Air Quality and Pollution Assessment Using Advanced Machine Learning Techniques

Md. Ashikuzzaman Ashik
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mdashikuzzaman004@gmail.com

Zakirul Islam
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
zakirul011@gmail.com

MD Sabbir Ahammed
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mdsa134867@gmail.com

Sakib Imtiaz
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
sakibimtiaz1998@gmail.com

Md. Musfiqur Rahman Mridha
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mdmridha100730@gmail.com

Md. Shahid Ahammed Shakil
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
shakil.ahammed074@gmail.com

***Abstract*—Air pollution is a critical environmental and public health issue, with significant impacts on human well-being and ecosystems. Accurately assessing air quality is essential for effective pollution mitigation and policymaking. This study proposes a machine learning-based approach to classify air quality levels using six classifiers: Logistic Regression, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, Random Forest, and K-Nearest Neighbors (KNN). The dataset was preprocessed through feature scaling, label encoding, and one-hot encoding, followed by hyperparameter optimization using grid search. Model performance was evaluated using metrics such as accuracy, precision, recall, F1 score, confusion matrices, and AUC values. Gradient Boosting emerged as the best-performing model with an accuracy of 96% and balanced performance across all metrics. The results highlight the effectiveness of machine learning methods in air quality classification tasks. Future work will explore incorporating additional features, advanced deep learning techniques, and real-time deployment for enhanced air quality monitoring and decision-making systems.**

***Index Terms*—Air Pollution, Machine Learning, Gradient Boosting, Classification Models, Environmental Monitoring**

I. Introduction

Air pollution is the contamination of the indoor or outdoor environment by chemical, physical, or biological agents that alter the natural characteristics of the atmosphere. [1] It is caused by pollutants such as particulate matter, carbon monoxide, sulfur dioxide, nitrogen oxides, and ozone, which are emitted from sources like motor vehicles, industrial facilities, and household combustion devices. Air pollution poses severe threats to human health, contributing to respiratory diseases, cardiovascular conditions, and premature deaths. According to the World Health Organization (WHO), nearly 99% of the global population breathes air that exceeds recommended pollution limits, [2] with low- and middle-income countries suffering the highest exposure levels 13. Furthermore, air pollution significantly impacts ecosystems and contributes to climate change through greenhouse gas emissions.

Given its widespread health and environmental consequences, accurately assessing air quality is essential for effective mitigation strategies and policymaking. Traditional methods for air quality monitoring often rely on expensive equipment and manual data analysis, which limits scalability and efficiency. Recent advancements in machine learning (ML) provide an opportunity to address these challenges by enabling automated, scalable, and accurate air quality classification.

This study explores the application of six machine learning classifiers—Logistic Regression, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, Random Forest, and K-Nearest Neighbors (KNN)—to classify air quality levels. The dataset was preprocessed through feature scaling, label encoding, and one-hot encoding to ensure consistency. Hyperparameter optimization using grid search was employed to enhance model performance. The models were evaluated using metrics such as accuracy, precision, recall, F1 score, confusion matrices, and ROC-AUC curves.

The primary objective of this research is to identify the most effective machine learning model for air quality classification while highlighting the strengths and limitations of each approach. The findings aim to contribute to the development of efficient air quality monitoring systems that can support real-time decision-making for pollution control.

II. Related Works

Mauro Castelli et al. [3] employed Support Vector Regression (SVR) with radial basis function kernels to predict pollutant levels and AQI categories in California, achieving an accuracy of 94.1%.

Farzaneh Mohammadi et al. [4] applied machine learning models like ANN, Random Forest, and SVM for PM2.5 prediction in Isfahan, Iran using meteorological datasets. ANN achieved the highest accuracy of 90.1%, followed by Random Forest at 86.1%.

N. Srinivasa Gupta et al. [5] applied Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression to predict AQI in Indian cities. The highest accuracy achieved was 90.97% for Kolkata, highlighting the effectiveness of regression models for AQI prediction.

Samayan Bhattacharya et al. [6] used a Support Vector Regression (SVR) model with a Radial Basis Function (RBF) kernel to predict pollutant levels and the Air Quality Index (AQI), achieving 93.4% accuracy.

SK Natarajan et al. [7] combined Grey Wolf Optimization (GWO) with Decision Tree regression to predict AQI in Indian cities, achieving a maximum accuracy of 94.48%.

AH Almaliki et al. [8] introduced machine learning models such as Exponential Boosted Adaptive Trees (EBAT) for AQI prediction in Makkah, achieving a maximum accuracy of 94.8%.

