Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

Modeling And Predicting Air Quality Index Of Bangladesh And India Using Machine Learning Approach

Nayema Sanzida Alam
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
nayemasanzida@gmail.com

Marufa Akter Mitu
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mitumarufa24@gmail.com

Md. Foysal Siddik
Department of Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
fx5894@gmail.com

Ahammad Hossain
Department of Computer Science and Engineering
Rajshahi, Bangladesh
ahammadstatru@gmail.com

Md. Mizanur Rahman
Department of Computer Science and Engineering
Rajshahi, Bangladesh
mizanur@vu.edu.bd

A.H.M. Rahmatullah Imon
Center for Interdisciplinary Research
Varendra University
Rajshahi, Bangladesh
director.cir@vu.edu.bd

***Abstract*—Air pollution is a growing concern, particularly in developing countries like Bangladesh and India, where rapid urbanization and industrialization have significantly affected air quality. Poor air quality impacts public health, contributes to global warming, and harms ecosystems. This paper focuses on modeling and predicting the Air Quality Index (AQI) using machine learning techniques to provide accurate and timely forecasts. By utilizing data analysis from environmental sensors, the study analyzes various factors of air pollution, such as location, time, and key pollutants like Carbon Monoxide (CO), Sulphur Dioxide (SO_2), and particulate matter. Machine learning models, including Random Forest, DNN, ARIMA, and SARIMA, are employed to predict AQI levels and pollutant concentrations effectively. The results demonstrate that Random Forest provides the best predictions, with a coefficient of determination of 0.88 for India, though it performs less effectively for Bangladesh (0.28). The DNN, ARIMA, and SARIMA models show poor performance, with negative or low coefficient of determination values, highlighting the challenges in predicting air quality in different regions. These findings emphasize the importance of model selection and provide valuable insights for policymakers in improving air quality management.**

Keywords—Air Quality Index, machine learning, air pollution, Sulphur Dioxide prediction, environmental monitoring.

I. Introduction

With rapid economic growth and technological advancements, urbanization has intensified environmental challenges, including air, water, and noise pollution. Among these, air pollution has increasingly become a critical concern due to its profound human health and the environment are significantly affected by pollution. Exposure to pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), and ozone (O_3) contributes to a wide range of respiratory and cardiovascular diseases, significantly increasing the global disease burden. There is a strong correlation between particulate matter and various other pollutants, including PM2.5, NO2, SO2, CO, PM10, O3, etc. [1] Bangladesh and India, with high population densities and rapid industrialization, face severe air quality degradation due to vehicular emissions, industrial activities, biomass burning, and construction dust. Seasonal factors like winter crop residue burning worsen the issue, with cities like Dhaka and Delhi often reporting "unhealthy" or "hazardous" AQI levels, stressing the need for effective management strategies. Higher CO/NO values indicate that mobile sources are major contributors to the increase in NO levels, with the HYSPLIT model tracing the pollutant's origin. The SARIMA model predicts a rise in PM2.5 concentrations next year, with values exceeding 100 μg/m³.[2]

Machine learning (ML) has become a vital tool for environmental monitoring, enabling the analysis of large datasets and complex pattern prediction. This study applies advanced ML techniques to model and predict AQI levels in Bangladesh and India, leveraging historical data to identify pollutant trends, forecast AQI, and offer actionable insights for mitigating pollution impacts.

II. Methodology

The methodology involves developing machine learning models to predict Bangladesh and India's Air Quality Index (AQI) using historical air quality and meteorological data. Air quality is influenced by various factors, such as location, time, and unpredictable variables.[3] Key steps include data preprocessing, model selection, training, and evaluation to ensure accurate and reliable predictions.

A. Dataset

Country: Bangladesh & India

Source: US Embassy, Dhaka, Bangladesh

Central Pollution Control Board (CPCB), India

Period: 2014-2024

Format: Monthly (Converted from daily values)

Public Availability: Yes.

B. Non-normalized Dataset Approach

A non-normalized dataset approach means working with raw data without applying scaling or standardization. The

original feature values and units are preserved, which maintains the data's interpretability. However, this can cause features with larger value ranges to dominate and potentially bias machine-learning models.

