Air Pollution Monitoring and Prediction using Big Data Analytics and Machine Learning

Md. Ferdous Rahman, Mahmud Bin Farid Hasan, Tamanna Akter, and Md. Solaiman Mia*

Department of Computer Science and Engineering, Green University of Bangladesh, Narayanganj-1461, Dhaka, Bangladesh
{ferdousrahman652, hasanf597, tamannatonni901}@gmail.com, solaiman@cse.green.edu.bd*

Abstract. Air pollution is a growing threat to the environment in many developing countries that impacts millions of lives. Rapid urbanization and industrial activity have led to a worse impact of air pollution in Bangladesh. Accurate monitoring and forecasting are essential for public safety. This research adopts recent advances in machine learning to predict real-time air quality in a District of Bangladesh named Narsingdi, using hourly pollution and meteorological data from 2020 to 2024. We have tested modern machine learning and deep learning models such as Support Vector Machine, Regression, XGBoost, Artificial Neural Networks, and a powerful stacked ensemble that blends Random Forest, XGBoost and LightGBM. Following the standard Air Quality Index and adding weather characteristics to the analysis, we have made our predictions more accurate and relevant. Our proposed model correctly classifies air quality with healthy accuracy of 99.6%. It can also pinpoint and predict the main pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 with high reliability. Our research shows that machine learning can help in predicting air quality quickly and affordably for health warnings and clean air. This approach can be adapted in other regions that face similar challenges.

Keywords: Air Quality Index, Machine Learning, Big Data, Air Pollution Prediction

1 Introduction

Air pollution is no longer a problem for mega-cities or industrial centers. It is a global problem and health crisis with long term consequences. According to the World Health Organization (WHO), approximately 7 million people die prematurely due to polluted air [1]. With the majority of these deaths occurring in low and middle economic countries like Bangladesh. Urban area's Air Quality Index (AQI) reaches hazardous level, particularly during the dry season when vehicular emissions, brick kiln operations, construction dust are most intense. Traditional air monitoring system in Bangladesh are limited in both scale and forecasting capabilities. These systems typically consist of a small number of

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_84

fixed location sensors. This provides historical data which cannot predict future pollution levels. On the other hand, manual nature reporting often introduce delays that reduces the effectiveness of real-time responses. This creates critical gaps between environmental data collection and public decision making.

In recent years, Machine Learning (ML) and big data analytics are being applied in environmental monitoring which offers a solution to these gaps. These technologies can process large amount of data which can identify patterns, make predictions with high accuracy. Comparative analysis of ML have demonstrated how ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), etc. outperform traditional statistical models when it comes to AQI prediction in urban areas of Bangladesh [2]. Similarly Deep Learning (DL) based forecasting showed that DL models are capable of not only forecasting air pollution level but also uncovering links between AQI and infant morality, highlighting the social impact of accurate AQI forecasting [3]. Also, data-driven approach for forecasting AQI shows strong results in AI-based air quality prediction across South Asia [4].

In this paper, we have built a comprehensive, scalable and robust AQI prediction framework tailored specifically for a district named Narsigndi in Bangladesh. We have used a rich dataset of over 43,000 hourly entries which covers the year from 2020 to 2024. The dataset includes pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3 and meteorological features such as temperature, humidity, wind speed and direction, solar radiation, rainfall and barometric pressure. This allowed us to experiment with multiple algorithms and architectures, including traditional ML, DL and hybrid ensemble models. The key contributions of our research are summarized below.

- We have developed a refined and imputed AQI dataset.
- We have implemented some ML and DL models in which the best performing model is the combination of RF, XGBoost and LightGBM with Ridge regression as a meta learner.
- Our models are evaluated using both regression and classification metrics. Our proposed model has achieved an outstanding accuracy of 99.6%.
- Beyond predicting AQI as a single value, we have evaluated model performance across individual pollutants. Our proposed model has achieved Root Mean Squared Error (RMSE) of 25.61 and R^2 of 0.77 for $PM_{2.5}$, which is better than the existing works in the literature.

In summary, our research bridges a critical gap between data availability and decision making capability in air quality management.

