Machine Learning-Based Air Quality Forecasting in Key Urban-Industrial Zones of Bangladesh

Md. Anamul Hossain Chowdhury
Department of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
anamulchowdhury@sec.ac.bd

Md Abdul Mutalib
Department of Computer Science and Engineering
North East University Bangladesh
Sylhet-3100, Bangladesh
mutalib.cs@gmail.com

Pujan Dey Anik
Department of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
pujandeyanik14@gmail.com

Rumi Chowdhury
Department of Civil Engineering
Barishal Engineering College
Barishal, Bangladesh
rumichowdhury412@gmail.com

Rayhan Mahmud Rakib
Department of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
rayhanrakib93415@gmail.com

Saeid Islam Tanvir
Department of Civil Engineering
Sylhet Engineering College
Sylhet-3100, Bangladesh
saeidislamtanvir@gmail.com

***Abstract*— Worldwide air contamination stands as a serious problem which produces huge effects on human health and the climate as well as the economy. This research explores the prediction of Air Quality Index across essential industrial and urban regions in Bangladesh through RF, DT, KNN and XGBoost machine learning algorithms. A time-span dataset at Kaggle from July 2013 until September 2022 presents information about CO, PM_{10}, NO_x, $PM_{2.5}$, and O_3 alongside other pollutants. We aim to choose the best ML model for predicting Air Quality Index and explore trends related to particulate matter in air quality. The research found that XGBoost performs thes best model by achieving MAE 0.349 and R^2 value 0.997. Our focus centers on PM2.5 and PM10 because they significantly worsen air quality and we have selected target-based pollution reduction approaches as our main strategy. The research demonstrates that advanced machine learning models are highly reliable in predicting air quality, making them useful tools for monitoring the environment and deciding on actions.**

Keywords*— *urban air pollution, air quality index, environmental monitoring, urban pollution forecasting, machine learning.

I. Introduction

A major worldwide concern, air pollution significantly jeopardizes human health, destabilizes natural environments, and imperils economic security. Urbanization stands alongside industrial emissions along with vehicular exhaust and energy production activities as main air pollution elements [1]. Human health faces adverse consequences through respiratory diseases and cardiovascular problems that deteriorate lung performance and trigger death rates to rise. Furthermore, the environmental impacts encompass global warming, acid rain, and ecosystem degradation [2]. AQI serves as the universal assessment mechanism to identify the pollution levels in addition to detailing expected safety hazards [3]. The pollutants responsible for modifying AQI levels are $PM_{2.5}$, CO, PM_{10}, NO_x, SO_2, and O_3. The AQI reference allows tracking of hazardous environmental situations that simultaneously endanger human health and ecological systems [4]. Accurate monitoring of the Air Quality Index along with forecasting requires immediate attention to achieve international sustainable development targets. Multiple prediction systems have been designed for AQI analysis through the use of statistical approaches together with deterministic and physics-based and deep learning and machine learning methods.

Advanced ML models evaluate and predict AQI levels across Bangladesh through analysis of different environmental regions in this research study. The study area includes urban and industrial zones, such as Dhaka Division's Gazipur and Narayanganj as well as Chattogram Division's TV Station and Agrabad and Sylhet City.

The data was colleted from Kaggle, encompassing Air Quality Index values from (2013 to 2022). Pollutants for example CO, NO_x, $PM_{2.5}$, O_3, and PM_{10} were analyzed to compute AQI based on the Department of Environment, Bangladesh's methodology. This research aims to forecast the trends in air pollution by particulate matter and to determine the most appropriate machine learning method to address this problem.

A. Objectives

- Collect and preprocess air quality data of major urban and industrial cities in Bangladesh.
- To implement and evaluate multiple ML algorithms for forecasting the Air Quality Index (AQI)
- Compare and assess model performance.
- Select the most precise and effective prediction mode.
- Develop and compare predictive ML models, including Random Forest, KNN, Decision Tree and XG Boost.

II. Literature Review

Air pollution is among the major environmental obstacles in the world, contributing very significantly not only to the human health atmosphere but also to ecosystems and global climate alike. In such a backdrop, AQI essentially supports convey information on air quality to the masses. The use of ML techniques for forecasting AQI has been a subject of extensive global inquiry. One such instance is the work by Fang and Yang, who applied four distinct machine learning models, notably random forest, to determine the most effective method for precisely and dependably estimating PM2.5 levels in Beijing. [5]. In another work, Seng et al. proposed LSTM networks to catch temporal dependencies in air pollution data these models outperformed significantly traditional statistical models [6].

