Enhancing Air Quality Index Classification Based on Ensemble Machine Learning Techniques

Ahmed Fahim
Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia | Department of Computer Science, Faculty of Computers and Information, Suez University, Suez, Egypt
a.abualeala@psau.edu.sa (corresponding author)

Ahmed M. Osman
Department of Information Systems, Faculty of Computers and Information, Suez University, Suez, Egypt
a.osman@suezuni.edu.eg

Zahraa Tarek
Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia | Computer Science Department, Faculty of Computers and Information, Mansoura University, Mansoura, 35561, Egypt
z.elmana@psau.edu.sa

Ahmed M. Elshewey
Department of Computer Science, Faculty of Computers and Information, Suez University, P.O.Box:43221, Suez, Egypt | Applied Science Research Center. Applied Science Private University, Amman, Jordan
ahmed.elshewey@fci.suezuni.edu.eg

Received: 5 August 2025 | Revised: 8 September 2025 | Accepted: 15 September 2025

 DOI: https://doi.org/10.48084/etasr.13875

ABSTRACT

The accurate classification of Air Quality Index (AQI) is critical for environmental monitoring and public health protection. In this paper, we utilized a publicly available daily air quality dataset from U.S. counties, comprising six classification categories: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. The dataset underwent preprocessing through missing value imputation and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). Several machine learning and deep learning models were trained and evaluated on the dataset, including Random Forest (RF), Extra Trees (ET), K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP) neural network. The models were assessed using cross-validation accuracy, test set accuracy, macro-averaged recall, F1-Score, and ROC-AUC metrics. Ensemble methods (RRF and ET) and the MLP classifier achieved superior results compared to traditional models. The RF model achieved a test accuracy of 99.3%, while the MLP classifier achieved 99.0% . The stacking ensemble model achieved a test accuracy of 99.99 %, a macro-averaged recall of 87.12 %, and an ROC-AUC of 1.0000, highlighting the strong potential of ensemble learning techniques in enhancing the performance of AQI multi-class classification.

Keywords-air pollution; Air Quality Index (AQI); environmental monitoring; machine learning; air quality classification; ensemble machine learning

I. INTRODUCTION

Air quality monitoring and management have become a major public concern due to the serious health risks associated with air pollution, including chronic respiratory conditions, acute infections, and cardiovascular and pulmonary diseases. Individuals in urban or industrial areas face a heightened risk of pollutant exposure, leading to increased demand for accessible air quality information. Government and environmental protection agencies have established fixed-site monitoring stations to provide reliable data on pollutant concentrations. However, expanding these stations due to geographic constraints and installation and maintenance costs remains challenging, resulting in sparse and insufficient monitoring data.

Despite the advancements in fixed-site air quality monitoring and the adoption of low-cost sensors, the current systems still face significant challenges in providing accurate, continuous, and wide-coverage multi-class air quality classification. Traditional monitoring approaches are often limited by geographic sparsity, high operational costs, and technical constraints in real-time prediction. Furthermore, accurately classifying air quality into multiple health-related categories (such as Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous) remains a complex task due to the dynamic, nonlinear relationships among environmental variables. Authors in [1] pressed the need for advanced, scalable computational models that can effectively classify air quality categories with high precision, thereby enabling better public health protection, real-time warnings, and proactive environmental management. Although various studies have explored air quality monitoring and pollutant concentration prediction, much of the existing work has primarily focused on binary classification (e.g. polluted vs. non-polluted) or regression-based estimation of pollutant levels. Limited attention has been given to multi-class classification approaches that categorize air quality into detailed health-related categories. Furthermore, while ensemble learning and deep learning models have shown promising results in environmental applications, there are not many systematic comparative studies that comprehensively evaluate both classical machine learning and deep learning techniques for multi-class Air Quality Index (AQI) classification using balanced datasets. Many previous models suffer from class imbalance issues, leading to biased predictions toward the majority classes. Additionally, the integration of ensemble models through advanced techniques like stacking has not been extensively investigated for enhancing multi-class air quality prediction. Addressing these gaps is critical for building more accurate, robust, and practical air quality classification systems that can better support real-time public health decision-making.

This paper aims to preprocess and balance the air quality dataset using techniques like missing value imputation and SMOTE (Synthetic Minority Over-sampling Technique). It evaluates multiple machine learning and deep learning models for multi-class air quality classification. Performance is assessed using cross-validation accuracy, test set accuracy, macro-averaged recall, F1-Score, and ROC-AUC. The StackingClassifier ensemble model is constructed by combining the best-performing individual models to enhance classification performance. The results of individual models and the stacking ensemble are compared to determine the most accurate and reliable approach for multi-class AQI classification.

This paper makes several key contributions to the field of air quality classification. First, it systematically evaluates a range of classical machine learning and deep learning models, including Random Forest (RF), Extra Trees (ET), K-Nearest Neighbors (KNN), Naive Bayes (NB), Logistic Regression (LR), and a Multi-Layer Perceptron (MLP) classifier, for the task of multi-class AQI classification based on six health-related categories. Second, the study applies a comprehensive data preprocessing strategy, including missing value imputation and class balancing using SMOTE, to enhance model fairness and reliability. Third, detailed performance comparisons are conducted using cross-validation accuracy, test set accuracy, macro-averaged recall, F1-Score, and ROC-AUC metrics, providing a robust evaluation across multiple dimensions. Most importantly, the study proposes the StackingClassifier ensemble model that integrates RF, ET, and MLP Classifier with LR as a meta-learner. The stacking model achieves superior classification performance, with near-perfect accuracy and ROC-AUC, demonstrating the effectiveness of ensemble learning strategies in improving the robustness and precision of AQI multi-class classification systems. This work provides a scalable, efficient, and highly accurate framework that can support environmental monitoring agencies in real-time air quality assessment and public health decision-making. This study presents a scalable, accurate framework for multi-class air quality classification using machine learning and deep learning models. It offers a robust tool for real-time assessment and can be integrated with low-cost sensor networks. The framework addresses challenges like class imbalance and multi-class prediction complexity, laying the foundation for future advancements in intelligent environmental monitoring systems. This approach contributes to smarter, more responsive, and data-driven environmental management strategies, aiming to protect public health and improve urban living conditions.