2 Literature Review

In recent years, researchers around the world have tried to use new technologies such as IoT sensors and ML models to improve monitoring and prediction of air quality. These modern methods can collect and process data much faster compared to the old manual ways. Many research works have used different models and data sources for air pollution monitoring, but there is still a need for solution that can handle messy data which will work well for different pollutants and will provide a good result that can help people to make real-time decisions.

The authors of [5] investigated air pollution across urban areas in Bangladesh using both IoT and ML techniques. Their study highlights how sensors can derive environmental data with classification and regression models and can detect pollution patterns. Their research shows continuous data stream in improving the timeliness and reliability of air quality assessment. While their system demonstrated the value of live data streams, the study focused mainly on model performance without evaluating how easily the approach could be scaled or maintained in real-world city operations. Issues such as integration with city management systems or ongoing maintenance of sensor networks, were not discussed.

A real-time dataset on monitoring AQI using IoT and ML has enabled broader collaboration and benchmarking in this field [6]. Their dataset was collected from multiple Bangladeshi cities, which includes hourly pollutant concentration and associated meteorological parameters providing a valuable resources for training, evaluating and comparing ML models. The dataset provides an important foundation for research, but lacks detailed documentation about sensor calibration and environmental context. Without this, it can be difficult for other researchers to compare results or understand potential sources of error in the measurements.

Islam et al. [7] presented a pioneering approach by integrating IoT-based sensor networks with ML models to monitor urban air quality in Bangladesh. Their system demonstrated real-time data collection and they combined supervised learning algorithms which significantly enhanced both the spatial coverage and accuracy of AQI prediction. Their approach allowed for immediate identification of pollution hotspots and supported more responsive decision making for the local authorities. But the study did not discuss handling challenges like missing sensor data or adapting the system for long term monitoring and prediction.

In [8], the authors showed that for the prediction of atmospheric $PM_{2.5}$ concentrations, ML algorithms and DL architectures are well suited for capturing the complex, non linear meteorological variables and particular pollutions. They improved the accuracy over traditional statistical approaches, particularly when models are trained on large, high frequency datasets. Their advanced models achieved strong results, but the study gave limited attention to model interpretability and how predictions could be explained to policymakers or to the public. This may limit real-world adoption, as users often require more transparent and explainable systems.

A recent work on ensemble-based classification approaches for $PM_{2.5}$ concentration forecasting by Saminathan and Malathy [9] has advanced the field. They combined multiple base learners such as Decision Tree (DT), Gradient Boosting (GB) and Neural Network (NN) in their research. Although their ensemble model achieved better accuracy, it requires more computing resources and longer processing times compared to simpler methods. This could make it harder to use for real-time air quality monitoring, especially in places where fast results or limited hardware are important.

Although the existing works in the literature showed good progress with IoT and ML in predicting and monitoring air pollution, but they have some limitations. Many of them used only one type of model or had small datasets for AQI prediction. Most of them did not combine several strong methods with messy data and multiple pollutants together. To solve these problems, in this paper, we have presented a new approach that uses a mix of advanced models which handles missing data and a full process for making accurate and useful AQI forecasting.

3 Proposed Methodology

In this section, we have discussed the complete workflow of our proposed system. We have described the data sources, preprocessing steps, feature engineering techniques, and different ML and DL techniques. The methodology also defines our proposed ensemble framework and explains the mathematical formulations behind each approach. Fig. 1 shows the methodology diagram, where the workflow starts with data preprocessing and missing values are handled. Then, data visualization and classification are carried out to understand distributions and to assign AQI labels. Input features include pollutant values (SO_2, NO, NO_2, NO_X, CO, O_3, $PM_{2.5}$, PM_{10}) and AQI values. The dataset is then split into training and testing sets. ML, DL and hybrid models are trained, and the proposed RF+XGBoost+LightGBM ensemble model produces the final results. The workflow ends with the evaluation of the model using metrics such as accuracy, precision, recall, F1-score, RMSE, and R^2.