In South Asia, where air pollution levels are alarmingly high, researchers have increasingly focused on ML

applications for AQI forecasting [7]. In India, Kumar and Pande found that XGBoost gave the best predictions, with the highest linear correlation between what was predicted and the actual data [8]. Ashraf and Farwa Akram applied these techniques in Pakistan to model the variation of AQI by considering the meteorological constant variables like temperature, humidity, and wind speed influencing the outcomes [9]. Further, Gond and Jamal checked the integration of satellite data with ML algorithms in Korba district, India's largest coal mining hub in Asia and it enhanced the predictability of the machine learning model very significantly [10]. Recently, Pavikars & Jansi applied deep learning techniques in Nepal for categorizing pollution and estimation of the severity through images of pollution [11]. The created web application allows any user to upload images and inform about the state of environmental health. Through this, one is able to come to perform the on-site assessment of the surrounding environmental factors.

Most of the air pollution problems in Bangladesh, especially in urban areas like Dhaka, are caused by rapid urbanization, industrial emissions, and vehicular exhaust [12]. Sakib and Sara used different methods such as SARIMAX, Auto-ARIMA, VAR and ARIMA to analyze the AQI in Dhaka. The study found that SARIMAX which considers seasonal cycles, changing trends and external effects, gave the most correct air quality estimates for Dhaka city [13]. The most recent work of Rowley & Karakuş presented the performance with the multi-modal ML approaches using satellite imagery data for AQI prediction. Internet-of-Things (IoT) sensor data is combined in an integrated framework with ML models for real-time AQI predictions, showing very promising potential applications for Bangladeshi smart cities as suggested in Islam and Jibon [14] [15].

III. MATERIALS AND METHODS

A. Study area

This study assessed the Air Quality Index (AQI) across Bangladesh, covering diverse urban, and industrial zones for comprehensive geographic representation in Fig. 1.

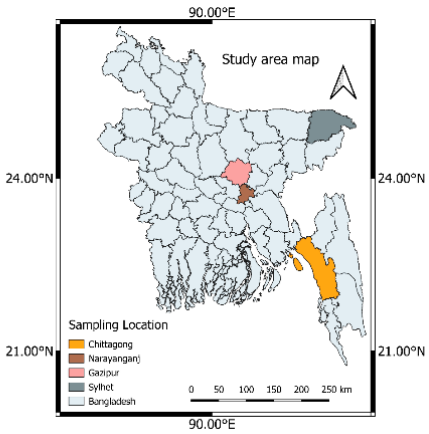


Fig. 1. Geographic map of the research area.

Key locations included Dhaka Division's Gazipur and Narayanganj, highlighting urban and industrial characteristics. In the Chattogram Division, data from Chattogram city's TV Station and Agrabad reflected urban and industrial air quality. Sylhet City provided insights from a minimally industrialized region. This extensive coverage ensured a robust dataset for analyzing AQI variations across regions with different environmental, industrial, and urbanization patterns, offering insights into air quality management challenges and opportunities in Bangladesh.

B. Air Quality Index (AQI)

The AQI was made for determine the severity levels of air pollutants and particulate matter. Table I presents the classification of major pollutants, categorized according to their AQI values [16]. The higher AQI value indicates more severe air pollution and risk to public health. The AQI for each pollutant was calculated using the methodology proposed by Tiwari and Ali [17]. The AQI was determined using the given formula:

$$Q = 100 \times \frac{v}{v_s} \qquad (1)$$

Where,

Q = Air quality levels

V = Field measured data

V_S = Standard value shown in Table II for different pollutants.

AQI estimation is based on the assumption that if there are 'n' pollutants being monitored for air quality, the mean value of these 'n' is assessed using respective air quality ratings (AQR) [18].

$$g = \text{antilog}\left(\frac{\log_a + \log_b + \log_c}{n}\right) \qquad (2)$$

Where,
g = mean value
n = number of pollutants
a, b, c = different AQR values.

TABLE I. CLASSIFICATION OF AMBIENT AIR QUALITY BASED ON AQI

Category	Range of AQI	Status
I	> 125	Extremely hazardous
II	100–125	Highly contaminated
III	75–100	Contaminated
IV	50–75	Moderately degraded
V	25–50	Faily clean
VI	10–25	Clean
VII	< 10	Fresh

TABLE II. AIR QUALITY STANDARDS VALUE FOR BANGLADESH ACCORDING TO WHO GUIDANCE.

Pollutants	Bangladesh Standard	WHO Guideline	Averaging time
Carbon Monoxide (CO)	10 mg/m^3	10 mg/m^3	8 hours (a)
	40 mg/m^3	30 mg/m^3	1 hour (a)
Oxides of Nitrogen (NO_x)	100 µg/m^3	-	Annual (a)
	50 µg/m^3	15 µg/m^3	Annual (b)
Particulates Matters (PM_{10})	150 µg/m^3	50 µg/m^3	24 hours (c)
	15 µg/m^3	10 µg/m^3	Annual
Particulates Matters ($PM_{2.5}$)	65 µg/m^3	25 µg/m^3	24 Hours
	235 µg/m^3	-	1 hour (d)
Ozone (O_3)	157 µg/m^3	100 µg/m^3	8 hours
	80 µg/m^3	-	Annual

IV. METHODOLOGY

The methodology sections start with collecting air quality datasets then data preprocessing, handling missing values and selecting the feature for data analysis which was performed including time series analysis, data correlation and data distribution. The data was separated into 80% for training and 20% for testing the model.