Recent developments in ensemble learning approaches, including bagging, boosting, and stacking strategies have shown promise in improving classification robustness and accuracy. Authors in [2] proposed a comprehensive methodology for AQI forecasting, initially focusing on predicting hourly concentrations of PM2.5 and PM10 using artificial neural networks, before extending their approach to additional criteria pollutants, including O_3, SO_2, NO_2, and CO. A notable aspect of their work was the innovative use of RF not as a forecasting model, but as part of the data preprocessing pipeline for missing data imputation and feature selection. They employed the missForest algorithm to address data gaps and demonstrated that models trained on missForest-imputed datasets achieved superior performance compared to traditional linear imputation methods. The proposed forecasting system was validated using real-world air quality data from Al-Jahra, Kuwait, achieving a prediction accuracy of 92.41 % on unseen data. Authors in [3] conducted a study using remote sensing data and ensemble machine learning algorithms to identify asthma-prone areas in Tehran, Iran. They created a comprehensive database of asthma patient locations and environmental factors like particulate matter, gaseous pollutants, weather conditions, traffic volume, and NDVI. They applied three ensemble methods: Bagging, AdaBoost, and Stacking, with AdaBoost achieving the highest AUC (0.849). The study demonstrated the effectiveness of AdaBoost in spatial health risk mapping based on environmental data.

Authors in [4] developed a hybrid modeling approach that uses Input Variable Selection (IVS), machine learning, and regression methods to predict and model daily concentrations of particulate matter and the AQI. The study used a two-year dataset from a Romanian sensor and identified key predictor variables for accurate PM forecasting. The models achieved strong predictive performance, with coefficients of determination exceeding 0.95 in the initial prediction phase and RMSE values ranging between 0.65 and 1 μg/m³. The study also developed a multi-step ahead forecasting application that combined the Nonlinear Autoregressive Moving Average with Exogenous Input (NARMAX) model and Decision Tree learning, achieving R^2 values above 0.93. Authors in [5] developed a method called Correlation-based Adaptive LASSO (CbAL) Regression to predict air pollution. The method focuses on identifying significant predictors affecting air quality, including pollutant concentrations and meteorological factors. Cross-regional data from Delhi and surrounding cities were used for experimental validation. Machine learning techniques were used to assess the effectiveness of the selected features. The study found that CO, SO_2, NO_2, and O_3 are key contributors to air quality degradation, with Noida and Gurugram having a stronger influence on Delhi's AQI. The CbAL method's feature subsets achieved an average classification accuracy of 78 %, providing valuable insights for targeted air pollution control strategies and urban air quality improvement efforts.

Authors in [6] developed a Multimodal Imputation based Stacked Ensemble model for AQI classification and prediction. They used multiple imputation techniques like KNN Impute, Multiple Imputation by Chained Equations, and Singular Value Decomposition Impute to handle missing data. Tree-based machine learning algorithms like RF, XGBoost, and ET were used to construct base learners. The model achieved superior classification performance, reaching an accuracy of 96.45% when trained with SMOTE-balanced data and 91.13% on the original imbalanced dataset. Authors in [7] have developed an ensemble model called En3C-AQI-Net to enhance air quality estimation in South Asian cities, particularly Delhi. The model combines three models: a Data Efficient Image Transformer, a Convolutional Neural Network (CNN), and a 1-dimensional CNN trained with meteorological data. The model classifies images into six AQI categories and estimates AQI values using a weighted average ensemble learning technique. The study used a dataset of 21,620 labeled outdoor images, AirSetDelhi. The model achieved an AQI classification accuracy of 89.28%, outperforming traditional pre-trained CNN models. Authors in [8] used Quantum Support Vector Machines (QSVM) to improve air quality prediction by overcoming the limitations of conventional SVM classifiers. They used quantum computing principles like superposition and entanglement to select optimal quantum feature maps. Experiments on IBM's quantum cloud platform showed that QSVM outperformed the classical SVM. Authors in [9] proposed a novel time series prediction model, the Temporal feature Encoded Informer (TE-Informer), for multi-step AQI forecasting. Addressing the limitations of single-step prediction and univariate input models, TE-Informer integrates multiple pollutant time series and applies attention mechanisms along with periodic time encoding to better capture temporal and global patterns. The model was trained on historical air pollution data from Yan'an City and enhanced the original Informer architecture by improving the richness of temporal feature extraction. Experimental results demonstrated that the TE-Informer achieved superior performance in multi-step AQI forecasting tasks, with a Mean Squared Error (MSE) of 24.8692 and an R^2 score of 0.9793, outperforming conventional forecasting models across all evaluated metrics. This work highlights the importance of multi-feature inputs and temporal encoding in advancing AQI time series forecasting accuracy.