3.1 Data Acquisition

The dataset for this research was obtained from the Department of Environment (DoE), under the Ministry of Environment, Forest and Climate Change, Bangladesh [10]. Data was collected through the Continuous Air Monitoring Station (CAMS) located in Narsingdi, a district of the Dhaka division. The dataset spans the years from 2020 to 2024 and contains 43,846 measurements of major air pollutants including SO_2, NO, NO_2, NO_X, CO, O_3, $PM_{2.5}$, and PM_{10}. In addition to pollutant levels, it also provides meteorological variables such as temperature, relative humidity, rainfall, solar radiation, wind speed, and wind direction. To make the dataset accessible for research and validation purposes, we have uploaded it to Kaggle [11]. The dataset has been released under the CC BY-SA 4.0 license, allowing it to be reused with proper citation of the original source.

3.2 Data Preprocessing

Environmental data often contains noise, missing values, and inconsistent intervals. To ensure data quality, we applied the following steps:

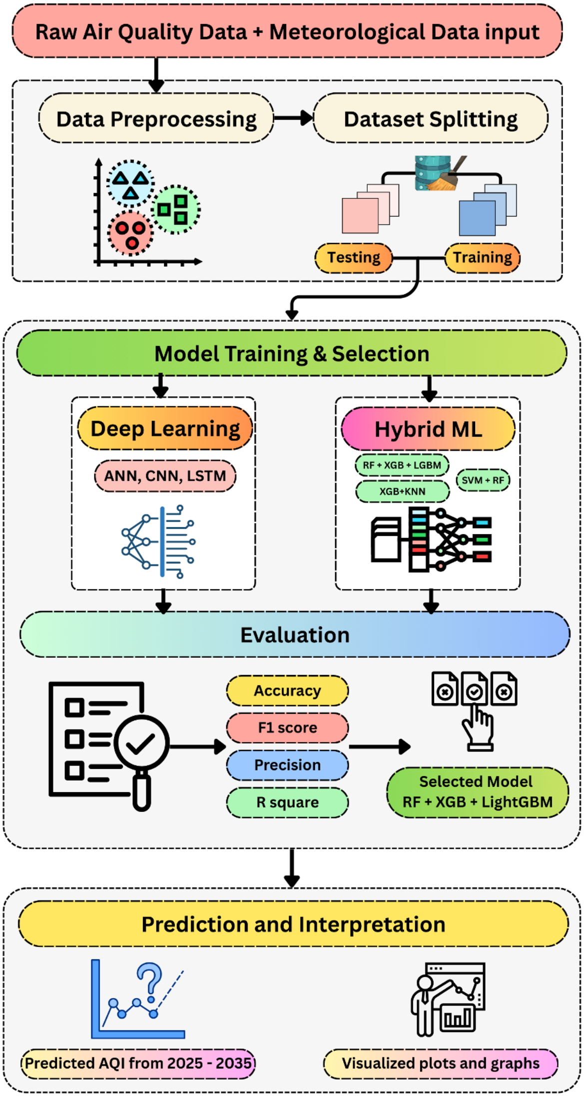

Fig. 1. Workflow of the Proposed Method for AQI Prediction

1. Missing Values: We filled short gaps using interpolation; longer missing blocks were removed.
2. Normalization: Standardization was applied to ensure uniform feature scales.
3. Outlier Handling: Extreme spikes due to sensor errors were excluded.
4. Resampling: Data was aggregated into daily averages to maintain temporal consistency.
5. AQI Labeling: Pollutant concentrations were converted into AQI classes.

3.3 Multi-Pollutant AQI Prediction System (Algorithm)

This subsection presents the step-by-step workflow of our AQI prediction pipeline. The process begins with raw pollutant data and meteorological variables, and concludes with the final ensemble model.

Data Preprocessing: Raw environmental data is first organized and cleaned which involves loading the dataset D and sorting by timestamp. Then, pollutant units are converted into standard forms ($PM_{2.5}$ and PM_{10} in $\mu g/m^3$, CO in ppm, gases in ppb). For handling missing values, short gaps are interpolated, while long gaps are removed. The values are aggregated into daily means by removing extreme outliers. Finally, for stable model training scaling features using z-score or min-max normalization is incorporated.

