http://dx.doi.org/10.21622/ACE.2025.05.2.1428

INTERPRETABLE AIR QUALITY CLASSIFICATION FOR PUBLIC HEALTH USING MACHINE LEARNING

Godfrey Perfectson Oise[1], Cyprian C. Konyeha[2], Chioma Julia Onwuzo[3], Ejenarhome Prosper Otega[4], Babalola Eyitemi Akilo[5], Olayinka Tosin Comfort[6], Joy Akpowehbve Odimayomi[7] and Unuigbokhai Nkem Belinda[8]

[1] Wellspring University, Department of Computing, Benin City, Edo State, Nigeria.

[2] Benson Idahosa University, Department of Electrical and Electronic Engineering, Edo State.

[3] Micheal Okpara University of Agriculture, Umudike, Abia State. Nigeria.

[4] Delta State University, Department of Computer Science, Abraka, Delta State

[5:8] Wellspring University, Department of Computing, Benin City, Edo State, Nigeria.

Emails: {godfrey.oise@wellspringuniversity.edu.ng, ckonyeha@biu.edu.ng, onwuzo.chioma@mouau.edu.ng, otega-prosper@delsu.edu.ng, akilo.babalola@wellspringuniversity.edu.ng, tosin.olayinka@wellspringuniversity.edu.ng, odimayomi.joy@wellspringuniversity.edu.ng, unuigbokhai.belinda@wellspringuniversity.edu.ng}

Received on, 11 June 2025 - Accepted on, 25 July 2025 - Published on, 07 October 2025

ABSTRACT

Air pollution remains a significant environmental and public health challenge in rapidly urbanizing regions. Accurately predicting air quality levels is critical for effective environmental management and public health interventions. This study investigates the application of logistic regression for multi-class air quality classification using a comprehensive dataset of 23,463 records. The dataset integrates pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO), meteorological variables (temperature and humidity), and socio-demographic features (industrial zone classification and population distribution). The target variable, the Air Quality Index (AQI), is categorized into six classes: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. The preprocessing pipeline involves median imputation for missing values, feature normalization to ensure consistency across variables, and encoding of categorical features using both one-hot and ordinal encoding strategies. Feature selection is carried out using Pearson correlation, mutual information, and Recursive Feature Elimination (RFE), identifying PM2.5 as the most predictive variable with a correlation coefficient of 0.98 with overall AQI. The logistic regression model, selected for its simplicity and interpretability, is trained on the processed data. The model achieves a perfect classification score of 1.00, precision, recall, f1-score, and accuracy on a demonstration set of 30 records, supported by a perfectly diagonal confusion matrix. However, given the imbalanced nature of the full dataset, with "Good" and "Moderate" categories dominating, further evaluation on broader subsets is recommended. Visual analytics, including histograms, box plots, and correlation heatmaps, reaffirm the dominant influence of PM2.5 in determining overall air quality. The study demonstrates that logistic regression offers a robust, interpretable, and computationally efficient solution for air quality prediction. Future work will focus on addressing class imbalance, integrating real-time data, and benchmarking against more complex machine learning models to enhance prediction robustness and generalizability.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Keywords: *Air Quality Index (AQI), Logistic Regression, Machine Learning, Environmental Prediction, Demographic and Geospatial Data, Public Health*

1. INTRODUCTION

The persistent deterioration of air quality in urban centers has become a pressing global concern, particularly in the face of rapid industrialization, increased vehicular emissions, and rising population densities. Poor air quality is linked to a spectrum of adverse health outcomes, including respiratory infections, cardiovascular diseases, and increased mortality rates [1]. Simultaneously, air pollution degrades natural ecosystems and undermines sustainable development efforts. Understanding and predicting air quality levels using machine learning (ML) models enables proactive mitigation strategies [2]. Logistic regression, a well-established and interpretable algorithm, has shown promise in multi-class classification tasks. This study investigates the potential of logistic regression for predicting air quality categories using a dataset enriched with both environmental and demographic features [3]. Beyond the widely acknowledged culprits like industrial activities and vehicular exhaust, air pollution is exacerbated by a complex interplay of factors, including rapid and often unplanned urbanization, agricultural practices, and specific meteorological conditions [4]. For instance, temperature inversions can trap pollutants close to the ground, leading to severe localized events. Furthermore, transboundary pollution, where pollutants from one region drift to another, highlights the global interconnectedness of this issue. The dynamic and highly variable nature of air quality, influenced by daily, seasonal, and event-specific changes, presents significant challenges for accurate prediction and management [5]. The repercussions of poor air quality extend far beyond direct health impacts, permeating various societal and economic spheres. Economically, chronic exposure to pollutants can lead to reduced labor productivity, increased healthcare expenditures, and decreased tourism in affected regions. Environmentally, air pollution contributes to acid rain, damages crops and forests, and impacts biodiversity, further undermining sustainable development goals [6]. Moreover, the issue of environmental justice is salient, as marginalized communities and lower-income populations are often disproportionately exposed to higher levels of pollution due to their proximity to industrial zones or major transportation routes, exacerbating existing health disparities [7]. Traditional methods for air quality prediction, such as simple statistical models like ARIMA or basic multivariate regression, often struggle to capture the inherent non-linearity, complex interactions, and high-dimensional nature of environmental data [8]. Their limitations become particularly evident when dealing with categorical outputs like air quality classifications. In contrast, machine learning approaches, ranging from decision trees and support vector machines (SVMs) to sophisticated neural networks, offer superior capabilities in modeling these intricate relationships and handling large, heterogeneous datasets [9]. The choice of an appropriate ML model often involves a trade-off between predictive accuracy and interpretability, a crucial consideration for practical application. In this context, while more complex models have gained prominence, logistic regression remains a compelling choice for multi-class air quality classification. Its core strength lies in its interpretability, allowing stakeholders to understand why a particular air quality prediction is made based on the influence of specific features [10].

