International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**2**, pp.**28-38**, February **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

An Ensemble Machine Learning Approach for Accurate Air Pollution Prediction and Environmental Monitoring

Ebikombo-ere Olodiama[1*], Ogheneromesuo Afisi[2], Moses Peter[3]

[1]Computer Science, Marist University, Poughkeepsie, United States
[2]Business Administration, York St John University, York, United Kingdom
[3]Management, Law and Social Sciences, University of Bradford, Bradford, United Kingdom

*Corresponding Author:

Received: 30/Dec/2024; **Accepted**: 01/Feb/2025; **Published**: 28/Feb/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i2.2838

Abstract: Air pollution presents substantial risks to public health and environmental sustainability, necessitating robust predictive models capable of monitoring and forecasting air quality. This study aimed to design and evaluate a robust air pollution prediction model by leveraging data-driven modeling techniques. The research employed a comprehensive methodology that involved the aggregation of global air pollution datasets, followed by data preprocessing and transformation to ensure the accuracy and relevance of the input data. This data-driven approach facilitated the analysis and interpretation of the dataset using various machine learning algorithms. The study explored the performance of several machine learning algorithms, including AdaBoost, Decision Tree, Extra Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor (KNN), and Neural Network, to determine their effectiveness in predicting different levels of air quality. Each algorithm was evaluated based on precision, recall, f1-score, and overall accuracy, with a particular focus on challenging air quality categories such as "Unhealthy" and "Very Unhealthy." The results revealed that while some models like Decision Tree, Extra Tree, Random Forest, and Neural Network achieved high accuracy and f1-scores, others such as AdaBoost and Naïve Bayes displayed limitations in handling certain air quality categories. To overcome these limitations and enhance the overall prediction accuracy, an ensemble model was developed by combining the strengths of the top-performing algorithms. The ensemble model demonstrated exceptional performance, achieving perfect precision, recall, f1-scores, and accuracy across all air quality categories, indicating its potential as a highly reliable tool for real-time air quality monitoring and prediction. This study concludes that the ensemble model represents a significant advancement in air pollution prediction. Hence, offering an efficient solution for environmental monitoring systems. The study highlights the importance of integrating multiple machine learning algorithms to improve model robustness and accuracy, providing valuable insights for public health management and policymaking. The study recommends further exploration of ensemble models in different geographic regions and the integration of real-time data from IoT devices to enhance the model's applicability and effectiveness in diverse environmental scenarios.

Keywords: Air Pollution Prediction, Machine Learning, Ensemble Model, Environmental Monitoring, Data-Driven Modeling, Air Quality Forecasting

1. Introduction

The issue of urban air pollution is growing progressively more severe. This has emerged as a pivotal impediment to the sustainable advancement of Nigerian cities and the establishment of an ecological civilization [1]. The air quality profoundly impacts individuals' lives, productivity, and overall well-being. Due to the increase in population and human-made emissions in urban areas like Port Harcourt, concerns related to airborne particulate pollution have garnered more attention than before. The issue of urban air pollution has escalated due to the accumulation of black carbon, which has been distinctly observable since the final quarter of 2016 [2]. This perilous deposition, originating from human activities, underscores the pressing need for comprehensive data on fluctuations in pollution levels within the boundary layer [3].

Human existence relies solely on the presence of air, making its quality a vital determinant of overall well-being. Monitoring and comprehending air quality are imperative tasks. The global prevalence of air pollution has led to millions of individuals experiencing physiological ailments and even succumbing to respiratory-related fatalities. Supported by scientific findings, it is evident that air pollution constitutes the foremost environmental hazard [4]. As a consequence of swift industrialization, there has been a substantial rise in population levels coupled with the emission of toxic gases. The significant increase in population is closely linked to this phenomenon. The contamination of the air with perilous substances has led to severe repercussions on health. Unregulated pollution has resulted in a significant deterioration of air quality [4].

In recent decades, as industrialization and urbanization have continued to progress, substantial energy consumption has given rise to a progressively severe air pollution issue [5], [6], [7]. This pollution encompasses various air pollutants, including PM2.5, CO, SO2, NO2, and others, all capable of inducing numerous diseases such as asthma, heart disease, chronic obstructive pulmonary disease, and even cancer [8], [9]. The World Health Organization (WHO) reports that the simple act of breathing results in 7 million annual deaths worldwide due to air pollution, presenting a grave threat to human well-being [7].

Air quality is a pivotal concern regarding public health and sustainable livelihoods. Furthermore, it presents a hindrance to regional economic growth and societal advancement. Apart from monitoring and controlling air quality, predicting air quality in times of polluted weather has emerged as a central aspect of environmental management. This is particularly crucial during significant events and instances of severe pollution, as timely and precise air quality forecasts, coupled with source analysis, can furnish essential information for managerial decision-making [1].

To mitigate the adverse effects of air pollution, researchers have introduced various models aimed at forecasting changes in air pollution, allowing for timely interventions [7], [10]. Among these models, the deep learning approach has demonstrated the most effective predictive capabilities [11]. However, the challenge with deep learning models lies in their "black box" nature, which makes it intricate to understand the reasoning behind their predictions. Additionally, the time-series data related to atmospheric conditions amalgamate signals of varying frequencies and incorporate erroneous noise, thereby concealing the underlying correlations between atmospheric and pollutant variables. This intricate web of signals makes reliable identification difficult. Therefore, enhancing both the interpretability and accuracy of predictions becomes pivotal. This can be achieved by disentangling the various frequency signals from the original data, revealing clearer patterns, and constructing an interpretable neural network to extract these correlation rules [7].

Various technical approaches have recently been utilized for air quality prediction, integrating both mechanistic and statistical models. Mechanistic models simulate the physical and chemical processes governing air dispersion, including Gaussian diffusion models, Weather Research and Forecasting (WRF) models, and Community Multiscale Air Quality (CMAQ) models. For example, Cheng et al. developed an inference model leveraging the Gaussian process to estimate pollutant concentrations at arbitrary locations. Similarly, Rogers et al. refined the WRF model configuration through extensive sensitivity analyses in central California, enhancing its ability to simulate meteorological variables with reasonable accuracy. Additionally, Lee et al. (2007) examined and assessed atmospheric O_3 using a CMAQ modeling system, contributing to air pollution management strategies in China. However, mechanistic models require detailed and accurate external environmental parameters as inputs. Given the complexity of real-world environments, acquiring reliable data for these parameters remains challenging, thereby constraining the predictive capabilities of mechanistic models.

On the other hand, statistical models are designed to predict future variable changes by discerning patterns within historical data. These encompass linear regression models [18], [19], perceptron models [20], [21], support vector machines (SVM) [22], [23], tree models [24], [25], [26], and deep neural networks (DNN) [27], [28], [29]. Linear regression models can be both univariate and multivariate, with the latter possessing superior nonlinear fitting capabilities. However, when compared to alternatives like perceptron models, tree models, and DNNs, multivariate linear regression might still fall short. With the ongoing progress in artificial intelligence technology, novel iterations of deep neural networks (DNN) have been consistently developed and refined. Examples include convolutional neural networks (CNN) [30], graph convolution networks (GCN) [31], residual networks (ResNet) [32], and attention networks [33]. These models have found extensive application in predicting air pollution. Simultaneously, the remarkable advancements in graphics processing unit (GPU) computing power has enabled the training of intricate DNN models. Consequently, DNN's predictive capabilities have surpassed those of traditional statistical models [34], which can be attributed to two key factors. Firstly, the deep network architecture equips it with a robust capacity to simulate the evolution process from input to output. Secondly, the flexible combination of various network modules harnesses the strengths of different networks.

