



Research Article

Machine Learning-based Spatial Mapping of Urban-Rural Air Quality: A Case Study of Mymensingh District in Bangladesh

Maksudul Amin Talukdar¹, Ananya Roy^{1✉}, Mst. Tanjina Akter¹, Muhammad Ashik-E-Rabbani² and Md. Bashirul Islam²¹Department of Environmental Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh-2224²Department of Farm Power and Machinery, Bangladesh Agricultural University, Mymensingh-2202.

ARTICLE INFO

ABSTRACT

Article history

Received: 05 September 2026

Accepted: 27 September 2026

Published: 30 September 2026

Keywords

Air quality,
PCA,
Machine Learning,
ANN,
Urban and Rural

Correspondence

Ananya Roy

✉: ananyaroy058@gmail.com

Air pollution is one of the most urgent environmental and social health issues in Bangladesh's rapidly urbanizing areas, but secondary cities still lack fine-scale evaluation and forecasting. This paper presents a combined PCA-machine learning-GIS model to quantify, model, and map air quality in urban and rural settings of the Mymensingh district. Department of Environment (DoE) monitoring records provided concentrations of SO₂, NO₂, CO, PM₁₀, and PM_{2.5}, supplemented with primary field measurements from an Aeroqual Series 500 sensor. To minimize multicollinearity, we used Principal Component Analysis (PCA) to create an Air Quality Index (PCA-AQI) composite as the target variable for machine learning modeling. We trained and evaluated Random Forest, Support Vector Machine, XGBoost, and Artificial Neural Network (ANN) models using standard performance measures. The ANN had the highest forecasting accuracy ($R^2 = 0.998$), suggesting strong potential for capturing nonlinear interactions among pollutants. We used the primary datasets to train an ANN model optimized to predict PCA-AQI, and mapped the spatial distribution in ArcGIS using Inverse Distance Weighting (IDW) interpolation. Findings indicate that there are mostly poor air quality conditions in the urban areas, which are fueled by traffic and business operations, and the rural areas were generally good to moderate, with hotspots on highways and in markets. It is shown in the proposed framework that the fusion of PCA-based composite indices and machine learning can be effective in assessing air quality in data-scarce areas. The proposed framework shows that fusion of PCA-based composite indices with machine learning can effectively assess air quality management tool in the urban-rural setting and evidence-based environmental policy within developing countries.

Copyright ©2026 by authors and BAURES. This work is licensed under the Creative Commons Attribution International License (CC By 4.0).

Introduction

Air pollution is one of the most significant, urgent environmental and societal issues of the 21st century that is propelled by increased urbanization, the spread of industries, the increase in population, and motorization (Lelieveld et al., 2015). The World Health Organization (WHO, 2022) claims that about 99 percent of the world population dwells in regions where the quality of the air has surpassed the recommended levels, leading to over seven million untimely fatalities every year (Chen & Kan, 2008). The major sources of this burden are fine particulate matter (PM_{2.5} and PM₁₀) and gaseous pollutants (NO₂, SO₂, and CO), which lead to serious respiratory, cardiovascular, neurological, and developmental conditions (Cohen et al., 2017; Lelieveld et al., 2015). In addition to health effects, air pollution has a huge socio-economic cost in terms of higher

healthcare costs, lower work efficiency, and worsening living conditions (Mannocci et al., 2019).

Air pollution in South Asia and especially Bangladesh, is very (Jabbar et al., 2022; Begum et al., 2013). Dhaka has always been one of the most polluted cities in the world, with the levels of PM_{2.5} exceeding the WHO guideline values by a factor of over ten each year (T. Islam et al., 2023). The key sources of pollution are the brick kilns, motor vehicles, industry, construction dust, road dust re-suspended, and uncontrolled waste burning, as well as biomass burning and seasonal agricultural burning (Hime et al., 2018; Lelieveld et al., 2015). The Department of Environment (DoE) implemented the Clean Air and Sustainable Environment project for Continuous Air Monitoring Stations (CAMS) in major cities. However, the

Cite This Article

Talukdar, M.A., Roy, A., Akter, M.T., Rabbani, M.A.E. and Islam, M.B. 2026. Machine Learning-based Spatial Mapping of Urban-Rural Air Quality: A Case Study of Mymensingh District in Bangladesh. *Journal of Bangladesh Agricultural University*, 24(3): 284-294. <https://doi.org/10.3329/jbau.v24i3.93901>

monitoring infrastructure remains limited and poorly distributed (Ahmed et al., 2026). Therefore, fast-growing secondary cities like Mymensingh lack proper monitoring, which limits the analysis of air pollution at a micro-level, hotspot detection, and targeted mitigation (T. Islam et al., 2023).

Air pollution is a multi-pollutant phenomenon by nature, which is dictated by a complicated interplay between the sources of emissions, atmospheric chemistry, and the weather (Jolliffe & Cadima, 2016). The single-pollutant analysis and conventional Air Quality Index (AQI) formulations (which in most cases are based on the maximum sub-index approach) do not often reflect the compounding and synergistic actions of various pollutants (Horn & Dasgupta, 2024; Plaia & Ruggieri, 2011). More integrated representation data-driven composite indices, which are based on multivariate statistical methods, most commonly Principal Component Analysis (PCA), are computed by transforming correlated pollutant variables into orthogonal components and capturing most of the variance (Jolliffe & Cadima, 2016). The first principal component is often employed as a continuous, variance-based air quality measure, which allows for stronger trend analysis, spatial mapping, and predictive modeling (Isokääntä et al., 2020).

Machine learning (ML) methods in recent years have transformed the process of air quality assessment/forecasting by modeling nonlinear relationships and high-dimensional interactions between pollutants and environmental drivers effectively (Karmoude et al., 2025; Liang et al., 2020). The models that have shown higher performance compared to the traditional statistical models in predicting $PM_{2.5}$, PM_{10} and composite air quality measures include: the Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN) (Ayus et al., 2023; Shahriar et al., 2020; Zamani et al., 2019). Combining machine learning allows further stability and interpretability of the model by alleviating the presence of multicollinearity and dimensionality, which is particularly important in the case of highly correlated air pollutants (Jolliffe & Cadima, 2016).