C. Data Pre-processing

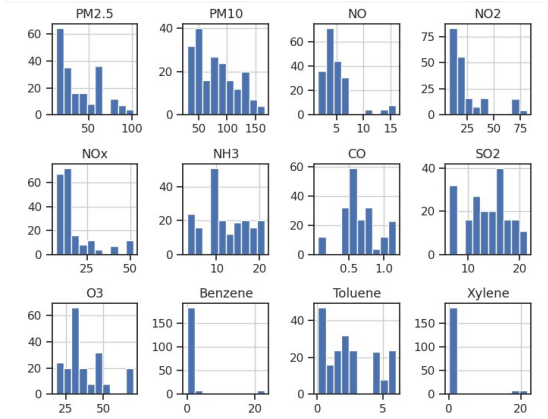
Data preprocessing is crucial in preparing the dataset for accurate analysis and model training. For this study, the dataset includes historical air quality data consisting of pollutant concentrations (e.g., PM2.5, PM10, NO_2, CO, SO_2, O_3 and etc) and meteorological parameters. The preprocessing process begins with handling missing values using appropriate imputation techniques to ensure data completeness. Outliers, often caused by sensor malfunctions or extreme events, are detected and addressed to prevent skewed results. Features are normalized to bring all variables to a common scale, ensuring uniformity during model training. Additional feature engineering is performed, such as deriving temporal features (e.g., month, day, hour) to capture seasonal patterns and encoding categorical variables into numerical formats. To ensure unbiased model evaluation, the dataset is shuffled and divided into training and testing sets in an 80:20 ratio. These steps ensure the data is clean, well-structured, and ready for use in machine learning models, leading to reliable and accurate AQI predictions.

D. Exploratory Data Analysis

In this working process, we cleaned some data from our dataset like missing values and outliers, and found the correlation in our data feature. ED analysis actually provides all the information about the dataset. Such as Pairplot, Histogram, number of feature, perform Correlation Matrix etc.

Fig 1: Histogram

The histogram visualizes the distribution of air pollutants,



showing their frequency across different concentration ranges and helping to identify imbalances, high points, and the spread of pollutant levels. For the AQI dataset of Bangladesh and India (2014-2024), the histogram reveals a high frequency of moderate to unhealthy air quality days, indicating frequent pollution events. The pairplot further illustrates strong correlations between pollutants such as CO, SO_2, and PM2.5 with AQI levels, highlighting the significant impact of particulate matter (PM2.5) on air quality in both countries. This combined analysis offers a comprehensive view of the factors influencing air quality in the region.

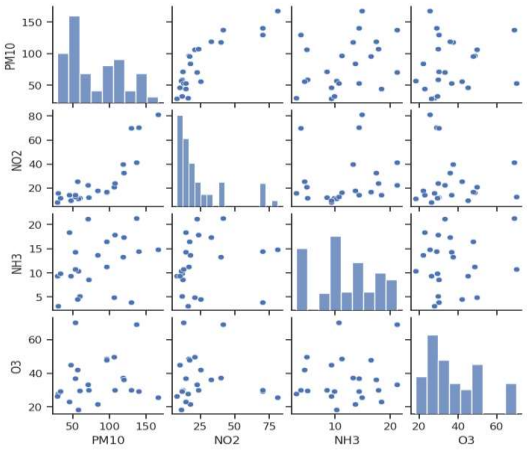


Fig 2 : Pairplot

The pairplot shows scatterplots and histograms to depict relationships and distributions among selected air quality variables. It provides insights into potential correlations and patterns between pollutant concentrations.