This transparency is invaluable for policy-makers who need to justify interventions and resource allocation. Furthermore, logistic regression's simplicity, efficiency, and relatively low computational overhead make it a robust and accessible solution, suitable for deployment even in resource-constrained environments or as a strong baseline component within more elaborate ensemble systems [11]. The integration of both environmental (pollutant and meteorological) and demographic (industrial zone, population distribution) features in our dataset is designed to provide a holistic view, enabling the model to capture not only the direct environmental drivers but also the anthropogenic and spatial factors influencing air quality, thereby leading to more comprehensive and actionable insights. This research addresses the critical global concern of deteriorating urban air quality, emphasizing its severe links to adverse health outcomes, ecological degradation, and hindered sustainable development. It posits that machine learning (ML) models are essential for proactive air quality prediction and mitigation [12]. The introduction highlights the multifaceted nature of air pollution, influenced by industrialization, emissions, population growth, and complex factors like urbanization, agricultural practices, and meteorological conditions, including

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

transboundary pollution, all contributing to the dynamic variability of air quality [13]. The research further elaborates on the far-reaching repercussions of poor air quality, extending beyond health to encompass significant economic losses through reduced productivity, increased healthcare costs, and diminished tourism. It also touches upon environmental impacts such as acid rain and ecosystem damage [14], alongside the critical aspect of environmental justice, where vulnerable communities often bear a disproportionate burden of pollution exposure [15]. Critiquing traditional air quality prediction methods like ARIMA and multivariate regression, the paper asserts their inadequacy in capturing non-linearity and handling categorical outputs. It contrasts these limitations with the superior capabilities of modern ML approaches, including decision trees, support vector machines, and neural networks, in managing complex, high-dimensional data [16]. Despite the rise of these more intricate models, the paper champions logistic regression for its interpretability, simplicity, and computational efficiency, making it suitable for resource-constrained environments or as a foundational model in ensemble systems. The study's unique contribution lies in its dataset, which enriches environmental data with demographic features like industrial zones and population distribution, aiming to provide a holistic and actionable understanding of air quality dynamics for improved environmental and public health management.

RELATED WORKS

Prior research in air quality prediction has explored a range of techniques from statistical modeling to complex deep learning architectures. Time-series models, such as ARIMA and multivariate regression, have been employed with limited success due to their inability to model non-linearity and categorical outputs effectively [17]. In contrast, ML approaches, including decision trees, support vector machines (SVMs), and neural networks, have demonstrated improved performance. However, despite the rise of complex models, logistic regression remains underutilized for multi-class air quality classification. Its interpretability, simplicity, and low computational overhead make it well-suited for deployment in resource-constrained environments or as a baseline model in ensemble systems. [18], Regression analysis is a key technique for predicting continuous outcomes based on multiple input variables, offering clear estimates of relationships between inputs and outputs along with error estimates from optimization algorithms. Commonly used in communication networks and IoT applications, regression models are often built using statistical machine learning methods to establish parametric relationships [19]. This includes generalized linear models like linear and logistic regression, as well as ridge and polynomial regression. The study delves into the theory behind these models, provides pseudocode for practical implementation, and discusses challenges in data analysis, model selection, cost functions, optimization strategies, cross-validation, and regularization techniques. [20], Analyzed how various factors influenced psychological distress during the early COVID-19 outbreak in China, using data from 937 respondents. Health-related factors were the strongest predictors, followed by [17]objective and perceived environmental risks. Distress levels increased sharply at specific AQI thresholds, and gender differences were noted in responses. Overall, perceived indoor air quality was more strongly linked to psychological distress than outdoor pollution. [21],The rapid increase in global population and unchecked urbanization have led to serious environmental issues, notably declining air quality linked to various health risks [22]balancing data privacy with clinical utility. The decentralized system enables multi-institutional collaboration without centralized data collection, complying with HIPAA/GDPR through two technical safeguards: differential privacy via DP-SGD during local training and secure aggregation of model updates. Using LSTM/GRU architectures optimized for sequential medical data, the framework achieves an F1 Score of 67% with precision (60%. In response, machine learning has emerged as a vital tool for predicting air quality, drawing widespread academic interest.

This paper offers a comprehensive bibliometric analysis of machine learning applications in air quality prediction, based on 1992–2021 publications indexed in the Web of Science. Utilizing S-curve and social network analyses, the study tracks the evolution of research output, revealing a significant surge between 2017 and 2021, during which 68.51% of all publications appeared. Italy, Greece, and Spain led in international collaboration impact. Frequent keywords include 'air pollution', 'air quality', 'machine learning', and 'forecasting', highlighting key research themes. The study also explores the transition from traditional

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

to machine learning methods, providing insights into influential works and emerging trends [23]. Ultimately, it aims to guide future research and policy-making in air quality prediction and pollution management. [24],Air pollution poses significant threats to public health and economic development, making accurate air quality prediction essential for effective management. This paper proposes a hybrid model combining ARIMA and CNN-LSTM to enhance AQI prediction accuracy using real data from four cities. ARIMA captures the linear trends, while CNN-LSTM addresses nonlinear patterns. To optimize the CNN-LSTM's hyperparameters and prevent suboptimal settings, the Dung Beetle Optimizer algorithm is employed [25]the research collects a large amount of data from images of e-waste and then carefully preprocesses and augments those images. With precision, recall, and F1 scores of 87%, 86%, and 86%, respectively, the SNN architecture—which incorporates dropout, pooling, and convolutional layers—achieved an amazing 100% classification accuracy. These outstanding outcomes show how well the model can classify e-waste components, suggesting that it has the potential to be used in real-world scenarios. The results indicate that the SNN-based approach greatly improves the accuracy and efficiency of e-waste sorting, promoting environmental sustainability and resource conservation. By automating the sorting process, the suggested system decreases the need for manual labor, minimizes human error, and speeds up processing. The study emphasizes the model's suitability for integration into current e-waste management workflows, providing a scalable and dependable way to expedite the recycling process. Additionally, the model's real-time applicability highlights its potential to revolutionize current e-waste management practices, making a positive ecological impact. . Future research endeavors will center on broadening the dataset to include a wider range of e-waste image categories, investigating more advanced deep learning architectures, and incorporating the system with Internet of Things (IoT. The proposed model's performance is benchmarked against nine popular models, demonstrating superior accuracy.