Dataset Enrichment: Each pollutant p with concentration C is converted into a sub-index I_p using linear interpolation which is defined in Equation (1) [12]. Here, C_{Hi} and C_{Lo} are the break-point concentrations that bound C, I_{Hi} and I_{Lo} which are the corresponding AQI scale values.

$$I_p = \frac{I_{Hi} - I_{Lo}}{C_{Hi} - C_{Lo}}(C - C_{Lo}) + I_{Lo} \tag{1}$$

The overall AQI is then determined as the maximum of all sub-indices using Equation (2). This ensures the most critical pollutant dictates the final AQI level.

$$AQI = \max(I_{p1}, I_{p2}, \ldots, I_{pn}) \tag{2}$$

Feature Engineering: Calendar and seasonal features are extracted as day of month, month, and day of week to capture natural seasonal patterns, weekday-weekend differences, and monthly pollution cycles. Also, wind direction (0-360°) is converted into smooth sine and cosine components to avoid circular discontinuity [13].

$$Wind_{\sin} = \sin\left(\frac{\pi \cdot WD}{180}\right) \tag{3}$$

$$Wind_{\cos} = \cos\left(\frac{\pi \cdot WD}{180}\right) \tag{4}$$

Equation (3) and Equation (4) help the model better understand wind-driven pollutant dispersion. In addition, somr lag features are created at 1h, 3h, 6h, 12h, and 24h, along with 3h, 6h, and 24h moving averages. For each pollutant X, the following lag features were created:

$$X_{\text{lag-1}},\ X_{\text{lag-3}},\ X_{\text{lag-6}},\ X_{\text{lag-12}},\ X_{\text{lag-24}}$$

These features provide short-term memory, smooth noisy readings, and help to keep track pollution build-up over time.

Dataset Splitting: In addition to the main train-test split, we also applied time-series cross-validation to ensure reliable evaluation. A walk-forward approach was used, where each fold trained on all data up to time t and validated on the next time segment. Metrics were averaged across folds to reduce variance, and the stability of errors across folds was monitored to confirm that the model was not overfitting. All pollutant and meteorological values were standardized using z-score normalization to ensure equal scale across features. Sensor outliers caused by faulty readings were removed using an interquartile-range filter. Missing data were handled using a combination of interpolation for short gaps and deletion for long gaps to maintain temporal continuity. For DL models, the ANN used $ReLU$ activation with a final $softmax$ layer, CNN models used $1D$ convolutional filters to learn temporal patterns, and LSTM models used 64 hidden units with $tanh$ activation and recurrent dropout to stabilize training.

The quality of pollutant measurements varied across sensors and required additional preprocessing. $PM_{2.5}$ and PM_{10} showed occasional sharp spikes caused by sensor drift and dust accumulation on the optical chamber. Gaseous pollutants such as NO_2 and SO_2 contained short missing intervals due to calibration delays, while CO and O_3 values were generally stable but exhibited noise during extreme humidity and wind conditions. To ensure reliable learning, short gaps were interpolated, long gaps were removed, and extreme outliers were filtered using the IQR method. These pollutant-specific corrections improved data consistency and reduced the impact of sensor noise on model performance.

Hyperparameter Optimization: All ML and DL models were tuned using randomized search with time-aware cross-validation. Only the most impactful hyper-parameters were explored within defined ranges, and the best settings were selected based on validation performance.

Evaluation: Two types of performance assessments were carried out which are regression evaluation and classification evaluation. RMSE, Mean Absolute Error (MAE) and coefficient of determination (R^2) are considered as regression evaluation, and accuracy, precision, recall, and F1-score are considered as classification evaluation.

Model Selection and Reporting: Finally, the results from all models are compared. The ensemble (RF+XGB+LightGBM) is chosen as the final model, since it consistently provides the best performance across both regression and classification tasks. The model outputs are saved and visualizations such as confusion matrices and error distributions are reported.

3.4 Mathematical Formulation

In this section, we have presented only the mathematical elements that are directly relevant to AQI computation and our proposed hybrid ensemble. To keep the focus on the application rather than textbook derivations, standard ML formulations (e.g., RF, XGBoost, LightGBM, ANN, CNN, LSTM) are intentionally omitted, as they follow well-established definitions in prior literature.