Although there have been methodological improvements, the air quality studies in Bangladesh have mainly been on Dhaka and large industrial areas, and little work has been done on the secondary cities (Begum & Hopke, 2018; T. Islam et al., 2023; Rahman et al., 2021). Mymensingh, a developing divisional city, is experiencing increasing traffic density, commercial expansion, and urban growth, creating complex urban-rural air pollution gradients that are not adequately

represented by the existing monitoring network (T. Islam et al., 2023). Furthermore, a majority of the research done in Bangladesh is based on secondary data from DoE, and little to no combination of primary real-time sensor data is integrated. Joint application of low-cost sensors and official data is uncommon, which restricts resolution and the detection of the hotspots of micro-scale pollution (Shahriar et al., 2020).

Another important research gap is the limited application of PCA-based composite AQI, machine learning, and GIS-based spatial mapping within a single analytical framework. Not many studies have comparatively performed an analysis of RF, SVM, XGBoost, and ANN models based on the standard performance measures (MAE, RMSE, R^2) to determine which predictor is the most suitable in AQI prediction, especially in urban and rural conditions. Moreover, visualization of predicted distributions of AQI is frequently missing in GIS based format, limiting the transfer of modeling results into policy implications.

To fill these gaps, this research applied an integrated PCA machine learning-GIS framework to measure, forecast, and map air quality in both urban and rural areas within the Mymensingh district in Bangladesh. In this study, a PCA-derived composite Air Quality Index (PCA-AQI) has been developed from primary sensor readings and DoE monitoring data (SO_2 , NO_2 , CO, PM_{10} , and $PM_{2.5}$) using RF, SVM, XGBoost, and ANN models. The best-performing model is then used to create finer-scale spatial distributions of air quality through interpolation using GIS. The objective of this research is to develop a powerful and scalable machine-learning approach for air quality assessment, while predicting PCA-based AQI and spatially comparing air pollution across urban and rural environments in Mymensingh district. It assists in evidence-based environmental management and policymaking in fast-paced urban environments.

Methodology

Study Area

This study was conducted in urban (Mymensingh Sadar) and rural areas (Trishal and Muktagacha) of Mymensingh district, shown in Figure 1, located in north-central Bangladesh (24°38'N -25°12'N, 90°12'E - 90°45'E) (Mili et al., 2022).

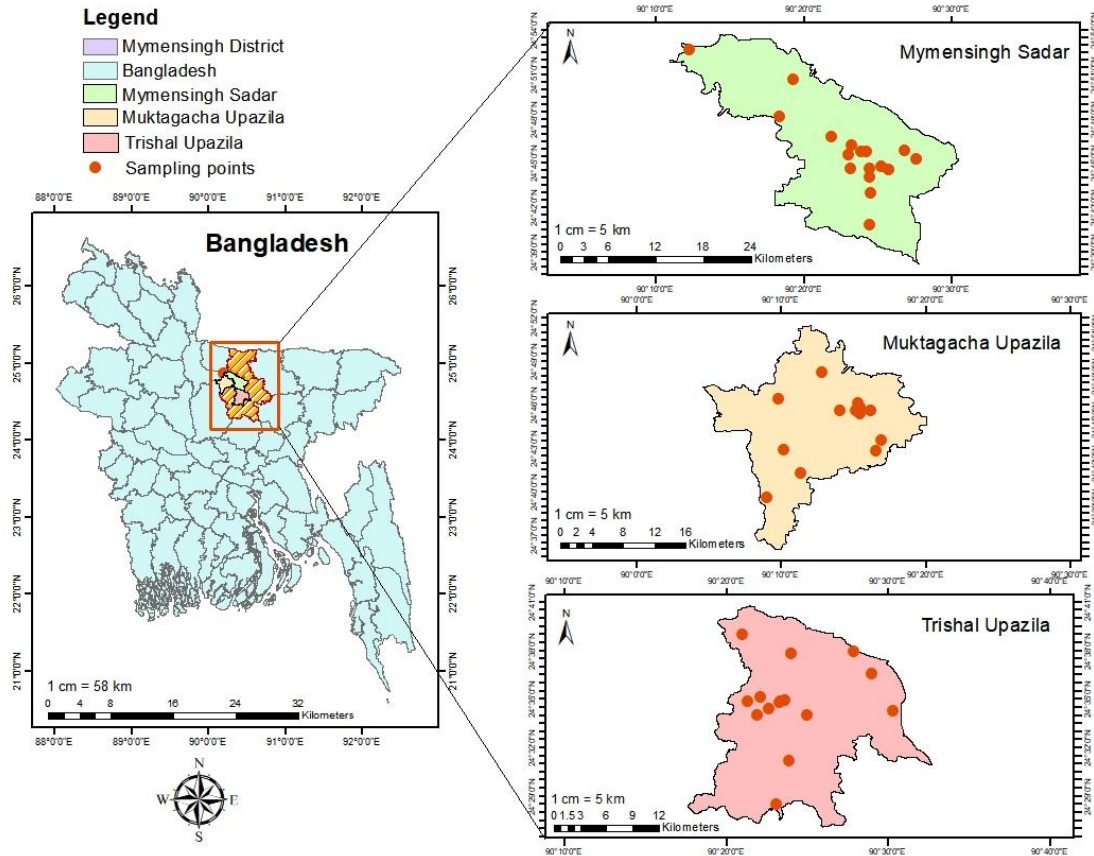


Fig. 1: Study Areas map along with sampling points

The climate there is subtropical monsoon, with a warm and wet monsoon (April- October) and a cold and dry winter (November- March). The amount of rainfall per year is 2000-2500 mm, and the temperature ranges between 12°C and 33°C in winter and summer (Mortuja & Uddin, 2021). Mymensingh comprises both urban and agriculture based rural areas located adjacent to each other, making it suitable for studying urban–rural air pollution differences (Hossain et al., 2025). Among the three selected location (Mymensingh City, Muktagacha and Trishal), the urban location is characterized by high population density, traffic congestion, small industries and commercial activities, whereas the rural locations are characterized by agricultural activities, biomass burning, brick kilns, and local transportation (Hossain et al., 2025; Sajan et al., 2017) These sites achieve a variety of pollution parameters (SO_2 , NO_2 , CO , PM_{10} , $\text{PM}_{2.5}$) due to emission and land use alteration (Guttikunda et al., 2014).