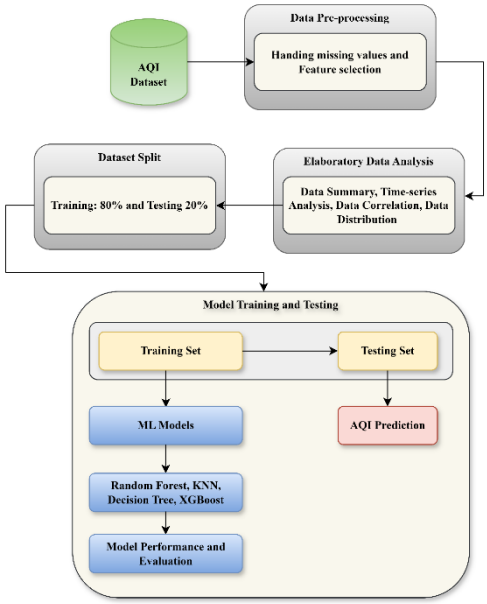


Fig. 2. Methodological framework of the study for air quality prediction.

Fig. 2 diagram represents the architecture of the air quality prediction workflow. Four ML models were utilized by using air quality parameter datasets. The datasets were then trained for machine learning models to predict AQI values. Different standard metricises were used to evaluate each model's performance.

A. *Data Pre-processing*

The reliability of the data was ensured by a number of pre-processing procedures. Records with more than 30% missing data were eliminated, and linear interpolation was used to manage missing pollutant concentrations, particularly $PM_{2.5}$ and NO_x. The researchers standardized the temporal data and then converted all pollutant measurement values into numerical format. Mild outliers received a set maximum value and extreme outliers were excluded from the data. Every observation received an AQI value through the application of the formula shown in (1). These processes created a standardized dataset which turned out to be ready for modeling and analysis.

B. *Machine learning models*

The prediction of AQI in particular urban areas was conducted using four ML models which included RF, DT along with KNN and XGBoost. We separated the data into training sections combined with testing portions. Model learning from Air Quality Index parameters occurred primarily through the 80% training set which enabled testing set analysis to determine model capability in predicting AQI values from new data.

C. *Random Forest (RF)*

The Random Forest (RF) model operates as a supervised machine learning algorithm through an efficient process to analyze AQI together with pollutant concentration predictions. The ensemble nature of RF builds multiple decision trees which are combined through prediction to provide accurate results for avoiding overfitting. RF shows advantages in extracting nonlinear relationships from high-dimensional data therefore making it a suitable method for air quality modeling [19], [20]. The heterogeneous variables including PM_{10}, NO_2, and $PM_{2.5}$ are effectively treated by RF to achieve accurate AQI prediction outcomes. "According to a prior study, Random Forest (RF) has achieved superior results compared to SVM, ANN, and traditional linear classifiers [21], [22]. The air pollution significance indicators of RF allow users to recognize industrial pollution as well as traffic patterns among prominent factors affecting pollution. The predictive power of the model increases through the addition of polynomial features. RF achieves vital use in protecting public health and evidence-based policy-making through its ability to deliver predictable pollutant forecasting results[21], [23].

D. *KNN*

K-Nearest Neighbors (KNN) is regarded as an easy approach for instance-based learning which provides efficient Air Quality Index (AQI) prediction. KNN determines closest matching instances (named nearest neighbors) in order to calculate average values thus providing high performance results in data prediction. Research conducted by KKN in environmental modeling of localized air quality variations and urban heat islands and industrial complexes showed successful results in pollution dispersion analyses for Gorakhpur City India where model effectiveness reached an R^2 value of 0.985 for AQI prediction [24]. High-dimensional data containing multiple pollutants together with meteorological variables can reduce KNN performance because of the curse of dimensionality effect. Applying PCA to reduce input data dimensionality resulted in a real-time air quality prediction accuracy of 95% [25]. Through its pattern recognition capacity in both time-based and geographical data, KNN functions superbly as a method for hotspot detection and health assessment and environmental prevention services.

E. *Decision Tree (DT)*

Research indicates that Decision Tree stands as one of the most widely applied tree-based ML algorithms which provides AQI forecasting abilities through its interpretability and capability to handle numerical and categorical features [26], [27]. The partitioning process of DT into homogeneous groups makes this method suitable for locating the key determinants behind AQI. DT presents overfitting as a significant challenge while dealing with high-dimensional along with noisy datasets. To overcome this challenge several techniques exist which include pruning alongside minimum sample splits and discretization [28]. DT achieves excellent predictive ability (93%–99.05%) when used to model air quality indices in Bangladesh and Indonesia [26],[28]. Environmental research and air quality management benefit from Decision Trees because their capabilities to detect non-linear connections along with providing interpretable results [27].