AQI Computation: The AQI is derived from the sub-indices of multiple pollutants. For a given pollutant concentration C, the sub-index I_p is obtained through linear interpolation between break-point values, as shown in Equation (1) [12]. Once sub-indices are computed, the overall AQI is defined as the maximum of the individual sub-indices, as given in Equation (2). This ensures that the pollutant with the most critical health impact determines the final AQI level.

Hybrid Ensemble Formulation: Our proposed ensemble integrates RF, XGBoost, and LightGBM as base learners. For an input feature vector x, each base model produces a prediction $f_i(x)$. The final ensemble prediction is computed using a ridge regression meta-learner [14] which is shown in Equation (5), where w_i are the learned weights.

$$\hat{y} = \sum_{i=1}^{3} w_i f_i(x) \tag{5}$$

The ridge regression objective minimizes the prediction error while controlling model complexity [14] which is shown in Equation (6), where λ is the regularization strength. This regularization helps stabilize the ensemble, especially when base-model predictions are correlated.

$$\mathcal{L} = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 + \lambda \sum_{i=1}^{3} w_i^2 \tag{6}$$

4 Results and Analysis

This section holds the main findings of our research by comparing the performance of different ML and DL models for AQI monitoring and prediction. We compared our proposed hybrid model against other model using metrics such as accuracy, precision, recall and F1-score. The results highlight the strength and weakness of each model and demonstrate the effectiveness of our proposed model for evaluating both classification and regression tasks.

4.1 Overall Model Performance

In this paper, some ML and DL models were evaluated for air quality and pollutant prediction on the Bangladesh AQI dataset [11]. As shown in Table 1, our proposed model (RF+XGBoost+LightGBM) consistently outperformed all the other models achieving an outstanding accuracy of 99.6%, precision of 99.7%, recall of 99.8% and F1-score of 99.7%. In comparison, classic ML models such as SVM+RF and XGBoost+KNN achieved accuracies of 97% and 98.3%, respectively. Consequently, DL models such as ANN, CNN, and LSTM lagged behind where ANN performed the weakest at just 65% of accuracy. These results indicate that our proposed ensemble approach can maintain the strength of individual learners to provide more robust and reliable predictions for AQI classification.

Table 1. Evaluation Comparison of Machine Learning and Deep Learning Models

Model	Accuracy	Precision	Recall	F1-Score
SVM+RF	97	98.1	98.4	98.3
XGBoost+KNN	98.3	98.31	99.2	98.8
ANN	65	66.0	68.0	67.4
CNN	85.2	85.9	84.2	84.4
LSTM	74.4	73.5	69.9	70.8
Proposed Model	**99.6**	**99.7**	**99.8**	**99.7**

4.2 Pollutant Wise Prediction Analysis

The performance in pollutant control, reported as MAE, RMSE, and R^2 for each model and pollutant, gives a deeper insight where lower values indicate the better predictive performance. For $PM_{2.5}$, the best RMSE was achieved by our proposed model (25.61), which was lower than all DL models, with LSTM at 39.80 and ANN at 38.46. Similar trends were observed for PM_{10}, NO_2, SO_2, CO, and O_3. For R^2, the proposed model achieved 0.77 for $PM_{2.5}$ and more than 0.80 for CO and O_3. ANN, CNN, and LSTM generally struggled in all pollutants, with much larger RMSE and lower values of R^2. The result confirms that our proposed model is more capable of capturing pollutant patterns in Bangladesh's atmospheric data. The heatmaps of MAE, RMSE, and R^2 are illustrated in Fig. 2, Fig. 3, and Fig. 4, respectively show a clear difference in model performance between individual pollutants. Our proposed model achieves the lowest error value and highest R^2 in almost every pollutants. On the other hand, DL models produce higher errors for PM_{10} and O_3, as shown by darker cells in MAE and RMSE heatmaps and lower in R^2.