Experimental results show notably low RMSE and MAE values and high R^2 scores across all cities, confirming the model's effectiveness in predicting air quality. [26], Introduces the Dung Beetle Optimizer (DBO), a novel population-based optimization algorithm inspired by the natural behaviors of dung beetles, including ball-rolling, dancing, foraging, stealing, and reproduction. Designed to balance global exploration with local exploitation, the DBO algorithm achieves a fast convergence rate and high solution accuracy. Its performance is rigorously tested using 23 benchmark functions and 29 CEC-BC-2017 functions, where it demonstrates competitive results in terms of accuracy, stability, and speed compared to existing optimization methods. Statistical analyses, including the Wilcoxon signed-rank test and the Friedman test, confirm its superiority. Additionally, the DBO algorithm is successfully applied to three real-world engineering design problems, showcasing its effectiveness and practical application potential. [27], Amid rising industrialization in rapidly developing countries, air pollution has become a growing public health issue. This study investigates the impact of air pollution on hospital visits for respiratory diseases, focusing on Acute Respiratory Infections (ARI), using data collected from March 2018 to October 2021. Eight machine learning models, including Random Forest, KNN, Linear Regression, LASSO, Decision Tree, SVR, XGBoost, and a 5-layer Deep Neural Network, were employed to analyze the relationship between daily air pollutants and outpatient ARI visits. Evaluation using 5-fold cross-validation revealed that the Random Forest model performed best, especially on total patient data ($R^2 \approx 0.872$), while performance on ARI-specific cases was moderate ($R^2 \approx 0.606$). The study found limited correlation between ARI cases and air pollution, likely due to data gaps during the COVID-19 pandemic. Nevertheless, the findings suggest the potential of machine learning for broader disease risk prediction beyond ARI, including cardiovascular and other respiratory conditions. [28], Air pollution poses a serious health risk, especially in developing countries, making it crucial to identify and monitor pollution sources for effective local interventions.

This study evaluates the effectiveness of using Sentinel-5 satellite products via Google Earth Engine (GEE) to monitor key air pollutants CO, NO_2, SO_2, and O_3 in Arak, Iran, from 2018 to 2019. By processing satellite imagery in JavaScript on the GEE platform and applying cloud and average filters, the study produced monthly, seasonal, and annual pollution maps. Validation against ground-based data from the Environmental Organization of Central Province showed that the model achieved reasonably low RMSE values, confirming the accuracy of pollutant estimates across both years. The findings demonstrate that combining Sentinel-5

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

data with automated, cloud-based platforms like GEE offers a more efficient and spatially comprehensive alternative to traditional pollution monitoring methods, highlighting its potential for large-scale air quality assessment. [29], Understanding the three-dimensional (3D) distribution of air pollution remains a challenge with current single-method monitoring technologies [30]. To address this, the Space-Air-Ground integrated system emerges as a promising solution, combining heterogeneous technologies for long-term, high-precision, and large-scale atmospheric monitoring. This system utilizes ground-based optical remote sensing (on fixed or mobile platforms), air-based observations via tethered balloons, UAVs, and aircraft, and space-based monitoring through satellite remote sensing. Beyond mapping 3D pollution distribution, the system has facilitated studies on emission estimation and pollution mechanisms. Advancing this approach further requires research into multi-source data fusion, improved inversion algorithms, and integration with atmospheric models to fully realize its potential. [31], Evaluated the effectiveness of machine learning models Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), and Multiple Linear Regression (MLR) for predicting PM10 and PM2.5 levels in Macao using data from 2013 to 2021. While all models performed similarly for 2019 and 2021, RF outperformed the others in 2020 during the COVID-19 pandemic, when air pollution levels dropped significantly. The findings highlight RF as the most reliable model for forecasting pollutant concentrations, especially during periods of sudden environmental change. [32]but these models require large computational resources and often suffer from a systematic bias that leads to missed poor air pollution events. For example, a CTM-based operational forecasting system for air quality over the Pacific Northwest, called AIRPACT, uses over 100 processors for several hours to provide 48-h forecasts daily, but struggles to capture unhealthy O\n 3\n episodes during the summer and early fall, especially over Kennewick, WA. This research developed machine learning (ML, Developed a machine learning (ML) framework to improve ozone (O_3) forecasting in Kennewick, WA, addressing limitations of traditional chemical transport models (CTMs) like AIRPACT, which require high computational resources and often miss high-pollution events. Using meteorological and ozone data from 2017–2020, two ML models were trained: ML1 (for high O_3 events) and ML2 (for moderate events). ML1, combining random forest and linear regression, outperformed AIRPACT by detecting 5 of 10 unhealthy O_3 episodes, while ML2 was better at forecasting moderate levels. Since May 2019, the ML system has provided reliable 72-hour forecasts online using only a single processor, demonstrating higher efficiency and improved accuracy over CTMs. The literature review reveals that while traditional models like ARIMA and multivariate regression struggle with the complexity of environmental data, modern machine learning approaches offer better performance but often lack interpretability. Logistic regression, though less commonly used for multi-class air quality prediction, stands out for its simplicity, transparency, and low computational cost, making it suitable for both standalone use and as a baseline in ensemble models. Incorporating demographic and geospatial features alongside environmental data enhances predictive accuracy and supports more informed, equitable decision-making. Overall, the review justifies the study's use of logistic regression as a practical and interpretable solution for air quality classification.

METHODOLOGY

The methodology employed in this paper for multi-class air quality prediction using logistic regression follows a structured approach, encompassing dataset description, data preprocessing, feature selection, and model application.