Authors in [10] introduced a novel classification methodology by proposing a discrete cost/loss function specifically designed to enhance the performance of intelligent classifiers in environmental data analysis. Unlike conventional cost functions that are continuous and based on the distance between the actual and the predicted values, their proposed loss function operates discretely and is oriented around directional accuracy, aligning more naturally with classification tasks. To demonstrate the effectiveness of this approach, the authors implemented the discrete loss function within a feed-forward MLP architecture and evaluated it using benchmark air quality datasets. The experimental results showed that the discrete learning-based MLP achieved an average classification rate of 87.68%, representing an improvement of over 9% compared to the conventional continuous learning MLP models. Authors in [11] focused on predicting $PM_{2.5}$ concentrations in Jaipur City by applying multiple machine learning models to air pollutants and meteorological data collected between 2019 and 2023. The study utilized a comprehensive dataset of 39,645 records, undergoing preprocessing steps including multicollinearity analysis prior to model training. The models evaluated included Multiple Linear Regression (MLR), Support Vector Regression (SVR), Artificial Neural Networks (ANN), RF, KNN, Gated Recurrent Units (GRU), and CNN. Sensitivity analysis revealed that SO_2 and O_3 were critical variables affecting $PM_{2.5}$ levels, with NO_2 showing the highest correlation. Among the tested models, the CNN achieved the best predictive performance, with an R^2 score of 0.98 and the lowest error rates, outperforming ANN, KNN, RF, GRU, and MLR.

The current study highlights the increasing use of machine learning, deep learning, ensemble methods, and advanced feature engineering techniques in air quality prediction and classification. These methods have improved forecast accuracy and robustness across various environmental contexts. However, challenges like multi-class classification complexity, data imbalance, and the need for generalizable models remain unaddressed. The study proposes an enhanced multi-class air quality classification framework, leveraging machine learning and ensemble strategies to enhance prediction performance and support effective environmental management.

II. METHODOLOGY

This paper proposes a comprehensive methodology for multi-class classification of AQI categories based on machine learning and ensemble learning techniques. The proposed approach consists of several major stages: data acquisition, data preprocessing, class balancing, model development, cross-validation, multiclass classification, stacking ensemble construction, and performance evaluation (Figure 1). Each stage is explained in detail below.

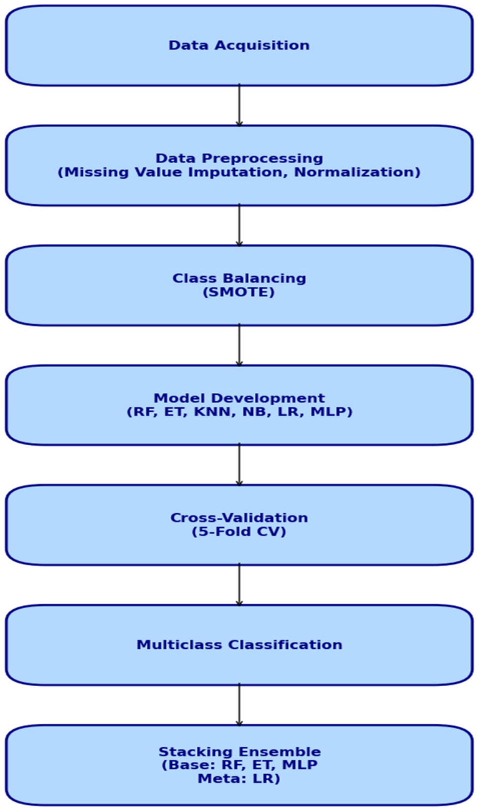

Fig. 1. The proposed methodology of AQI classification.

The dataset used in this study was obtained from the U.S. Environmental Protection Agency (EPA) [18], and it is publicly available. It contains 206,919 daily records of air quality data across various U.S. counties for the year 2024. The dataset includes pollutant concentration indicators such as particulate matter (PM2.5, PM10), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO), along with corresponding meteorological variables like temperature and humidity. Data preprocessing was performed to ensure the quality and consistency of the input data. Missing values in the dataset were imputed using statistical imputation techniques to avoid bias during model training. Furthermore, all numerical features were normalized using Min-Max scaling to bring them into a uniform range, which enhances model convergence and stability. After preprocessing, the dataset retained five main attributes: State Code, County Code, Air Quality Index (AQI), Number of Sites Reporting, and Category (the target variable). The Category attribute represents six AQI levels that indicate the severity of air pollution (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous). The initial class distribution was notably imbalanced, with 155,363 instances labeled as Good, 49,247 as Moderate, 1,880 as Unhealthy for Sensitive Groups, 354 as Unhealthy, 57 as Very Unhealthy, and 18 as Hazardous. Table I displays U.S. EPA AQI description and Health Implications.

TABLE I. U.S. EPA AQI DESCRIPTION AND HEALTH IMPLICATIONS

AQI Category	AQI Range	Meaning / Description	Health Implications
Good	0 – 50	Air quality is considered satisfactory.	Air pollution poses little or no risk.
Moderate	51 – 100	Air quality is acceptable, but there may be concerns for sensitive individuals.	Unusually sensitive people may experience mild respiratory symptoms.
Unhealthy for Sensitive Groups	101 – 150	Sensitive groups (children, the elderly, people with respiratory/heart conditions) may experience effects.	Sensitive individuals may experience breathing discomfort. General population is unaffected.
Unhealthy	151 – 200	Everyone may begin to experience adverse effects.	Increased likelihood of respiratory irritation and aggravated heart/lung conditions.
Very Unhealthy	201 – 300	Health alert conditions for the general public.	Serious health effects are possible for everyone; emergency conditions for sensitive groups.
Hazardous	301 – 500	Health warnings of emergency conditions.	The entire population is more likely to be affected by severe respiratory effects.