4.3 Interpretability and Key Insights

To make our proposed ensemble model more useful for real-world decision making, we examined which features contributed most to its predictions. The im-

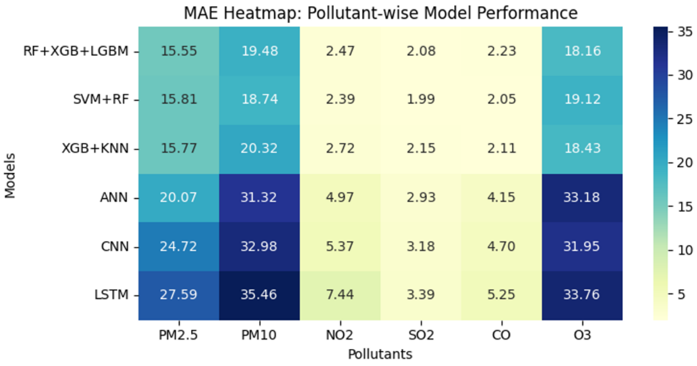

Fig. 2. MAE Heatmap for Different Models and Pollutants

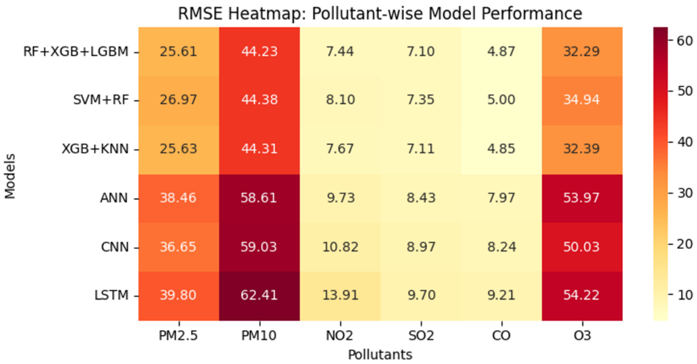

Fig. 3. RMSE Heatmap for Different Models and Pollutants

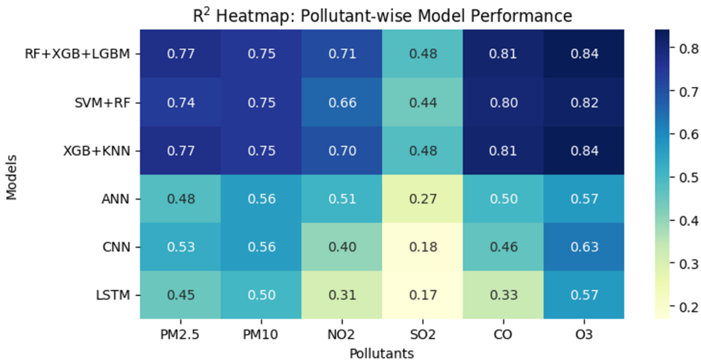

Fig. 4. R^2 Heatmap for Different Models and Pollutants

portance of features and SHAP-based interpretation showed that $PM_{2.5}$, PM_{10}, temperature, humidity, and wind direction were the strongest drivers of the changes in AQI. These variables consistently appeared as the top predictors across all the base models in the ensemble. The model does not treat AQI as a black box. Instead, it highlights the environmental factors that most strongly influence pollution levels. For example, high $PM_{2.5}$ combined with low wind speed consistently pushed AQI into unhealthy categories, while higher humidity and strong wind helped reduce particulate concentration. Such insights allow city authorities to understand when and why pollution peaks occur, and they support more targeted actions such as traffic control, industrial monitoring, or early public warnings.

4.4 AQI Prediction and Trend Analysis

Fig. 5 shows the projected monthly AQI for the next 10 years till 2035 as generated by our proposed model. The predicted value shows gradual increase of AQI. The AQI is expected to rise steadily that highlights the warning air quality in Bangladesh, if no significant intervention is made. Our model can also be used to calculate monthly and seasonal AQI averages that offers valuable insights about how air quality varies throughout the year. This allows more precise identification of pollution peaks during specific months or season, enabling targeted policy interventions and public health planning to present air pollution.

