Real-Time Air Quality Prediction Using AI: A Machine Learning-Based Dashboard for Urban Pollution Monitoring

Nur Liyana Binti Sazali1, a) and S Prabha Kumaresan1, 2, b)

1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

2Center of Natural Language Processing (NLP), CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

*b) Corresponding author: prabha.kumaresan@mmu.edu.my*

a) nur.liyana.sazali@student.mmu.edu.my

**Abstract.** Pollution poses significant environmental and health challenges, particularly in metropolitan and industrial areas. Accurate prediction of air quality levels is crucial for mitigating health risks and informing policy decisions. This work develops a machine learning-based Air Quality Index (AQI) prediction system that estimates pollution levels using historical and environmental data. By integrating key pollutants (PM2.5, PM10, NO₂, CO, and O₃) with meteorological parameters (temperature, humidity, and wind speed), the model enhances predictive accuracy. The methodology involves data collection from publicly available sources, including the IQ Air World Air Quality Database (between 2021 and 2024) and the EPA Air Quality System (AQS), preprocessing to handle missing values, and feature selection to identify significant air quality determinants. Several machine learning models, including Random Forest, Gradient Boosting, and Neural Networks, were evaluated using air quality data from multiple global cities. Performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score. Experimental results indicate that the Random Forest model achieved the highest accuracy, with an MAE of 0.0033, RMSE of 0.0422, and R² of 0.9982 on recent AQI data from Beijing and Los Angeles (between 2022 and 2025). The selected model was deployed in an interactive real-time air quality dashboard, providing early warnings for poor air conditions. This system offers practical benefits for urban planners, policymakers, and the public by enabling proactive air quality management.

# Introduction

As concerns about air quality gain more and more presence in the public consciousness, both the natural environment and human health are being increasingly compromised in urban settings, heavily polluted industrial niches of society, and urban developments surrounded by heavy emissions. Health effects are both respiratory and cardiovascular, but there are other ramifications of air pollution, including health, life span, economy, and the current state of the environment. The World Health Organization (WHO) estimates that emissions of airborne pollutants, particularly fine particulate matter (PM2.5), trigger about seven million premature deaths throughout the world. The uncertainty caused by needing real-time air quality data, prediction, and estimation will be paramount to future urban and environmental health policy development and regulation.

Traditional monitoring stations offer precise measurements with bad coverage, costly operation, and tardy reporting, thereby rendering it challenging to keep track of sudden air quality level changes. As a response, machine learning approaches offer a promising solution through the dynamic prediction of AQI values from a subset of environmental parameters.

Machine learning models can learn pollutant levels such as PM2.5, PM10, NO₂, SO₂, CO, O₃, meteorological parameters such as temperature, humidity, wind speed, atmospheric pressure, and time variables to identify complex patterns that are otherwise undetectable using classical statistical analysis. Implementation on real-time dashboards enables residents, policymakers, and city planners to implement precautionary measures such as alerts and regulation of pollution sources.

This study suggests the development of an AQI prognosis system based on machine learning by utilizing past pollutants and meteorological data. Random Forest, Gradient Boosting, and Neural Networks are checked on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score to ascertain the best model. This model is embedded in an interactive real-time dashboard and is a green solution that can easily be scaled up for better urban air quality management as industrial growth and environmental deterioration take hold.

# RELATED WORKS

Air quality forecasting is critical to ensuring public safety and environmental sustainability. Recent improvements in machine learning have enabled the creation of complex models using IoT devices, meteorology, and satellite data, therefore improving accuracy and facilitating real-time monitoring. However, data inconsistency, computational complexity, and geographical variability continue to restrict performance.

Regression and ensemble models have worked successfully. [1] used Gradient Boosting and Random Forest to forecast PM2.5 in Taiwan, utilizing high R² values. However, time variability reduced dependability. For AQI prediction, [2] used Cat Boost, which has a low RMSE but a large computational cost and limited real-time use. Ensemble approaches like Random Forest outperform traditional regressions, but they are very data-dependent, making them difficult to deploy in low-resource contexts.