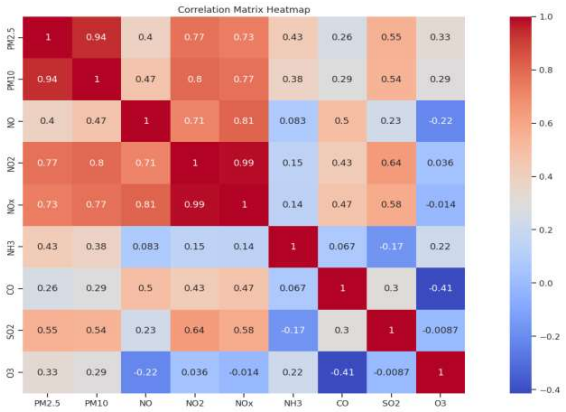


Fig 3: Correlation Matrix

The correlation matrix heatmap illustrates the strength and direction of relationships between pollutants. High values (red) indicate strong positive correlations, while low values (blue) indicate negative correlations or weak relationships.

E. Working Details

In this study, we compare ML and statistical methods to model and predict the Air Quality Index (AQI) for Bangladesh and India. Specifically, we evaluate the performance of Deep Neural Networks (DNN), Random Forest (RF), ARIMA (Auto-Regressive Integrated Moving Average), and SARIMA (Seasonal Auto-Regressive Integrated Moving Average) in capturing the complexities and temporal dependencies in air quality data. These models are assessed for their ability to detect both linear and non-linear patterns, as well as seasonal fluctuations in air quality. The study also incorporates key environmental factors such as pollutant concentrations and meteorological variables to improve prediction accuracy. By evaluating these models, our goal is to determine the most efficient method for providing reliable forecasts of air quality and supporting better environmental management in these regions.

1) Deep Neural Networks (DNN): DNNs are employed to model and capture complex non-linear patterns and relationships within the dataset, enabling more accurate predictions. DNNs use multiple hidden layers to learn complex patterns between air pollutant levels and meteorological factors. DNNs effectively capture non-linear relationships in large datasets by utilizing historical data and backpropagation for optimization. Their ability to adapt to multi-dimensional data makes them particularly suited for predicting AQI by considering a variety of pollutant and weather variables.

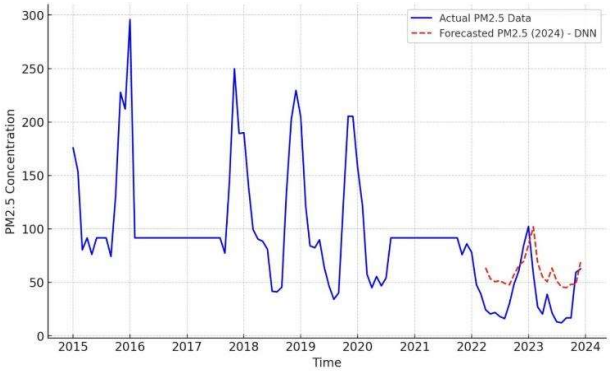


Fig 4: DNN Forecasting PM2.5

The graph illustrates the actual PM2.5 concentration from 2015 to 2023 and the forecasted values for 2024 using a DNN model.

2) Random Forest: Random Forest is a versatile ensemble learning technique that predicts AQI by merging outputs from numerous decision trees, ensuring precision and resilience. Each tree is trained on a randomly selected portion of the data, which helps capture diverse aspects of the underlying features. RF reduces overfitting and ensures reliable predictions by averaging the outcomes of individual trees. Random Forest is an ensemble learning method that aggregates the outputs of multiple decision trees to enhance prediction accuracy and reliability. Its ability to handle large datasets, manage missing values.

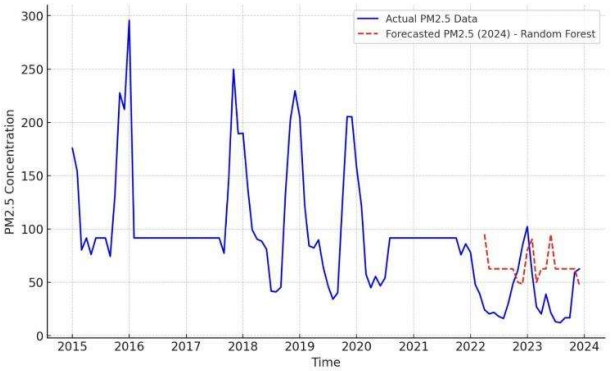


Fig 5: RF Forecasting PM2.5

In this graph, I have applied a Random Forest model to forecast PM2.5 levels for 2024. The graph shows the actual PM2.5 data in blue and the Random Forest forecast in red (dashed line).