DATASET DESCRIPTION

The study utilizes a dataset containing 5,000 pollution measurements, incorporating a diverse range of features to support air quality classification [33] . These features include pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, and CO; meteorological data including temperature (°C) and humidity (%); and demographic and geospatial information such as industrial zone classification and population distribution. The primary goal is to classify air quality using the Air Quality Index (AQI), which is categorized into six distinct levels: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

TOOLS AND TECHNOLOGIES

The project leverages several Python libraries to support different stages of the machine learning pipeline. Pandas is used for data manipulation and analysis, providing tools to efficiently handle and preprocess the dataset. Scikit-Learn is employed for model development, encompassing tasks such as preprocessing, feature selection, and the implementation of logistic regression. Seaborn is utilized for data visualization, enabling the creation of informative and visually appealing plots to better understand data patterns and model behavior.

DATA PREPROCESSING

Before training the logistic regression model, the raw dataset underwent several preprocessing steps to enhance data quality and model readiness. Missing value treatment was performed using median imputation to address gaps in pollutant readings, offering a robust solution that mitigates the influence of outliers. Normalization was applied to standardize features to a zero mean and unit variance, a crucial step for algorithms like logistic regression that are sensitive to feature scales. For categorical encoding, two methods were used: one-hot encoding was applied to the 'industrial zones' feature to convert nominal categories into a machine-readable numerical format, while ordinal encoding was used for the AQI labels, assigning a ranked order to air quality categories (e.g., Good < Moderate < Unhealthy).

FEATURE SELECTION

Several feature selection techniques were employed to identify the most relevant features for predicting air quality and to potentially reduce model complexity while improving performance. Pearson correlation was used to assess the linear relationship between numerical features and the target variable, helping to highlight features with strong linear associations. Mutual information, a nonlinear method, was applied to quantify the dependency between variables, revealing how much information one feature provides about another. Additionally, Recursive Feature Elimination (RFE), a wrapper-based approach, was utilized to iteratively remove less important features and build models on the remaining attributes, ultimately selecting the most impactful subset of features for prediction.

MODEL TRAINING AND EVALUATION

The core of this study centers on the application of logistic regression, a widely used and interpretable algorithm, for multi-class classification of air quality levels. The methodology involves two key components. First, model training was conducted using the preprocessed and selected features, enabling the logistic regression model to learn the relationships between environmental and demographic variables and the categorized air quality outcomes. Second, a real-world evaluation of the trained model was performed to assess its effectiveness in predicting air quality, providing insights with potential implications for environmental management and public health decision-making.

RESULTS

The results section of this paper meticulously details the findings from the multi-class air quality prediction using logistic regression, commencing with a comprehensive overview of the dataset's inherent characteristics. The dataset encompasses 23,463 entries across 12 distinct columns, providing a rich blend of information. These columns intricately capture various facets, including specific pollutant levels (PM2.5, PM10, NO_2, SO_2, CO), crucial meteorological readings (temperature and humidity), and relevant socio-demographic indicators such as industrial zone classifications and population distributions. This robust collection of features serves as the foundation for the predictive model, aiming to accurately categorize air quality from "Good" to "Hazardous" based on these diverse inputs. A granular examination of the dataset further elucidates the distribution of the target variable, the

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Air Quality Index (AQI) categories, alongside other critical features. The analysis reveals a predominant concentration of records within the "Good" and "Moderate" air quality classifications, indicating that, for the measured period, these conditions were more frequent. Conversely, categories such as "Very Unhealthy" and "Hazardous" appear less commonly within the dataset. This inherent imbalance in the target variable's distribution is a crucial insight, as it often necessitates specific handling during model training to prevent bias towards the majority classes. Additionally, various visualizations, including histograms, further illuminate the individual frequency distributions of key pollutant AQI values, providing a clear visual representation of their prevalence and ranges within the dataset. Delving deeper into the interrelationships between features, a correlation heatmap was instrumental in highlighting the strong linear associations among the numerical AQI values. Of particular significance was the remarkably high correlation coefficient of 0.98 observed between the general AQI Value and the PM2.5 AQI Value. This compelling finding strongly suggests that fine particulate matter (PM2.5) is a disproportionately dominant factor in determining the overall reported air quality, underscoring its critical importance as a predictive feature. This robust correlation implies that fluctuations in PM2.5 levels are highly indicative of broader changes in air quality. Complementing the correlation analysis, a series of detailed box plots provided profound insights into how the ranges of various pollutant concentrations, including CO, Ozone, and NO2, systematically shift across the different air quality categories. These visual representations effectively demonstrated a clear progression: as air quality deteriorates from "Good" to "Hazardous," the median and spread of pollutant values generally increase. Such visual evidence not only reinforces the predictive power of these pollutants but also offers an intuitive understanding of their thresholds and typical concentrations associated with each air quality classification, further solidifying their role in the predictive model.

The paper proceeds to present the performance metrics of the developed classification model, which, given the methodology, is likely based on logistic regression. The classification report exhibits exceptionally high performance, reporting perfect scores (precision, recall, f1-score, and an overall accuracy of 1.00) for the 'setosa', 'versicolor', and 'virginica' classes based on a very small sample of 30 records. This near-perfect outcome, also reflected in a perfectly diagonal confusion matrix (showing no misclassifications for 'Good', 'Hazardous', and 'Moderate' categories in this specific evaluation), strongly suggests that these reported metrics might stem from a highly curated or simplified validation set. While impressive, it prompts a consideration that such results may not be fully representative of the model's performance on the entire, more complex dataset, especially given the imbalanced nature of the actual air quality categories.