III. Methodology

This study focuses on assessing air quality and pollution using machine learning (ML) techniques. The methodology involves data preprocessing, model training, hyperparameter optimization, evaluation of multiple ML classifiers, etc. as shown in Fig. 1 The steps are detailed as follows:

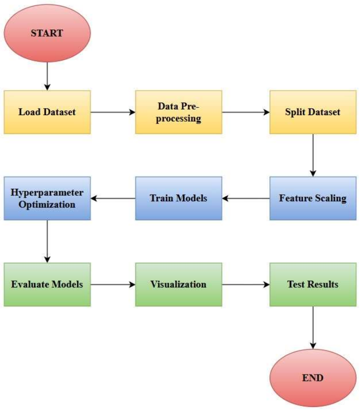

Fig. 1. Flowchart of the Overall Experiment.

A. Dataset and Preprocessing

The dataset Air Quality and Pollution Assessment used in this study, sourced from Kaggle, focuses on air quality assessment across various regions and contains 5,000 samples. It captures critical environmental and demographic factors influencing pollution levels. The key features are shown in Table I to provide a comprehensive representation of the factors affecting air quality.

This dataset was preprocessed to encode the target variable into numerical labels for machine learning models, where ”Good” was encoded as 0, ”Moderate” as 1, ”Poor” as 2, and ”Hazardous” as 3. Key preprocessing steps included:

- Label Encoding: The target variable, *Air Quality*, was encoded into binary classes using the LabelEncoder function.
- Handling Categorical Variables: Categorical features were converted into one-hot encoded variables using the get_dummies function.
- Data Splitting: The dataset was split into training, validation, and test sets in a 60:20:20 ratio using the train_test_split function.
- Feature Scaling: Features were standardized using StandardScaler to ensure uniform scaling across all input variables.

TABLE I
Environmental Metrics and Demographic Insights for Predicting Air Quality

Key Features	**Description**
Temperature (°C)	Average temperature of the region.
Humidity (%)	Relative humidity recorded in the region.
PM2.5 Concentration (µg/m^3)	Fine particulate matter levels.
PM10 Concentration (µg/m^3)	Coarse particulate matter levels.
NO_2 Concentration (ppb)	Nitrogen dioxide levels.
SO_2 Concentration (ppb)	Sulfur dioxide levels.
CO Concentration (ppm)	Carbon monoxide levels.
Proximity to Industrial Areas (km)	Distance to the nearest industrial zone.
Population Density (people/km^2)	Number of people per square kilometer in the region.

B. Machine Learning Models

Six machine learning classifiers were employed to predict air quality classes: Logistic Regression, Support Vector Machine (SVM), Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier and K-Nearest Neighbors (KNN). These models were implemented using the Scikit-learn library.

To optimize model performance, a grid search with cross-validation (GridSearchCV) was conducted for each classifier.

The best hyperparameters were selected based on accuracy scores obtained during cross-validation.

C. Model Evaluation

The trained models were evaluated on the validation and test datasets using two metrics:

1) Performance Metrics: The experiment was evaluated using important metrics like accuracy, precision, recall, f-1 score and AUC values.
2) Curves: This study was further evaluated using confusion matrix of each model, training and validation accuracy graph and ROC curve.

IV. Results and Analysis

This section presents the results of the machine learning models used for air quality and pollution assessment. The performance of each classifier is analyzed using metrics such as accuracy, ROC-AUC scores, and confusion matrices. Additionally, visualizations are provided to compare the models' effectiveness.

The class-wise performance of Gradient Boosting algorithm is shown in Table II.

TABLE II
CLASS-WISE PERFORMANCE METRICS FOR GRADIENT BOOSTING ALGORITHM

Algorithm	Class	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0 (Good)	96%	1.00	1.00	1.00
	1 (Moderate)		0.90	0.85	0.87
	2 (Poor)		0.96	0.97	0.97
	3 (Hazardous)		0.89	0.90	0.90

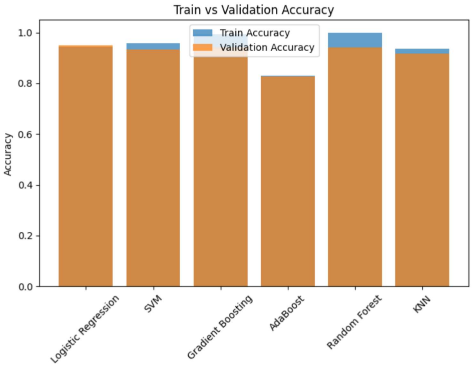

Fig. 2. Train vs Validation Accuracy

A. Train vs Validation Accuracy Analysis

This bar chart in Fig. 2 compares the training and validation accuracies of all classifiers. The blue bars represent training accuracy, while the orange bars represent validation accuracy for each model. This visualization helps assess whether a model is overfitting (high training accuracy but low validation accuracy) or generalizing well (similar training and validation accuracies). Most models exhibit comparable accuracies, suggesting good generalization.