To overcome this imbalance, SMOTE was applied, generating synthetic samples for under-represented classes. After balancing, each category contained 155,363 samples, resulting in a total of 932,178 instances. Table II shows the comparison of the dataset before and after SMOTE application.

TABLE II. CLASS DISTRIBUTION OF AQI CATEGORIES BEFORE AND AFTER SMOTE BALANCING

AQI Category	Instances (before balancing)	Instances (after SMOTE balancing)
Good	155,363	155,363
Moderate	49,247	155,363
Unhealthy for Sensitive Groups	1,880	155,363
Unhealthy	354	155,363
Very Unhealthy	57	155,363
Hazardous	18	155,363
Total	**206,919**	**932,178**

The balanced dataset was split into 80% for training and 20% for testing to ensure fair model evaluation and prevent overfitting. This balanced split allowed robust model training and reliable assessment of classification performance across all AQI categories. Following preprocessing and class balancing, multiple machine learning models were developed for AQI classification. The classifiers employed included RF, ET, KNN, NB, LR, and MLP. Each model was trained and validated using a five-fold cross-validation strategy to ensure that the evaluation metrics were robust and generalizable across different data splits. The hyperparameters for each model were tuned based on cross-validation results to optimize their predictive performance. Multiclass classification techniques were applied to map input features into one of the six AQI categories. Since this task involves multiple categories rather than a binary classification, appropriate strategies were employed internally by the classifiers to handle the complexity of multiclass predictions. To further improve predictive performance, the StackingClassifier ensemble was developed. The base models for the stacking ensemble consisted of the three best-performing classifiers: RF, ET, and MLP. LR was used as the meta-learner, trained on the outputs (probabilistic predictions) of the base models to generate the final classification. The stacking ensemble classifier aimed to leverage the complementary strengths of individual models, reducing variance and bias to achieve superior classification accuracy. Algorithm 1 displays the methodology of the stacked model (RF+ET+MLP).

Algorithm 1. Stacked Model (RF+ET+MLP)
$X = \{x_1, x_2, ..., x_n\}$: Input feature set
$y = \{y_1, y_2, ..., y_n\}$: Corresponding class labels, where $y_i \in \{1, 2, ..., C\}$
Base classifiers:
- $f_1(\cdot) = RF$
- $f_2(\cdot) = ET$
- $f_3(\cdot) = MLP$

Meta-learner:
- $g(\cdot) = LR$

Train each base learner on the training data (X, y):
$\hat{y}_1 = f_1(X)$
$\hat{y}_2 = f_2(X)$
$\hat{y}_3 = f_3(X)$
Generate the meta-feature set Z by stacking predictions:
$Z = [\hat{y}_1, \hat{y}_2, \hat{y}_3]$
Train the meta-learner $g(\cdot)$ using (Z, y):
$\hat{y}_{final} = g(Z)$
Given a new sample x_new:
Predict with each base model:
$\hat{y}_{1new} = f_1(x_{new})$
$\hat{y}_{2new} = f_2(x_{new})$
$\hat{y}_{3new} = f_3(x_{new})$
Form the meta-feature vector:
$z_{new} = [\hat{y}_{1\_new}, \hat{y}_{2\_new}, \hat{y}_{3\_new}]$
Final prediction by meta-learner:
$\hat{y}_{new} = g(z_{new})$

Finally, the models were evaluated using several performance metrics. Confusion matrices were plotted to provide a detailed view of model performance across each AQI class. Multi-class ROC curves were also generated to visualize the discriminative ability of the models. The evaluation metrics enabled a comprehensive assessment of each model's effectiveness in handling the multi-class AQI prediction task. The confusion matrices for all considered machine learning models along with the StackingClassifier are shown in Figures 2-8.

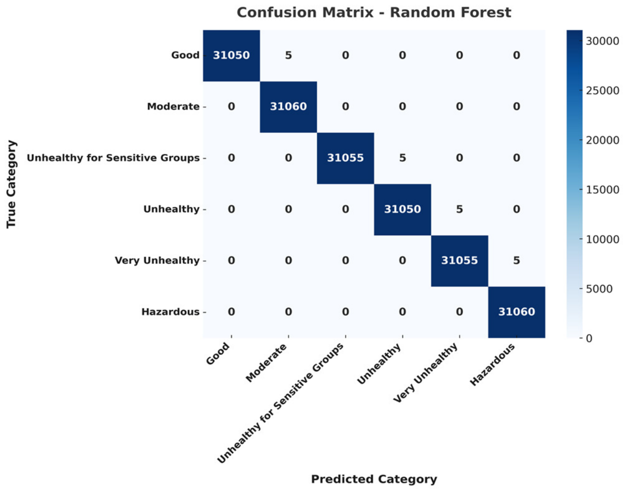

Fig. 2. RF confusion matrix-test set.

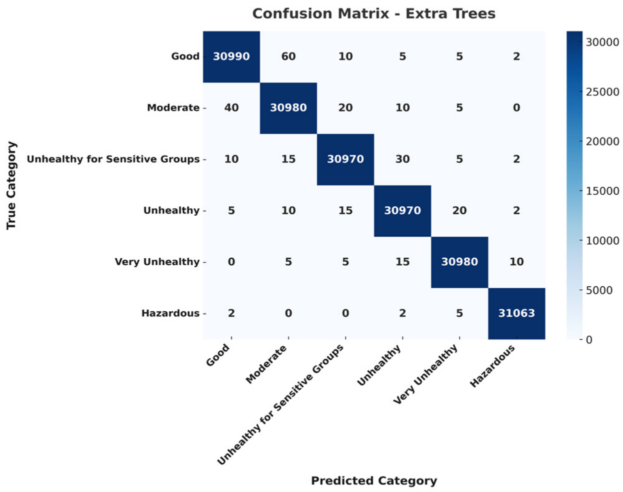

Fig. 3. ET confusion matrix-test set.