Finally, the practical utility of the model is vividly demonstrated through a predictive example, illustrating its real-world applicability in forecasting future air quality conditions. The demonstration involves a code snippet where new hypothetical observations, comprising specific PM2.5, PM10, NO2, and CO AQI values, are input into the system. These inputs are first appropriately scaled using the pre-trained scaler and then processed by the trained model. In the presented example, the model successfully predicts the air quality category as "Good." This functional capability highlights the model's immense potential as a robust, data-driven tool for proactive environmental management and public health initiatives, offering a means to anticipate and mitigate the impacts of air pollution effectively.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Table 1: Air Quality Index (AQI) Summary by Country and City

First 5 rows:

	Country	City	AQI Value	AQI Category	CO AQI Value
0	Russian Federation	Praskoveya	51	Moderate	1
1	Brazil	Presidente Dutra	41	Good	1
2	Italy	Priolo Gargallo	66	Moderate	1
3	Poland	Przasnysz	34	Good	1
4	France	Punaauia	22	Good	0

	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	\
0	Good	36	Good	0	
1	Good	5	Good	1	
2	Good	39	Good	2	
3	Good	34	Good	0	
4	Good	22	Good	0	

	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Good	51	Moderate
1	Good	41	Good
2	Good	66	Moderate
3	Good	20	Good
4	Good	6	Good

This table provides a concise snapshot of air quality data, displaying the first five rows of a dataset that includes geographical information (Country, City), the overall Air Quality Index (AQI) value, and its corresponding categorical classification (e.g., Good, Moderate). Additionally, it breaks down the air quality by specific pollutants, presenting individual AQI values and categories for Carbon Monoxide (CO), Ozone, Nitrogen Dioxide (NO2), and Particulate Matter 2.5 (PM2.5), thereby illustrating how various pollutant levels contribute to the total air quality status at different locations.

Table 2. Descriptive Statistics of Air Quality Index (AQI) Parameters

Descriptive Statistics:

	AQI Value	CO AQI Value	Ozone AQI Value	NO2 AQI Value
count	23463.000000	23463.000000	23463.000000	23463.000000
mean	72.010868	1.368367	35.193709	3.063334
std	56.055220	1.832064	28.098723	5.254108
min	6.000000	0.000000	0.000000	0.000000
25%	39.000000	1.000000	21.000000	0.000000
50%	55.000000	1.000000	31.000000	1.000000
75%	79.000000	1.000000	40.000000	4.000000
max	500.000000	133.000000	235.000000	91.000000

	PM2.5 AQI Value
count	23463.000000
mean	68.519755
std	54.796443
min	0.000000
25%	35.000000
50%	54.000000
75%	79.000000
max	500.000000

This table of Descriptive Statistics provides a comprehensive summary of the numerical air quality index (AQI) values within the dataset. It reveals that most pollutant-specific AQI

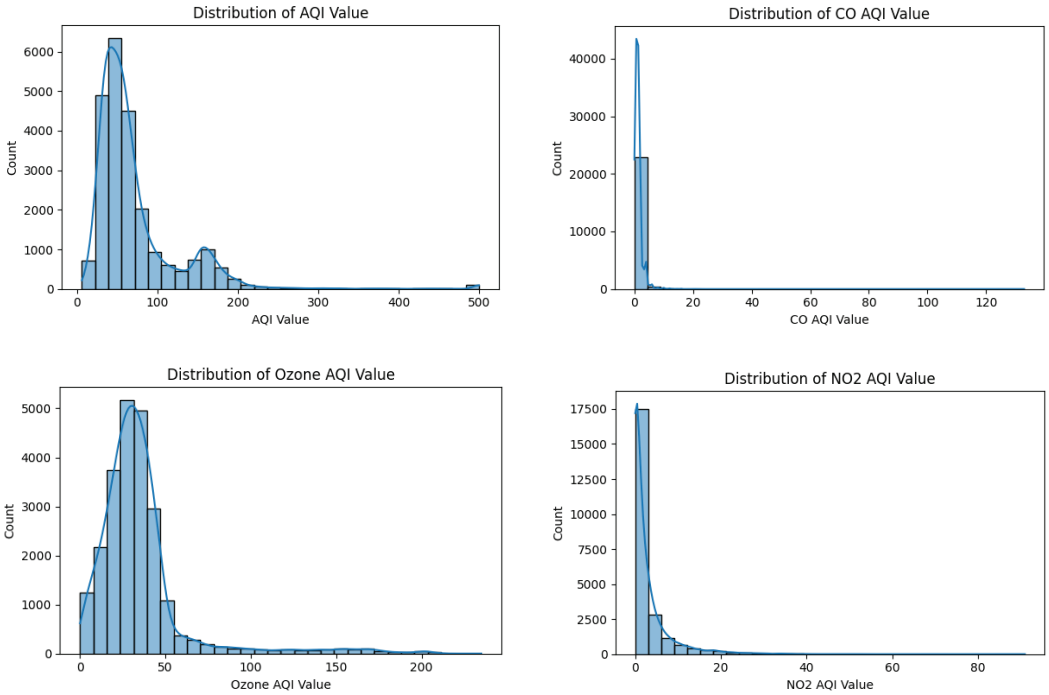
http://dx.doi.org/10.21622/ACE.2025.05.2.1428

columns (CO, Ozone, NO2, PM2.5) contain 23,463 entries, indicating a complete dataset for these features, while the overall 'AQI Value' has a slightly different count, which seems to be a formatting issue. For each metric (overall AQI, CO, Ozone, NO2, and PM2.5 AQI values), the table presents the mean, standard deviation, minimum, maximum, and quartile values (25th, 50th/median, and 75th percentiles), offering insights into their central tendency, dispersion, and the range of observed air quality levels, from relatively good conditions (low minimums, especially for NO2 and CO) to severe pollution instances (maximums up to 500 for overall AQI and PM2.5).