F. *XGBoost*

Extreme gradient boosting goes by the acronym XGBoost which operates as a strong ensemble learning algorithm. The algorithm constructs decision trees one after another thus each tree functions to adjust the mistakes made in preceding trees. XGBoost demonstrates effective capabilities for Air Quality Index (AQI) prediction through its successful integration with missing value processing together with high processing speed and scalability features [29]. XGBoost implements regularization techniques to stop overfitting which enables it to process datasets containing many features. Environmental modeling utilizes XGBoost to achieve high accuracy levels when predicting air pollution concentrations as per Wibowo et al with 99.516% accuracy and 99.528% F1-score [30]. XGBoost scoring identifies the critical contributors to air quality degradation by ranking NO_x, $PM_{2.5}$ and PM_{10} amongst other factors [31]. The algorithm makes better predictions through its support for inputting missing values by using the Expectation-Maximization algorithm. The combination of XGBoost model with meteorological data has resulted in substantial lower error rates compared to alternative models thus becoming a vital instrument for reducing air quality prediction and management expenses[32].

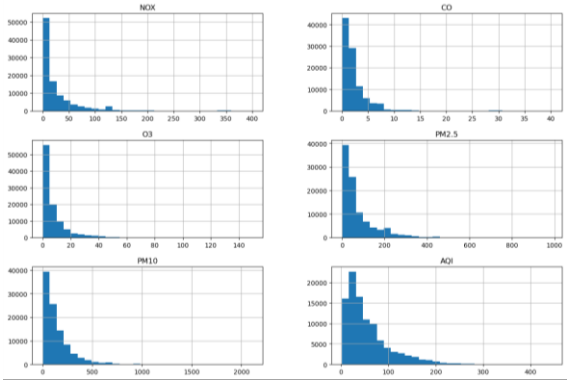


Fig. 3. Histograms depicting the distribution of key air quality indicators

V. EXPLORATORY DATA ANALYSIS

A. *Data analysis*

The data was obtained from Kaggle which contains 100,000 observations of air quality taken over a span of multiple years [33]. It consists of significant pollutants including CO, NO_x, $PM_{2.5}$, O_3, and PM_{10}, along with AQI and temporal variables as year and hour. In Table III the statistical summary reflects a broad variation in pollutant concentration, ranging from AQI 2.14 to 443.78.

TABLE III. DESCRIPTIVE STATISTICS OF AIR QUALITY PARAMETER DATASETS

	NO_x	**CO**	**O_3**	**$PM_{2.5}$**	**PM_{10}**	**AQI**
Count	100000	100000	100000	100000	100000	100000
Mean	31.6	2.5	7.90	74.6	141.7	56.4
Std	50.8	3.4	10.92	80.8	139.8	48.1
Max	399	40.1	149.4	985	2106.2	443.7
Min	0	0.01	0.02	0.02	0.09	2.1
25%	5.5	0.8	1.5	21.1	46.7	22.3
50%	13.1	1.6	4.1	46.1	98.8	40.7
75%	35.0	2.9	9.5	95.4	186	73.4

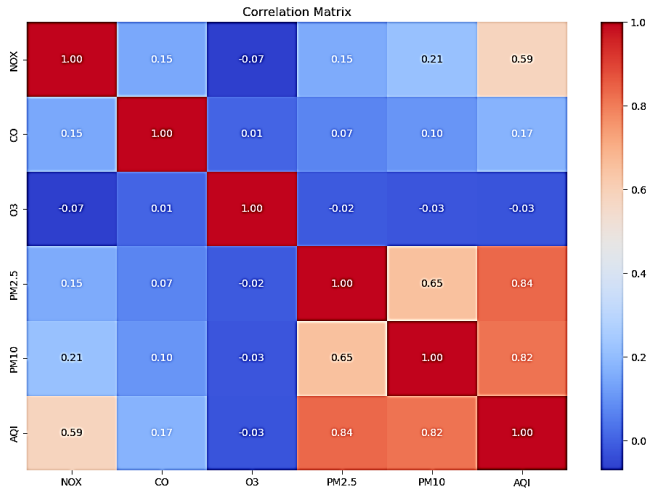


Fig. 4. Correlation between air pollutants and environmental factors.

The average AQI value is 56.49, reflecting the moderate air pollution level on average. The concentrations of NO_x range between 0 and 399.6, and the concentrations of $PM_{2.5}$ and PM_{10} range up to 985 and 2106.29, respectively. Variation is also reflected in CO and O_3 concentrations by the data, ranging up to 40.13 and 149.42, respectively.

