ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

Air Pollution Monitoring and Prediction System

Ansh Jaiswal
CSIT Dept.
Acropolis Institute of Technology and Research, M.P, India
anshjaiswal210287@acropolis.in

Anmol Soni
CSIT Dept.
Acropolis Institute of Technology and Research, M.P, India
anmolsoni211029@acropolis.in

Aaysha Khan
CSIT Dept.
Acropolis Institute of Technology and Research, M.P, India

Shruti Lashkari
CSIT Dept.
Acropolis Institute of Technology and Research, M.P, India
shrutilashkari@acropolis.in

Ashish Anjana
CSIT Dept.
Acropolis Institute of Technology and Research, M.P, India
ashishanjana@acropolis.in

***Abstract*——Air pollution has been a growing concern in recent years, making it essential to measure and analyze air quality. Previous research has leveraged machine learning algorithms to forecast the Air Quality Index (AQI) for specific locations. While these models have achieved reliable results, they still face challenges such as low accuracy and insufficient data analysis. In this paper, we propose a web-based air quality prediction system using Java technologies, including Spring Boot, JSP, and Servlets. The system integrates machine learning models such as Random Forest, XGBoost, and Neural Networks to predict AQI for various cities in India. The backend is developed using Spring Boot for efficient data processing and API management, while the frontend leverages React to display real-time AQI predictions. The system39;s performance is evaluated using Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2) to ensure accurate predictions. Additionally, data analysis techniques are applied to enhance model accuracy. This paper demonstrates how Java-based technologies can be effectively utilized to develop a scalable and robust air quality prediction platform.**

***Index Terms*—Spring Boot, Spring Security, React, Machine Learning, AQI, Data Analysis**

I. Introduction

Recent economic and social developments have significantly impacted various environmental factors, including land, water resources, and air quality. As a result, air quality monitoring using wireless sensor networks has emerged as a crucial research area. According to the World Health Organization (WHO), air pollution is responsible for approximately seven million deaths annually. To quantify air pollution levels, an Air Quality Index (AQI) is computed based on the concentration of pollutants harmful to human health. The AQI ranges from 0 to 500, where higher values indicate poorer air quality.

Several methodologies exist for calculating AQI, including mathematical formulas and machine learning techniques. In 2018, research led by Samir

ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

Lemes et al. analyzed different AQI estimation approaches, comparing results across standards such as US AQI, EU AQI, and SAQI 11. Their study demonstrated the variations in AQI computations when applied to different regions.

On the other hand, machine learning (ML) has proven to be a powerful tool by integrating knowledge from statistics, artificial intelligence, and computer science. In recent years, ML-based AQI prediction models have gained traction among researchers due to their ability to provide more accurate forecasts. However, certain challenges remain, including handling missing data, feature significance analysis, and feature engineering to fully utilize the dataset. This paper presents a Spring Boot, JSP, and Servlet-based web application for AQI prediction. Our system leverages machine learning models—Random Forest, XGBoost, and Neural Networks—to analyze environmental data and forecast air quality. The performance of these models is evaluated using Root Mean Square Error (RMSE) and Coefficient of Determination (R2). By integrating Java technologies, this study aims to develop an efficient and scalable AQI prediction system that enhances air quality monitoring and decision-making processes.

II. AQI Calculation And Dataset

A. AQI Calculation

In order to get the AQI value, there are several different methods to calculate it worldwide. For example, the AQI formula for China, which is based on the National Ambient

The Air Quality Standard of China (NAAQS-1996) differs from the AQI calculation method defined by the US Environmental Protection Agency (1994) and from the method developed by India (NAAQS Dependent Air Quality Index)[4]. Therefore, in this work, we used the estimating formula for China as the reference in comparison with using the machine learning approach.