B. Confusion Matrix Analysis

The confusion matrices for selected models are presented in the following figures. These matrices provide insight into the performance of each model.

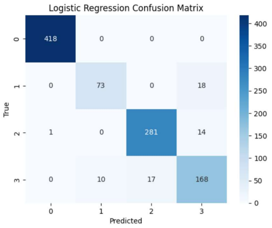

Fig. 3. Logistic Regression Confusion Matrix

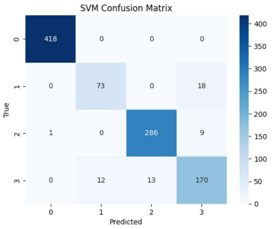

Fig. 4. SVM Confusion Matrix

Fig. 3 illustrates the performance of the Logistic Regression model on the test dataset. The diagonal cells represent correct predictions, while off-diagonal cells indicate misclassifications. The model performs well for class 0 and class 2, with minimal misclassifications. However, there are some errors in predicting classes 1 and 3, as seen in the off-diagonal cells.

The confusion matrix (Fig. 4) for the SVM classifier indicates strong performance, particularly for class 0 and class 2. The model has fewer misclassifications compared to Logistic Regression, especially for class 2. However, some errors remain in predicting classes 1 and 3.

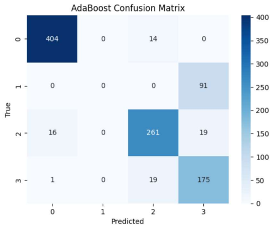

Fig. 5. AdaBoost Confusion Matrix

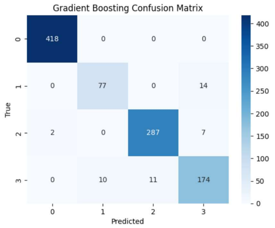

Fig. 6. Gradient Boosting Confusion Matrix

The AdaBoost classifier (Fig. 5) shows good performance for classes 0 and 3 but struggles more with class 2, as evidenced by higher misclassification rates. The off-diagonal values indicate that some samples from class 2 are misclassified into other classes.

The Gradient Boosting classifier (Fig. 6) demonstrates excellent performance across all classes, with minimal misclassifications. Class 0 is predicted perfectly, while classes 1 and 3 show slight errors. This model achieves better balance across all classes compared to AdaBoost.

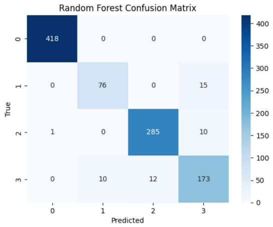

Fig. 7. Random Forest Confusion Matrix

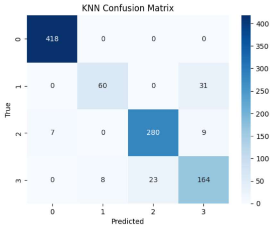

Fig. 8. KNN Confusion Matrix

The Random Forest classifier (Fig. 7) demonstrates strong performance across all classes. Class 0 is perfectly predicted with no misclassifications. For class 1, a small number of samples are misclassified into class 3. Class 2 has minimal errors, with only a few samples misclassified into classes 0 and 3. Similarly, class 3 shows slight misclassifications into class 2.

Fig. 8 illustrates the performance of the K-Nearest Neighbors (KNN) classifier. The model performs exceptionally well for class 0, with all predictions correct. However, for class 1, there are notable misclassifications, with some samples being predicted as class 3. Class 2 is predicted reasonably well, but a few samples are misclassified into classes 0 and 3. Class 3 also sees some misclassifications into class 2.