Hybrid techniques that combine ML with domain expertise are promising. [3] suggested a hybrid Particle Swarm Optimization-LSTM algorithm that improved time series prediction but had problems with scalability. [4] combined LSTM and VGG16 with picture data and achieved high accuracy but encountered interpretability issues common in deep learning.

IoT technology improved air quality monitoring. [5] combined IoT sensors and TensorFlow models for urban areas, achieving high accuracy but with limits due to sensor dependability in adverse weather. [6] highlighted the low-cost sensor calibration issues in comparison to standard equipment.

Accurate data collection and preparation are still necessary. [7] found that combining meteorological and IoT data in Seoul improved model dependability, although there was still inconsistency across data sources [8] used complex imputation approaches to address missing data, which may have introduced bias in certain cases. [9] investigated spatiotemporal feature engineering for input quality enhancement, albeit at a significant computational cost.

Regional variation is an important aspect in model performance. [10] found regional anomalies in PM2.5 prediction over the US, whereas [8] struggled to simulate NO₂ hotspots over Alberta due to unsupervised emission sources. [11] revealed that, while ML models accurately reflected urban pollution, they performed poorly for unexpected occurrences like dust storms.

These findings are also supported by recent research. [12] found that ANN beat LSTM and CNN for PM2.5 forecasting in Chiang Mai, based on median absolute error. [13] discovered that by applying multiple linear regression, environmental consequences are predicted by public knowledge and the use of renewable energy. However, the use of artificially generated information diminished from the study's authenticity. The promise of Random Forest models for environmental forecasting was highlighted by their great accuracy (98.66%) when forecasting tropical weather. [14] discovered that for predicting rainfall and temperature, basic linear regression models might perform better than more intricate models, highlighting the significance of dependent on context model selection.

Lastly, [15] created a machine learning model to predict net environmental effects by blending physical indicators such as air quality, carbon emissions, land use, and social indicators such as public awareness and environmental attitudes in a multiple linear regression framework. Although the model performed well in predictive accuracy (R² = 0.67) with simulated data, it emphasized the positive effect of renewable energy consumption and public awareness on environmental performance. These findings affirm that social behavior metrics need to use social behavior measures within forecasting models but also acknowledge the limits of synthetic data and linear assumptions.

# RESEARCH METHODOLOGY

The full process for a real-time air quality forecast system based on machine learning is depicted in Figure 1. The first step in this process is gathering data, such as meteorological and air quality data from IQ Air and EPA AQS. Data preparation is a step in the process that includes cleaning the data, handling missing information, and organizing it for more analysis. Then, during the train/test split, the data is divided into training and testing datasets. Several machine learning methods that are used to produce prediction models are Random Forest, Gradient Boosting, and Neural Networks. The models are evaluated using MAE, RMSE, and R² score to verify the accuracy of the models. The best-performing model is then selected during the model selection step based on this evaluation.

## Data Collection

Data will be collected from publicly available sources, such as open databases, governmental environmental organizations, and government-funded weather monitoring programs, as a part of the data collection procedure. Because those sources are public, they ensure that the collected data are reliable and accessible for study, discussion, and dissemination. The source selection procedure includes searching for specific urban and industrial regions with significant monitoring of air quality and making historical AQI data. The chosen sites were selected purposely because of high air pollution concerns and the fact that the data was extensively available. The historical data will capture many years of data with measured climate variables such as temperature, humidity, wind speed, and air pressure, including particulate matter like PM2.5, PM10, NO2, and SO2. The data can be processed and utilized by storing the data in structured files like excel or comma-separated values. Implementation is performed using Python and relevant libraries for data analysis and modeling, such as Pandas, NumPy, and Scikit-learn, as well as libraries Matplotlib and Seaborn for visualization of graphical representations of results.

Data Collection

Data Preprocessing

Train/Test Split

Model Development

Evaluation Metrics

Model Selection

**FIGURE 1.** Research methodology flowchart

## Data Preprocessing

The subsequent process is data preprocessing, which simply means cleaning and formatting the datasets that have been collected. Missing values need to be handled using imputation approaches, gaps need to be filled in, and the data needs to be standardized (or simply scaled) to meet the requirements of the machine learning models. Feature selection approaches will identify the most relevant parameters that influence the estimates of the AQI, mainly concentrations of pollutants and meteorological conditions. The data will then be divided into three sets, training (70%), validation (15%), and test (15%) for model development.