Based on the method proposed by NAAQS-2012, the components used in the formula include 6 pollutants (PM_{10}, $PM_{2.5}$, SO_2, Ozone, NO_2, and CO) and 7 indexes, including the maximum 8-hour Ozone concentration (mg/m^3), the maximum 1-hour Ozone concentration, and the daily average concentration of SO_2, NO_2, CO, PM_{10}, and $PM_{2.5}$. The calculation is shown in eq. 1 for each individual pollution-

$$AQI = \frac{BP_{\mathrm{h}} - BP_{\mathrm{l}}}{C_Q - BP_{\mathrm{l}}} + AQI_{\mathrm{l}} \quad (1)$$

Where, C_Q is the pollutant Q's daily mean value. The pollution levels for substance Q are, respectively, AQI_h and AQI_l, with the corresponding estimated highs and lows being BP_h and BP_l. The final AQI value is the largest value in the AQI series by Eq. (2), obtained after completing each AQI math operation:

$$AQI = \max(AQI_0, AQI_1, \ldots, AQI_n) \quad (2)$$

Where n is the number of pollutants considered. In this experiment, we decided to use the Machine Learning approach to predict the AQI value because the first method is quite time consuming, and complicated. Most importantly, the dataset is not always available to be calculated by the formula, which requires information on pollutant concentrations both daily and hourly.

III. Dataset

This research employed a publicly accessible dataset con- taining 29531 instances of Indian air quality. This data set was collected over a six-year period (January 2015 to June 2020), allowing us to evaluate proposed air quality calculation methods. Each instance has had the average daily AQI and some other pollutants from different stations in cities across India. The Central Pollution Control Board [12], the official website of the Government of India, provides the dataset. There are 12 features that have been recorded, including some significant air pollution contaminants like particulate matter (PM 2.5 and PM 10), ozone (O3), nitrogen oxides (NO), NOx, nitrous dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO) emissions, amoniac NH3 and other chemical occurrences (benzen, toluene, xylene). However, there are some important features that contribute mostly to the value of the AQI, which are particulate matter (PM 2.5 and PM

ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

10), CO, NO2, and SO2. On top of that, NOx, which is associated with acid rain, photochemical smog, and tropospheric ozone destruction, is another indicator for AQI prediction [13]. In the dataset, time plots are significant for some analysis related to changes in AQI over months and years, which helps us choose an effective method to predict the AQI value along with time series. We did some analysis regarding the changes in all data features and the AQI value according to year in Fig. 1. The total value is the sum of all pollutants recorded in all cities at different times throughout the given period. It is evident to note that there is an upward trend when it comes to the pollutants and AQI values throughout the 6-year period. The last 3 years from 2018 to 2020 witnessed the highest figures of these pollutants. According to this, 2019 and 2020 are the most polluted years recorded, in which the AQI value and particulate matter peaked in October, November, and December. According to [4], there are some main pollutants that lead to high degrees of air pollution. Thus, we used the total value recorded in five main indexes, including AQI, PM 10, visualization is displayed in Fig. 2. Among the five most polluted cities above, they all recorded high levels of five pollutants, which are AQI, PM 10, PM 2.5, CO, and NO2. On average, Ahmedabad is the most polluted city when it comes to the AQI value, at almost 450 on average. Second in terms of pollution are Delhi, Patna, Gurugram, Lucknow, and so on, which had a high degree of pollutants including AQI, PM 10, PM 2.5, and NO2. However, the CO value was record-high in only Ahmedabad, with more than 20, whereas in other cities this substance only ranged from 0 to approximately 2.

IV. DATA PREPROCESSING AND METHOD

The data preprocessing is the first and most important step, which not only results in a good validation result but also improves the predictive performance of the model later on. This stage often includes missing data imputation, removing strange datapoints, feature engineering techniques, and feature selection. The two first steps help us have a full set of data, improving the accuracy of the models. Meanwhile, selecting useful features can reduce running time, minimize overfitting while running the model.

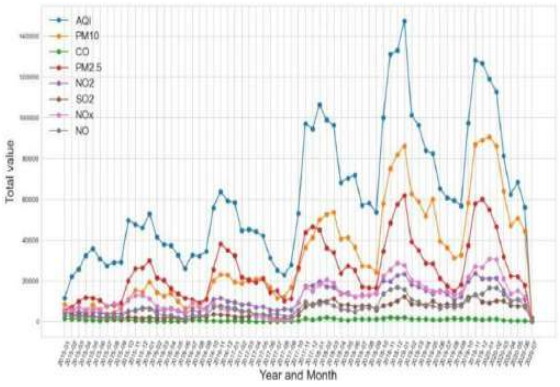

Fig. 1: Air pollutants depicted by time.

