outperforms the existing Linear Regression approach in terms of prediction accuracy and robustness. The proposed system achieves superior results by leveraging the strengths of ensemble learning, making it well-suited for handling complex urban air quality data. The study contributes to the development of scalable, efficient, and reliable air quality monitoring systems utilizing ML approaches for LPWAN data analysis. The findings can aid policymakers and environmental authorities in implementing proactive measures to mitigate air pollution and enhance urban sustainability.

Key words: Air Quality Monitoring, Urban Pollution, Low-Power Wide-Area Networks (LPWAN), Data Analytics, Air Pollution Prediction

1. INTRODUCTION

Energy consumption and its consequences are inevitable in modern age human activities. The anthropogenic sources of air pollution include emissions from industrial plants; automobiles; planes; burning of straw, coal, and kerosene; aerosol cans, etc. Various dangerous pollutants like CO, CO2, Particulate Matter (PM), NO2, SO2, O3, NH3, Pb, etc. are being released into our environment every day. Chemicals and particles constituting air pollution affect the health of humans, animals, and even

doi: 10.48047/ijiee.2025.15.4.24

plants. Air pollution can cause a multitude of serious diseases in humans, from bronchitis to heart disease, from pneumonia to lung cancer, etc. Poor air conditions lead to other contemporary environmental issues like global warming, acid rain, reduced visibility, smog, aerosol formation, climate change, and premature deaths. Scientists have realized that air pollution bears the potential to affect historical monuments adversely [1]. Vehicle emissions, atmospheric releases of power plants and factories, agriculture exhausts, etc. are responsible for increased greenhouse gases. The greenhouse gases adversely affect climate conditions and consequently, the growth of plants [2]. Emissions of inorganic carbons and greenhouse gases also affect plant-soil interactions [3]. Climatic fluctuations not only affect humans and animals, but agricultural factors and productivity are also greatly influenced [4].

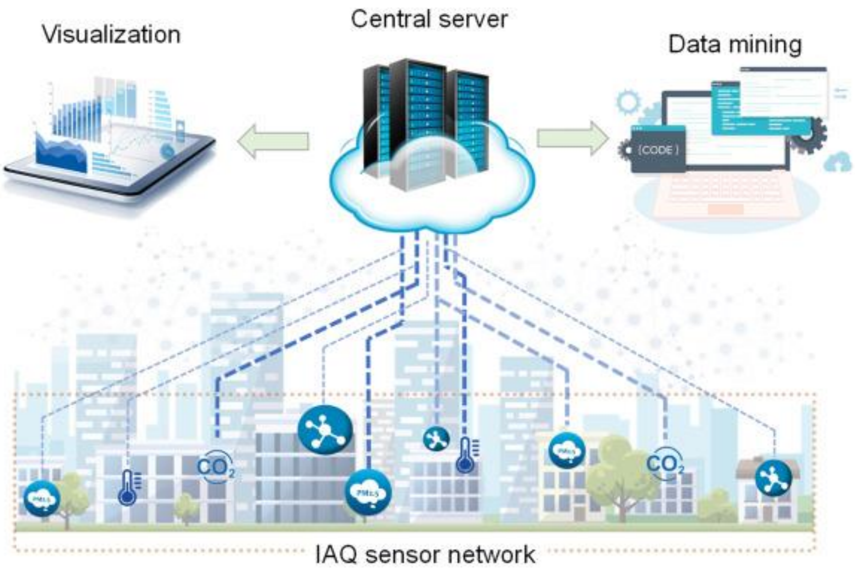


Fig 1: IoT enabled Air quality monitoring.