Fig. 3 illustrates the distributions of six major air quality parameters NO_x, CO, O_3, $PM_{2.5}$, PM_{10}, and AQI. All of these values exhibit strong positive skewness which indicates that most data points are concentrated at lower values.

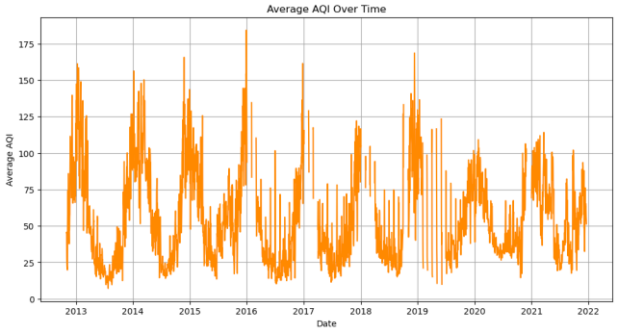


Fig. 5. Average air quality over time from (2013 – 2022)

The air quality data analysis reveals trends and relationships between pollutants. As exhibited in Fig. 5, the average AQI of 2013-2022 reveals a wavering trend, which is caused by multiple factors such as PM_{10}, $PM_{2.5}$, and NO_x etc. Fig. 4, the correlation matrix, reveals positive correlations of $PM_{2.5}$ (0.84) and PM_{10} (0.82) with AQI, which reflects their primary contribution to air quality. NO_x too shows moderate correlation (0.59), while CO and O_3 show poorer correlations (0.17 and -0.03, respectively). They emphasize the superiority of particulate matter in AQI variability, and therefore, pertinent controls should predominantly target the attenuation of $PM_{2.5}$ and PM_{10}.

VI. Results and Discussion

This section presents model assessment metrics which include four essential indicators MAE, MSE, RMSE and R^2 in Table IV. By analyzing different model accuracy and actual vs. predicted graphs represent in Fig. 6, the study found that XGBoost performed the best in terms of predictive power, with the lowest MAE = 0.349, MSE = 5.11 and R^2 = 0.997. The results suggest that XGBoost accurately predicts the AQI. In contrast, K-Nearest Neighbors regression demonstrated only moderate performance. Notably, both RF and DT models produced identical results, MAE = 0.947, MSE = 7.58, and RMSE = 120.2. The relatively low R^2 = 10.9 indicates potential overfitting or difficulties in modeling high-dimensional feature interactions, which may have limited the models' generalization capabilities. In summary, our outcome presents XGBoost as the highest performing AQI predicting model with greater stability and accuracy. The outcome reflects on how selection of proper models is imperative while dealing with environmental data for processing in order to receive right information that can prove useful. Therefore, the perfect parameters selection is crucial for getting appropriate AQI prediction.

TABLE IV. Performance Evaluation of Distinct Machine Learning Methods

Models	MAE	MSE	RMSE	R^2
Random Forest	0.947	7.58	120.2	10.9
XGBoost	0.349	5.11	2.261	0.997
KNN	0.997	1.029	4.723	2.173
Decision Tree	0.947	7.58	120.2	10.9

A. *Model- Wise Comparative Analysis*

The line graph shown in Fig 7 the R^2 scores of different models and their comparative performance. Several observations emerge when comparing the performance model:

- **XGBoost,** showed the most accurate model by displaying small error metrics and achieving an R^2 close to 1 which made its predictions the most reliable. The data handling capabilities and regularization abilities of XGBoost result from its known success in working with structured information.
- **K-Nearest Neighbors,** performed with average error measures but produced an invalid R^2 value above 2 which indicates probable errors during residual variation computation or model instability on this dataset.
- **RF and DT,** the error results from RF and DT models (RMSE = 120.2, R^2 = 10.9) suggest either configural errors or evaluation leakage due to the highly similar results. Before applying any conclusions the abnormal test scores need close attention for debugging to be performed to identify misinterpretations.

The research results emphasize why proper model optimization and verification methods matter. Under present circumstances XGBoost proved to be the sole model producing consistent and interpretable results which were replicable.