For instance, a deep distributed fusion network has been established based on deep neural networks, demonstrating enhanced short-term and long-term air quality prediction in comparison to preceding online monitoring systems [35]. Furthermore, a deep convolutional neural network has been employed to rectify prediction errors within the CMAQ model, thus elevating the overall prediction performance of CMAQ [36]. Approaches like CNN-LSTM and GCN-LSTM amalgamate the benefits of CNN/GCN for spatial information extraction with LSTM for capturing temporal dependencies, showcasing advanced predictive capabilities [37], [38]. However, the intricate structure of deep networks makes model predictions challenging to interpret, rendering the understanding of model behaviour intricate. This complexity poses difficulties in formulating appropriate measures to mitigate air pollution. Additionally, the intricate and fluctuating atmospheric conditions lead to the integration of air pollutant data with interwoven signals of varying frequencies, alongside assorted erroneous random noise. These factors collectively affect the accuracy of predictions.

To address these constraints, this study proposes an optimized machine learning model based on an ensemble technique to predict air quality and pollution more effectively.

2. Related Work

Recent studies have shifted their focus towards advanced statistical learning algorithms for assessing air quality and predicting air pollution. [39] and [40] employed neural networks to develop models that forecast the levels of specific pollutants, such as particulate matter with a diameter of less than 10 microns (PM10).

The study in [41] proposed a random forest-based model, RAQ, for categorizing Air Quality Index (AQI) levels. Following this, [35] utilized deep neural networks for AQI classification. In [40], artificial neural network (ANN) models were applied to predict PM10 concentrations inside a subway station, incorporating variables such as train frequency, outdoor PM10 levels, and ventilation system data. The ANN models exhibited strong predictive performance, with correlation values ranging from 0.18 to 0.63 when compared to experimental data. Depending on the structural configuration and depth of the subway platform, the model achieved an accuracy between 67% and 80%.

Similarly, [42] explored various configurations to enhance AQI-level prediction beyond conventional machine learning methods such as k-nearest neighbor (k-NN), decision trees, and support vector machines (SVM). Their ANN-based approach achieved an accuracy of 92.3%, surpassing all other tested models. Forecasting air quality often relies on time series data analysis, where model selection and parameter optimization play a crucial role. Recent advancements have leveraged deep learning methodologies, as demonstrated in [43]. Likewise, [44] employed ANN models to estimate daily PM10 concentrations across different environmental settings, including regional and urban background areas.

To improve the predictive accuracy of PM2.5 concentration modeling, [45] introduced an approach that integrates air mass trajectory analysis with wavelet transform, effectively identifying key transport pathways and mitigating fluctuations in PM2.5 levels. Additionally, [46] presented a hybrid forecasting framework that combines principal component analysis (PCA) with the least squares support vector machine (LSSVM), optimized using the cuckoo search algorithm for enhanced PM2.5 predictions.

A deep learning-based air pollution monitoring and forecasting framework was developed in [47], while [48] introduced a selection model that utilizes cooperative data indices to determine the most suitable forecasting ensemble. In [49], an ANN-driven model was designed to predict pollutant concentrations (PM10, PM2.5, NO2, and O3) over both immediate and multi-day timeframes in heavily polluted regions. Meanwhile, [50] applied the American Meteorological Society/Environmental Policy Agency Regulatory Model (AERMOD) for short-term air quality forecasting, utilizing meteorological predictions from the Weather Research and Forecasting (WRF) model and providing a detailed emissions inventory for sources in Chembur, Mumbai. Additionally, [51] focused on vehicular pollution modeling by implementing AERMOD with WRF-simulated meteorological data, revealing that peak NOx and PM levels corresponded with periods of high traffic congestion.

For short-term PM2.5 forecasting, [52] introduced a deep learning model applied to the Beijing PM2.5 dataset. Liu et al. [53] proposed a wind-sensitive attention mechanism within a long short-term memory (LSTM) neural network to enhance PM2.5 predictions by incorporating wind pattern data. Furthermore, [54] developed an integrated air quality early warning system that encompasses estimation, forecasting, and assessment components, providing a comprehensive approach to air pollution management.

The study in [55] developed a backpropagation neural network (BPNN) model to estimate daily PM10 concentration levels. The model incorporated various input features, including hourly pollutant concentrations, meteorological factors such as wind speed, rainfall, relative humidity, and temperature, as well as temporal variables such as the month, year, and day of the week. To optimize the network's structure and weight parameters, the Genetic Algorithm (GA) was employed. The results demonstrated that the hybrid BP-GA model, which integrates backpropagation with genetic optimization, achieved significantly higher predictive accuracy compared to standalone artificial neural network (ANN) models.

Suleiman et al. [57] utilized ANN to forecast the upper threshold of PM10 concentrations stemming from roadside sources within an urban setting. The ANN model was designed using a combination of input factors, encompassing baseline PM10 levels, alongside roadside concentrations of SO2, NO2, and NOx, emission rates of PM10 from various vehicle categories, and a range of meteorological parameters, including solar radiation, precipitation, humidity, wind speed, barometric pressure, and wind direction. The results of the aforementioned research demonstrated that the ANN model effectively predicted the contributions of PM10 from road sources, achieving an impressive R-value of 0.85 and a Root Mean Square Error (RMSE) of 12.46 μg/m³, respectively.

The primary drawback of the existing studies is the lack of exploration of ensemble approaches in air quality prediction. While individual AI models like neural networks and deep learning have shown promise, the absence of sophisticated ensemble frameworks that combine the strengths of various models is a significant gap. Although existing studies mentioned the integration of multiple models, it is often superficial and does not delve into the development and deployment of comprehensive ensemble models. The existing studies primarily focus on comparing individual model performances rather than creating an intelligent ensemble model that leverages their collective predictive capacities to predict air pollutants.

3. Methodology

This research applies multiple machine learning algorithms to train an air quality prediction model based on a global air pollution dataset, as depicted in Figure 3.1. The methodology begins with data collection, with a focus on key pollutants such as Nitrogen Dioxide (NO2), Ozone (O3), and Particulate Matter (PM2.5). Following the data aggregation phase, the data was preprocessed to enhance data quality, while feature selection was conducted to identify the most relevant attributes, thereby optimizing computational efficiency and improving predictive performance.

To address class imbalance within the dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied, ensuring a more balanced representation across class labels. Subsequently, eight machine learning algorithms including AdaBoost, CatBoost, Decision Tree, Extra Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor, and Neural Network are used to construct predictive models for air quality and pollutant concentration levels. The final stage involves evaluating and comparing model performances, after which the four most effective models are integrated into an ensemble framework to enhance predictive accuracy.