Economic losses are the allied consequences too. The Air Quality Index (AQI), an assessment parameter is related to public health directly. higher level of AQI indicates more dangerous exposure for the human population. Therefore, the urge to predict the AQI in advance motivated the scientists to monitor and model air quality. Monitoring and predicting AQI, especially in urban areas have become a vital and challenging task with increasing motor and industrial developments. Mostly, the air quality-based studies and research works target the developing countries, although the concentration of the deadliest pollutant like PM2.5 is found to be in multiple folds in developing countries [5]. A few researchers endeavoured to undertake the study of air quality prediction for Indian cities. After going through the available literature, a strong need had been felt to fill this gap by attempting analysis and prediction of AQI for India. Various models have been exercised in the literature to predict AQI, like statistical, deterministic, physical, and Machine Learning (ML) models. The traditional techniques based on probability, and statistics are very complex and less efficient. The ML-based AQI prediction models have been proved to be more reliable and consistent. Advanced technologies and sensors made data collection easy and precise. The accurate and reliable predictions through such huge environmental data require rigorous analysis which only ML algorithms can deal with efficiently.

2. LITERATURE SURVEY

doi: 10.48047/ijiee.2025.15.4.24

In [6], Gopalakrishnan combined Google's Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable. The author developed a web application to predict air quality for any location in the city neighbourhood. Sanjeev [7] studied a dataset that included the concentration of pollutants and meteorological factors. The author analysed and predicted the air quality and claimed that the Random Forest (RF) classifier performed the best as it is less prone to over-fitting. Castelli et al. [8] endeavoured to forecast air quality in California in terms of pollutants and particulate levels through the Support Vector Regression (SVR) ML algorithm. The authors claimed to develop a novel method to model hourly atmospheric pollution. Doreswamy et al. [9] investigated ML predictive models for forecasting PM concentration in the air. The authors studied six years of air quality monitoring data in Taiwan and applied existing models. They claimed that predicted values and actual values were very close to each other.

In [10], Liang et al. studied the performances of six ML classifiers to predict the AQI of Taiwan based on 11 years of data. The authors reported that Adaptive Boosting (AdaBoost) and Stacking Ensemble are most suitable for air quality prediction, but the forecasting performance varies over different geographical regions. Madan et al. [11] compared twenty different literary works over pollutants studied, ML algorithms applied, and their respective performances. The authors found that many works incorporated meteorological data such as humidity, wind speed, and temperature to predict pollution levels more accurately. They found that the Neural Network (NN) and boosting models outperformed the other eminent ML algorithms. Madhuri et al. [12] mentioned that wind speed, wind direction, humidity, and temperature played a significant role in the concentration of air pollutants. The authors employed supervised ML techniques to predict the AQI and found that the RF algorithm exhibited the least classification errors.

Monisri et al. [13] collected air pollution data from various sources and endeavoured to develop a mixed model for predicting air quality. The authors claimed that the proposed model aims to help people in small towns to analyze and predict air quality. Patil et al. [14] presented some literary works on various ML techniques for AQI modeling and forecasting. The authors found that Artificial Neural Network (ANN), Linear Regression (LR), and Logistic Regression (LogR) models were exploited by most of the scholars for AQI prediction. Bhalgat et al. [15] applied the ML technique to predict the concentration of SO2 in the environment of Maharashtra, India. The authors concluded that being highly polluted, some cities of this Indian province require grave attention. The authors mentioned that their model was not capable of exhibiting expected outputs. Mahalingam et al. [16] developed a model to predict the AQI of smart cities and tested it in Delhi, India. The authors reported that the medium Gaussian Support Vector Machine (SVM) exhibited maximum accuracy. The authors claim that their model can be used in other smart cities too.

Soundari et al. [17] developed a model based on NNs to predict the AQI of India. The authors claimed that their proposed model could predict the AQI of the whole county, of any province, or of any geographical region when the past data on concentration of pollutants were available. Sweileh et al. [18] came up with a very interesting study about the analysis of global peer-reviewed literature about air pollution and respiratory health. The authors extracted 3635 documents from the Scopus database published between 1990 and 2017. They observed that there was a substantial increase in publications from 2007 to 2017. The authors reported active countries, institutions, journals, authors, international collaborations in the realm and concluded that research works on air pollution and respiratory health had been receiving a lot of attention. They suggested securing public opinions about mitigation of outdoor air pollution and investment in green technologies. From the literature, it has been observed

doi: 10.48047/ijiee.2025.15.4.24

that research works in air quality analysis and prediction for Indian cities acquired lesser attention from scholars. In spite of the fact that out of the ten most polluted cities in the world, nine cities are Indian [19], very few researchers investigated AQI prediction from the Indian perspective.