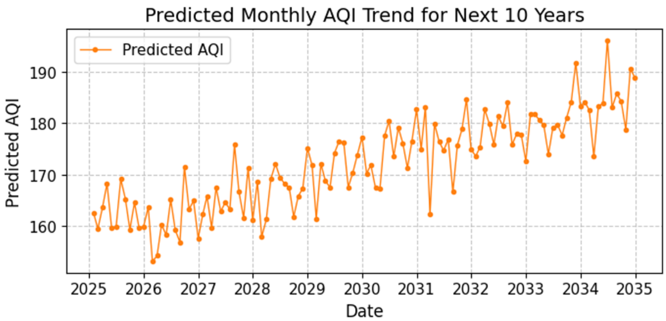

Fig. 5. Predicted Monthly AQI Trend for the Next 10 Years

4.5 Confusion Matrix of Our Proposed Model

Our proposed model achieved strong result for challenging pollutants and was consistent across both training and test datasets which indicates good generalization and reliability. Further evaluation using confusion matrix shown in Fig. 6 confirms models robustness in AQI classification. Most prediction fall along diagonal which reflect accurate categorization of air quality in the majority of cases.

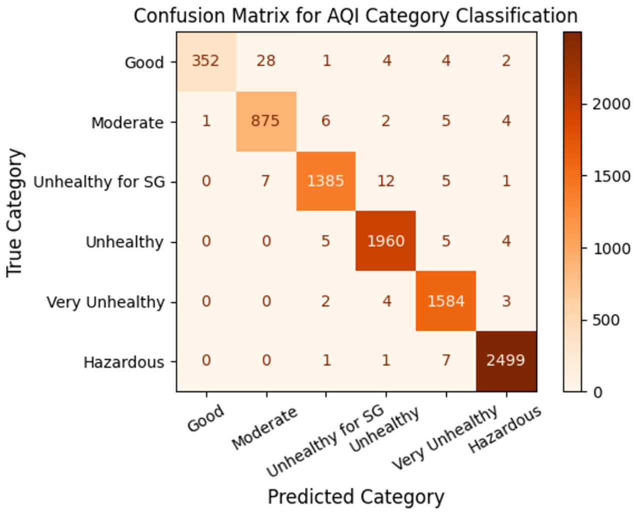

Fig. 6. Confusion Matrix for AQI Category classification

4.6 Comparative Results

To evaluate the effectiveness of our proposed framework, a comparison with recent state-of-the-art approaches for the AQI prediction is presented in Table 2. Table 2 demonstrates that our proposed model (RF+XGBoost+LightGBM) outperforms all the existing works in the literature.

Table 2. Comparison with the Existing Works

Existing Works	Dataset	Used Models	Accuracy
[5]	Dhaka; IoT AQI	RF	97.2%
[6]	Dhaka, Gazipur	RF, KNN, DT, NB, LR	97.0%
[7]	Dhaka, Gazipur	RF, KNN, DT, NB, LR	97.2%
[8]	Isfahan, Iran	ANN, RF, SVM, KNN	90.1%
[9]	India; AQ+Met	LR, SVM, RF, XGB, MLP	97.0%
Proposed Work	**BD_AQI [11]**	**RF + XGBoost + LightGBM**	**99.6%**

4.7 Model Generalization and Transferability

Although the model was trained on data from Narsingdi, its design makes it suitable for use in other Bangladeshi districts that share similar seasonal patterns and meteorological conditions. Regions such as Dhaka, Gazipur, and Narayanganj often show comparable pollution behavior, meaning the same features and model structure can be applied with little adjustment. Because the dataset spans several years (2020–2024), the model also learns recurring seasonal and yearly trends, which helps it generalize to future periods. In locations with very different climates or emission sources, a small amount of local retraining may be needed to maintain the accuracy.

5 Conclusion

This paper shows how well our proposed model can predict air pollution and AQI using real data from a district of Bangladesh named Narsingdi. Our model performs better than the conventional ML and DL methods by giving accurate results for different pollutants and AQI categories. We made sure our model works well by evaluating not only for classifying air quality but also for predicting exact AQI values. The model can even spot patterns by season and month which is useful for planning and public health. The results prove that ML can help to track air quality in real-life and guide smart decisions for a cleaner environment. One of our biggest contributions is building a powerful ensemble model that blends the strengths of RF, XGBoost and LightGBM. We also used smart data cleaning steps like filling missing values with KNN and labeling AQI so that our model works well even if data is messy. Our approach can be helpful to guide others working on air quality or environmental monitoring. Looking ahead, we hope to include more types of sensor data and try models that look at both time and location to make prediction even more accurate. In the future, our system might be used in different cities for tracking the pollutants that will help our community.