The hybrid methodological framework used secondary monitoring data (2022-2025) from the Department of Environment (DoE) along with field measurements to evaluate machine learning-based air quality in 2025. A

purposive site-selection approach was used to represent contrasting urban and rural air-quality environments. A calibrated Aeroqual Series 500 portable air-quality monitor was used to measure primary air-quality data at three predetermined sites, namely Mymensingh (urban), Muktagacha (rural-1), and Trishal (rural-2), selected based on degree of urbanization, population and human activity, traffic intensity and potential emission sources, surrounding land use, accessibility and safety, and suitability for repeated field measurements. The sampling frame for primary field measurements consisted of accessible monitoring points within the selected urban and rural environments of Mymensingh. Sampling was performed twice a day (day and night) to capture the day-day variability, with sample sizes of 17 at urban and 15 at each rural location. A total of 47 field observations represents repeated air-quality measurements at three selected monitoring sites rather than independent random samples.

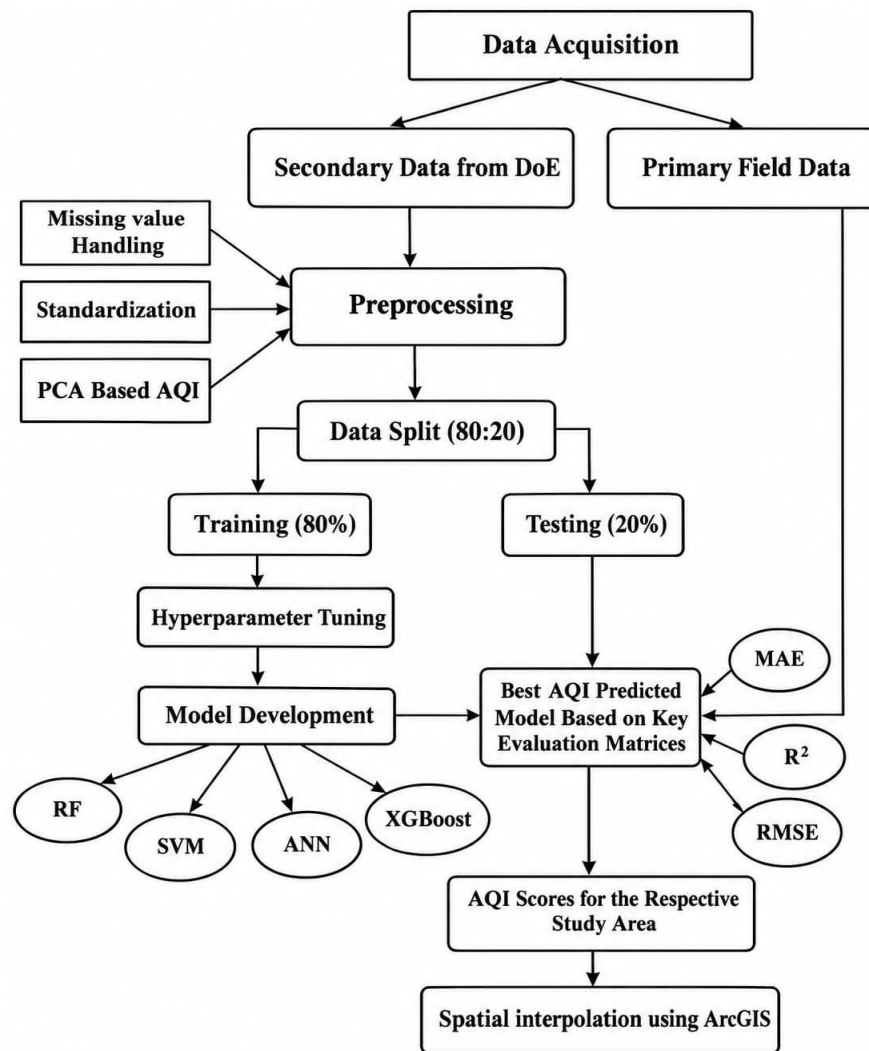


Fig. 2: Workflow of the study

Software and Computational Framework

All the computational analysis presented in Figure 2 was conducted using a Python virtual environment with the following libraries: pandas, numpy, matplotlib, scikit-learn, and joblib (for modelling). Also, ArcGIS 10.8 was used to perform spatial interpolation and mapping, and RStudio was used to visualize graphs and charts.

Data Preprocessing

Preprocessing of the data involved cleaning, harmonization, normalization, and standardization of the units (Kotsiantis et al., 2006). All gaseous pollutants were transformed to the form of ppm and particulate matter to $\mu\text{g m}^{-3}$ and standardized with the use of z-score normalization to eliminate the effect of the scale (Jain et al., 2005):

$$Z = \frac{X_i - \mu}{\sigma} \quad (1)$$

where Z = standardized variable, X_i = observed value, μ = mean of the variable, σ = standard deviation of the variable.

The last dataset included a total of five predictor variables that are SO_2 , NO_2 , CO , PM_{10} , and $\text{PM}_{2.5}$.

Principal Component Analysis (PCA) Computational Procedure

Principal Component Analysis (PCA) was performed for the transformation of correlated pollutants to a PCA-based Air Quality Index (PCA-AQI), which minimizes multicollinearity (Isokääntä et al., 2020). Correlation-based PCA was conducted on the standardized pollutant matrix of SO_2 , NO_2 , CO , PM_{10} , and $\text{PM}_{2.5}$. The variables were all standardized to z-scores to give equal weight to

all variables by making all the variables have a zero mean and unit variance. To assess interrelationships between pollutants, covariance and correlation matrices were calculated, and then eigen decomposition was done to obtain eigenvalues (λ_i) and eigenvectors (e_i). The eigenvector with the largest eigenvalue is denoted by PC_1 , which is the most dominant pattern of pollution.

Principal component scores were calculated (Choi et al., 2015) as:

$$PC_1 = e_1^T Z \quad (2)$$

where, e_1 = is the eigenvector of PC_1 and Z is the standardized pollutant matrix.