A. *Missing Data Imputation*

Tab. I shows the missing value percentage for each col- umn in the dataset. From the given result, the missing data proportion is distributed mostly in xylene, PM 10, NH3, and toluene, which are 61.3%, 37.7%, 35.0%, and 27.2%, respectively. Data loss rates in other situations range from 12% to 19%. This issue can be resolved in a number of ways, including by eliminating dropped data points or by adding the most common value from each case to the missing data. In this study, the K-Nearest

TABLE I: Statistics of Null Values

Pollutants	**Value**	**Percentage (%)**
Xylene	18109	61.3
PM 10	11140	37.7
NH_3	10328	35.0
Toluene	8041	27.2
Benzene	5623	19.0
AQI	4681	15.9
PM 2.5	4598	15.6
NO_x	4185	14.2
O_3	4022	13.6
SO_2	3854	13.1
NO_2	3585	12.1
NO	3582	12.1
CO	2059	7.0

Neighbors Imputer (KNNImputer) technique is used to overcome losing information [14]. At this stage,

ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

each sample's missing values are imputed by the mean value, calculated from 3-neighbors nearest data points in the dataset. This technique was used because it is easy to use and works well. It is also more accurate than simple imputation.

B. Feature Engineering

New features that are created based on the features in the dataset can be very helpful to improve the performance of the model. In this step, we used the mathematical transform, which groups some existing features into a new one that has a good association with the target. In the dataset, we came up with 3 new features by using this method.

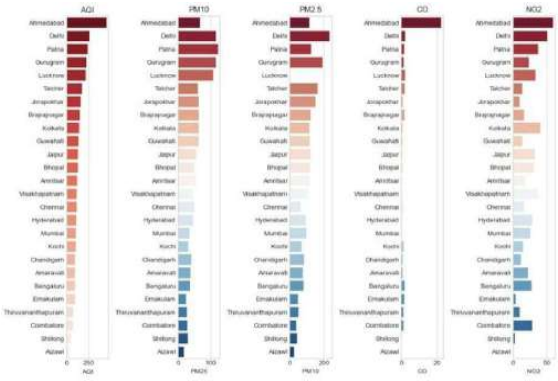

Fig. 2: Most polluted cities towards five pollutants.

The first one is "ParticulateMattersi", which was made by adding the value of PM 10 to PM 2.5 together. The second new feature is "Nitoi", obtained by the sum of NO2, NOx, and NO. Finally, the attribute average Ni is the average value of N previous AQI data points. In addition, year and month are two time features extracted from data information. Three new numerical features are depicted by Eqa. (3)-(5) below:

$$\text{ParticulateMatters}_i = \text{PM}_{10_i} + \text{PM}_{2.5_i} \tag{3}$$

$$\text{Nito}_i = (\text{NO}_2)_i + \text{NO}_i + (\text{NO}_x)_i \tag{4}$$

$$\text{average\_N}_i = \frac{1}{N} \sum_{k=i-N}^{i-1} \text{AQI}_k \tag{5}$$

C. Method

Our method to estimate the AQI value used nine main features, which were reached in the previous stages. We implemented steps in order to get the AQI prediction. Firstly, after being preprocessed as well as experienced data selection and data engineering, selected features were divided into 2 subsets called Training set and Test set. In which Training set accounted for 80% of the total dataset while Test set held the remaining volume, at 20%. The aim of the division is to validate the model's performance later on. In this work, we used three machine learning algorithms, namely Random Forest Regression, Gradient Boosted Regression, and Neural Network Regression, to train three models on the training set. Then, the trained models could be applied to the test set to deal with the unknown data, and get the target prediction. Finally, we used some criteria to assess its effectiveness with regard to model prediction. Fig 3 shows the steps we undertook in our study:

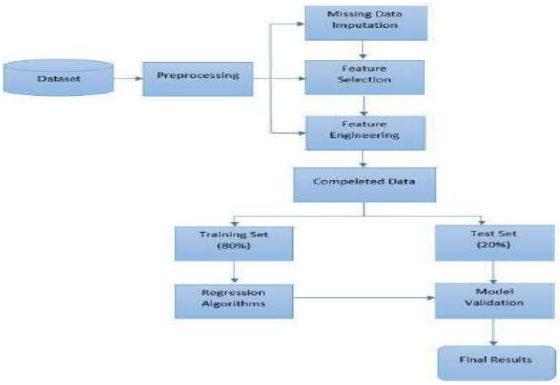

Fig. 3: Machine Learning prediction steps.