## Model Development

Three supervised regression models were created and tested to predict the Air Quality Index. Random Forest Regressor, Gradient Boosting Regressor, and Neural Network are all implemented as Multi-Layer Perceptron (MLP) Regressors. Every model was trained on the preprocessed data and subsequently tested on an independent hold-out set to determine generalization. Random Forest Regressor, an ensemble method that builds a multitude of decision trees and averages them, was used as it is robust and provides feature importance scores, which helps in making the model more interpretable. Gradient Boosting Regressor builds the trees in a sequence where each tree learns from the errors of the previous tree and thus is suited for learning subtle patterns with finely tuned hyperparameters, such as learning rate and tree depth. The Neural Network (MLP Regressor), comprising multiple layers of connected neurons, utilizes nonlinear mappings to model intricate relations in data and performs optimally with sufficient data and training. Hyperparameter optimization for best results was achieved via Grid Search CV, systematically exploring a specified parameter grid while invoking 3-fold cross-validation for mitigating overfitting and enhancing generalization performance.

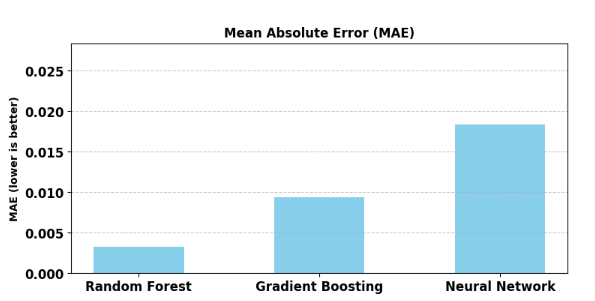
## Evaluation Metrics

The performance of the models was evaluated based on three common regression metrics. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score (Coefficient of Determination). MAE records the average magnitude of errors of predictions in either direction, where lower scores denote greater accuracy. RMSE puts greater weight on larger errors, thus, it is well-placed to identify and punish large differences in forecasts. The R² score describes how well the model can describe the variance in the target variable, with near 1.0 being high predictive capability. The Random Forest Regressor performed the best among the other models with the lowest MAE at 0.0033, RMSE at 0.0422, and R² score at 0.9982. The results demonstrate the model's excellent predictive power and real-time AQI prediction feasibility in this study.

# RESULTS

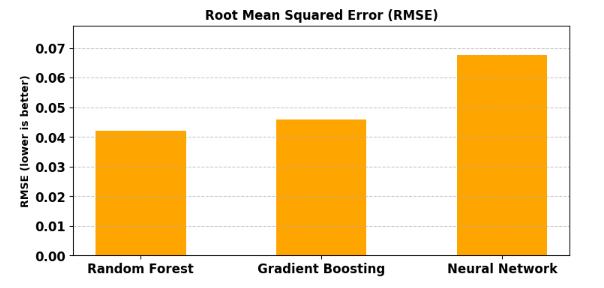
Random Forest, Gradient Boosting, and Neural Network are the three machine learning models that have been shown to perform the best, with the Random Forest Regressor producing the best AQI forecasts. With the smallest Mean Absolute Error (MAE) of 0.0033, it showed that the average difference between the actual and predicted AQI readings was minimal. Additionally, its Root Mean Squared Error (RMSE) was the lowest at 0.0422, suggesting that the predictions were highly consistent and had little variation. With a maximum R² value of 0.9982, the model appears to be nearly perfectly fitted and explains more than 99.8% of the variance in the AQI. Despite being a bit less accurate than Random Forest, the Gradient Boosting Regressor also did well, with an MAE of 0.0094, RMSE of 0.0460, and R² score of 0.9979. Despite its reliability, the Neural Network model's prediction errors were larger, with an R² score of 0.9955, RMSE of 0.0676, and MAE of 0.0184. Among them, Random Forest model turned out to be the best and most reliable for AQI prediction in the future.