The sign of PC_1 was set to match the concentrations of $PM_{2.5}$ in order to make it interpretable, where the high value represented an indication of worse air quality. The scores on the PCA were then scaled to a 0-500 scale, equivalent to more traditional scales of AQI (Choi et al., 2015), as:

$$PCA \text{ Based AQI (scaled)} = 500 \times \frac{PC_1 - PC_{(min)}}{PC_{(max)} - PC_{(min)}} \quad (3)$$

The machine learning (ML) methods were used to simulate nonlinear correlations between air pollutant concentrations and meteorological factors and the calculated PCA-based Air Quality Index (PCA-AQI). To develop the four supervised ML algorithms, namely: Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), Python (v3.13) and scikit-learn and xgboost were used. PCA-AQI was the dependent (target) variable in all the models, and pollutant concentrations (SO_2 , NO_2 , CO , PM_{10} , $PM_{2.5}$)

Training and Validation Framework

The DoE dataset from 2022-2025 was split randomly into 80% training and 20% testing datasets. The coefficient of determination (R^2), root mean square error (RMSE) and minimum absolute error (MAE) were used to assess model performance. The optimized artificial neural network model was then used on the primary field dataset for AQI prediction.

Model Optimization and Accuracy Assessment

All the models were adjusted to provide the best generalization:

Machine Learning Modeling

Table 1: The best values of user-defined parameters of various regression models

Used Algorithm	User-defined Parameters
RF	Number of trees ($n_{estimators}$ = 100–500) and tree depth (max_depth = 5–15) tuned using grid search.
ANN	Hidden layer structure, learning rate, and epochs optimized through iterative experimentation (epochs = 200–500, learning rate = 0.001–0.01).
SVM	Regularization parameter (C) and kernel width (γ) optimized using cross-validation.
XGBoost	Regularization parameters (λ , α), learning rate (η = 0.05–0.3), and number of boosting rounds adjusted using early stopping.

Each model's predictive accuracy was measured based on the following performance indicators:

Mean Absolute Error (MAE)

MAE is a commonly used statistical metric that measures the average magnitude of prediction errors without considering their direction (Chai & Draxler, 2014; Willmott & Matsuura, 2005).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

where, n = Number of samples; y_i = Actual value; $\hat{y}_i$ = Predicted value

MAE denotes the mean size of errors regardless of the direction.

Root Mean Square Error (RMSE)

RMSE is applied to compare the overall model performance, measured by the square root of the average squared differences between actual and predicted values. It gives higher weight to larger errors and is sensitive to outliers (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

where, n = Number of samples; y_i = Actual value and $\hat{y}_i$ = Predicted value

Coefficient of Determination (R^2)

The coefficient of determination (R^2) evaluates how well the predicted values explain the variability of the observed data and indicates the goodness of fit of the model (Renaud & Victoria-Feser, 2010).

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (6)$$

where, n = Number of samples; y_i = Actual value; $\hat{y}_i$ = Predicted value and $\bar{y}$ = Mean of actual values
 R^2 measures the extent to which the model accounts for variation in the measured PCA_AQI.

Best Model Selection and Implementation

ANN & SVM models were always highly predictive ($R^2 > 0.98$) with low values of RMSE (Özüpak et al., 2025). RF was relatively poorer in performing in rural validation because there was poor variability in training data, whereas XGBoost was consistent in its work overall, whether in an urban or rural setting (Cengil, 2025; Sidhu et al., 2024). The optimized ANN model has been used for the prediction of AQI from measured data based on

the learned nonlinear correlations between pollutants and air quality (Liang et al., 2020).

Spatial Interpolation and Visualization

The AQI values predicted by ANN were interpolated by employing the IDW technique in ArcGIS 10.8 software for visualization of the spatial pattern of air quality in the study area. These maps were instrumental in highlighting pollution zones and spatial patterns of air quality between urban and rural areas.

Results

Machine Learning Models Performance

The PCA-based AQI was used as the target variable, and the predictive performance of SVM, Random Forest (RF), XGBoost, and Artificial Neural Network (ANN) models was determined. ANN discovers complex nonlinear relationships using interconnected neural networks while SVM-RBF recognize nonlinear correlations by converting input data into a higher-dimensional feature space. RF integrate projections from multiple decision trees using bootstrap sampling and random feature selection, whereas XGBoost develop decision trees to fix errors from previous trees through gradient boosting.

Table 2: Statistical performance of various models

Model	MAE	RMSE	R^2
RF	0.133	0.205	0.955
ANN	0.026	0.037	0.998
SVM	0.023	0.060	0.996
XGB	0.096	0.143	0.978

Scatter plots of observed and predicted variables of the PCA-AQI in Figure 3 indicate that all models may reflect the general trend of the air quality.

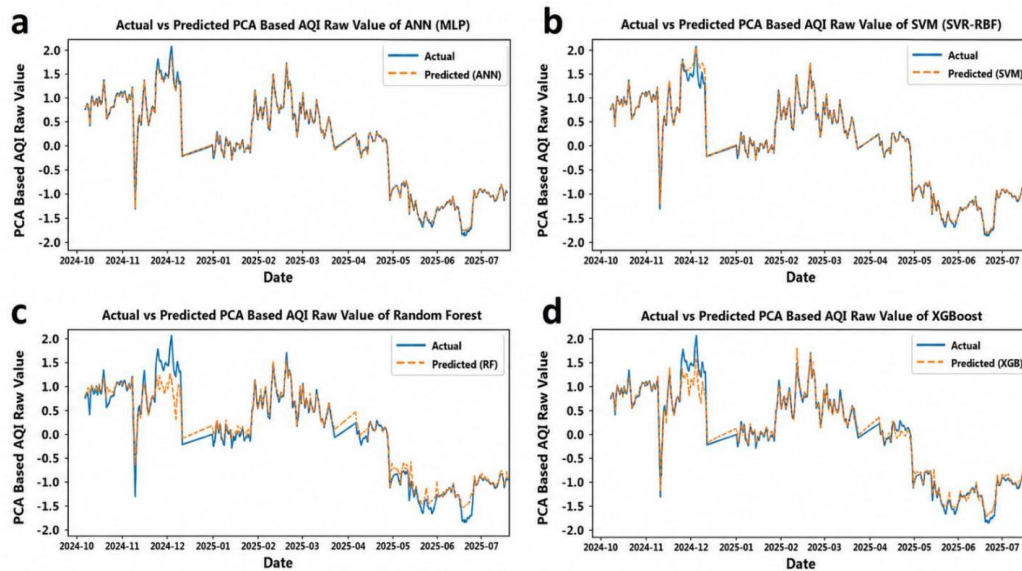


Fig. 3: Performance validation of ensemble learning models (a) ANN (b) SVM (SVR-RBF) (c) Random Forest, and (d) XGBoost in AQI estimation

The two models that exhibited the maximum degree of agreement between the observed and predicted values are ANN and SVM. According to table 2, ANN recorded the highest R^2 (0.998) and a low RMSE (0.037), while SVM achieved the lowest MAE (0.023) with an R^2 of 0.996. Both models showed close agreement between observed and predicted PCA-AQI values, with predictions clustered near the 1:1 line. XGB was good at predicting but with a higher dispersion at extreme values of AQI. Both ensemble-based models were comparatively more scattered, particularly at the high levels of pollution, suggesting that it was not strong in the high variability cases.