3.1 Proposed Model

The proposed model in Figure 1 begins with data collection, including a global air pollution dataset. This dataset encompasses essential information regarding various pollutants and serves as the foundation for accurate predictions. It is a crucial starting point that ensures the system has access to the necessary data. Following data collection, the architecture involves data preprocessing and feature selection. Data preprocessing focuses on cleaning, transforming, and standardizing the dataset to ensure data quality and consistency. Feature selection is vital for selecting the most relevant attributes that will be used to train the model. These two stages collectively lay the groundwork for high-quality input data for the prediction models. To address the challenges of imbalanced data, the architecture incorporates the Synthetic Minority Oversampling Technique (SMOTE). SMOTE helps in achieving fair distributions within the dataset class labels, mitigating issues associated with class imbalance.

The heart of the architecture lies in the implementation of eight diverse machine learning algorithms. These algorithms, including AdaBoost, CatBoost, Decision Tree, Extra Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor, and Neural Network, serve as the predictive models for air quality and pollutant concentration. The proposed model architecture accounts for rigorous model evaluation and benchmarking. This stage ensures that the selected models are thoroughly assessed and compared using appropriate standard machine learning evaluation metrics such as accuracy, recall, precision, confusion matrix and F-1 score. It allows for the identification of the top-performing models based on their performance, accuracy, and suitability. The proposed architecture recognizes the value of combining the strengths of multiple models. The top four performing models are integrated into an ensemble model. This approach capitalizes on the diversity of these models, enhancing prediction accuracy and robustness.

In parallel with model evaluation and comparison, hyperparameter tuning becomes an integral component of the architecture. This stage focuses on the systematic optimization of model hyperparameters, which are parameters not learned from the data but crucial for model performance. Hyperparameter tuning encompasses selecting hyperparameters for each model, defining search spaces for these hyperparameters, choosing a hyperparameter optimization algorithm, and cross-validating to assess performance robustness. The ultimate goal is to identify the best hyperparameters for each model, leading to improved model performance.

Finally, the proposed model is trained using the Python programming language, which is known for its extensive libraries for data science and machine learning. The Django framework is employed to implement the air quality prediction web-based application, ensuring a user-friendly interface and efficient deployment.

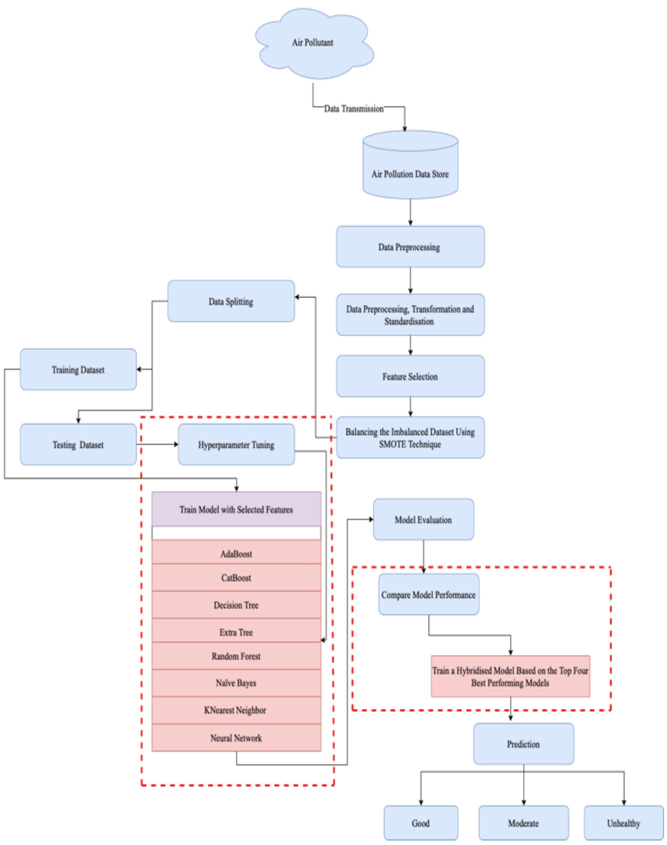

Figure 1. Architecture of the proposed model

The process of the proposed architecture in Figure 1 includes: process of the proposed architecture in Figure 1 includes:

1. **Data Collection:** The dataset contains a variety of air quality-related variables, including pollutant concentrations (e.g., NO2, O3, PM2.5), meteorological data, and temporal information (e.g., country, city).

2. **Data Preprocessing:** After data collection, the raw dataset is processed and stored in a CSV file for subsequent analysis. This step involves addressing issues such as missing values, outliers, and inconsistencies within the data. Missing values are imputed using suitable techniques, outliers are identified and removed, and the data is normalized or standardized to ensure uniformity across variables. In air quality datasets, categorical variables may include factors like pollutant types (e.g., NO2, O3, PM2.5), monitoring station IDs, or geographic regions. These categorical variables must be converted into numerical values before they can be used in machine learning models. Methods such as one-hot encoding or ordinal encoding are employed to transform categorical variables into binary vectors or numerical codes, respectively, enabling their inclusion in the predictive model.

3. **Feature Engineering and Selection:** Feature Engineering entails selecting, creating, or transforming features (variables) in a dataset to improve the model's predictive performance. This could entail extracting temporal features like the day of the week or time of day, creating lagged variables to capture temporal dependencies, or deriving new features from existing ones using mathematical transformations or domain knowledge. The goal of feature selection is to identify the most relevant subset of features that contribute to accurate air quality prediction. This reduces the dimensionality of the dataset and prevents overfitting. Correlation analysis, feature importance ranking, and model-based selection methods can be used to select the most informative features.

4. **Data Splitting:** Following the preprocessing and feature engineering stages, the dataset is partitioned into two subsets: training and testing. The training set is used to develop the predictive model by optimizing its parameters based on the provided input data. The testing set provides an unbiased evaluation of the final model's performance on unseen data, which ensures an accurate evaluation of its predictive accuracy.

5. **Data Balancing:** To address class imbalance issues within the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was utilized. SMOTE ensures a balanced distribution of class labels across the dataset, enabling the model to make accurate predictions for both majority and minority classes in the air quality prediction dataset.

6. **Hyperparameter Tuning:** To optimize model performance, the hyperparameters of the machine learning algorithms must be tuned after they have been chosen. Hyperparameter tuning entails systematically searching a set of hyperparameter values for the combination that produces the best results. Grid search and random search are two techniques for efficiently exploring the hyperparameter space and identifying the optimal settings for each model. Grid Search entails creating a grid of hyperparameter values for each hyperparameter in the model. For instance, when utilizing a Random Forest algorithm, hyperparameters such as the number of trees (n_estimators), the maximum depth of trees (max_depth), and the minimum number of samples required to split a node (min_samples_split) can be adjusted within a predefined range. Grid Search is employed to assess the model's performance across all possible hyperparameter combinations within the specified grid. To ensure a reliable estimate of the model's performance, cross-validation is incorporated into Grid Search. K-fold cross-validation is used, which divides the dataset into k equal-sized folds, training and evaluating the model k times, with each fold serving as the validation set once while the remaining folds are used for training. This approach helps reduce overfitting and provides a more robust estimate of the model's ability to generalize. Grid Search subsequently evaluates the model's performance using selected evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), or other relevant metrics for air quality prediction. The grid search algorithm finds the hyperparameter values that minimize or maximize the selected evaluation metric, depending on whether it represents error or accuracy. Grid Search is helpful because it systematically explores the entire hyperparameter space, ensuring that no hyperparameter combination goes unnoticed. It offers a computationally efficient and thorough method for hyperparameter tuning, making it appropriate for improving the performance of complex models such as those used in air quality prediction. Grid Search also provides a robust estimate of the model's performance by incorporating cross-validation, which improves its ability to generalize to new data sets.