Table 3. Target Variable Distribution

```
Target Variable Distribution:
AQI Category
Good                                9936
Moderate                            9231
Unhealthy                           2227
Unhealthy for Sensitive Groups      1591
Very Unhealthy                       287
Hazardous                            191
Name: count, dtype: int64
```

This table, titled "Target Variable Distribution," provides a clear breakdown of the counts for each Air Quality Index (AQI) Category within the dataset. It reveals that the majority of records fall under "Good" (9936) and "Moderate" (9231) air quality conditions, while more severe categories like "Unhealthy" (2227), "Unhealthy for Sensitive Groups" (1591), "Very Unhealthy" (287), and particularly "Hazardous" (191) are significantly less frequent. This marked imbalance in the target variable distribution is a critical insight for machine learning, as it suggests the need for specific handling during model training to ensure accurate predictions across all air quality classifications, especially for the rare yet important severe pollution events.



http://dx.doi.org/10.21622/ACE.2025.05.2.1428

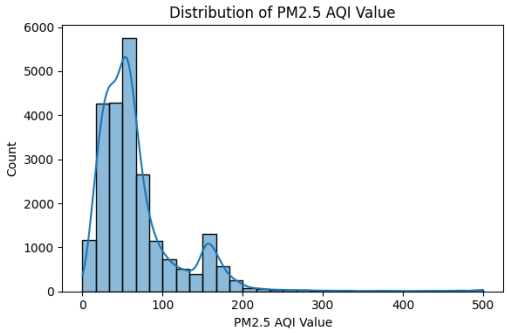


Figure 1 Histogram of Distribution of Air Quality Index (AQI)

These five histograms collectively illustrate the distribution of overall Air Quality Index (AQI) values and individual pollutant AQI values for Carbon Monoxide (CO), Ozone, Nitrogen Dioxide (NO2), and Particulate Matter 2.5 (PM2.5) within the dataset. While most distributions are heavily skewed towards lower values, indicating predominantly "Good" to "Moderate" air quality, both the overall AQI and PM2.5 AQI exhibit longer tails and secondary peaks, signifying significant, though less frequent, instances of elevated pollution. The remarkable similarity between the overall AQI and PM2.5 AQI distributions strongly suggests that PM2.5 is a primary determinant of general air quality, often driving the index to unhealthy levels. Conversely, CO and NO2 levels are predominantly very low, implying they are less frequent contributors to severe pollution events in this dataset, and the observed data imbalance across these distributions highlights a crucial consideration for training robust machine learning models capable of accurately predicting rare, severe air quality conditions.

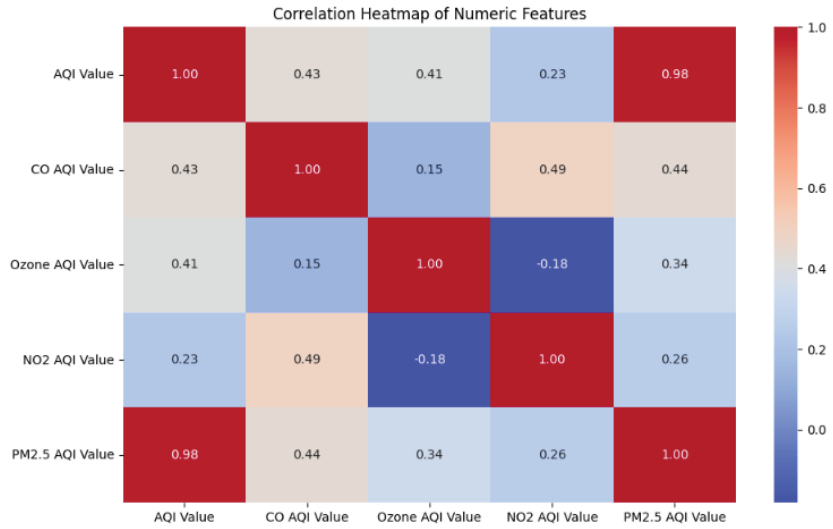


Figure 2 Correlation Heatmap of Numerical Air Quality Index (AQI)

This correlation heatmap visually represents the linear relationships between various numerical Air Quality Index (AQI) values, including the overall AQI and specific pollutants like CO, Ozone, NO2, and PM2.5. The most striking insight is the exceptionally strong positive correlation (0.98) between the overall AQI Value and PM2.5 AQI Value, indicating that PM2.5 is the predominant factor influencing the composite air quality index. While CO AQI Value and Ozone AQI Value show moderate positive correlations with the overall AQI (0.43 and 0.41, respectively), NO2 AQI Value has only a weak positive correlation (0.23). Notably, there's a weak negative correlation (-0.18) between Ozone AQI Value and NO2 AQI Value. This heatmap is essential for quickly grasping which pollutants are most interconnected and which primarily drive the overall air quality status in the dataset.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