1) Random Forest Regression (RFR)

This algorithm is a synthetic prediction algorithm that integrates many different models to create more efficient models. Random Forest (RF) consists of many decision trees, each of which predicts a certain object well and is different from the others[9][10][15]. By averaging the results, we were able to significantly reduce the number of overfit-

ISSN (Online):3048-5088
International Journal of Advancement and Innovation in Technology and Research (IJAITR)
Volume 3, Issue 1, (Jan-Jun) 2026

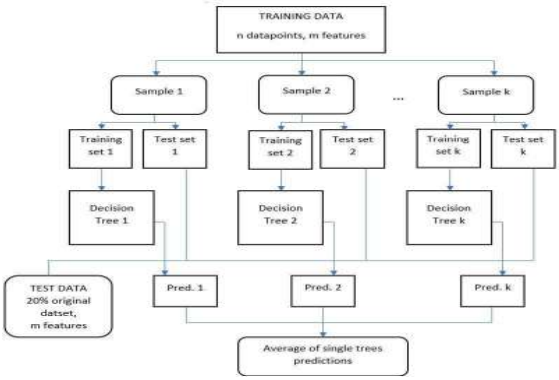

Fig. 4: Flow chart of Random Forest Regression.

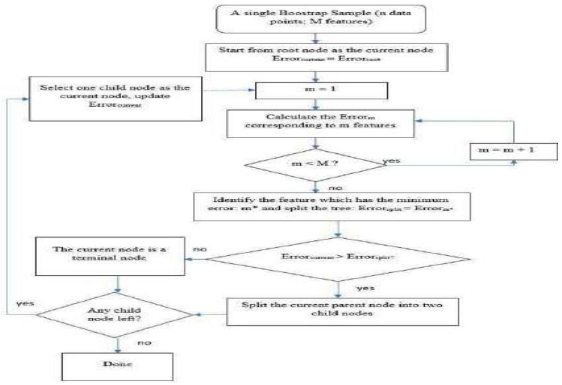

Fig. 5: Flowchart of each bootstrap Sample in Random Forest.

ting while maintaining the model's good predictive score. The steps are as follows:

a) Step 1.: From the initial dataset, we need to build several subsets of data. The technique used to do this task is called the bootstrapping method, in which, from n samples data points, we repeatedly choose random data points with replacement. The result is n datasets called bootstrap samples, which have the same size as the original dataset but in which some data points will be absent or some will be repeated[16]. This method guarantees that each bootstrap sample is modestly different from the others.

b) Step 2.: For each new dataset, build a decision tree with a slight modification: instead of choosing the best test for a specified node, in each node we randomly choose k features (out of the total n features, where $k < n$), and choose the best split among these chosen attributes. By doing this, each tree will perform differently on k distinct selected features, leading to different performances each time.

c) Step 3.: To make a prediction on the unknown dataset, the algorithm uses the predictions obtained in step 2 and averages the results to get the final prediction.

Fig. 4 and Fig. 5 demonstrate the flowchart of the Random Forest Model based on the research of Lingjian Yang in 2017[16], and the flowchart of Decision Tree Model which is a part of the Random Forest Algorithm based on the work done in 2017 by Ibrahim A Ibrahim[11].

2) Gradient Boosted Regression (XGBoost)

In this study, Gradient Boosted Regression (XGBoost)[13] was implemented as a variant of Gradient Boosted Regression Trees. Similar to RF, Gradient Boosted (GB) models are built by many simple decision trees (weak learners), with a depth of one to five. However, the idea behind this model is that each tree can better predict and correct the mistakes of the previous ones. This results in the overall performance of the GB model being improved by adding more trees, and it can make more accurate predictions than the RF model if the parameters are set up meticulously. Therefore, XGBoost requires high accuracy and reliability from datasets. However, it requires careful tuning of the parameters and takes a long time to run.