3) ARIMA(Auto-Regressive Integrated Moving Average): ARIMA is a statistical method designed to handle univariate time series data. It models AQI by identifying trends and patterns in historical data, focusing on temporal autocorrelations. ARIMA is particularly useful for short term AQI forecasting. It is a statistical model best suited for analyzing and forecasting univariate time series data. It captures trends and temporal dependencies, making it effective for short-term AQI prediction in stable conditions.

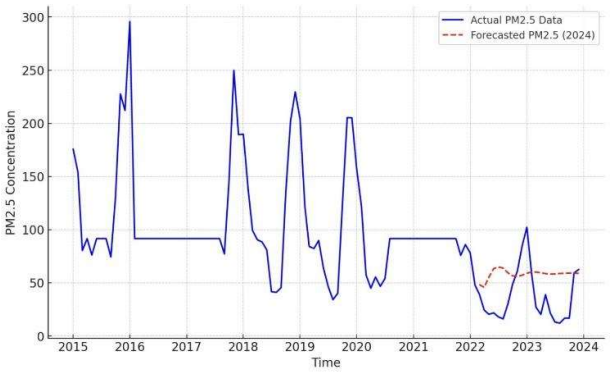


Fig 6: ARIMA Forecasting PM2.5

The graph illustrates the actual PM2.5 concentration from 2015 to 2023 and the forecasted values for 2024 using an ARIMA model.

4) SARIMA (Seasonal Auto-Regressive Integrated Moving Average): SARIMA extends ARIMA by incorporating seasonal components, making it ideal for handling AQI data with periodic variations. It captures long-term trends and seasonal fluctuations, ensuring accurate predictions during recurring pollution events. SARIMA expands on ARIMA by accounting for seasonal patterns, enabling it to model data with periodic fluctuations effectively. This makes it ideal for capturing seasonal fluctuations in AQI due to recurring events like crop burning or weather changes.

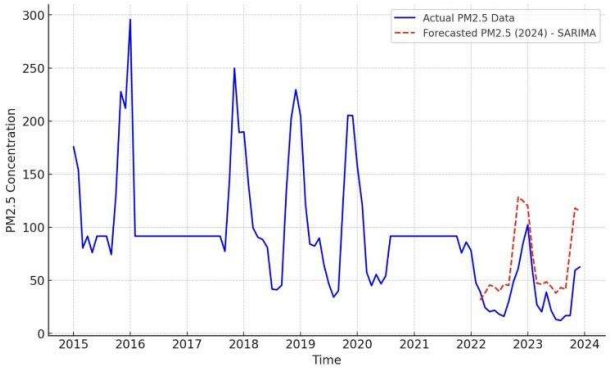


Fig 7: SARIMA Forecasting PM2.5

The graph illustrates the actual PM2.5 concentration from 2015 to 2023 and the forecasted values for 2024 using a Seasonal Autoregressive Integrated Moving Average (SARIMA) model.

III. RESULTS AND DISCUSSIONS

We evaluated the performance of these models on the dataset using the Coefficient of Determination and Mean Squared Error which have been shown along with accuracy rates. Performance measures for the DNN, RF, ARIMA, and SARIMA models can be generated using R^2. The Coefficient of Determination counts how many of the model's predictions were accurate and inaccurate in visualizing and evaluating the performance of a DNN, RF, ARIMA, and SARIMA model.

A. Performance measurements

R^2 (Coefficient of Determination) indicates the proportion of variance explained by the model.

MSE, or Mean Squared Error, quantifies the average squared deviation between predicted and actual values.