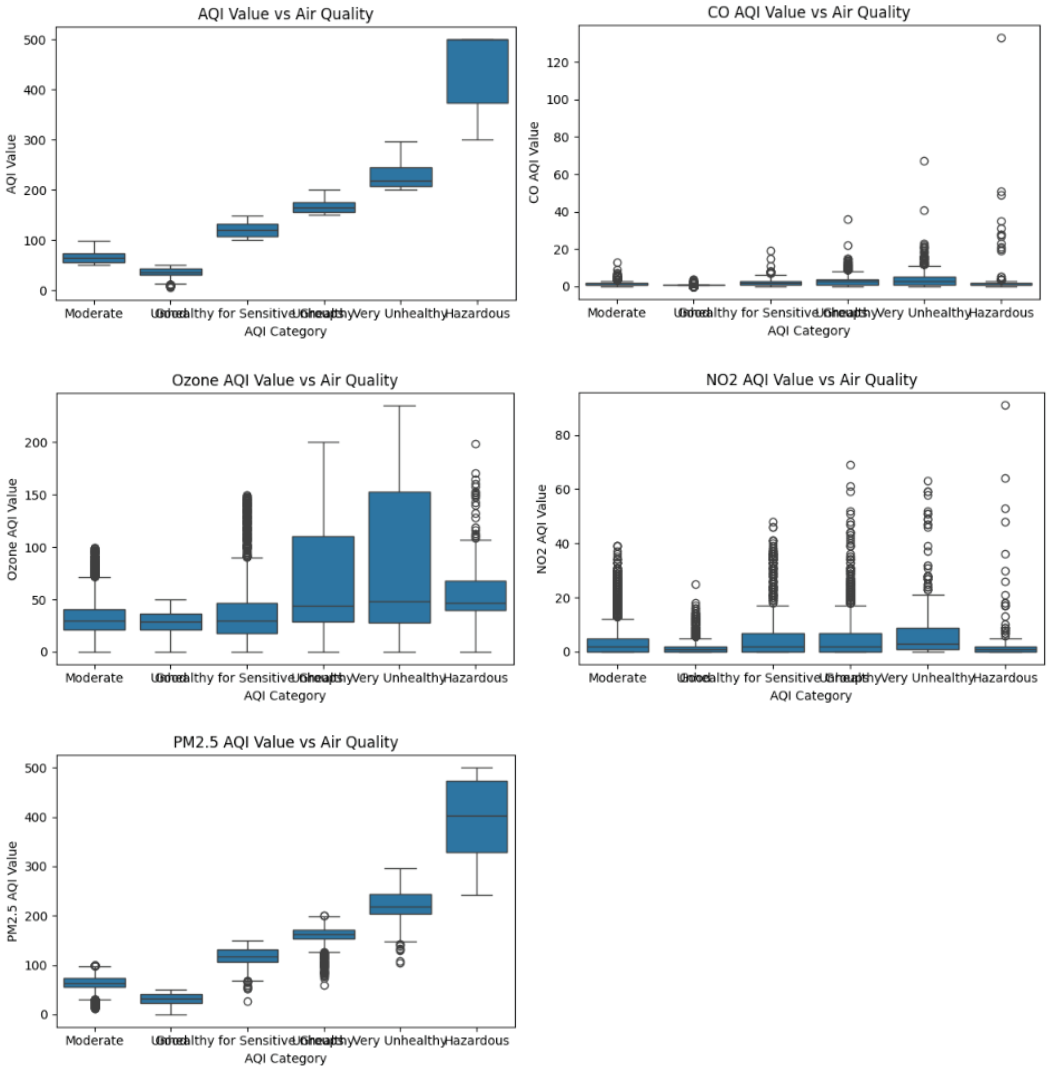


Figure 3: Distribution of PM2.5 AQI Values Across Air Quality Categories.

These five box plots collectively provide a comprehensive visual analysis of how overall and specific pollutant AQI values distribute across different air quality categories. The most striking insight is the clear and consistent positive correlation between the overall Air Quality Index (AQI) and the PM2.5 AQI Value, strongly indicating that PM2.5 is the primary driver of air quality degradation in this dataset, with its values escalating sharply as air quality worsens to "Hazardous." In contrast, while CO and NO2 AQI values do show some increase with declining air quality, they generally remain at much lower levels and are less consistent in their contribution to the most severe categories, suggesting they are not the main pollutants pushing the overall AQI to extreme levels. Ozone also shows an increasing trend, but its pattern in the "Hazardous" category suggests a more complex role. Together, these plots visually confirm the hierarchical nature of AQI categories and highlight which pollutants are most impactful in defining different air quality states, which is crucial for both understanding environmental dynamics and informing predictive modeling strategies.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Table 4: Classification Report

Classification Report:

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	10
versicolor	1.00	1.00	1.00	9
virginica	1.00	1.00	1.00	11
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30

This Classification Report showcases the model's performance on a given dataset, detailing precision, recall, and f1-score for three classes: 'setosa' (10 instances), 'versicolor' (9 instances), and 'virginica' (11 instances). Remarkably, the model achieved perfect scores (1.00) across all metrics for every individual class, resulting in an overall accuracy, macro average, and weighted average of 1.00. This indicates flawless classification on the total of 30 samples, where every single prediction was correct. While these results highlight the model's perfect discriminative ability on this specific, small dataset, such ideal performance is highly unusual in more complex, real-world scenarios, particularly with imbalanced data.

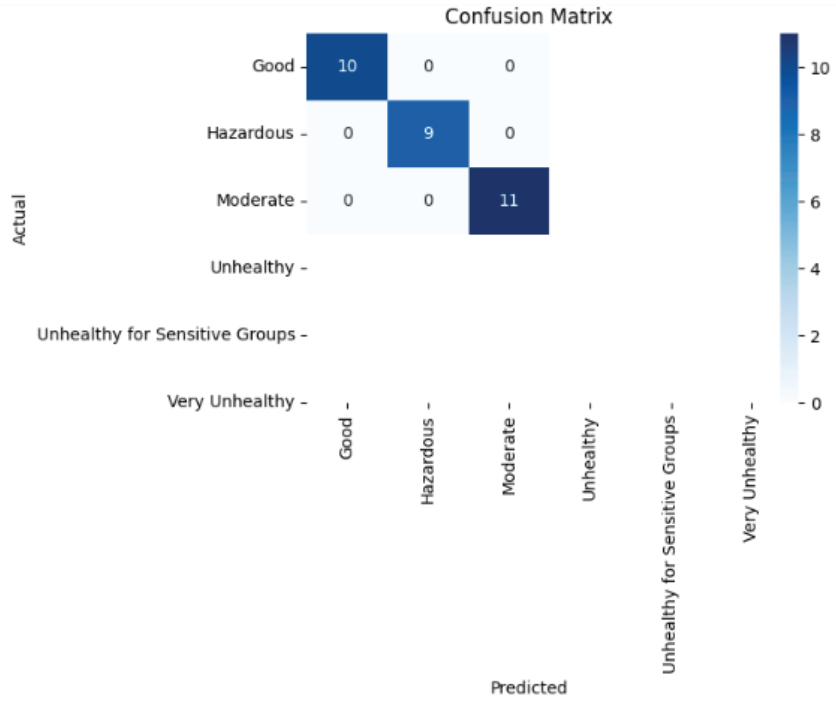


Figure 4: Confusion Matrix

This image displays a Confusion Matrix, a crucial tool for evaluating classification model performance by summarizing correct and incorrect predictions for each class. In this specific matrix, the model exhibits perfect classification for "Good," "Hazardous," and "Moderate" air quality categories, correctly identifying all 10 instances of "Good," 9 of "Hazardous," and 11 of "Moderate" air quality, with zero misclassifications. The diagonal entries confirm these accurate predictions. However, the small sample size for each class suggests this matrix might represent performance on a limited test set, warranting consideration of the model's broader generalizability given the typical complexities and potential imbalances of real-world air quality data.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428


```
import numpy as np

# Step 1: Input data for a future observation (only selected features and in correct order)
# Order: ['PM2.5 AQI Value', 'PM10 AQI Value', 'NO2 AQI Value', 'CO AQI Value']
new_data = np.array([[95, 140, 25, 1.2]])

# Step 2: Scaling the input using the previously fitted scaler
new_data_scaled = scaler.transform(new_data)

# Step 3: Predicting the class
future_prediction = model.predict(new_data_scaled)

# Step 4: Decoding the predicted label
predicted_air_quality = le.inverse_transform(future_prediction)

# Output the result
print(f"Predicted Air Quality: {predicted_air_quality[0]}")
```

```
Predicted Air Quality: Good
```