3. PROPOSED METHODOLOGY

Air quality prediction using IoT sensor data is a critical application that leverages technology to monitor, assess, and forecast air quality conditions in various environments. This process involves collecting real-time data from a network of IoT sensors deployed in different locations, analyzing this data, and using it to make predictions about air quality. In the process of collecting and managing data from IoT sensors, the information gathered is carefully stored within a centralized database or cloud-based platform. This data is marked with timestamps, providing details about the location of the sensors, the specific type of sensors used, and the actual measurements recorded. This meticulous record-keeping ensures that we have a comprehensive dataset to work with. Prior to delving into data analysis, there is a crucial step known as data preprocessing. During this phase, the data undergoes a series of operations aimed at refining it for further analysis. These operations include addressing missing data points, handling outliers, and, if necessary, converting data into standardized formats. This step ensures that the data is in its best possible condition for accurate analysis. Once the data is preprocessed, the next step involves feature engineering. Here, we extract and create relevant features from the raw sensor data.

Moving forward, this research employs machine learning and statistical models to scrutinize the data and construct predictive models. These models draw insights from historical data, which is often used to train them. A range of algorithms, such as regression, time series analysis, or neural networks, can be applied in this context. The primary objective is to build models capable of forecasting future air quality conditions based on both historical patterns and the latest sensor readings. With the trained models in place, the proposed model equipped to make real-time predictions about upcoming air quality conditions, leveraging the most recent sensor readings. These predictions encompass a variety of valuable information, including AQI values, pollutant concentrations, and air quality forecasts tailored to specific time intervals, such as hourly or daily periods.

To ensure that the air quality information is accessible and comprehensible, it is often visualized through intuitive mediums like dashboards, maps, or graphs. This facilitates easy understanding for the general public, environmental agencies, and policymakers. Additionally, mechanisms are put in place to issue alerts and warnings in cases where air quality levels exceed safety thresholds. The insights drawn from air quality predictions can guide important actions and mitigation strategies. For instance, these predictions can inform decisions related to traffic management, industrial emissions control, and the issuance of health advisories to safeguard public well-being.

doi: 10.48047/ijiee.2025.15.4.24

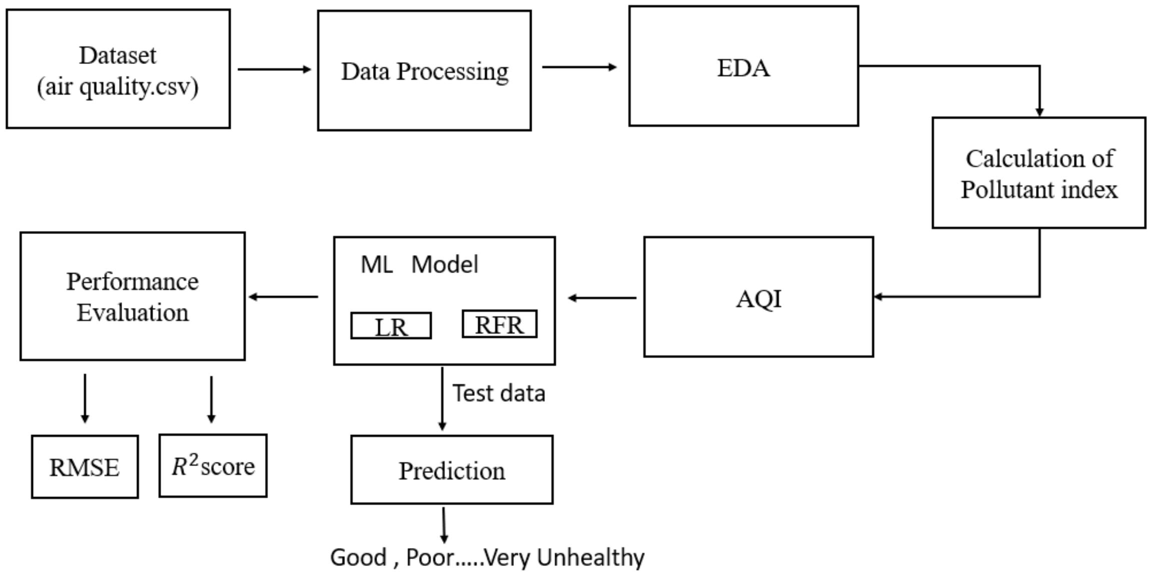


Figure 2 Overall design of proposed air quality prediction.