TABLE I: FOR BANGLADESH DATASET

Model Name	Training		Testing		Overall	
	R^2	MSE	R^2	MSE	R^2	MSE
DNN	-0.35	1787	-1.71	3466	-0.58	2123
RF	0.79	256.1	-1.92	3732	0.28	958.7
ARIMA	-0.24	1644	-0.55	1979	-0.28	1711
SARIMA	-0.54	2032	-0.44	1842	-0.49	1994

TABLE II : FOR INDIA DATASET

Model Name	Training		Testing		Overall	
	R^2	MSE	R^2	MSE	R^2	MSE
DNN	-0.46	3853	-17.01	10620	-0.75	5231
RF	0.94	131.6	-1.04	1203	0.88	349.8
ARIMA	0.33	1757	-0.82	1078	0.45	1619
SARIMA	0.27	1929	-0.93	1141	0.40	1768

TABLE III: COEFFICIENT OF DETERMINATION

Model Name	Bangladesh Dataset (R^2)	India Dataset(R^2)
DNN	-0.59	-0.75
RF	0.28	0.88
ARIMA	-0.55	0.45
SARIMA	-0.44	0.40

Random Forest performs the best, especially for India, where it achieves a high coefficient of determination is 0.88. In contrast, its performance for Bangladesh is lower, with a coefficient of determination is 0.28. The DNN model shows poor performance, with a coefficient of determination values of -0.59 for Bangladesh and -0.75 for India, indicating that it fails to capture the patterns in the data. The ARIMA model achieves a coefficient of determination of -0.55 for Bangladesh and 0.45 for India, suggesting it performs moderately better on the Indian dataset but still struggles with both regions. Similarly, SARIMA performs poorly with coefficient of determination values of -0.44 for Bangladesh and 0.40 for India. These results highlight the importance of model selection. Further optimization of the models and feature engineering may help improve their performance.

IV. CONCLUSION & FUTURE WORK

This study presents a comprehensive approach to modeling and predicting the Air Quality Index (AQI) for Bangladesh and India using machine learning algorithms, including Deep Neural Networks (DNN), Random Forest (RF), ARIMA, and SARIMA. Forecasting air quality is challenging because of the fluctuating, unpredictable, and highly variable nature of pollutants and particulates across both time and space.[3] The models capture complex relationships and seasonal variations by effectively preprocessing diverse air quality and meteorological data, enabling accurate AQI forecasting. The findings highlight the potential of these models to support air quality management, inform timely interventions, and improve public health. Future work could expand to additional metropolitan areas and explore the correlation between air quality and health impacts, leveraging advanced time-series models like SARIMA to address seasonal variations in pollution-related diseases.

Future research can focus on implementing advanced deep learning models, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), to capture long-term dependencies and temporal patterns in air quality data more effectively. Additionally, incorporating larger and more diverse datasets from multiple cities across Bangladesh and India can help improve the generalizability of the models. Future work could also focus on analyzing the correlation between air quality and public health outcomes to assess the health burden of air pollution.

V. REFERENCES

[1] Hossain, A., Bhuiya, R. A., & Ali, M. Z. (2022). Particulate matter forecasting in Bangladesh, India, China, and USA: A machine learning approach. LAP Lambert Academic Publishing. Republic of Moldova. ISBN: 978-620-5-50137-5.

[2] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020(1), 8049504.

[3] 3. Kang, G. K., Hoon, J., Kim, H., & Lee, J. (2018). Air quality prediction: Big data and machine learning approaches. International Journal of Environmental Science and Development, 9(1), 8-16.

[4] Sakib, S. R., Rahman, M., Mahmud, I., & Sayeed, A. (2023). Time series analysis and forecasting of air quality index of Dhaka city of Bangladesh. In 2023 world AI IoT Congress (AIIoT) (pp. 1-6). IEEE.

[5] Kumar A., & Goyal P. (2013). Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure and Applied Geophysics, 170, 711-722.

[6] Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports.

[7] Ramadan, M. S., Abuelgasim, A., & Al Hosani, N. (2024). Advancing air quality forecasting in Abu Dhabi, UAE using time series models. Frontiers in Environmental Science, 12, 1393878.

[8] Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air quality prediction using machine learning algorithms. International Journal of Computer Applications Technology and Research, 8(9), 367–370.

[9] Maltare, N., et al. (2023). Air quality index prediction using machine learning for Ahmedabad city. Digital Chemical Engineering.

[10] Haq, M. A. (2022). SMOTEDNN: A novel model for air pollution forecasting and AQI classification. Computers, Materials & Continua, 71(1).