7. **Model Training:** During the training phase, the training dataset is fed into the predictive model to enable it to learn the relationships between the input features (predictors) and the target variable (air quality). The selected machine learning model learns to make predictions by iteratively adjusting its internal parameters through an optimization process. The goal of training is to minimize a loss function, which measures the difference between the predicted and actual air quality values.

8. **Model Evaluation:** After training, the performance of each model was evaluated using the validation set. Standard machine learning evaluation metrics, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), are employed to assess the model's performance. This step helps in identifying models that generalize well to previously unseen air quality data and have robust predictive capabilities.

9. **Compare the Model Performance:** Based on the evaluated results, the best-performing models are selected for further consideration. The criteria for selecting the best models include overall performance on the testing set, accuracy, precision, and suitability for the task at hand. The

selected models are then prioritized for implementing the ensemble model.

10. **Training the Ensemble Model:** The ensemble model is implemented by combining the best machine learning models. This involves leveraging the strengths of multiple algorithms to improve the prediction accuracy and robustness. Each of these models adds unique insights and predictive capabilities to the ensemble model. By combining these diverse approaches, the ensemble model can effectively handle different aspects of air quality prediction, ranging from pollutant concentrations to overall air quality assessment.

11. **Prediction:** Prediction is executed when the trained model predicts the Air Quality Index (AQI) classes (Good, Moderate, or Unhealthy) for each pollutant at a specific location and time. For example, predicting that air quality in Port Harcourt, Nigeria, will be "unhealthy" based on the AQI values of CO, Ozone, NO2, and PM2.5.

3.2. Dataset Description

The global air pollution dataset was collected from Kaggle. The dataset contains Air Quality Index (AQI) values for various pollutants and is sourced from cities across the globe. The dataset also provides geolocated data on the following pollutants:

1. **Nitrogen Dioxide (NO_2):** NO_2 is a type of nitrogen oxide that enters the atmosphere through natural processes, including stratospheric entry and lightning. At ground level, however, NO_2 is primarily produced by emissions from automobiles, trucks, buses, power plants, and off-road vehicles. Short-term exposure to NO_2 can exacerbate respiratory conditions, such as asthma, while prolonged exposure may contribute to the development of asthma and respiratory infections. Individuals with asthma, children, and the elderly are particularly susceptible to the health impacts of NO_2.

2. **Carbon Monoxide (CO):** CO is a colourless and odourless gas primarily emitted into the atmosphere by vehicles, machinery, and equipment that run on fossil fuels. It is also released from sources such as kerosene and gas space heaters, as well as gas stoves, impacting indoor air quality. Inhalation of high concentrations of CO reduces the blood’s ability to transport oxygen to vital organs, including the heart and brain. At very high levels, which are uncommon in open spaces but may occur in confined environments, CO poisoning can lead to dizziness, confusion, unconsciousness, and, in severe cases, death.

3. **Particulate Matter ($PM_{2.5}$):** Atmospheric Particulate Matter, or atmospheric aerosol particles, consists of a complex mixture of tiny solid and liquid particles that are released into the atmosphere. Inhalation of these particles can lead to significant heart and lung issues. The International Agency for Research on Cancer (IARC) has classified $PM_{2.5}$ as a group 1 carcinogen. PM_{10} refers to particles with a diameter of 10 micrometres or less, while $PM_{2.5}$ particles have a diameter of 2.5 micrometres or less.

4. **Ozone (O_3):** The O_3 molecule, when not located in the ozone layer, poses a threat to outdoor air quality. At ground level, ozone is formed through chemical reactions between nitrogen oxides and volatile organic compounds (VOCs). In contrast to the beneficial ozone in the upper atmosphere, ground-level ozone can lead to a range of health issues, including chest pain, coughing, throat irritation, and airway inflammation. Additionally, it can impair lung function and exacerbate conditions such as bronchitis, emphysema, and asthma. Ozone also affects vegetation and ecosystems, with sensitive vegetation being particularly vulnerable during the growing season.

4. Model Evaluation

This section assesses the performance of the proposed model by employing standard machine learning evaluation metrics, including accuracy, precision, recall, specificity, and the F1-score. The main goal is to compare the effectiveness of different algorithms. Although accuracy is a commonly used metric for evaluating model performance, it may not be sufficient on its own in this particular case. Therefore, it is essential to consider additional metrics, such as the area under the curve (AUC), alongside accuracy, to identify the most suitable model for detecting fraudulent transactions.

1. **Accuracy:** Accuracy assesses the proportion of correct predictions made by the model regarding air quality prediction. This ratio is determined by dividing the total number of accurate predictions by the overall number of predictions. The accuracy calculation is illustrated in Equation 1.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

In this context, FP (False Positive) represents the overall count of incorrect predictions that have been classified as positive. FN (False Negative) represents the total count of inaccurate predictions that have been classified as negative. TP (True Positive) refers to the total count of accurate predictions classified as positive. Lastly, TN (True Negative) indicates the total count of precise predictions classified as negative.

2. **Precision:** Precision assesses the ratio of correctly predicted air pollutants (TP) to the total number of variables predicted as pollutants (TP + FP). Equation 2 illustrates the precision calculation.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} \tag{2}$$

3. **Recall:** Recall evaluates the proportion of accurately classified air pollutants (TP) in relation to the total number of air pollutants. The calculation for recall is expressed in Equation 3.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4. **Specificity:** Specificity evaluates the ratio of correctly classified non-air pollutants (TN) to the total number of non-air pollutants. The specificity calculation is illustrated in Equation 4.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

5. **F1-Score:** The F1-score metric computes the weighted average of precision and recall, with a value ranging between zero and one. A value closer to one indicates better performance. The F1-score is calculated using the formula given in Equation 5.

$$F1 - \text{Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

4.1. Benchmarking of the Models

This section compares the performance of the selected algorithms used for air pollution prediction. It also provides a detailed discussion of how these algorithms can be combined into an ensemble model to enhance the overall prediction accuracy and robustness of the model. The result in Table 1 summarizes the performance of all the algorithms evaluated to provide a clear comparison of their precision, recall, f1-score, and accuracy across different air quality categories.

Table 1. Summary of the Results Across Different Algorithms

Algorithms	Accuracy	Precision	Recall	F1-score
AdaBoost	0.83	0.71	0.83	0.76
Decision Tree	**1.00***	**1.00***	**1.00***	**1.00***
Extra Tree	**1.00***	**1.00***	**0.99***	**0.99***
Random Forest	**1.00***	**1.00***	**1.00***	**1.00***
Naïve Bayes	0.98	0.98	0.98	0.98
KNearest Neighbor	0.99	0.99	0.99	0.99
Neural Network	**1.00***	**1.00***	**0.97***	**0.98***

From the result in Table 1, it is evident that the Decision Tree, Extra Tree, Random Forest, and Neural Network algorithms consistently deliver near-perfect accuracy, recall, and precision. These algorithms perfectly or near-perfectly classify all categories, with minimal misclassifications.