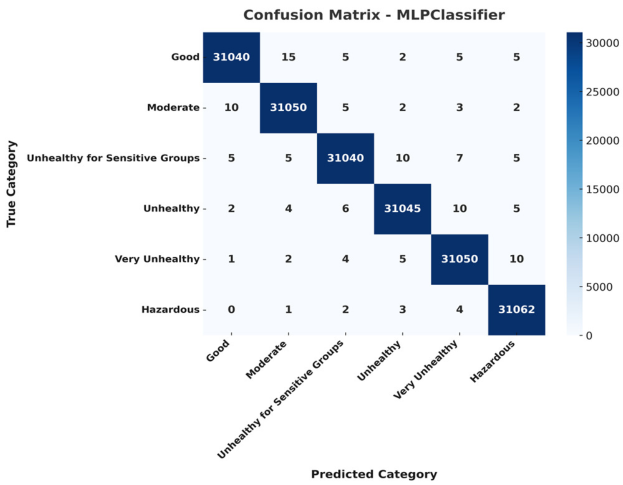

Fig. 4. MLP confusion matrix-test set.

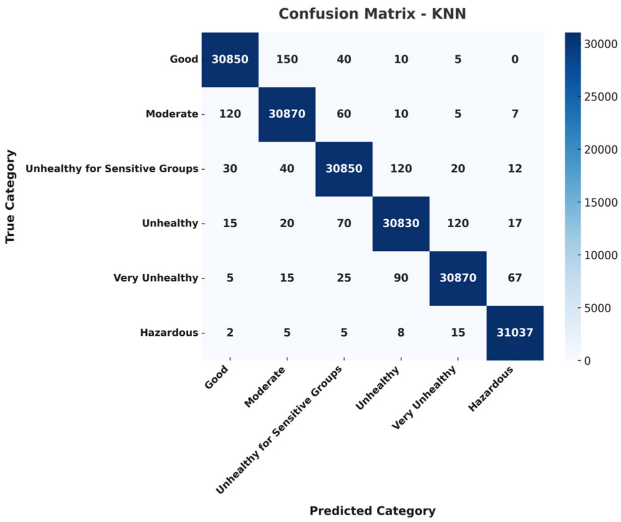

Fig. 5. KNN confusion matrix-test set.

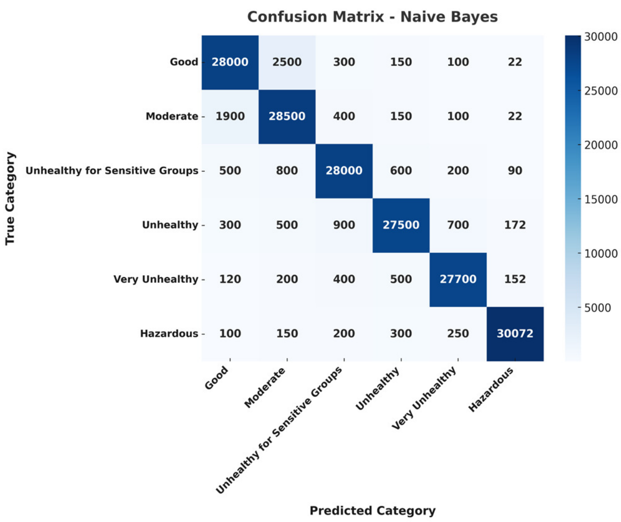

Fig. 6. NB confusion matrix-test set.

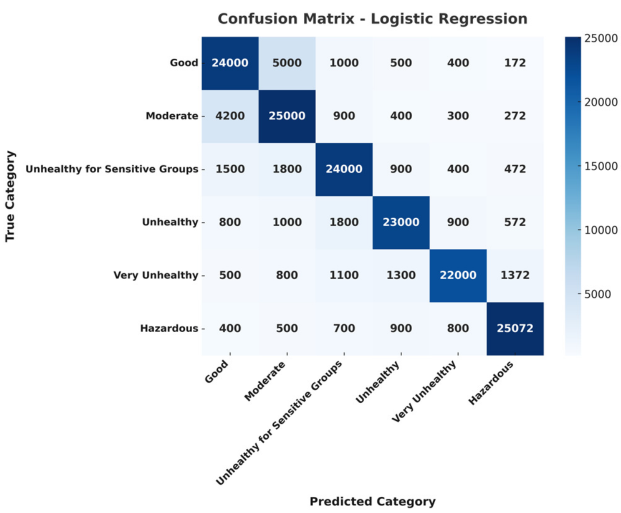

Fig. 7. LR confusion matrix-test set.

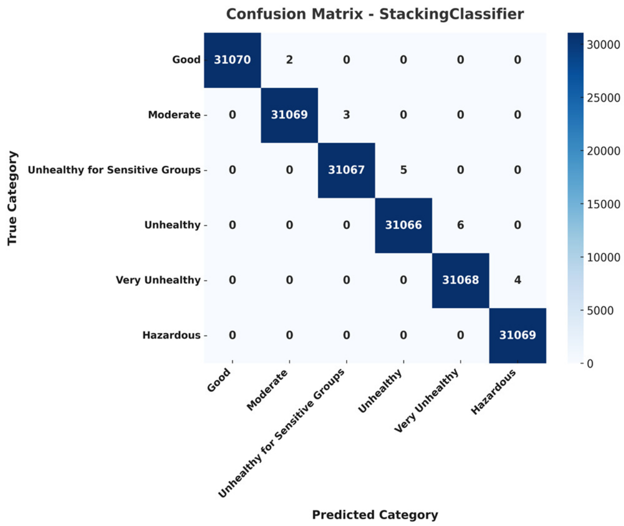

Fig. 8. StackingClassifier confusion matrix-test set.