ANN-AQI prediction of primary field data

The tuned ANN was tested using field data collected from Mymensingh Sadar (urban), Muktagacha (rural), and Trishal (rural). It was found that there were significant differences in AQI in the three different

regions illustrated in Figure 4. The air quality was categorized based on the Air Quality Index (AQI) value into six distinct classifications: Good (0–50), Moderate (51–100), Unhealthy for Sensitive Groups (101–150), Unhealthy (151–200), Very Unhealthy (201–300), and Hazardous (301–500). AQI in Mymensingh Sadar is high, and some AQI values exceed 300 and fall into the Hazardous level which is prominently concentrated in the eastern urban center (Mymensingh City Corporation area), the southeastern industrial zones of Bhabkhali and Ghagra unions (traditional brick kilns and cluster of small manufacturing units) and Shambhuganj Bridge Moor and adjacent bypass intersections. On the other hand, AQI in Muktagacha and Trishal is relatively low, but unhealthy and hazardous AQI values have been reported from some regions. Hazardous and Very Unhealthy AQI values are reported from 35% and 18%, respectively, in Mymensingh, whereas the values are 8% and 23% in Muktagacha and 27% and 0% in Trishal.

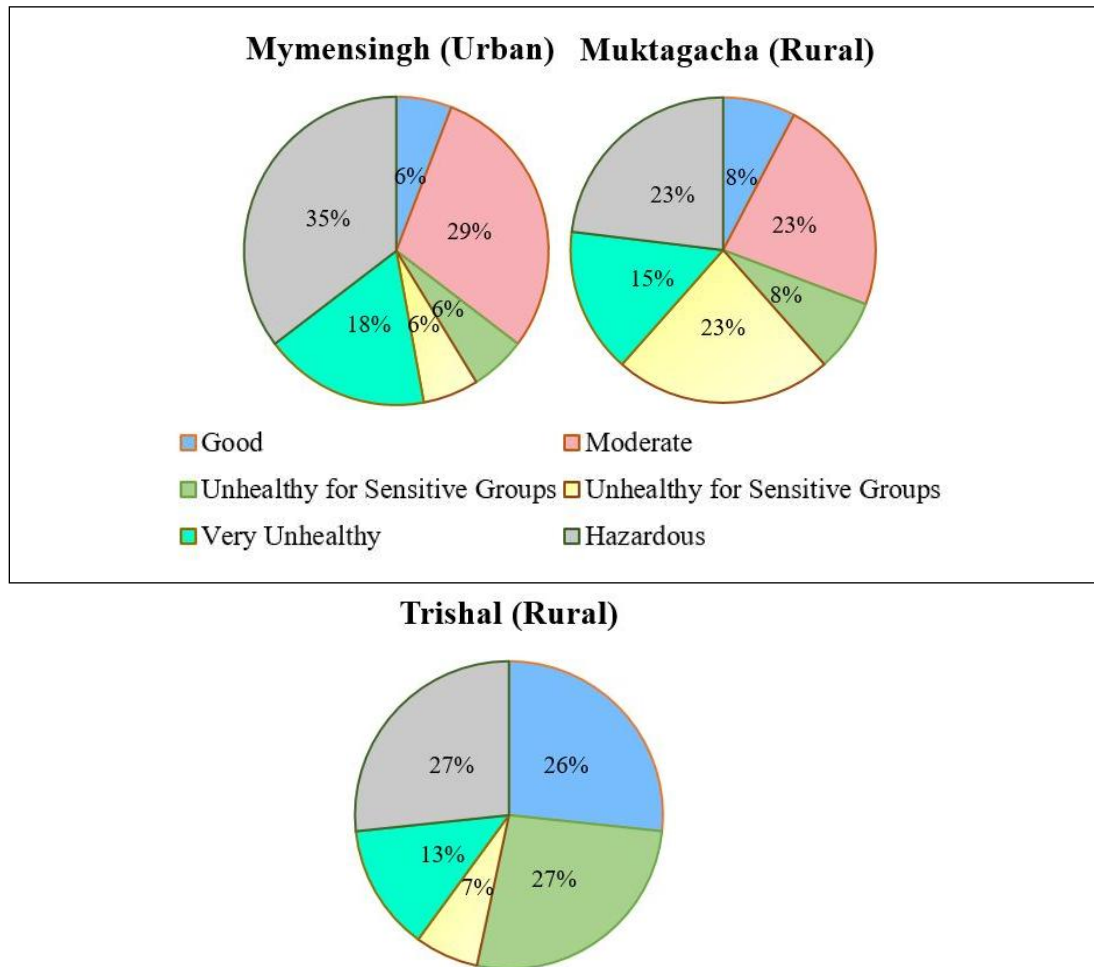


Fig. 4: The pie chart represents the Air Quality in three distinct Areas in Mymensingh district

Air Quality Comparison: Urban to Rural

A strong gradient of air quality between urban and rural areas has been seen in Figure 5. It exhibits the predicted AQI values for 47 sampling locations across Mymensingh Sadar (S1–S17), Muktagacha (S18–S32), and Trishal (S33–S47). The predicted AQI values display considerable variation among the sampling locations. The predicted AQI was relatively higher in Mymensingh Sadar, with the highest mean (219.20) and median (220.76), while Muktagacha and Trishal showed lower

mean and median values of 199.26 and 150.05, and 164.80 and 148.43, respectively. Mymensingh Sadar had the maximum value for the AQI score, where predicted values over 300 indicate very poor to hazardous air quality. On the other hand, most of the AQI scores in Muktagacha fell into the moderate category but sometimes exceeded 300, meaning higher levels of pollution. Thus, the urban area faced significantly worse air quality compared to rural areas.