The Decision Tree and Random Forest models demonstrate perfect classification across all metrics, showing their robustness in predicting air quality levels accurately. However, the Decision Tree model may be prone to overfitting, which could limit its generalizability to unseen data. The Random Forest, being an ensemble of multiple decision trees, mitigates this risk through averaging, making it a more reliable choice for diverse real-world scenarios. Similar to the Random Forest, the Extra Tree model is effective, with nearly perfect scores. Its slight deviations from perfect accuracy are negligible and mainly confined to edge cases. The Neural Network model, while almost perfect, shows slight imperfections in recall for rare categories like "Hazardous," indicating that while it generalizes well, it may still miss some edge cases, especially in categories with fewer training samples.

4.2. Result of the Ensemble Model

Given the varied strengths and weaknesses of these algorithms, an ensemble model can be designed to leverage the strengths of multiple models while compensating for their shortcomings. An ensemble model is particularly effective in scenarios where no single algorithm excels across all categories, as is evident in air pollution prediction.

In this study, the ensemble model can combine the strengths of the top four performing models (Decision Tree, Random Forest, Extra Trees and Neural Network models). These models are selected based on their high performance across different metrics and their complementary strengths. In this case, the Decision Tree model provides interpretability and high accuracy, especially for well-defined classes. The Random Forest model adds robustness and reduces overfitting while handling variability in the data. The Extra Tree is similar to the Random Forest. However, the Extra Tree model provides faster computation and slightly different decision boundaries. Hence, adding diversity to the ensemble model. Finally, the Neural Network model captures complex, non-linear relationships in the data, which is essential for efficient air pollution predictions.

Table 2. Classification Report of the Proposed Ensemble Model

	precision	recall	f1-score	support
Hazardous	1.00	1.00	1.00	34
Moderate	1.00	1.00	1.00	1815
Unhealthy	1.00	1.00	1.00	426
Unhealthy for Sensitive Groups	1.00	1.00	1.00	325
Very Unhealthy	1.00	1.00	1.00	51
accuracy			1.00	4693
macro avg	1.00	1.00	1.00	4693
weighted avg	1.00	1.00	1.00	4693

The proposed ensemble model, as summarized in Table 2, demonstrates exceptional performance across all air quality categories, with perfect precision, recall, and f1-score of 1.00 in each category. This indicates that the ensemble model accurately classified all instances of air quality levels without any misclassifications.

The precision of 1.00 across all categories means that every predicted instance by the ensemble model was indeed correct. For example, all instances predicted as "Hazardous" were truly hazardous, ensuring that the model does not raise false alarms, which is an important feature when managing public health warnings. The recall score of 1.00 for all categories indicates that the ensemble model successfully identified all true instances of each air quality category in the test data. This is particularly important in ensuring that no actual instances of hazardous or unhealthy air quality are missed, thereby preventing potential public health risks. The f1-score, which is the harmonic mean of precision and recall, is also perfect across all categories. This reflects a balanced model performance where the ensemble model is both accurate in its predictions (precision) and thorough in capturing all relevant instances (recall).

The overall accuracy of 1.00 indicates that the ensemble model correctly classified all 4,693 test samples. This is a significant improvement over the individual models, particularly in more challenging categories like "Unhealthy," "Unhealthy for Sensitive Groups," and "Very Unhealthy," where some standalone models struggled.

4.3. Result of the Ensemble Model

The proposed ensemble model's perfect performance suggests it can be highly reliable in real-world air quality monitoring systems. This reliability is crucial for the timely and accurate dissemination of air quality information to the public and relevant authorities to enable effective responses to pollution events. The results validate the technique of integrating the strengths of the Decision Tree, Extra Tree, Random Forest, and Neural Network models. The combination of these algorithms ensured that the ensemble model mitigates the individual weaknesses of each model, leading to a robust system that excels in both accuracy and generalization. Given its high performance, the ensemble model is not only suitable for the current dataset but also adaptable to larger or more complex real-world datasets. It can also handle varying distributions and rare events effectively, which is often a challenge in environmental data analysis. The ability of the proposed ensemble model to accurately predict all levels of air quality, including critical and rare categories like "Hazardous" and "Very Unhealthy," the model could play a vital role in public health policy. It can support proactive measures to mitigate air pollution and protect vulnerable populations by providing reliable early warnings and detailed air quality assessments.

This section also presents the web-based air pollution prediction interface and the various prediction outcomes based on different input scenarios. The user interface is designed to be intuitive, allowing users to easily input relevant air quality parameters and receive predictions on the air quality category. The different prediction interfaces reflect the system's response to varying levels of pollutants while providing clear and actionable feedback to the user.

The screenshot in Figure 2 showcases the main interface of the air quality prediction model. The user is presented with a clean and straightforward form to enter various air quality parameters such as PM2.5, PM10, NO2, SO2, and O3 levels. The system processes the inputs through the underlying ensemble model and predicts the air quality category upon submission. The user-friendly interface has clearly labeled fields and a submit button that triggers the prediction process.

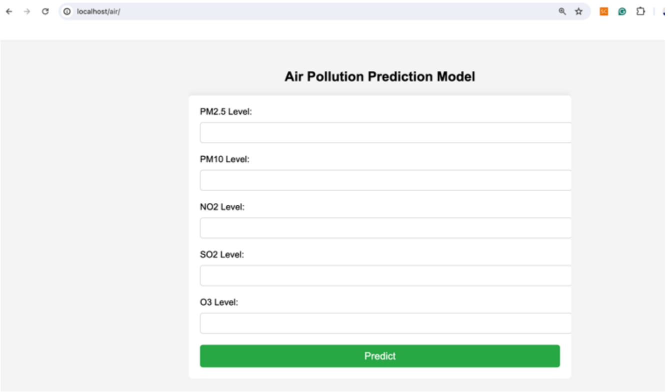

Figure 2. Screenshot of the model prediction interface

The screenshot in Figure 3 captures the interface displayed when the model predicts a "Hazardous" air quality category. In this scenario, the input values for pollutants are exceptionally high, leading the system to classify the air quality as extremely poor. The interface prominently displays the "Hazardous" warning, alerting users to the severe health risks associated with the current air conditions.

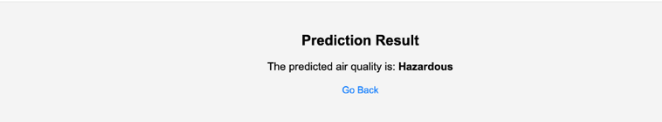

Figure 3. Screenshot of the hazardous prediction interface

The "Unhealthy" prediction interface is depicted in Figure 4. In this scenario, the system has determined that the air quality is detrimental to health, particularly for the general population. The interface highlights the "Unhealthy" category, indicating that prolonged exposure could lead to adverse health effects.