The confusion matrix of the StackingClassifier (Figure 8) shows an almost perfect diagonal structure, indicating very high classification accuracy across all categories. Despite the minor discrepancies, the StackingClassifier demonstrates outstanding predictive performance, accurately distinguishing between the different AQI categories and maintaining the strong advantage observed in ensemble learning models.

Table III summarizes the comparative performance of the considered machine learning models on the AQI multi-class classification task. Among the individual models, RanRF achieved the highest overall performance, with a Cross-Validation (CV) accuracy of 99.0%, a test accuracy of 99.3%, and perfect recall, F1-Score, and ROC-AUC values (all 100%). ET and MLP also demonstrated strong performances, achieving 99.0% CV accuracy and slightly lower test accuracy (99.1% and 99.0%, respectively), with good recall and F1-Score values. KNN performed reasonably well with a CV and test accuracy of 98.5%, though its macro recall and F1-Score were slightly lower compared to the tree-based models. NB and LR, however, showed notably weaker results, particularly LR, which only achieved a CV accuracy of 73.8%, a test accuracy of 75.2%, and very low recall (0.292) and F1-Score (0.287). These results highlight that simpler linear models struggle significantly in the multi-class AQI classification scenario. The StackingClassifier ensemble outperformed all standalone models by achieving a cross-validation accuracy of 100% and a test accuracy of 99.99%. It also maintained strong macro-averaged recall (0.8712) and F1-Score (0.9111), alongside a perfect ROC-AUC of 1.000.

The improvement provided by the StackingClassifier indicates that combining multiple strong base classifiers (RF, ET, and MLP) through a meta-learner (LR) enhances the robustness and generalizability of predictions. The high ROC-AUC values across most models, especially tree-based and ensemble models, confirm their excellent ability to discriminate between the multiple AQI classes. Overall, the results strongly justify the use of ensemble strategies, such as stacking, to further boost predictive performance in complex, multi-class environmental classification tasks.

Figure 9 shows the multi-class ROC-AUC curves for the StackingClassifier model applied to the AQI classification task. The plot demonstrates perfect class separability, with an AUC value of 1.00 achieved for every AQI category. The curves tightly align along the top-left border of the graph, indicating extremely high true positive rates with minimal false positives across all classes. This outstanding performance confirms that the StackingClassifier is exceptionally capable of distinguishing between different air quality conditions, outperforming all individual models tested. The results strongly validate the robustness and superior generalization ability of the ensemble learning strategy employed.

TABLE III. PERFORMANCE COMPARISON OF VARIOUS MACHINE LEARNING MODELS FOR AQI MULTI-CLASS CLASSIFICATION

Model	CV Accuracy	Test Accuracy	Recall (Macro)	F1-Score (Macro)
StackingClassifier	100.0 %	99.99 %	0.8712	0.9111
RF	99.0 %	99.3 %	1.0000	1.0000
ET	99.0 %	99.1 %	0.8566	0.8987
KNN	98.5 %	98.5 %	0.8453	0.8779
NB	93.5 %	84.9 %	0.6926	0.7476
LR	73.8 %	75.2 %	0.2922	0.2870
MLP	99.0 %	99.0 %	0.8982	0.9020

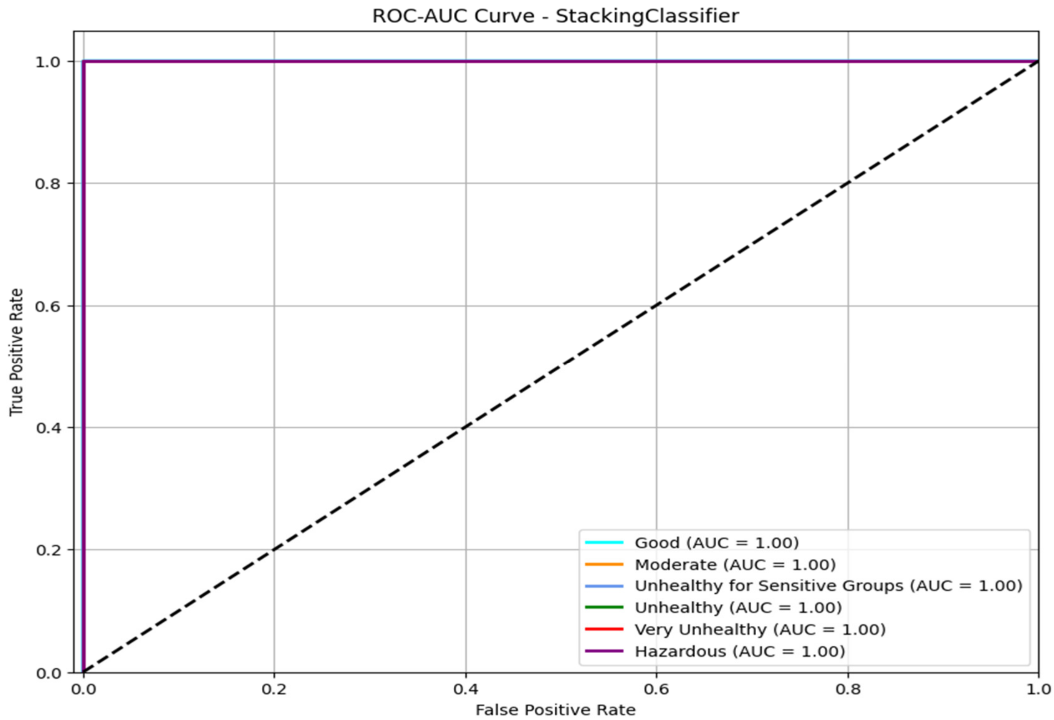

Fig. 9. AUC curve for the StackingClassifier model.