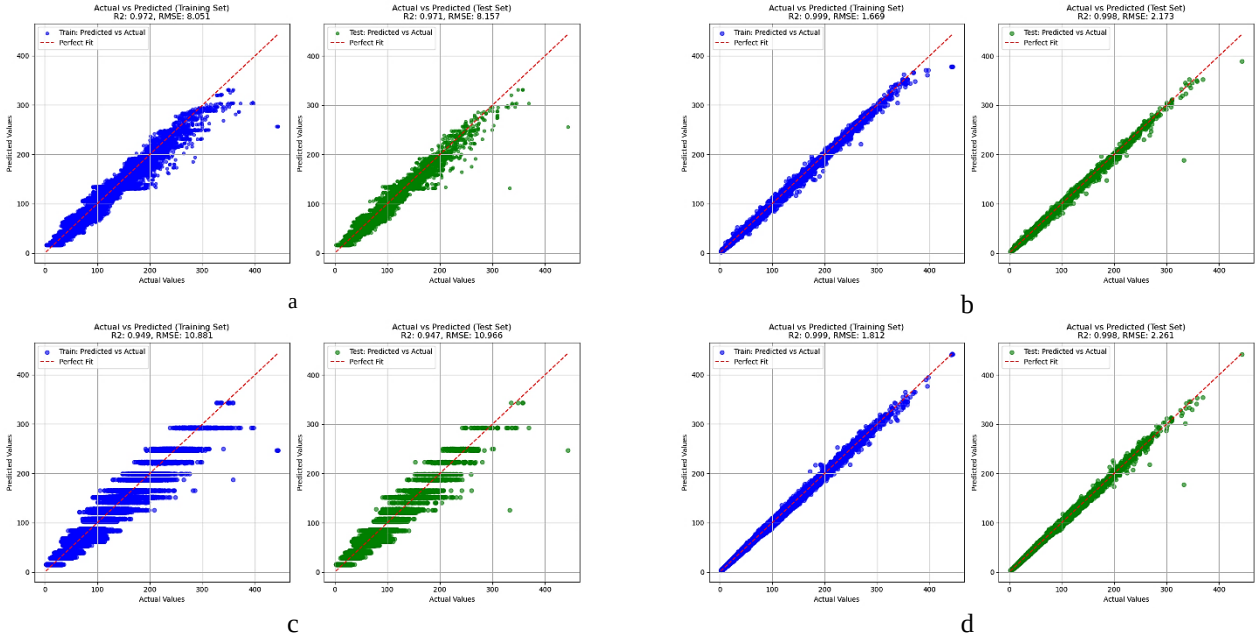


Fig. 6. Plots of actual vs. predicted values of machine learning models. Figures labelled a, b, c, and d are for Random Forest, KNN, Decision Tree, and XGBoost respectively.

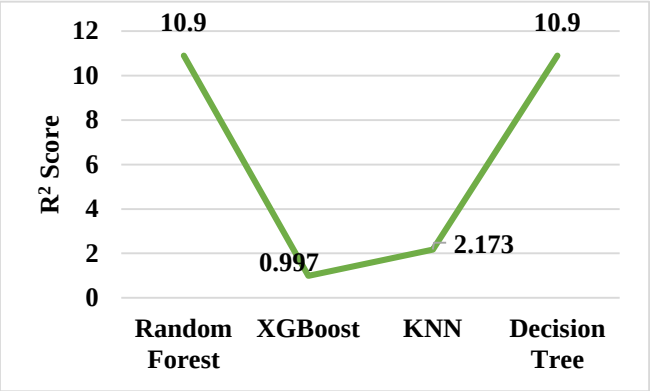


Fig. 7. Line Graph Showing R^2 Score Comparison of Machine Learning Model

B. Practical Implication and Future Directions

The findings extend their relevance from academic reach because valid prediction models serve as crucial sources for both environmental intervention decisions and public environmental awareness programs. XGBoost proved to be the most accurate model among four experimental versions because it reached 0.997 R^2 while producing 5.11 value MSE and 2.261 value RMSE thus displaying excellent correspondence to real observations. Two models namely RF and DT showed $R^2 = 10.9$ but their RMSE measurement stood at 120.2 which points to possible implementation mistakes or incorrect metric understanding. The KNN model demonstrated lower performance than XGBoost because it produced an R^2 value of 2.173 alongside an RMSE score of 4.723 accompanied by increased error variability.

XGBoost seems to be the best choice for prediction applications because of its high performance capabilities. The model applied in prediction demonstrates good environmental decision-making capabilities because of both its low RMSE and high R^2. Random Forest and Decision Tree provided results that seem unreliable according to the test results. The peculiar results emphasize the need for advanced evaluation methods and warrant checking the calculations for metrics.

Future research needs to focus on validation techniques besides preprocessing methods as well as developing performance error assessment approaches. A review of test-train split procedures together with leakage detection and residual analysis needs to be performed to understand the reasons behind unrealistic score reports from specific models. Using SHAP tools would help analysts analyze feature contributions to boost trust in model prediction results.

VII. Conclusion

The research focuses on comparing various ML algorithms that forecast the AQI measurements in main industrial and urban Bangladeshi zones. The predictive model with the most accuracy was XGBoost with the minimum MAE (0.349) and maximum R2 value (0.997). This finding shows that the model successfully detects intricate patterns which exist within air quality information. AQI deterioration stems mostly from particulate matter that displays exceptionally strong positive relationships with AQI. The relative associations between CO and O3 with air quality are weak compared to PM_{10}. and $PM_{2.5}$. The obtained results indicate that $PM_{2.5}$ and PM_{10} reduction should be the main focus because it will enhance air quality conditions. Machine learning demonstrates exceptional capabilities for environmental monitoring and can establish a sustainable basis to conduct research and policy work aimed at obstructing air pollution while protecting public health alongside environmental integrity. Bigger temporal datasets will enable better accuracy and reliability measurements of AQI predictions in future utilization. The study of both industrial zones and urban communities enables the investigation of various pollution distribution styles across multiple geographical areas. Combining LSTM with CNN offers an effective method for capturing both time-based and location-based changes in air quality. These methods reflect the increasing application of ML and DL in environmental studies.