3) Neural Network

Neural Network or Multilayer Perceptrons is a type of linear model that uses various stages of processing to get the final output. A multilayer model can perform efficiently with a large dataset, constructing a very complex model[16]. However, the models require quite a bit of running time as well as meticulous fine-tuning of the parameters. There are some parameters in this model that we have to take into account while implementing. First, hidden layer sizes (HLS), which is the number of

ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

hidden layers in the models; the number of units in each hidden layer; and the regularization, which is used to control the model's complexity.

D. Validation

In this study, overall performance was assessed using three indexes:

a) 1) Mean Absolute Error (MAE): is the absolute difference between the observed value (y_i) and the predicted result ($\hat{y}_i$). The lower the MAE, the closer the predicted result is to the actual value, and $MAE = 0$ is the ideal value.

b) 2) Root Mean Square Error (RMSE): [11] is the average of the difference between $\hat{y}_i$ and y_i. The lower the RMSE, similar to MAE, the closer $\hat{y}_i$ is to y_i. The higher the RMSE, the more dispersed the $\hat{y}_i$ values are over a wider range.

c) 3) Coefficient of Determination (R^2): [14] has a value range of 0 to 1, indicating how close the predictions $\hat{y}_i$ are to the true value y_i of the model. When $R^2 = 1$, the ideal prediction is understood because it perfectly fits the real data and maximizes performance. In contrast, as R^2 approaches zero, the model becomes less reliable. In summary, a good model is satisfied when the RMSE, MAE, and R^2 are low.

Equations (6)-(8) determine the above three indexes:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

Where $\bar{y}_i$ is the average AQI value at data point i.

V. RESULTS AND DISCUSSION

In the study, the target of the model is the AQI in various Indian cities. In order to get the most precise prediction, we split the dataset into 2 parts: a training set containing 80% of the total data, which was used to train the model; and a test set holding the rest of the data points, which was used in the validation step. In order to choose the parameters for each model, we used a method called Grid Search[6], which tried different values of parameters in the model and then chose the optimized set containing the best ones.

TABLE II: Performance of Models Based on Three Criteria

Methods	MAE	R^2	RMSE
Random Forest (RFR)	19.18	0.94	33.22
XGBoost	18.98	0.942	32.6
Neural Network	22.36	0.928	36.39

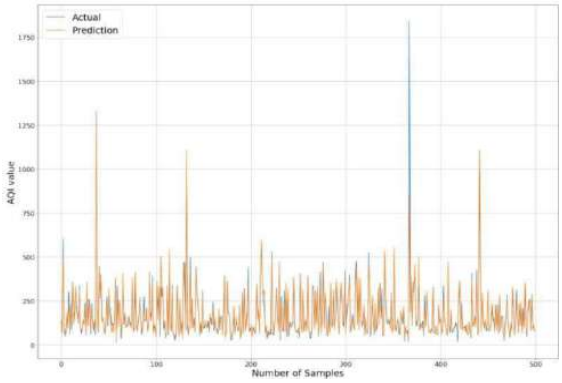

Fig. 6: Correlation between AQI values predicted and measured using the XGboost algorithm.

A. Prediction of the AQI Applying Random Forest Regression

In this research, we employed two values: n estimators and n features. The number of decision trees included in the model is n estimators, and the subset of features in each decision tree is n features. We applied Grid Search to locate the n estimators parameter. Meanwhile, n features were obtained by taking the square root of the total features. According to that, the best combination of parameters used in the work was n estimators = 500 and n features = 3. The criteria for validating the model were: MAE = 19.18, R^2 = 0.94, and RMSE = 33.22.

ISSN (Online):3048-5088
International Journal of Advancement and Innovation in Technology and Research (IJAITR)
Volume 3, Issue 1, (Jan-Jun) 2026

B. AQI Prediction Applying XGBoost Regression

For the XGBoost model, we also used Grid Search to choose two parameters, n estimators and learning rate, which are the number of trees and the rate at which a tree can fix the mistakes of the previous ones. Meanwhile, the third parameter n jobs, the selected set of parameters was n estimators = 300, learning rate = 0.02, and n jobs = 4. The model's statistical criteria were: MAE = 18.98, R^2 = 0.942, and RMSE = 32.6.

C. AQI Prediction of Neural Network Model

According to the previous section, we had two parameters in Neural Network models: hidden layer sizes (HLS) and the number of units in each hidden layer (α). Using the Grid Search method, we obtained the optimal values for the two parameters with HLS = 50, and $\alpha = 0.5$. The results of the validation criteria were MAE = 22.36, R^2 = 0.928, and RMSE = 36.39.