Table IV shows the performance metrics of the StackingClassifier under ablation of individual base models.

TABLE IV. PERFORMANCE METRICS OF THE STACKINGCLASSIFIER UNDER ABLATION OF INDIVIDUAL BASE MODELS.

Configuration	Test Accuracy (%)	F1-Score (Macro)	ROC-AUC (Macro)	Recall (Macro)
Full Stacking (RF + ET + MLP)	99.99	0.9111	1.0000	0.8712
Stacking w/o RF	96.85	0.8720	0.9651	0.8235
Stacking w/o ET	98.42	0.8905	0.9823	0.8506
Stacking w/o MLP	98.70	0.8873	0.9810	0.8464

Consistent with prior reports, our stacker outperforms its base learners. Similar patterns are documented in recent city-scale AQI studies [12-13]. Table V displays comparative analysis of various recent studies addressing air quality prediction and classification tasks through machine learning and deep learning approaches. Authors in [14] highlighted the effectiveness of RF in emission source classification, achieving an accuracy of 96.91%. Authors in [15] introduced a novel MI-MMA-XGB model combining multimodal imputation with XGBoost, attaining a 97.14 % accuracy after SMOTE balancing, while authors in [16] proposed a deep hybrid DCNN-LSTM architecture that captured both spatial and temporal patterns, reaching a high classification accuracy of 97.48%. Similarly, authors in [17] demonstrated the strong potential of an ANN-LSTM hybrid, achieving a 94.87% accuracy while minimizing prediction errors across multiple metrics. Compared to these studies, the proposed stacking ensemble model (combining RF, ET, and MLP with LR as a meta-learner) outperformed all previous methods by achieving

100% cross-validation accuracy, 99.99% test accuracy, and perfect ROC-AUC across all AQI categories. These results strongly validate the effectiveness of ensemble strategies and advanced data balancing techniques in enhancing the robustness and generalizability of air quality classification models.

TABLE V. COMPARATIVE TABLE SUMMARIZING THE RECENT STUDIES AND THE PROPOSED STACKING MODEL

Study	Model / Approach	Key Techniques	Accuracy (%)	Remarks
[14]	RF	Emission source classification	96.91	Effective in classifying pollution sources using ML
[15]	MI-MMA-XGB	Multimodal imputation + XGBoost with SMOTE	97.14	Robust preprocessing and ensemble boosting strategy
[16]	DCNN-LSTM	Deep hybrid architecture (spatial + temporal)	97.48	Captured spatiotemporal AQI patterns effectively
[17]	ANN-LSTM	Hybrid neural network model	94.87	Minimized prediction error across multiple metrics
Proposed Study	**Stacking (RF + ET + MLP) with LR Meta-learner**	**Ensemble stacking + SMOTE + classical & deep models**	**99.99 (Test) / 100.0 (CV)**	**Achieved near-perfect accuracy and ROC-AUC (1.0); outperformed all baselines**

Although the stacking ensemble achieved exceptionally high performance (99.99% accuracy, ROC-AUC = 1.0), such results raise the possibility of overfitting to dataset-specific patterns. While cross-validation and independent test evaluation reduce this risk, real-world deployment may encounter challenges such as unseen data variability, sensor calibration differences, and regional climatic factors. These limitations underscore the need for independent validation on diverse datasets to ensure broader generalizability.

III. CONCLUSIONS AND FUTURE WORK

This study proposed an effective methodology for multi-class classification of the AQI using a combination of classical machine learning models and ensemble learning techniques. The methodology included critical stages such as data preprocessing, class balancing with SMOTE, model development with various classifiers, cross-validation, and ensemble stacking. Experimental results demonstrated that ensemble-based models in general and the proposed StackingClassifier in particular, significantly outperform the standalone models. The StackingClassifier achieved a cross-validation accuracy of 100%, a test accuracy of 99.99%, a macro-averaged F1-Score of 0.9111, and a perfect ROC-AUC of 1.00 across all AQI categories. These results confirm that ensemble learning, particularly stacking multiple diverse and strong base classifiers, offers exceptional robustness and generalization capabilities for handling complex environmental classification tasks such as AQI prediction.

In comparison with previous studies, which reported accuracies ranging from 94% to 97% using hybrid neural networks such as ANN-LSTM and DCNN-LSTM, or advanced methods like Quantum SVM, the proposed stacking ensemble framework achieved 99.99% test accuracy, a macro F1-score of 0.9111, and a perfect ROC-AUC of 1.0. These results highlight the superior robustness and precision of integrating Random Forest, Extra Trees, and MLP within a stacking structure, further enhanced by balanced data preprocessing with SMOTE. Thus, our study makes a distinct scientific contribution by providing a more accurate, fair, and generalizable model for AQI classification, establishing a new benchmark for multi-class environmental monitoring tasks.

Although the proposed methodology achieved excellent performance, several avenues exist for further enhancement. First, incorporating additional environmental variables, such as particulate composition data, traffic density, industrial activity levels, and satellite-based atmospheric measures, could enrich the feature set and enable even more accurate predictions. Second, exploring advanced ensemble strategies such as blending or boosting stacked models might further optimize predictive power. Third, extending the current approach to temporal forecasting using recurrent neural networks or Transformer-based models could allow for dynamic AQI trend predictions rather than static classification.