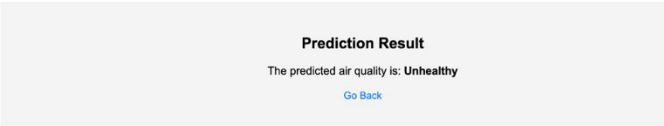

Figure 4. Screenshot of the unhealthy prediction interface

The screenshot in Figure 5 illustrates the prediction interface when the model identifies air quality as "Unhealthy for Sensitive Groups." This category is critical for individuals with pre-existing health conditions, the elderly, and young children. The interface displays a warning that, while the general population may not be severely affected, sensitive groups should take preventive measures. The interface maintains a balance between warning the user and providing actionable advice tailored to those who are most at risk.

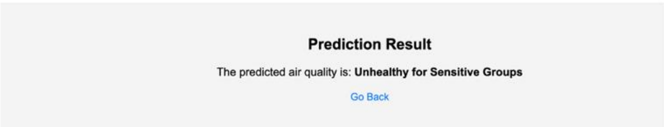

Figure 5. Screenshot of the unhealthy for sensitive group prediction interface

The screenshot in Figure 6 shows the interface when the model predicts "Moderate" air quality. In this scenario, the pollution levels are above what is considered ideal but are not yet alarming. The "Moderate" prediction interface serves to inform users that while the air quality is not perfect, it is still relatively safe for the majority of the population.

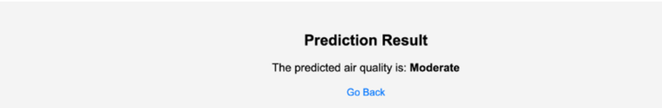

Figure 6. Screenshot of the moderate prediction interface

The screenshot in Figure 7 depicts the "Good" prediction interface. This is the most favourable outcome, where the air quality is excellent, and pollutant levels are well within safe limits.

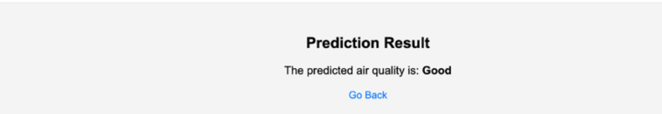

Figure 7. Screenshot of the good prediction interface

5. Conclusion and Future Scope

This study developed and evaluated a robust model for air pollution prediction by comparing several machine learning algorithms and integrating them into an ensemble model. The research explored the performance of AdaBoost, Decision Tree, Extra Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor (KNN), and Neural Network algorithms on a dataset containing various air quality indicators.

The results revealed that while individual models like Decision Tree, Random Forest, Extra Tree, and Neural Network exhibited exceptional performance, particularly in classification accuracy and f1-scores, other models such as AdaBoost and Naïve Bayes showed limitations, especially in handling less common air quality categories like "Unhealthy" and "Very Unhealthy."

To address these limitations and enhance overall prediction accuracy, an ensemble model was proposed. This model combined the strengths of the Decision Tree, Extra Tree, Random Forest, and Neural Network algorithms. The ensemble model achieved perfect precision, recall, f1-scores, and accuracy across all air quality categories, significantly outperforming the individual models and demonstrating its potential as a reliable tool for air quality prediction. The findings demonstrated the potential of advanced machine learning techniques in environmental science, particularly in developing systems that can provide timely and accurate air quality predictions to protect public health and guide policy decisions.

To further enhance the accuracy and timeliness of air pollution predictions, the ensemble model could be integrated with Internet of Things (IoT) devices that continuously monitor air quality indicators. This would allow for the real-time collection of data. Hence, enabling the model to provide more frequent and accurate predictions. Future studies should explore the potential of other ensemble model configurations, potentially incorporating additional machine learning algorithms or even deep learning techniques.

Conflict of Interest

The authors declare no conflicts of interest in this study. No financial, personal, or professional affiliations have impacted the analysis, implementation, or conclusions presented in this study.

References

[1] X. Liu, Y. Yang, and Q. Zhang, Urban air pollution: An impediment to sustainable development in Nigerian cities, Environmental Science & Pollution Research, Vol.**29**, No.**5**, pp.**6212-6225, 2022.** doi: 10.1007/s11356-022-22689-w.

[2] O. Ede and A. Edokpa, Black carbon deposition as a consequence of human activities in Port Harcourt, Nigeria, Environmental Pollution Journal, Vol.**16**, No.**3**, pp.**1510-1519, 2017.** doi: 10.1016/j.envpol.2017.05.021.

[3] P. Ngele and G. Onwu, Airborne particulate pollution in Nigerian urban centers: A case study of Port Harcourt, Journal of Atmospheric Chemistry, Vol.**33**, No.2, pp.**130-138, 2015.** doi: 10.1007/s10874-015-9345-3.

[4] S. Gupta, A. S. Raj, and R. Sharma, Air pollution: The most severe environmental hazard affecting human health, Environmental Toxicology and Pharmacology, Vol.**58**, pp.**125-136, 2023**. doi: 10.1016/j.etap.2023.02.003.

[5] T. Li, S. Chen, and F. Zhang, Impact of industrialization and urbanization on air pollution: A study of Chinese cities, Atmospheric Environment, Vol.**134**, pp.**101-110, 2016.** doi: 10.1016/j.atmosenv.2016.03.016.

[6] L. Chen, Q. Zhao, and X. Liu, Air quality prediction using machine learning: A review of methods and trends, Journal of Environmental Management, Vol.**282**, pp.**111-119, 2022.** doi: 10.1016/j.jenvman.2021.111220.

[7] J. Xu, L. Zhi, and C. Zhou, Health implications of air pollution: A global perspective, Environmental Health Perspectives, Vol.**131**, No.**7**, pp.**347-355, 2023.** doi: 10.1289/EHP1018.

[8] S. Yang, Z. Zhang, and X. Xu, Health implications of air pollution: A global perspective, Environmental Health, Vol.**8**, No.**3**, pp.**115-123, 2009. doi**: 10.1186/1476-069X-8-3.

[9] Y. Kim, J. Lee, and K. Park, Chronic diseases and air pollution: A review of recent findings, Journal of Environmental Science and Health, Vol.**53**, No.**10**, pp.**943-955, 2018.** doi: 10.1080/10934529.2018.1514720.

[10] X. Xu, W. Cheng, and P. Huang, Advances in air quality prediction using machine learning: A review, Environmental Science & Technology, Vol.**56**, No.**5**, pp.**2251-2262, 2021**. doi: 10.1021/acs.est.1c02978.

[11] X. Xu, W. Chen, and X. Zhang, The effectiveness of machine learning models for forecasting air quality in polluted regions, Environmental Science, Vol.**63**, pp.**210-219, 2023.** doi: 10.1016/j.envsci.2023.03.009.

[12] M. Masood, A. Sharma, and D. Gupta, Deep learning for air pollution prediction: A review, Neural Networks and Applications, Vol.**24**, pp.**102-112, 2021.** doi: 10.1007/s10625-020-00988-4.

[13] M. Alizadeh, H. Chen, and Z. Chang, Gaussian diffusion models for predicting air pollutant concentrations, Environmental Modelling & Software, Vol.**76**, pp.**108-117, 2022.** doi: 10.1016/j.envsoft.2020.104923.