Table II displays the comparison of the three models' performance over the three corresponding criteria.

As can be seen from the result while the Random Forest and XGBoost got performances that are approximately the same in both three criteria, Neural Network, however, performed less efficiently with the same conditions. The MAE and RMSE of this model are much higher, at 23.36 and 36.39, respectively, yet R2 = 0.928 is lower than that of the two other algorithms. As a result, XGBoost is the most effective among the three models when it comes to the statistical criteria, with MAE = 18.98, R2 = 0.942, and RMSE = 32.6. Fig. 6 shows the comparison between the result of the XGBoost model's prediction and actual AQI values with 500 samples. Look at this diagram. The predicted value line (orange) closely follows the actual value line (blue). The distinction is insignificant. This is consistent with the MAE = 18.98 value that we measured. Huixiang Liu et al. [10] used two indices to examine the difference in AQI prediction in Beijing, China: correlation (R2) and mean of difference (RMSE). They used Support Vector Regression (SVR) and Random Forest Regression (RFR) in their study and obtained two sets of indices (R2 = 0.9760, RMSE = 94.4918) and (R2 = 0.8401, RMSE = 83.6716). These two indexes are also used by Chao Song and Xiaoshuang Fu in their paper [[24]. They integrated a set of algorithms into the one called Combination Forecasting Model

(CFM) to get the predictions of AQI in Zengzhou and Shang- hai, China. Their results finally reached RMSE = 36.89 and

R2 = 0.86 for the dataset collected in Zhengzhou, and (RMSE = 35.32, R2 = 0.72) for the other location – Shang- hai. Even though these works are different from our research because of the dataset, Machine Learning algorithms, and some other criteria used to evaluate the models, it is suggested that our models achieved quite good results compared to those of other research when assessed using the same criteria.

VI. Literature Review

Air quality prediction has become increasingly important due to rising pollution levels and their impact on public health. Multiple research efforts have utilized machine learning models to predict the Air Quality Index (AQI), aiming to provide early warnings and promote timely interventions. This section explores key studies and technologies related to AQI forecasting, machine learning applications, data analysis, web development, and alert systems.

1. Traditional AQI Prediction Methods

Earlier approaches used statistical models such as linear regression and ARIMA for AQI time-series forecasting. These models could model simple patterns but lacked the capacity to capture complex relationships among pollutants, weather variables, and human activity.

2. Machine Learning-Based Forecasting

Recent advancements have seen the application of sophisticated machine learning models:

- **Random Forest** has been widely adopted for its ensemble-based accuracy and resistance to overfitting.
- **XGBoost**, known for its speed and accuracy, improves upon gradient boosting by adding

ISSN (Online):3048-5088
International Journal of Advancement and Innovation in Technology and Research (IJAITR)
Volume 3, Issue 1, (Jan-Jun) 2026

regularization. It has proven highly effective in structured AQI datasets.

- **Neural Networks**, especially deep learning models, can capture intricate nonlinear dependencies and temporal trends in pollution data. While powerful, they require substantial training data and computational resources.

3. Data Preprocessing and Feature Engineering

Techniques like normalization, missing value imputation, PCA, and correlation-based feature selection have significantly improved model performance by reducing noise and dimensionality. Several studies emphasize that thoughtful feature selection is as important as model choice in AQI forecasting.

4. Web-Based Systems for AQI Monitoring

Previous systems often relied on Python-based backends (e.g., Flask, Django) for AQI visualization. However, Java-based platforms using Spring Boot offer improved scalability and enterprise-grade robustness. Combining React for dynamic UI and Spring Security for access control ensures a modern, secure, and responsive web solution.

5. Alert Systems for Public Safety

Integrating alert mechanisms (via SMS, email, or web notifications) has proven effective in prior environmental monitoring solutions. Studies highlight the importance of timely alerts in reducing exposure during high-pollution events. An automated threshold-based alert system that notifies users when AQI crosses predefined limits is particularly valuable in urban settings. Some systems incorporate customizable alerts per user preferences, enhancing usability and engagement.