3.3 RFR Modelling

The RFR model is a powerful machine learning algorithm employed for tasks. It is a versatile and robust algorithm, well-suited for air quality prediction. Its ensemble nature, coupled with randomization in data sampling and feature selection, makes it effective at capturing complex relationships and delivering accurate predictions based on historical and environmental data.

doi: 10.48047/ijiee.2025.15.4.24

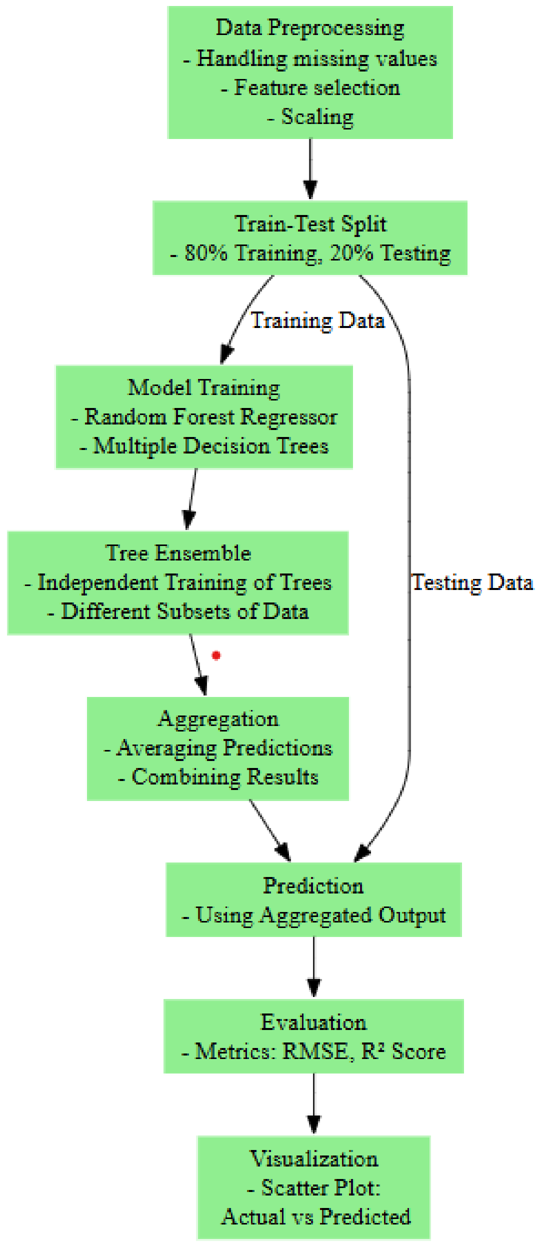


Fig 3: Workflow of RFR

It operates by constructing an ensemble of decision trees, and its functionality can be comprehensively explained as follows:

Step 1: Preparing the Data (Feature Extraction for X_train and y_train)

Before training the Random Forest Regressor, the dataset needs to be preprocessed to ensure consistent numerical representation of input features. The dataset comprises various environmental parameters that influence air quality.

- **X_train:** Contains sensor-based environmental data where each row represents a timestamped observation, and each column represents a feature such as temperature, humidity, particulate matter concentration (PM2.5, PM10), CO2 levels, NOx levels, and other relevant metrics.

doi: 10.48047/ijiee.2025.15.4.24

Feature extraction involves statistical measures like mean, median, variance, skewness, kurtosis, and rolling averages to enhance prediction accuracy.

- **y_train:** The corresponding target variable representing the Air Quality Index (AQI) calculated based on the sensor data. The AQI is a continuous numerical value indicating the overall air quality level.

The model is trained on X_train and y_train, allowing it to learn complex patterns between input features and the AQI.