Figure 2 depicts the Mean Absolute Error (MAE) of three machine learning models, which are Random Forest, Gradient Boosting, and Neural Network, utilized to estimate air quality levels. MAE is a crucial statistic, and lower is better as it signifies more accurate predictions. Among the machine learning models, Random Forest had the lowest MAE, which implies that it provides the most accurate air quality level estimations. Gradient Boosting performed well, but the Neural Network model made most mistakes and therefore reflected less accurate predictions.



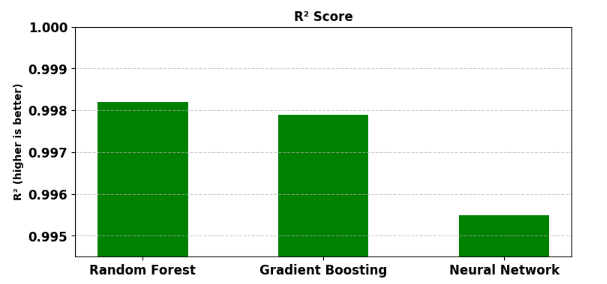
**FIGURE 2.** Model performance comparison for MAE

Figure 3 shows the Root Mean Squared Error (RMSE) of the machine learning models for forecasting air quality which are Random Forest, Gradient Boosting, and Neural Network. The RMSE is a commonly used metric for evaluating regression models, generally lower values suggesting more forecast accuracy and less variation from actual air quality data. In this evaluation, the Random Forest approach has the smallest RMSE, followed by Gradient Boosting, nevertheless, the Neural Network model exhibits the highest error, showing that the model performs the least accurately.



**FIGURE 3.** Model performance comparison for RMSE

Figure 4 displays the R² values of three machine learning techniques, Random Forest, Gradient Boosting, and Neural Network, used to anticipate air quality. The R², also known as the coefficient of determination, indicates whether the model's predictions match the actual data. Values approaching 1 indicate improved efficiency. The Random Forest model has the greatest R² value, followed by Gradient Boosting, suggesting their effectiveness in estimating variation in air quality data. The Neural Network model has the lowest R² score, indicating a poor fit to real data.



**FIGURE 4.** Model performance comparison for R² Score

Table 1 demonstrates that the random forest regression model was the most accurate predictor of the three models. The exceptionally low MAE and RMSE suggest that the model correctly predicted AQI values with a small deviation from real observations. The model's high R² value of 0.9982 demonstrates nearly 99.8% of AQI change, demonstrating its dependability in detecting complex relationships between variables in the environment and air quality indicators. This is especially useful in AQI prediction, where a consistent and dependable forecast is critical for real-time public health warnings and environmental surveillance.

**TABLE 1.** Evaluation metrics for Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R2 SCORE** |
| Random Forest | 0.0094 | 0.0460 | 0.9979 |

Table 2 shows that the Gradient Boosting Regressor worked as effectively, but with somewhat more prediction errors than Random Forest. The R² value was close to 1, showing that it may still explain a significant portion of the AQI variance. Gradient Boosting builds additive models in a forward stage-wise manner, making it effective at capturing nonlinear interactions between contaminants and meteorological factors that affect AQI. This advantage makes it a feasible alternative for AQI prediction, particularly in scenarios with moderate noise data.

**TABLE 2.** Evaluation metrics for Gradient Boosting

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R2 SCORE** |
| Gradient Boosting | 0.0094 | 0.0460 | 0.9979 |

Table 3 shows that the Neural Network model, though equally accurate, had greater error rates and a considerably smaller explanation. Neural networks are extremely effective for nonlinear pattern modeling, but they can be more difficult to adjust and require more datasets to beat ensemble models such as Random Forest. While the neural network adapted well in this case, its AQI prediction consistency was lower than the tree-based models, presumably due to its sensitivity to data scaling or network complexity.

**TABLE 3.** Evaluation metrics for Neural Network

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **R2 SCORE** |
| Neural Network | 0.0184 | 0.0676 | 0.9955 |

## DISCUSSION

This study utilized pollution and meteorological data to create an AQI prediction system based on machine learning models such as Random Forest, Gradient Boosting, and Neural Network. The Random Forest Regressor achieved the highest accuracy (MAE: 0.0033, RMSE: 0.0422, R²: 0.9982) because of its adaptability and ensemble learning capabilities. While Neural Networks and Gradient Boosting both performed well, they were notably less precise and more susceptible to data variability.