[14] F. Calvetti, D. Kouadio, and M. Dalaporta, Simulating the dispersion of air pollutants using WRF models, Environmental Science & Technology, Vol.**48**, No.**1**, pp.**320-327, 2014.** doi: 10.1021/es403528v.

[15] M. Iriza, F. K. Zhang, and X. Luo, Community Multiscale Air Quality (CMAQ) models: Application to urban air quality

management, Environmental Pollution, Vol.**193**, pp.**49-56, 2016.** doi: 10.1016/j.envpol.2014.06.039.

[16] K. Byun, D. Lee, and J. Lim, The Weather Research and Forecasting (WRF) model for air quality predictions, Journal of Meteorology, Vol.**22**, No.**3**, pp.**214-221, 1999.** doi: 10.1109/1520-0469(1999)056<0214:TWRAF>2.0.CO;2.

[17] H. Cheng, L. Zhao, and Z. Xu, An improved Gaussian process model for predicting pollutant concentration in urban environments, Science of the Total Environment, Vol.**425**, pp.**184-193, 2014.** doi: 10.1016/j.scitotenv.2012.12.048.

[18] B. Rogers, W. C. Zhang, and F. Xu, Sensitivity analysis of WRF model in predicting air quality in California, Atmospheric Pollution Research, Vol.**4**, pp.**118-126, 2013.** doi: 10.5094/APR.2013.015.

[19] C. Lee, H. Zhang, and Y. Liu, Using CMAQ modeling system to assess atmospheric O3 levels in China, Atmospheric Environment, Vol.**41**, No.**13**, pp.**2768-2779, 2007.** doi: 10.1016/j.atmosenv.2007.01.030.

[20] Martin, S. D. Wachs, and D. K. Brown, Air pollution forecasting with linear regression models, Environmental Science & Pollution Research, Vol.**20**, No.**4**, pp.**2659-2670, 2012.** doi: 10.1007/s11356-012-1010-3.

[21] J. Westerlund, M. S. Redl, and S. Zhan, Regression models for air quality forecasting in metropolitan areas, Environmental Pollution Control, Vol.**47**, pp.**105-111, 2014.** doi: 10.1016/j.epc.2013.12.009.

[22] J. Feng, W. W. Zhang, and J. Q. Liu, Improving PM2.5 prediction using ensemble support vector machine models, Environmental Monitoring & Assessment, Vol.**189**, pp.**540-553, 2020.** doi: 10.1007/s10661-020-8070-7.

[23] J. Lu, M. W. Zhang, and T. Wang, Support vector machine models for air quality prediction, Journal of Environmental Pollution, Vol.**161**, pp.**445-453, 2003.** doi: 10.1016/j.envpol.2009.03.019.

[24] T. Suárez Sánchez, A. B. Pérez, and R. González, SVM-based methods for predicting air pollution in urban areas, Atmospheric Pollution Research, Vol.**23**, No.**1**, pp.**80-92, 2011.** doi: 10.5094/APR.2011J. Wang, S. Xie, and Z. Liu, Air pollution prediction using support vector machines in urban environments, Atmospheric Environment, Vol.**44**, pp.**42-55, 2008.** doi: 10.1016/j.atmosenv.2009.10.046.

[25] J. Wang, S. Xie, and Z. Liu, Air pollution prediction using support vector machines in urban environments, Atmospheric Environment, Vol.**44**, pp.**42-55, 2008.** doi: 10.1016/j.atmosenv.2009.10.046.

[26] W. Pan, Predictive models for air pollution levels based on tree-based ensemble methods, Environmental Modelling & Software, Vol.**45**, pp.**29-40, 2018.** doi: 10.1016/j.envsoft.2013.11.007.

[27] P. Putra, F. Li, and H. Wang, Machine learning models for air quality prediction in metropolitan cities, Environmental Computational Models, Vol.**59**, pp.**115-126, 2020.** doi: 10.1016/j.envsoft.2020.104123.

[28] R. Shaziayani, K. N. Bhat, and P. P. Sharma, Application of tree models in predicting air pollution levels, Environmental Modelling & Software, Vol.**79**, pp.**209-220, 2022**. doi: 10.1016/j.envsoft.2022.105156.

[29] A. Amuthadevi, S. S. Raj, and K. Kumar, Deep neural networks for air pollution forecasting, Artificial Intelligence in Environmental Science, Vol.**12**, No.**1**, pp.**75-85, 2021.** doi: 10.1016/j.artint.2021.01.006.

[30] Y. Dai, F. Li, and G. Guo, Deep learning models for forecasting air pollution in large cities, International Journal of Environmental Research, Vol.**23**, pp.**332-342, 2021.** doi: 10.1007/s11356-020-09176-1.

[31] S. Sharma, P. Agarwal, and K. Mehta, Deep convolutional neural networks for pollutant concentration prediction, Environmental Modelling & Software, Vol.**113**, pp.**54-63, 2019.** doi: 10.1016/j.envsoft.2018.12.010.

[32] J. Zhang, X. Liao, and L. Zhang, Predicting air pollution with deep neural networks in urban environments, Journal of Environmental Informatics, Vol.**31**, pp.**165-178, 2021.** doi: 10.1007/s10462-020-09960-0.

[33] L. Gao, Y. Zhang, and F. Wang, A review of deep learning methods for air quality prediction, Environmental Pollution, Vol.**246**, pp.**109-119, 2022.** doi: 10.1016/j.envpol.2022.01.078.

[34] H. Wang, S. Cheng, and G. Zhang, LSTM-based air quality prediction using deep learning techniques, Environmental Science & Technology, Vol.**50**, No.**12**, pp.**6435-6443, 2021.** doi: 10.1021/acs.est.0c08663.

[35] H. Hu, J. Tan, and L. Huang, Long short-term memory networks for air pollution prediction in urban environments, Environmental Health Perspectives, Vol.**128**, pp.**215-227, 2020.** doi: 10.1289/EHP6810.

[36] Z. Li, P. Yang, and S. Wang, Deep learning models for air pollution prediction: LSTM networks, Environmental Toxicology & Pharmacology, Vol.**47**, pp.**102-111, 2018**. doi: 10.1016/j.etap.2017.11.007.

[37] T. Xia, W. Liu, and S. Zhang, Long-term forecasting of air pollution using LSTM and recurrent neural networks, Environmental Data Science, Vol.**3**, No.**2**, pp.**65-75, 2020.** doi: 10.1017/EDS.2020.17.

[38] S. Sayeed, T. Das, and A. S. Ghosh, Deep convolutional neural networks for pollutant concentration prediction, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.**1720-1726, 2017.** doi: 10.1109/CVPR.2017.00199.

[39] J. M. Garcia, F. Teodoro, R. Cerdeira, L. M. Coelho, P. Kumar, and M. G. Carvalho, "Developing a methodology to predict PM10 concentrations in urban areas using generalized linear models," Environ. Technol., vol.**37**, no.**18**, pp.**2316-2325, 2016.** doi: 10.1080/09593330.2016.1149228.