Step 2: Training the Random Forest Regressor Model

Once the dataset is prepared, we train the Random Forest Regressor to establish robust relationships between environmental features and AQI. The training process involves:

- **Creating an Ensemble of Decision Trees:**

  The Random Forest Regressor constructs multiple decision trees during training, where each tree learns from a randomly selected subset of the training data (Bootstrap Sampling).

- **Feature Randomization:**

  At each split in a tree, only a random subset of features is considered, increasing model robustness and reducing overfitting.

- **Aggregation of Predictions (Averaging):**

  The final prediction is obtained by averaging the predictions from all individual decision trees, which enhances accuracy and robustness against noise.

- **Hyperparameter Tuning:**

  Parameters such as the number of trees (n_estimators), maximum depth of trees (max_depth), and minimum samples required for splitting (min_samples_split) are optimized for enhanced performance.

The model identifies intricate patterns within the sensor data, providing a more comprehensive understanding of how different environmental variables influence AQI.

Step 3: Testing the Model with X_test (New Sensor Data for Prediction)

After training, the model is tested on new, unseen data points stored in X_test, which are preprocessed similarly to X_train.

- **X_test:** Contains new environmental data gathered from IoT sensors, transformed into numerical feature vectors for prediction.
- The trained Random Forest Regressor processes each observation in X_test and predicts the corresponding AQI by averaging predictions from all decision trees.

The model's ability to generalize well to new data is evaluated through prediction accuracy on real-time air quality monitoring scenarios.

Step 4: Generating Predictions and Evaluating y_test (Output Labels)

Once the model processes X_test, it generates predicted AQI values stored in y_test. These predictions are compared against actual AQI measurements to assess the model's performance.

doi: 10.48047/ijiee.2025.15.4.24

- Evaluation metrics commonly used include:
  - **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors.
  - **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual AQI values.
  - **R^2 Score (Coefficient of Determination):** Indicates the proportion of variance explained by the model. A value closer to 1 implies better performance.

4. RESULTS AND DISCUSSION

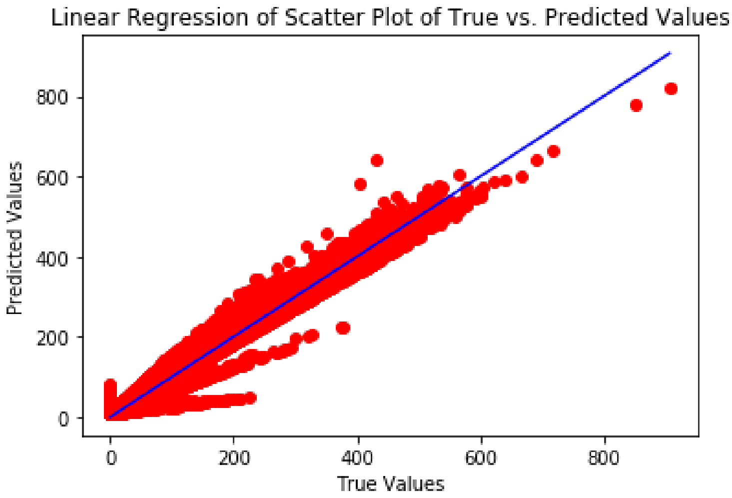
4.1 Dataset description

The dataset captures air quality measurements collected from various monitoring stations across different locations and dates. Each row represents a specific instance of air quality data recorded at a particular time and place. Key columns include stn_code, which uniquely identifies the monitoring station, and sampling_date along with date, which denote when the measurements were taken. State, location, and location_monitoring_station provide geographical context, identifying where the data was collected. The agency column specifies the organization responsible for monitoring, while type indicates the category or method of air quality assessment. The dataset includes measurements of several air pollutants such as sulphur dioxide (SO_2), nitrogen dioxide (NO_2), respirable suspended particulate matter (RSPM), suspended particulate matter (SPM), and fine particulate matter (PM2.5). These features are crucial for evaluating environmental health and pollution levels in different regions. Overall, the dataset is structured to support comprehensive analysis and modeling of air quality trends and pollution forecasting.