References

[1] E. Gladkova and L. Saychenko, “Applying machine learning techniques in air quality prediction,” in *Transportation Research Procedia*, Elsevier B.V., 2022, pp. 1999–2006. doi: 10.1016/j.trpro.2022.06.222.

[2] N. N. Maltare and S. Vahora, “Air Quality Index prediction using machine learning for Ahmedabad city,” *Digital Chemical Engineering*, vol. 7, Jun. 2023, doi: 10.1016/j.dche.2023.100093.

[3] T. S. Anggraini, H. Irie, A. D. Sakti, and K. Wikantika, “Machine learning-based global air quality index development using remote sensing and ground-based stations,” *Environmental Advances*, vol. 15, Apr. 2024, doi: 10.1016/j.envadv.2023.100456.

[4] 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*, vol. 338, Oct. 2023, doi: 10.1016/j.chemosphere.2023.139518.

[5] Z. Fang, H. Yang, C. Li, L. Cheng, M. Zhao, and C. Xie, “Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR,” *Archives of Environmental Protection*, vol. 47, no. 3, pp. 98–107, 2021, doi: 10.24425/aep.2021.138468.

[6] D. Seng, Q. Zhang, X. Zhang, G. Chen, and X. Chen, “Spatiotemporal prediction of air quality based on LSTM neural network,” *Alexandria Engineering Journal*, vol. 60, no. 2, pp. 2021–2032, Apr. 2021, doi: 10.1016/j.aej.2020.12.009.

[7] H. Li *et al.*, “Climate-driven deterioration of future ozone pollution in Asia predicted by machine learning with multi-source data,” *Atmos Chem Phys*, vol. 23, no. 2, pp. 1131–1145, Jan. 2023, doi: 10.5194/acp-23-1131-2023.

[8] K. Kumar and B. P. Pande, “Air pollution prediction with machine learning: a case study of Indian cities,” *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

[9] M. U. Ashraf, Farwa Akram, and Sardar Usman, “Comparative Analysis of Machine Learning Techniques for Predicting Air Pollution,” *Lahore Garrison University Research Journal of Computer Science and Information Technology*, vol. 6, no. 2, pp. 40–54, Jun. 2022, doi: 10.54692/lgurjcsit.2022.0602270.

[10] A. K. Gond, A. Jamal, and T. Verma, “Developing a machine learning model using satellite data to predict the Air Quality Index (AQI) over Korba Coalfield, Chhattisgarh (India),” *Atmos Pollut Res*, vol. 16, no. 2, Feb. 2025, doi: 10.1016/j.apr.2024.102398.

[11] M. M. Pavikars and R. Jansi, “A Deep Learning Approach for Air Pollution Classification Using InceptionV3 with Transfer Learning,” *Aerosol Science and Engineering*, Jan. 2025, doi: 10.1007/s41810-025-00285-5.

[12] M. S. Hassan, R. F. L. Gomes, and M. A. H. Bhuiyan, “Seasonal distribution of AOT and its relationship with air pollutants in central Bangladesh using remote sensing and machine learning tools,” *Case Studies in Chemical and Environmental Engineering*, vol. 8, Dec. 2023, doi: 10.1016/j.cscee.2023.100399.

[13] S. R. Sakib *et al.*, “Time Series Analysis and Forecasting of Air Quality Index of Dhaka City of Bangladesh,” *2023 IEEE*

World AI IoT Congress, AIIoT 2023, pp. 63–69, 2023, doi: 10.1109/AIIOT58121.2023.10174539.

[14] A. Rowley and O. Karakuş, "Predicting air quality via multimodal AI and satellite imagery," *Remote Sens Environ*, vol. 293, Aug. 2023, doi: 10.1016/j.rse.2023.113609.

[15] M. M. Islam, F. A. Jibon, M. M. Tarek, M. H. Kanchan, and S. U. Perbhez Shakil, "A real-time dataset of air quality index monitoring using IoT and machine learning in the perspective of Bangladesh," *Data Brief*, vol. 55, Aug. 2024, doi: 10.1016/j.dib.2024.110578.

[16] D. Grover and S. Chaudhry, "Ambient air quality changes after stubble burning in rice–wheat system in an agricultural state of India," *Environmental Science and Pollution Research*, vol. 26, no. 20, pp. 20550–20559, Jul. 2019, doi: 10.1007/s11356-019-05395-5.