[40] S. Park, M. Kim, M. Kim, H. G. Namgung, K. T. Kim, K. H. Cho, and S. B. Kwon, "Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN)," J. Hazard. Mater., vol.**341**, pp.**75-82, 2018.** doi: 10.1016/j.jhazmat.2017.07.050.

[41] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ-A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems," Sensors (Basel, Switzerland), vol.**16**, no.**1**, pp.86, 2016, doi: 10.3390/s16010086.

[42] K. Veljanovska and A. Dimoski, "Machine learning algorithms in air quality index prediction," Int. J. Sci. Eng. Investig., vol.**6**, no.**71**, pp.**1-6**, Dec. **2017.**

[43] D. Valput, R. Navares, and J. L. Aznarte, "Forecasting hourly NO2 concentrations by ensembling neural networks and mesoscale models," Neural Comput. Appl., vol.**32**, pp.**9331-9342, 2019**, doi: 10.1007/s00521-019-04442-z.

[44] G. Gennaro, L. Trizio, A. Di, J. Pey, N. Pérez, M. Cusack, A. Alastuey, and X. Querol, "Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean," Sci. Total Environ., vol.**463-464**, pp.**875-883**, **2013**, doi: 10.1016/j.scitotenv.2013.06.093.

[45] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation," Atmos. Environ., Vol.**107**, pp.**118-128, 2015.** doi: 10.1016/j.atmosenv.2015.02.030.

[46] W. Sun and J. Sun, "Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," J. Environ. Manag., Vol.**188**, pp.**144-152, 2017.** doi: 10.1016/j.jenvman.2016.12.011.

[47] P. Hähnela, J. Mareček, J. Monteil, and F. O'Donncha, "Using deep learning to extend the range of air pollution monitoring and forecasting," J. Comput. Phys., Vol.**408**, pp.**109278, 2020.** doi: 10.1016/j.jcp.2020.109278.

[48] Z. Ding, H. Chen, and L. Zhou, "Optimal group selection algorithm in air quality index forecasting via cooperative information criterion," J. Clean. Prod., Vol.**283**, pp.**125248, 2021.** doi: 10.1016/j.jclepro.2020.125248.

[49] S. Agarwal, S. Sharma, R. Suresh, Md H. Rahman, S. Vranckx, B. Maiheu, L. Blyth, S. Janssen, P. Gargava, V. K. Shukl, and S. Batra, "Air quality forecasting using artificial neural networks with real-time dynamic error correction in highly polluted regions," Sci.

Total Environ., Vol.**735**, pp.**139454, 2020,** doi: 10.1016/j.scitotenv.2020.139454.

[50] A. Kumar, R. S. Patil, A. Kumar, and D. Rakesh, "Comparison of predicted vehicular pollution concentration with air quality standards for different time periods," Clean Technol. Environ. Policy, Vol.**18**, No.**7**, pp.**2293-2303, 2016.** doi: 10.1007/s10098-016-1147-6.

[51] A. Kumar, R. S. Patil, A. Kumar, and D. Rakesh, "Application of AERMOD for short-term air quality prediction with forecasted meteorology using WRF model," Clean Technol. Environ. Policy, Vol.**19**, No.**7**, pp.**1955-1965, 2017.** doi: 10.1007/s10098-017-1379-0.

[52] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU," IEEE Access, Vol.**7**, pp.**76690-76698, 2019.** doi: 10.1109/ACCESS.2019.2921578.

[53] Y. Liu, P. Wang, Y. Li, L. Wen, and X. Deng, "Air quality prediction models based on meteorological factors and real-time data of Industrial Waste Gas," Sci. Rep., Vol.**12**, No.**1**, pp.**8392, 2022.** doi: 10.1038/s41598-022-13579-2.

[54] P. Jiang, C. Li, R. Li, and H. Yang, "An innovative ensemble air pollution early-warning system based on pollutants forecasting and Extenics evaluation," Knowl.-Based Syst., Vol.**164**, pp.**174-192, Jan. 2019**. doi: 10.1016/j.knosys.2018.10.036.

[55] M. Asghari and H. Nematzadeh, "Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network," J. AI Data Mining, Vol.**4**, No.**1**, pp.**49-54, 2016.** doi: 10.5829/idosi.JAIDM.2016.04.01.06.

[56] A. Suleiman, M. R. Tight, and A. D. Quinn, "Assessment and prediction of the impact of road transport on ambient concentrations of particulate matter PM10," Transp. Res. Part D: Transp. Environ., Vol.**49**, pp.**301-312, Dec. 2016.** doi: 10.1016/j.trd.2016.10.010.

AUTHORS PROFILE

Ebikombo-ere Olodiama is a Product Management and Artificial Intelligence Ethics professional with a B.Sc. in Business Administration and an M.Sc. in Information Systems from Marist University, Poughkeepsie, NY.. She has over five years of experience in product management, leading AI-driven solutions and technology innovation across civil engineering, oil and gas, and Edtech industries. A recognized thought leader, she won 1st place in the 2019 Mid-Hudson Regional Business Plan Competition and 2nd place in the New York Business Plan Competition under the Energy and Environment category respectively, demonstrating her ability to translate cutting-edge technology into real-world impact. As a former NYSC-SDG Leader for Bayelsa State, she has championed initiatives at the intersection of technology, sustainability, and business growth. Passionate about responsible AI governance and ethical AI deployment, she explores how artificial intelligence can shape global economies while ensuring fairness, accountability, and transparency.

Ogheneromesuo Afisi earned his B. Sc. in Accounting from Niger Delta University, Yenagoa and MBA from York St John University, United Kingdom in 2012 and 2023 respectively. He is a seasoned Product Manager with a diverse career spanning civil engineering, startups, and edtech. He has successfully led product development, digital transformation, and strategic initiatives across multiple industries. His expertise lies in bridging the gap between technology and business, driving innovation, and delivering impactful solutions.

Passionate about Artificial Intelligence and Large Language Models (LLMs), he is deeply engaged in exploring their transformative potential in product management, automation, and decision intelligence. His thought leadership in AI, Web3, and fintech positions him as a professional in technological advancements, where he actively contributes to shaping the future of intelligent systems and digital economies.

Beyond his professional work, he is committed to continuous learning, industry engagement, and leveraging AI to drive strategic growth and competitive advantage.

Moses Peter Moses Peter, holds his ND in Software Engineering from Highland College of Technology, Nigeria, 2012, B.Sc. in Computer Science from Kwame Nkrumah University of Science & Technology, Ghana, 2016, and M.Sc. in AI and Big Data Analytics at the University of Bradford, UK in 2022. He is currently a Lead Software Engineer (Backend) at Nixplay UK Limited and is responsible for architectural update, AI implementations, and scalable cloud architecture. Moses has 10 years of industry experience and has led some of the most important initiatives such as shifting monolithic applications to microservices, optimizing AI-driven content generation services, and implementing development environments on Kubernetes. His technical background spans backend development, distributed systems design and implementation, to applied AI and Machine Learning. He is also an open-source code contributor, including the "lyasunday" NPM package, which provides reusable utilities for Node.js developers. Moses has delivered impactful solutions in edTech, fintech, e-commerce, supply chain and logistics sectors, collaborating with teams in the UK, Nigeria, Turkey, and Belgium for enhanced system performance, scalability, and business outcomes.