6. Evaluation Metrics

Model performance is commonly evaluated using Root Mean Square Error (RMSE) for error quantification and Coefficient of Determination (R^2) for variance explanation. These metrics guide model tuning and allow for fair comparison across studies.

VII. Conclusions

Air quality has a direct impact on human life and society as a whole. As a result, not only the government, but also individuals and organizations, must work together to prevent environmental pollution, particularly air pollution. As a result, the AQI index is needed to evaluate air quality, and it can also be used to design and produce intelligent meteorological monitoring devices. In this paper, the air pollution indicators in many Indian cities were analyzed and predicted in this study using real data on pollutants provided by the Indian government. The study's findings demonstrated that, while all three models provide good predictive results, the XGB model outperforms the others. Meanwhile, the Neural Network model, which requires careful tuning parameters and much operating time, was not as effective as XGB and RFR. To conclude, the study showed the data analysis and transformation, then built three models for AQI predictions, which attempted to improve the performance in terms of each model's accuracy. In the future, we expect to develop algorithms on devices that use low-power microcontrollers to predict air quality remotely[6][8].

References

[1] D.-C. Nguyen, T. Duc-Tan, and D.-N. Tran, "Application of compressed sensing in effective power consumption of WSN for landslide scenario," in *2015 Asia Pacific Conference on Multimedia and Broadcasting*, 2015, pp. 1–5.

[2] D.T. Tran, "Development of a Wireless Sensor Network for Indoor Air Quality Monitoring," in *The 2015 International Conference on Integrated Circuits, Design, and Verification*, Vietnam, 2015, pp. 178–183.

[3] H. Gu, W. Yan, E. Elahi, and Y. Cao, "Air pollution risks human mental health: an implication of two-stages least squares estimation of interaction effects," *Environmental Science and Pollution Research*, vol. 27, no. 2, pp. 2036–2043, 2020.

[4] S. Lemes, "Air Quality Index (AQI)—comparative study and assessment of an appropriate model For B&H," in *2th Scientific/Research Symposium with International Participation 'Metallic And Nonmetallic Materials'*, MNM, 2018, pp. 282–291.

[5] N. H. Van, P. Van Thanh, D. N. Tran, and D.-T. Tran, "A new model of air quality prediction using lightweight machine learning," *International Journal of Environmental Science and Technology*, 2022. [Online]. Available: https://doi.org/10.1007/s13762-022-04185-w

ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

[6] N. C. Minh, T. H. Dao, D. N. Tran, Q. H. Nguyen, T. T. Nguyen, and D. T. Tran, "Evaluation of Smartphone and Smartwatch Accelerometer Data in Activity Classification," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021, pp. 33–38.

[7] N. T. Thu, T.-h. Dao, B. Q. Bao, D.-n. Tran, P. V. Thanh, and D.-T. Tran, "Real-Time Wearable-Device Based Activity Recognition Using Machine Learning Methods," *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 321–333, 2022. [Online]. Available: https://dx.doi.org/10.12785/ijcds/120126

[8] J. K. Sethi and M. Mittal, "A new feature selection method based on machine learning technique for air quality dataset," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 697–705, 2019. [Online]. Available: https://doi.org/10.1080/09720510.2019.1609726

[9] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences (Switzerland)*, vol. 9, no. 19, 2019.

[10] D. Rajasekhar, M. Rafi D, S. Chandre, J. Prasad and A. Gopatoti, "An Improved Machine Learning and Deep Learning based Breast Cancer Detection using Thermographic Images," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1152-1157, doi: 10.1109/ICEARS56392.2023.10085612.

[11] P. Shukla, "Multiple Classifier Framework System for Fast Sequential Prediction of Breast Cancer using Deep Learning Models," 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9030368.

[12] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," *Complexity*, vol. 2020, pp. 1–23, 2020. [Online]. Available: https://doi.org/10.1155/2020/8049504

[13] P. Bhawan and E. A. Nagar, "Central Pollution Control Board," pp. 1–93, 2019.

[14] R. R. Dickerson, D. C. Anderson, and X. Ren, "On the use of data from commercial NOx analyzers for air pollution studies," *Atmospheric Environment*, vol. 214, no. June, p. 116873, 2019. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2019.116873

[15] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2012.05.073

[16] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," *Lecture Notes in Computer Science*, vol. 7473 LNCS, pp. 246–252, 2012.