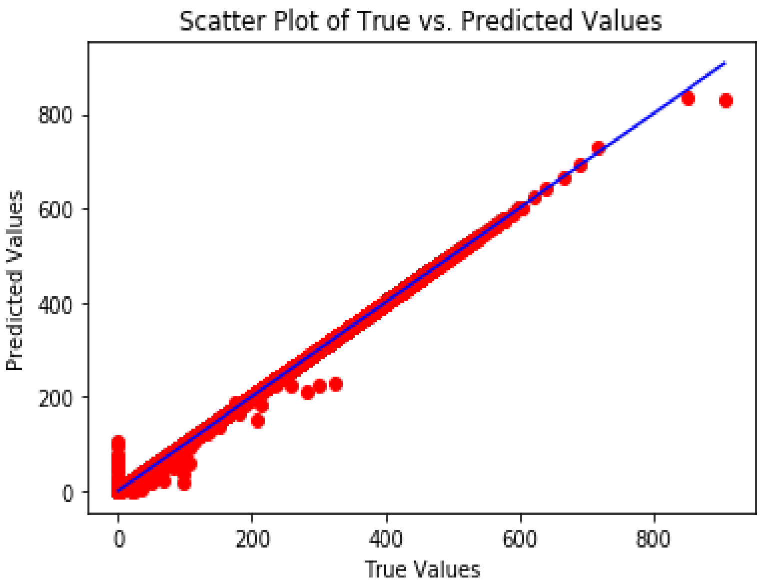
4.2 RESULTS AND DISCUSSION

Figure 4 is a scatter plot likely visualizes the performance of a LR model. In this plot, each point represents a data instance. The x-axis represents the true values (actual observations) of the target variable, while the y-axis represents the predicted values of the target variable made by the LR model. Each point on the plot corresponds to a data instance, where its position relative to the diagonal line (which represents a perfect prediction) indicates how well the model's predictions align with the actual data. If the points are close to the diagonal line, it suggests that the model's predictions are accurate. On the other hand , the scatter plot illustrates the performance of a Random Forest regressor model. Each point on the plot represents a data instance, where the x-axis shows the true values of the target variable, and the y-axis shows the predicted values made by the Random Forest model. The positioning of points relative to the diagonal line helps assess the accuracy of the model's predictions. From Figure 4, it indicates that the points cluster around the diagonal line are closely matches the actual values.

doi: 10.48047/ijiee.2025.15.4.24



(a)



(b)

Fig. 4: Confusion matrices obtained using (a)Existing Linear Regression. (b)Proposed RFR

The figure 5 shows Moderate" prediction typically means that the pollution levels are acceptable for most individuals, but there may be a minor health concern for a small number of sensitive individuals, such as those with respiratory issues. If your dataset is related to concrete strength, safety risk, or predictive maintenance, then "Moderate" would similarly suggest an intermediate condition—neither too high nor too low—indicating balanced or cautionary levels that are not extreme. The outcome reflects how the Random Forest model has aggregated predictions from multiple decision trees to reach a consensus, and it generally suggests that while conditions are manageable, close monitoring or slight adjustments might be beneficial depending on the application.

doi: 10.48047/ijiee.2025.15.4.24

```
Enter the value of SOI:25
Enter the value of NOI:80
Enter the value of RPI:65
Enter the value of SPM:25
Predicted Value: 80.0
AIR Quality is predicated as --> Moderate
```

Fig 5: Prediction on test data with proposed RFR

Table 1 provides a comparison of two different machine learning models used for air quality prediction based on two evaluation metrics: Root Mean Squared Error (RMSE) and R-squared (R^2) score. RMSE (Root Mean Squared Error): The RMSE is a metric used to measure the average magnitude of the errors between predicted values and actual (observed) values. It quantifies how well the predictions align with the actual data. A lower RMSE value indicates better predictive performance, as it means the model's predictions are closer to the actual values. From Table 1:

For the "LR" model, the RMSE is 13.67.

For the "Random Forest Regressor" model, the RMSE is 1.16.

A lower RMSE for the Random Forest Regressor suggests that it has smaller prediction errors compared to the LR model.