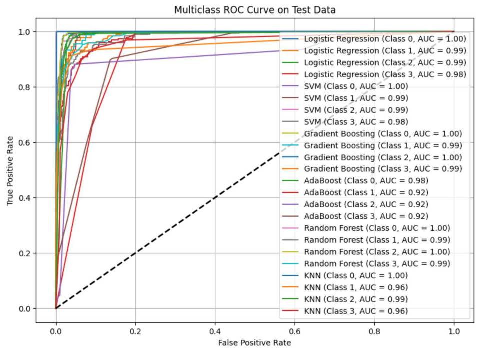

Fig. 9. Multiclass ROC Curve on Test Data

C. *Multiclass ROC Curve on Test Data*

Fig. 9 illustrates the Receiver Operating Characteristic (ROC) curves for all classifiers across multiple classes on the test dataset. Each classifier's performance is represented by separate curves for each class, with their respective Area Under the Curve (AUC) values displayed in the legend. The diagonal dashed line represents a random classifier's performance. The closer a curve is to the top-left corner, the better the model's performance. The figure demonstrates that most models achieve high AUC scores, indicating strong classification capabilities.

D. *Overall Performance*

Table III presents the performance metrics—accuracy, precision, recall, and F1 score—of various machine learning algorithms.

TABLE III
PERFORMANCE METRICS OF ALGORITHMS

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
LR	94%	0.92	0.90	0.91	0.99
SVM	95%	0,92	0.91	0.91	0.99
AB	84%	0.62	0.69	0.64	0.93
GB	96%	0.93	0.93	0.93	0.99
RF	95%	0.93	0.92	0.92	099
KNN	92%	0.90	0.86	0.88	0.97

Among them, Gradient Boosting stands out as the best-performing model, achieving the highest accuracy (96%) and consistently superior precision, recall, and F1 score (all 0.93), highlighting its effectiveness in classifying air quality.

TABLE IV
PERFORMANCE COMPARISON WITH RELATED WORKS

Models	**Accuracy (%)**
SVR [3]	94.1
ANN, RF, SVM [4]	90.1
SVR, RFR, CatBoost Regression [5]	90.97
SVR [6]	93.4
GWO, DTR [7]	94.48
EBAT [8]	94.8
Gradient Boosting (Our Model)	**96**

Table IV shows the comparison of our results compared to related studies of recent experiments.

V. CONCLUSION AND FUTURE WORKS

This study presented a machine learning-based approach for air quality and pollution assessment using six classifiers: Logistic Regression, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, Random Forest, and K-Nearest Neighbors (KNN). The evaluation metrics, including accuracy, precision, recall, F1 score, and confusion matrices, revealed that ensemble methods such as Gradient Boosting and Random Forest outperformed other models. Gradient Boosting emerged as the best-performing model with an accuracy of 96% and balanced precision, recall, and F1 scores across all classes. Random Forest followed closely with comparable results. Logistic Regression and SVM also demonstrated strong performance but exhibited slight weaknesses in distinguishing between certain classes. KNN provided good accuracy but struggled with inter-class misclassifications, while AdaBoost showed the lowest performance among the models. Overall, the results highlight the effectiveness of machine learning techniques for robust air quality classification.

Future research can focus on enhancing model performance by incorporating additional features such as meteorological data or pollutant-specific concentrations. Advanced deep learning techniques like convolutional or recurrent neural networks could also be explored for more complex datasets. Additionally, deploying these models in real-time air quality monitoring systems can provide actionable insights for pollution mitigation.

REFERENCES

[1] World Health Organization, "Air pollution," 2025, accessed: January 08, 2025. [Online]. Available: https://www.afro.who.int/node/5526

[2] ——, "Billions of people still breathe unhealthy air: New who data," 2022, accessed: January 08, 2025. [Online]. Available: https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data

[3] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in california," *Complexity*, vol. 2020, no. 1, p. 8049504, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/8049504

[4] F. Mohammadi, H. Teiri, Y. Hajizadeh, A. Abdolahnejad, and A. Ebrahimi, "Prediction of atmospheric pm2.5 level by machine learning techniques in isfahan, iran," *Scientific Reports*, vol. 14, no. 1, p. 2109, Jan. 2024, © 2024. The Author(s).

[5] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of air quality index using machine learning techniques: a comparative analysis," *Journal of Environmental and Public Health*, vol. 2023, no. 1, p. 4916267, 2023.

[6] S. Bhattacharya and S. Shahnawaz, "Using machine learning to predict air quality index in new delhi," *arXiv preprint arXiv:2112.05753*, 2021.

[7] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in india," *Scientific Reports*, vol. 14, no. 1, p. 6795, 2024.

[8] A. H. Almaliki, A. Derdour, and E. Ali, "Air quality index (aqi) prediction in holy makkah based on machine learning methods," *Sustainability*, vol. 15, no. 17, p. 13168, 2023.