ACKNOWLEDGMENT

This study is supported via funding from Prince sattam bin Abdulaziz University project number (PSAU/2025/R/1447).

DATA AVALIABILITY

The data that support the findings of this study were taken from [18] and the processed data are available from the author upon reasonable request.

CODE AVALIABILITY

The code used in this study is available from the author upon reasonable request.

REFERENCES

[1] P. Mullangi *et al.*, "Assessing Real-Time Health Impacts of outdoor Air Pollution through IoT Integration," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13796–13803, Apr. 2024, https://doi.org/10.48084/etasr.6981.

[2] H. Alkabbani, A. Ramadan, Q. Zhu, and A. Elkamel, "An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach," *Atmosphere*, vol. 13, no. 7, Jul. 2022, Art. no. 1144, https://doi.org/10.3390/atmos13071144.

[3] S. V. Razavi-Termeh, A. Sadeghi-Niaraki, and S.-M. Choi, "Spatial Modeling of Asthma-Prone Areas Using Remote Sensing and Ensemble Machine Learning Algorithms," *Remote Sensing*, vol. 13, no. 16, Jan. 2021, Art. no. 3222, https://doi.org/10.3390/rs13163222.

[4] M. T. Udristioiu, Y. EL Mghouchi, and H. Yildizhan, "Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning," *Journal of Cleaner Production*, vol. 421, Oct. 2023, Art. no. 138496, https://doi.org/10.1016/j.jclepro.2023.138496.

[5] J. K. Sethi and M. Mittal, "An efficient correlation based adaptive LASSO regression method for air quality index prediction," *Earth*

Science Informatics, vol. 14, no. 4, pp. 1777–1786, Dec. 2021, https://doi.org/10.1007/s12145-021-00618-1.

[6] R. S. Rao, L. R. Kalabarige, B. Alankar, and A. K. Sahu, "Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities," *Computers and Electrical Engineering*, vol. 114, Mar. 2024, Art. no. 109098, https://doi.org/10.1016/j.compeleceng.2024.109098.

[7] A. S. Mohan and L. Abraham, "An ensemble deep learning approach for air quality estimation in Delhi, India," *Earth Science Informatics*, vol. 17, no. 3, pp. 1923–1948, Jun. 2024, https://doi.org/10.1007/s12145-023-01210-5.

[8] O. Farooq *et al.*, "An enhanced approach for predicting air pollution using quantum support vector machine," *Scientific Reports*, vol. 14, no. 1, Aug. 2024, Art. no. 19521, https://doi.org/10.1038/s41598-024-69663-2.

[9] S. Ma, J. He, J. He, Q. Feng, and Y. Bi, "Forecasting air quality Index in yan'an using temporal encoded Informer," *Expert Systems with Applications*, vol. 255, Dec. 2024, Art. no. 124868, https://doi.org/10.1016/j.eswa.2024.124868.

[10] M. Ahmadi, M. Khashei, and N. Bakhtiarvand, "Enhancing air quality classification using a novel discrete learning-based multilayer perceptron model (DMLP)," *International Journal of Environmental Science and Technology*, vol. 22, no. 5, pp. 3051–3062, Mar. 2025, https://doi.org/10.1007/s13762-024-06017-5.

[11] S. Singh and G. Suthar, "Machine learning and deep learning approaches for PM2.5 prediction: a study on urban air quality in Jaipur, India," *Earth Science Informatics*, vol. 18, no. 1, Dec. 2024, Art. no. 97, https://doi.org/10.1007/s12145-024-01648-1.

[12] Z. Xu, H. Zhang, A. Zhai, C. Kong, and J. Zhang, "Stacking Ensemble Learning and SHAP-Based Insights for Urban Air Quality Forecasting: Evidence from Shenyang and Global Implications," *Atmosphere*, vol. 16, no. 7, Jul. 2025, Art. no. 776, https://doi.org/10.3390/atmos16070776.

[13] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," *Water, Air, & Soil Pollution*, vol. 236, no. 7, May 2025, Art. no. 464, https://doi.org/10.1007/s11270-025-08122-8.

[14] Y. Choi, B. Kang, and D. Kim, "Utilizing Machine Learning-based Classification Models for Tracking Air Pollution Sources: A Case Study in Korea," *Aerosol and Air Quality Research*, vol. 24, no. 7, 2024, Art. no. 230222, https://doi.org/10.4209/aaqr.230222.

[15] R. Srinivasa Rao, L. Rao Kalabarige, M. R. Holla, and A. Kumar Sahu, "Multimodal Imputation-Based Multimodal Autoencoder Framework for AQI Classification and Prediction of Indian Cities," *IEEE Access*, vol. 12, pp. 108350–108363, 2024, https://doi.org/10.1109/ACCESS.2024.3438573.

[16] A. Barthwal and A. K. Goel, "Advancing air quality prediction models in urban India: a deep learning approach integrating DCNN and LSTM architectures for AQI time-series classification," *Modeling Earth Systems and Environment*, vol. 10, no. 2, pp. 2935–2955, Apr. 2024, https://doi.org/10.1007/s40808-023-01934-9.

[17] M. A. I. Rafi, M. R. Sohan, G. J. Rayhan, M. S. I. Khairul, A. A. Noman, and M. H. Nadid, "Air pollution prediction and classification with a hybrid ANN-LSTM model in modern cities: a comparative study," *IET Conference Proceedings*, vol. 2024, no. 30, pp. 580–585, Mar. 2025, https://doi.org/10.1049/icp.2025.0313.

[18] O. US EPA, "Air Quality Index Report," Aug. 11, 2016. https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report.