Acknowledgments

This work was supported in part by the Center for Research, Innovation, Transformation, and Sustainability (CRITS) of Green University of Bangladesh (GUB).

References

1. World Health Organization, *Ambient (Outdoor) Air Pollution*, 2021. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Accessed: 2025-11-24.
2. M. Sakib, T. Sarkar, S. Das, M. Jannat, "Real-Time IoT-Based Toxic Gas Monitoring and Comparative Analysis of Machine Learning Techniques for AQI Prediction in Dhaka", in *IEEE Conf. Innovations in Environmental Technology*, 2024. DOI: https://doi.org/10.1109/ICISET62123.2024.10939534.
3. R. Islam, Y. Choi, S. A. Shahriar, S. R. Shams, "Deep learning-based PM forecasting and post-infant mortality assessment in urban areas: A case study in Bangladesh", *Air Quality, Atmosphere & Health*, 2025. DOI: https://doi.org/10.1007/s11869-025-01739-y.
4. S. Waghmare, G. Ghadvir, "Assessing urban air quality of Pune city using AI-based predictive model", *Asian Journal of Civil Engineering*, 2025. DOI: https://doi.org/10.1007/s42107-025-01334-7.

5. S. U. P. Shakil, M. A. Kashem, M. M. Islam, N. M. Nayan, J. Uddin, "Investigation of Air Effluence Using IoT and Machine Learning", in *International Conference for Emerging Technologies in Computing*, Springer, Cham, 2023, pp. 183–202. DOI: https://doi.org/10.1007/978-3-031-50215-6_12.
6. M. M. Islam, F. A. Jibon, M. M. Tarek, M. H. Kanchan, S. U. P. Shakil, "A real-time dataset of air quality index monitoring using IoT and machine learning in Bangladesh", *Data in Brief*, vol. 55, p. 110578, 2024. DOI: https://doi.org/10.1016/j.dib.2024.110578.
7. M. U. Islam, M. A. R. Sarkar, M. H. Rahman, S. S. Hoque, S. Haque, S. Mahbub, "Air Pollution Monitoring Using IoT and Machine Learning in Bangladesh", *AETiC*, vol. 8, no. 3, pp. 11–22, 2024. Available: https://aetic.theiaer.org/archive/v8/v8n3/p2.html.
8. F. Mohammadi, H. Teiri, Y. Hajizadeh, A. Abdolahnejad, A. Ebrahimi, "Prediction of atmospheric $PM_{2.5}$ level by machine learning techniques in Isfahan, Iran", *Scientific Reports*, vol. 14, no. 1, 2024. DOI: https://doi.org/10.1038/s41598-024-52617-z.
9. S. Saminathan, C. Malathy, "Ensemble-based classification approach for $PM_{2.5}$ concentration forecasting", *Frontiers in Big Data*, vol. 6, p. 1175259, 2023. DOI: https://doi.org/10.3389/fdata.2023.1175259.
10. Department of Environment, Ministry of Environment, Forest and Climate Change (Bangladesh), *Official Website*, 2025. Available: https://doe.gov.bd/. Accessed: 2025-11-25.
11. M. F. Rahman, M. B. F. Hasan, T. Akter, *Narsingdi AQI Dataset (Dhaka, Bangladesh)*, 2025. Available: https://www.kaggle.com/datasets/thesis321/narsingdi-aqi-dataset-dhaka-bangladesh. Accessed: 2025-11-24.
12. U.S. Environmental Protection Agency, *Technical Assistance Document for Reporting the Daily AQI*, 2024. Available: https://www.airnow.gov/publications/air-quality-index/technical-assistance-document-for-reporting-the-daily-aqi/. Accessed: 2025-11-25.
13. L. Contreras and C. Ferri, "Wind-sensitive interpolation of urban air pollution forecasts", in *Procedia Computer Science (ICCS 2016)*, vol. 80, pp. 313–323, 2016. DOI: https://doi.org/10.1016/j.procs.2016.05.343.
14. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed., Springer, 2021. Available: https://trevorhastie.github.io/ISLR/. Accessed: 2025-11-25.