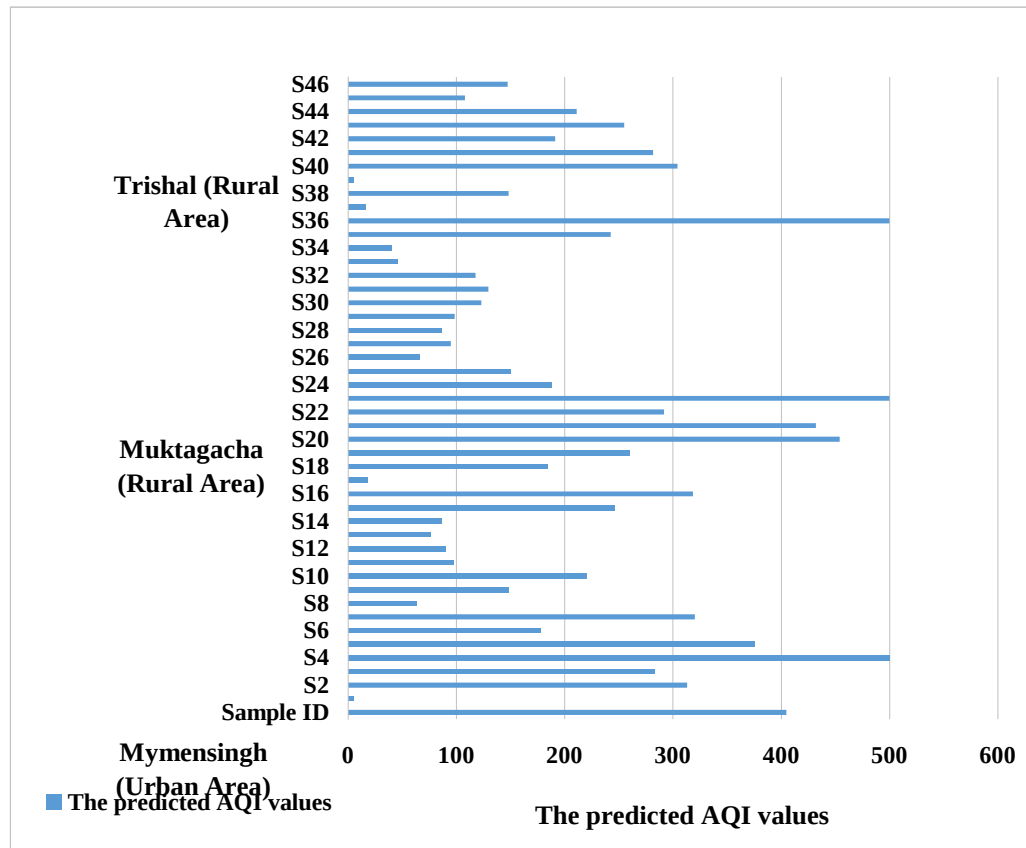


Fig. 5: The AQI Values of Urban and Rural Areas in Mymensingh district.

Spatial Distribution of PCA-AQI

The AQI values obtained from ANN predictions were subjected to spatial interpolation by the IDW technique, and it was found that the study area possessed great spatial heterogeneity in Figure 6. For Mymensingh Sadar, AQI values greater than 400 were mostly found in regions with high activity in the MCC zone such as Chorpara Medical Gate, Mashkanda-patgudam bus terminal, and Ganginarpar, where intense human and transport activities generate substantial emissions. Chorpara Medical Gate experiences frequent stop-and-go traffic, ambulance and patient-transport vehicle idling, and dense commercial activity, while Mashkanda–Patgudam is a central transit gateway

experience emissions of heavy black smoke and unburnt diesel particulates. Ganginarpar is characterized by gridlock of thousands of diesel-powered trucks, motorized rickshaws, auto-rickshaws, and non-motorized vehicles releasing raw, unburnt carbon. In the contrast AQI<150 were found in peripheral zones such as Parangonj and Begunbari. For Muktagacha, AQI values greater than 350 were mostly found in the vicinity of the Upazila Health Complex and Bus Stand, while those less than 350 were mostly found in and around Iswargram and Baliapara. For Trishal, AQI values were generally lower compared to the other two Upazilas, but there were hotspots with AQI values

greater than 300 found around Trishal Bazar and Boilor northern and western parts of Trishal Bazar. Lesser AQI values were mostly found in the

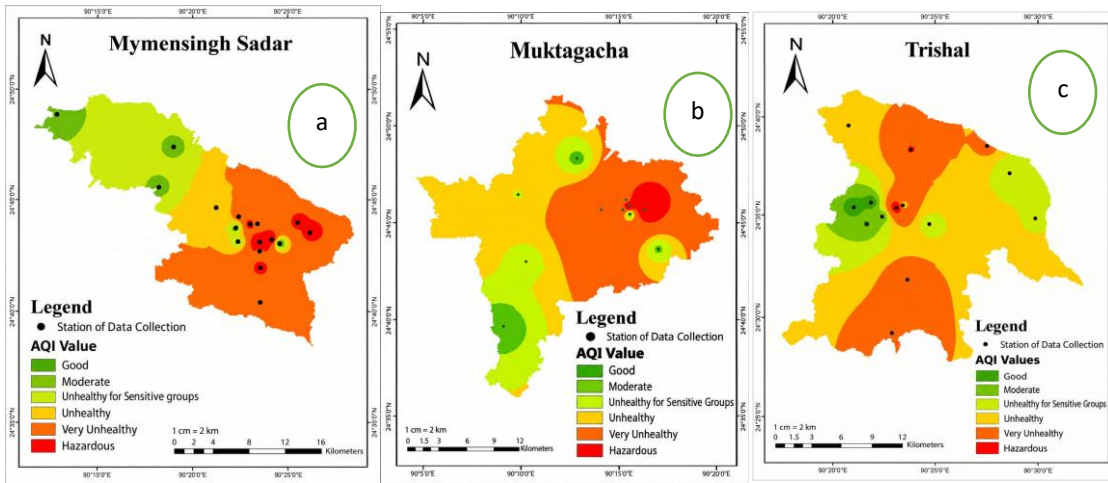


Fig. 6: Predicted AQI values of (a) Mymensingh Sadar, (b) Muktagacha and (c) Trishal

Discussion

High performance of ANN and SVM models proves their ability to represent complex and nonlinear correlations between air pollutants and the PCA-based AQI. In contrast to other models, the ANN model proved to be the most effective in terms of air quality prediction, because this model shows the highest value of R^2 (0.998) and the lowest value of RMSE (0.037).