[17] V. L. Rao, "An Estimation of Air Quality Index of a Coastal Station A case study," 2014. [Online]. Available: http://www.ijcmas.com

[18] S. Chaudhuri and A. R. Chowdhury, "Air quality index assessment prelude to mitigate environmental hazards," *Natural Hazards*, vol. 91, no. 1, pp. 1–17, Mar. 2018, doi: 10.1007/s11069-017-3080-3.

[19] K. Goyal and S. Goyal, "Predicting PM2.5 Air Quality Using Random Forest Regression Enhanced with Polynomial Features," *International IEEE Conference proceedings, IS*, no. 2024, 2024, doi: 10.1109/IS61756.2024.10705219.

[20] X. Wang, D. Li, M. Jiang, Z. Pei, and L. Xu, "K Nearest Neighbor Algorithm Coupled with Metabonomics to Study the Therapeutic Mechanism of Sendeng-4 in Adjuvant-Induced Rheumatoid Arthritis Rat," *Evidence-based Complementary and Alternative Medicine*, vol. 2018, 2018, doi: 10.1155/2018/2484912.

[21] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara, and B. Kuhaneswaran, "Machine Learning Approach for Predicting Air Quality Index," *2021 International Conference on Decision Aid Sciences and Application, DASA 2021*, pp. 622–626, 2021, doi: 10.1109/DASA53625.2021.9682221.

[22] O. Kothari, N. K. Sah, K. V. S. H. Kumar, P. C. Nair, and N. Sampath, "Forecasting India's Air Quality: A Machine Learning Approach for Comprehensive Analysis and Prediction," *2024 4th International Conference on Intelligent Technologies, CONIT 2024*, 2024, doi: 10.1109/CONIT61985.2024.10625932.

[23] J. Li, X. Shao, and H. Zhao, "An Online Method Based on Random Forest for Air Pollutant Concentration Forecasting," *Chinese Control Conference, CCC*, vol. 2018-July, pp. 9641–9648, Oct. 2018, doi: 10.23919/CHICC.2018.8483621.

[24] Mandvi, Hrishikesh Kumar Singh, and Vipin Kumar, "Air quality index prediction for Gorakhpur city using k-nearest neighbors: Model evaluation and analysis," *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 444–454, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2373.

[25] S. Sharma, V. Paranjape, and A. Singh, "Improving Real-Time Pollution Prediction Using Dimensionality Reduction and KNN Algorithms," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 11, no. 11, pp. 11779–11785, Nov. 2023, doi: 10.15680/ijircce.2023.1111059.

[26] M. Hussain, S. Afrin, A. Irin, and S. K. Park, "Applying Decision Tree Algorithm for Air Quality Prediction in Bangladesh," in *2021 5th International Conference on Electrical Information and Communication Technology, EICT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/EICT54103.2021.9733443.

[27] S. Naveen, M. S. Upamanyu, K. Chakki, M. Chandan, and P. Hariprasad, "Air Quality Prediction Based on Decision Tree Using Machine Learning," *International Conference on Smart Systems for Applications in Electrical Sciences, ICSSES 2023*, 2023, doi: 10.1109/ICSSES58299.2023.10200535.

[28] N. Eliyati, M. Rahmayani, S. Wijaya, D. A. Zayanti, E. S. Kresnawati, and Y. Resti, "PREDICTION OF AIR QUALITY INDEX USING DECISION TREE WITH DISCRETIZATION," *Indonesian Journal of Engineering and Science*, vol. 3, no. 3, pp. 061–067, Nov. 2022, doi: 10.51630/ijes.v3i3.82.

[29] H. Jing and Y. Wang, "Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost," in *E3S Web of Conferences*, EDP Sciences, May 2020. doi: 10.1051/e3sconf/202016502014.

[30] W. Wibowo, H. Al Azies, S. A. Wilujeng, and S. Abdul-Rahman, "Harnessing the XGBoost Ensemble for Intelligent Prediction and Identification of Factors with a High Impact on Air Quality: A Case Study of Urban Areas in Jakarta Province, Indonesia," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 191, pp. 319–334, 2024, doi: 10.1007/978-981-97-0293-0_24.

[31] A. A. Varghese, J. Krishnadas, and A. M. Antony, "Robust Air Quality Prediction Based on Regression and XGBoost," *Proceedings of the ACCTHPA 2023 - Conference on Advanced Computing and Communication Technologies for High Performance Applications*, 2023, doi: 10.1109/ACCTHPA57160.2023.10083379.

[32] S. Selvaraj, "Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted Xgboost," 2019. [Online]. Available: https://www.researchgate.net/publication/340032755

[33] "Air Quality Index of Bangladesh Dataset." Accessed: Apr. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/tahmidmir/air-quality-index-of-bangladesh-dataset/