R^2-score (Coefficient of Determination): The R^2 score is a statistical measure that represents the proportion of the variance in the dependent variable that's explained by the independent variables in a regression model. It ranges from 0 to 1, where higher values indicate that the model's predictions closely match the actual data. An R^2 score of 1 indicates a perfect fit. From Table 1:

For the "LR" model, the R^2 score is 0.9847.

For the "Random Forest Regressor" model, the R^2 score is 0.999.

The R^2 scores for both models are quite high, indicating that they both provide excellent fits to the data. However, the Random Forest Regressor's score of 0.999 suggests an almost perfect fit, meaning that it captures the variability in the data extremely well. Finally, the Random Forest Regressor outperforms the LR model in terms of both RMSE and R^2 score, indicating its superior predictive capability and ability to explain the variance in air quality data.

Table 1: Comparison of ML models.

Model name	RMSE	R^2-score
LR	13.67	0.9847
Random Forest Regressor	1.16	0.999

5. CONCLUSION

The work has successfully demonstrated how IoT technologies combined with machine learning can be effectively utilized to monitor and predict air quality in urban environments. By collecting and preprocessing data from low-power wide-area networks (LPWANs), the system accurately calculated pollutant-specific indices for common urban air pollutants such as sulfur dioxide (SO_2), nitrogen

doi: 10.48047/ijiee.2025.15.4.24

dioxide (NO_2), suspended particulate matter (SPM), and respirable suspended particulate matter (RSPM). These were used to derive the overall Air Quality Index (AQI), which was further categorized into qualitative labels to simplify interpretation for end users. The application of machine learning models, including Linear Regression and Random Forest Regression, enabled precise predictions of AQI based on the pollutant concentrations. Among these, Random Forest Regression demonstrated higher accuracy due to its robustness in handling complex, non-linear relationships and noisy data. The implementation of user-input-based predictions and visual analytics provided intuitive insights into environmental conditions, indicating the practical utility of the system in real-world applications. This work not only addresses the pressing need for continuous, cost-effective air quality monitoring but also establishes a foundation for scalable, intelligent systems that can support smarter environmental governance and health-focused decision-making.

REFERENCES

[1] Rogers CD (2019) Pollution's impact on historical monuments pollution's impact on historical monuments. SCIENCING. https://sciencing.com/about-6372037-pollution-s-impact-historical-monum ents.html

[2] Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013

[3] Fahad, S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021b) Sustainable soil and land management and climate change (1st ed.). CRC Press. https://doi.org/10.1201/9781003108894

[4] Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. https://doi.org/10.1201/9781003108931

[5] Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys Res Lett. https://doi.org/10.1029/2020GL091202

[6] Gopalakrishnan V (2021) Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascie nce.com/hyperlocal-air-quality-prediction-using-machine-learn ing-ed3a661b9a71.

[7] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538.

[8] Castelli M, Clemente FM, Popoviˇc A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/2020/8049504.

[9] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[10] Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning-based prediction of air quality. Appl Sci 10(9151):1–17. https:// doi.org/10.3390/app10249151

[11] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ ICACCCN51052.2020.9362912

[12] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123.

doi: 10.48047/ijiee.2025.15.4.24

[13] Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J Adv Sci Technol 29(5):6934–6943 Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840.
[14] Patil RM, Dinde HT, Powar SK (2020) A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms 5(8):1148–1152.
[15] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J Comput Appl Technol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006
[16] 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). IEEE 452–457. https://doi.org/10.1109/WiSPNET45539.2019. 9032734.
[17] Soundari AG, Jeslin JG, Akshaya AC (2019) Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 14(11):181–186.
[18] Sweileh WM, Al-Jabi SW, Zyoud SH, Sawalha AF (2018) Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidiscip Respiratory Med. https://doi.org/10.1186/s40248-018-0128-5.
[19] Deshpande T (2021) India Has 9 Of World's 10 most-polluted cities, but few air quality monitors. India spend. https://www.india spend.com/pollution/india-has-9-of-worlds-10-most-pollutedcities-but-few-air-quality-monitors-792521.

doi: 10.48047/ijiee.2025.15.4.24