A distinct urban-rural gradient in air quality was revealed, since Mymensingh had a considerably higher AQI compared to Muktagacha and Trishal. The high AQI value in urban settlements can be attributed to intensive traffic, commercial activity, construction dust, and local emissions. At the same time, rural areas have low emission intensity and large open space, which allow the dispersion of pollutants more easily. Besides that, spatial interpolation identified the persistent hotspots around traffic and commercial sites, such as Chorpara Medical Gate, Bus Terminal, Ganginarpar, Trishal Bazar, and Boilor Bazar, while the low AQI values were characteristic of peripheral, less developed areas.

These findings are in line with previous research on higher particulate pollution in urban Bangladesh and good predictability of PCA-based and machine learning models (Begum et al., 2011; Hossain et al. 2025). The findings emphasize the need for local air pollution management measures related to traffic regulation, restrictions on commercial activity and hotspot-based approaches. On the whole, the proposed PCA-ANN-GIS approach is promising in terms of applicability to air quality estimation and mapping in sparsely populated urban and rural regions.

Conclusion

The current study proposes an innovative PCA-machine learning-GIS framework to assess and visualize the air quality index in urban and rural locations of Mymensingh district in Bangladesh. Using PCA techniques, this research has efficiently integrated five pollutants, such as SO_2 , NO_2 , CO, PM_{10} , and $PM_{2.5}$, to formulate a new composite AQI value. Based on predictive capability, the artificial neural network model exhibited the highest accuracy with an R^2 of 0.998 and RMSE of 0.037. Results indicated a significant urban-rural gradient, where the highest average AQI was observed at Mymensingh Sadar, followed by Muktagacha and Trishal. High AQI (>400) concentrations were mostly distributed in and around Chorpara Medical Gate, Mushkanda-Patgudam Bus Terminal, and Ganginarpar owing to dense traffic emission sources and industrial activities. On the other hand, low AQI concentrations (<150) were mainly identified in Parangonj and Begunbari areas. Muktagacha witnessed pockets of high AQI concentrations (>350) in and around Upzila Health Complex and Bus Stand locations.

Spatial interpolation also indicated recovery of persistent pollution hotspots in traffic- and commercial-intensive areas, and rural areas had different but spatially distributed patterns of pollution. These findings are important in highlighting the prevailing influence of urbanization, motorization emissions, and high anthropogenic activity in influencing air quality deterioration in fast-growing secondary cities. This model is most appropriate for developing nations where there are few dense monitoring networks. The methodology can assist with evidence-based city

planning, specific emission regulation policies, and emphasis on pollution reduction measures. By applying this approach to other secondary cities, policymakers can improve the quality of the air, protect people's health, and lead to the development of sustainable cities.

Authors Contribution

Maksudul Amin Talukdar: Design, Methodology, Data collection, Data analysis, Writing. Ananya Roy: Design, Methodology, Demonstration, Reviewing Manuscript. Mst. Tanjina Akter: Data Collection, Data Processing, Data Analysis, Writing. Muhammad Ashik-E-Rabbani: Supervision, Reviewing Manuscript. Md. Bashirul Islam: Mapping, Data Analysis, Writing.