The implementation of the most effective machine learning model in a real-time dashboard highlights its practical application in public health and urban development, allowing for early warnings before pollution surges. However, issues persist in data quality, missing values, and model interpretability. Future research should focus on improving spatial-temporal analysis, integrating satellite or geographical information, and creating more interpretable models for policy decision-making.

## CONCLUSION

This study developed a machine learning based AQI prediction system using pollution and meteorological data. Of the many models tested, the Random Forest Regressor produced the best and most accurate results and therefore, was used to create a real-time interactive dashboard to deploy in the field in urban and industrial settings. The system provides relevant air quality information that decision makers can leverage in informing the public and improving public health. Although the study was successful, problems remained in data quality, model interpretability, and scale. Future research should focus on improving data integration, exploring distributed learning and develop explainable AI that is accepted in regulation.

# References

1. Doreswamy, H. K S, Y. Km, and I. Gad, “Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models,” Procedia Computer Science **171**, 2057–2066 (2020).
2. G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, “Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam,” Chemosphere **338**, 139518 (2023).
3. S. Dalal, U. K. Lilhore, N. Faujdar, S. Samiya, V. Jaglan, R. Alroobaea, M. Shaheen, and F. Ahmad, “Optimising air quality prediction in smart cities with hybrid particle swarm optimization‐long‐short term memory‐recurrent neural network model,” IET Smart Cities **6**(3), 156–179 (2024).
4. Y. Wu, X. Wang, M. Wang, X. Liu, and S. Zhu, “Time-Series Forecasting of PM2.5 and PM10 Concentrations Based on the Integration of Surveillance Images,” Sensors **25**(1), 95 (2024).
5. C. Banciu, A. Florea, and R. Bogdan, “Monitoring and Predicting Air Quality with IoT Devices,” Processes **12**(9), 1961 (2024).
6. I. Gryech, C. Asaad, M. Ghogho, and A. Kobbane, “Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review,” Engineering Applications of Artificial Intelligence **137**, 109182 (2024).
7. J. Lee, C. A. M. Barquilla, K. Park, and A. Hong, “Urban form and seasonal PM2.5 dynamics: Enhancing air quality prediction using interpretable machine learning and IoT sensor data,” Sustainable Cities and Society **117**, 105976 (2024).
8. E. Hajiparvaneh, H. Alizadeh, V. Hosseini, and C. R. Koch, “Machine Learning Classification of Air Quality Monitoring Stations to Achieve Ambient NO2 Objectives Using Emission Scenarios and Chemical Transport Model,” Atmospheric Environment **331**, 120582 (2024).
9. D. Tang, Y. Zhan, and F. Yang, “A review of machine learning for modeling air quality: Overlooked but important issues,” Atmospheric Research **300**, 107261 (2024).
10. P. P. Vignesh, J. H. Jiang, and P. Kishore, “Predicting PM2.5 Concentrations Across USA Using Machine Learning,” Earth and Space Science **10**(10), e2023EA002911 (2023).
11. S. J. Livingston, S. D. Kanmani, A. S. Ebenezer, D. Sam, and A. Joshi, “An ensembled method for air quality monitoring and control using machine learning,” Measurement: Sensors **30**, 100914 (2023).
12. M. Ratchagit, “Forecasting PM2.5 Concentrations in Chiang Mai using Machine Learning Models,” IJORAS **6**(2), 37–41 (2024).
13. S. Palaniappan, R. Logeswaran, A. Velayutham, and N.D. Bui, “Predicting Short-Range Weather in Tropical Regions Using Random Forest Classifier,” JIWE **4**(1), 18–28 (2025).
14. A. Subramanian, N. Palanichamy, K. -W. Ng, and S. Aneja, “Climate Change Analysis in Malaysia Using Machine Learning,” Journal of Informatics and Web Engineering **4**(1), 307–319 (2025).
15. S. Palaniappan, R. Logeswaran, S. Khanam, and Y. Zhang, “Machine Learning Model for Predicting Net Environmental Effects,” Journal of Informatics and Web Engineering  **4**(1), 243–253 (2025).