Reference

- Abdul Jabbar, S., Tul Qadar, L., Ghafoor, S., Rasheed, L., Sarfraz, Z., Sarfraz, A., Sarfraz, M., Felix, M., & Cherrez-Ojeda, I. (2022). Air Quality, Pollution and Sustainability Trends in South Asia: A Population-Based Study. *International Journal of Environmental Research and Public Health*, 19(12), 7534. <https://doi.org/10.3390/ijerph19127534>
- Ahmed, S., Pasha, A., Kumer, T., Akhtaruzzaman, M., & Hossain, M. (2026). Advancements in Air Quality Monitoring Systems: A Comprehensive Review of Emerging Technologies for Enhancing Environmental Health. *Indoor Air*, 2026(1), 3080684. <https://doi.org/10.1155/ina/3080684>
- Ayus, I., Natarajan, N., & Gupta, D. (2023). Comparison of machine learning and deep learning techniques for the prediction of air pollution: A case study from China. *Asian Journal of Atmospheric Environment*, 17(1), 4. <https://doi.org/10.1007/s44273-023-00005-w>
- Begum, B. A., Biswas, S. K., & Hopke, P. K. (2011). Key issues in controlling air pollutants in Dhaka, Bangladesh. *Atmospheric Environment*, 45(40), 7705–7713. <https://doi.org/10.1016/j.atmosenv.2010.10.022>
- Begum, B. A., & Hopke, P. K. (2018). Ambient Air Quality in Dhaka Bangladesh over Two Decades: Impacts of Policy on Air Quality. *Aerosol and Air Quality Research*, 18(7), 1910–1920. <https://doi.org/10.4209/aaqr.2017.11.0465>
- Begum, B. A., Hopke, P. K., & Markwitz, A. (2013). Air pollution by fine particulate matter in Bangladesh. *Atmospheric Pollution Research*, 4(1), 75–86. <https://doi.org/10.5094/APR.2013.008>
- Cengil, E. (2025). The Power of Machine Learning Methods and PSO in Air Quality Prediction. *Applied Sciences*, 15(5), 2546. <https://doi.org/10.3390/app15052546>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, B., & Kan, H. (2008). Air pollution and population health: A global challenge. *Environmental Health and Preventive Medicine*, 13(2), 94–101. <https://doi.org/10.1007/s12199-007-0018-5>
- Choi, J., Park, Y. S., & Park, J. D. (2015). Development of an Aggregate Air Quality Index Using a PCA-Based Method: A Case Study of the US Transportation Sector. *American Journal of Industrial and Business Management*, 05(02), 53–65. <https://doi.org/10.4236/ajibm.2015.52007>
- Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., ... Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082), 1907–1918. [https://doi.org/10.1016/S0140-6736\(17\)30505-6](https://doi.org/10.1016/S0140-6736(17)30505-6)
- Guttikunda, S. K., Goel, R., & Pant, P. (2014). Nature of air pollution, emission sources, and management in the Indian cities. *Atmospheric Environment*, 95, 501–510. <https://doi.org/10.1016/j.atmosenv.2014.07.006>
- Hime, N. J., Marks, G. B., & Cowie, C. T. (2018). A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution from Five Emission Sources. *International Journal of Environmental Research and Public Health*, 15(6), 1206. <https://doi.org/10.3390/ijerph15061206>
- Horn, S. A., & Dasgupta, P. K. (2024). The Air Quality Index (AQI) in historical and analytical perspective a tutorial review. *Talanta*, 267, 125260. <https://doi.org/10.1016/j.talanta.2023.125260>
- Hossain, Md. M., Islam, Md. T., Sikder, S. K., Hemstock, S. L., Islam, Md. A., Faruquee, M. H., & Hossain, Md. Z. (2025). The urban environment in South Asia: Studying the ambient air quality in a mid-sized city in Bangladesh. *Frontiers in Sustainable Cities*, 7, 1497768. <https://doi.org/10.3389/frsc.2025.1497768>
- Islam, A. R. Md. T., Al Awadh, M., Mallick, J., Pal, S. C., Chakraborty, R., Fattah, Md. A., Ghose, B., Kakoli, Most. K. A., Islam, Md. A., Naqvi, H. R., Bilal, M., & Elbeltagi, A. (2023). Estimating ground-level PM2.5 using subset regression model and machine learning algorithms in Asian megacity, Dhaka, Bangladesh. *Air Quality, Atmosphere & Health*, 16(6), 1117–1139. <https://doi.org/10.1007/s11869-023-01329-w>
- Isokääntä, S., Kari, E., Buchholz, A., Hao, L., Schobesberger, S., Virtanen, A., & Mikkonen, S. (2020). Comparison of dimension reduction techniques in the analysis of mass spectrometry data. *Atmospheric Measurement Techniques*, 13(6), 2995–3022. <https://doi.org/10.5194/amt-13-2995-2020>
- Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12), 2270–2285. <https://doi.org/10.1016/j.patcog.2005.01.012>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Karmoude, M., Munhungwarwa, B., Chiraira, I., Mckenzie, R., Kong, J., Smith, B., Ayana, G., Njara, N., Mathaha, T., Kumar, M., & Mellado, B. (2025). Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks. *Science of The Total Environment*, 1002, 180593. <https://doi.org/10.1016/j.scitotenv.2025.180593>
- Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data Preprocessing for Supervised Learning. 1(1).
- Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569), 367–371. <https://doi.org/10.1038/nature15371>
- Liang, Y.-C., Maimury, Y., Chen, A. H.-L., & Juarez, J. R. C. (2020a). Machine Learning-Based Prediction of Air Quality. *Applied Sciences*, 10(24), 9151. <https://doi.org/10.3390/app10249151>
- Mannocci, A., Ciarlo, I., D'Egidio, V., Del Cimmuto, A., De Giusti, M., Villari, P., & La Torre, G. (2019). Socioeconomic Deprivation Status and Air Pollution by PM10 and NO2: An Assessment at Municipal Level of 11 Years in Italy. *Journal of Environmental and Public Health*, 2019, 1–11.

- <https://doi.org/10.1155/2019/2058467>
- Mili, S. A., Islam, Md. S., Al Momen Sabuj, A., Haque, Z. F., Pongit, A., Hossain, Md. G., Hassan, J., & Saha, S. (2022). A Cross-Sectional Seroepidemiological Study on Infectious Bursal Disease in Backyard Chickens in the Mymensingh District of Bangladesh. *Veterinary Medicine International*, 2022, 1–8. <https://doi.org/10.1155/2022/9076755>
- Mortuja, M., & Uddin, M. (2021). Impact of Climatic Variables on Major Field Crops in Char Areas of Mymensingh District. *Journal of Environmental Science and Natural Resources*, 12(1–2), 51–55. <https://doi.org/10.3329/jesnr.v12i1-2.51985>
- Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, 236(7), 464. <https://doi.org/10.1007/s11270-025-08122-8>
- Plaia, A., & Ruggieri, M. (2011). Air quality indices: A review. *Reviews in Environmental Science and Bio/Technology*, 10(2), 165–179. <https://doi.org/10.1007/s11157-010-9227-2>
- Rahman, Md. S., Azad, Md. A. K., Hasanuzzaman, Md., Salam, R., Islam, A. R. Md. T., Rahman, Md. M., & Hoque, M. Md. M. (2021). How air quality and COVID-19 transmission change under different lockdown scenarios? A case from Dhaka city, Bangladesh. *Science of The Total Environment*, 762, 143161. <https://doi.org/10.1016/j.scitotenv.2020.143161>
- Renaud, O., & Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140(7), 1852–1862. <https://doi.org/10.1016/j.jspi.2010.01.008>
- Sajan, D., Md., S. Q. H., Rumana, A., Sumaiya, H., Sumana, K., Md., Z. H. G., & Mohammad, S. (2017). Socioeconomic conditions and health hazards of brick field workers: A case study of Mymensingh brick industrial area of Bangladesh. *Journal of Public Health and Epidemiology*, 9(7), 198–205. <https://doi.org/10.5897/JPHE2017.0927>
- Shahriar, S. A., Kayes, I., Hasan, K., Salam, M. A., & Chowdhury, S. (2020). Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh. *Air Quality, Atmosphere & Health*, 13(10), 1247–1256. <https://doi.org/10.1007/s11869-020-00878-8>
- Sidhu, K. K., Balogun, H., & Oseni, K. O. (2024). Predictive Modelling of Air Quality Index (AQI) Across Diverse Cities and States of India using Machine Learning: Investigating the Influence of Punjab's Stubble Burning on AQI Variability. *International Journal of Managing Information Technology*, 16(1), 15–35. <https://doi.org/10.5121/ijmit.2024.16102>
- Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. <https://doi.org/10.3354/cr030079>
- Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, 10(7), 373. <https://doi.org/10.3390/atmos10070373>