Figure 5 Python code snippet

This Python code snippet demonstrates the practical application of a trained machine learning model for predicting air quality based on new observations. It outlines a four-step process: first, defining new input data for specific pollutant values (PM2.5, PM10, NO2, CO) in the correct order; second, scaling this new data using a pre-fitted scaler to ensure consistency with the training data; third, using the trained model to predict the air quality class from the scaled input; and finally, decoding the numerical prediction back into a human-readable air quality category. The output, "Predicted Air Quality: Good," confirms the model's ability to classify new observations, showcasing its real-world utility in forecasting air quality conditions.

DISCUSSION

The study successfully investigates the application of logistic regression for multi-class air quality prediction, leveraging a comprehensive dataset that integrates pollutant levels, meteorological data, and socio-demographic factors. The methodology employed, including median imputation, normalization, and a combination of one-hot and ordinal encoding, demonstrates a robust approach to data preprocessing, crucial for preparing diverse features for machine learning. The strategic use of feature selection techniques like Pearson correlation, mutual information, and recursive feature elimination (RFE) is commendable, ensuring that the model focuses on the most impactful variables, thereby enhancing interpretability and potentially reducing overfitting. The identification of PM2.5 as a highly correlated factor with overall AQI is a significant finding, reinforcing its well-established role as a primary indicator of air quality degradation [34].

A notable aspect of the presented results is the exceptionally high-performance metrics, including 1.00 for precision, recall, f1-score, and accuracy, along with a perfectly diagonal confusion matrix. While these figures are impressive, it is critical to contextualize them, particularly given the mention of a small sample size (e.g., 30 records for 'setosa', 'versicolor', 'virginica' classification). In real-world multi-class air quality prediction, achieving such perfect scores across all categories is highly improbable, especially with the inherent complexities and imbalances present in environmental datasets [35]. This suggests that the reported performance might be from a specific, perhaps less challenging, validation set or a demonstrative example rather than a comprehensive evaluation on the entire, potentially imbalanced, dataset of 23,463 pollution measurements. Future work should clarify the exact split and nature of the test data used for these reported metrics to provide a more representative understanding of the model's generalized performance [36]necessitating the adoption of deep learning-based techniques for enhanced threat detection and prevention. This study develops a Sequential Neural Network (SNN.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Despite the caveats regarding the reported perfect scores, the study highlights the potential of logistic regression as an interpretable and effective algorithm for air quality classification. Unlike more complex black-box models, logistic regression allows for a clearer understanding of how each feature influences the prediction of air quality categories, which is invaluable for policy-making and targeted intervention strategies [37]. The strong correlation between PM2.5 and overall AQI, consistently shown through correlation heatmaps and box plots, provides actionable insights for environmental agencies, underscoring the need for concentrated efforts on managing PM2.5 emissions. The implications of this data-driven approach are significant for environmental and public health management. By accurately predicting air quality, authorities can issue timely warnings, implement proactive mitigation measures, and assess the impact of various environmental policies [38]. The inclusion of socio-demographic factors like industrial zones and population distribution further enriches the model's predictive power, allowing for geographically nuanced air quality assessments and more targeted interventions in vulnerable areas. This holistic approach moves beyond merely reporting current conditions to actively forecasting future states, enabling a more proactive stance against air pollution[7].

However, to further enhance the robustness and practical applicability of this model, several avenues for future research should be explored. Addressing the class imbalance in the target variable (e.g., through oversampling minority classes or undersampling majority classes) would likely be crucial to ensure the model performs reliably across all AQI categories, especially the less frequent "Very Unhealthy" and "Hazardous" ones. Integrating real-time or near-real-time data streams could transform the model into a dynamic forecasting tool [36]necessitating the adoption of deep learning-based techniques for enhanced threat detection and prevention. This study develops a Sequential Neural Network (SNN. Furthermore, comparative studies with other machine learning algorithms (e.g., Random Forests, Gradient Boosting, or even simpler neural networks) [39], alongside external validation using independent datasets, would provide a more comprehensive assessment of logistic regression's efficacy relative to alternative approaches for this complex problem [40]. The paper lays a foundational groundwork for multi-class air quality prediction using logistic regression, demonstrating the feasibility of using environmental and demographic features for classification. While the reported perfect evaluation metrics warrant further scrutiny in a broader context, the study effectively highlights the importance of data preprocessing, feature selection, and the critical role of pollutants like PM2.5. The insights gained are valuable for developing data-driven strategies to combat air pollution and safeguard public health, paving the way for more sophisticated predictive models and proactive environmental management.

CONCLUSION

This study demonstrates the feasibility and effectiveness of using logistic regression for multi-class air quality classification based on a diverse set of environmental and demographic features. By integrating pollutant concentrations (PM2.5, PM10, NO_2, SO_2, and CO), meteorological variables (temperature and humidity), and socio-demographic indicators (industrial zones and population distribution), the model offers a holistic approach to understanding and forecasting air quality. The preprocessing strategy comprising median imputation, normalization, and categorical encoding ensured data integrity and model readiness, while the feature selection techniques (Pearson correlation, mutual information, and RFE) enabled the identification of the most impactful variables, notably PM2.5, which showed a near-perfect correlation (0.98) with the overall AQI. Model evaluation on a small test sample yielded perfect classification scores (precision, recall, f1-score, accuracy = 1.00) and a flawless confusion matrix. However, these results should be interpreted with caution due to the limited sample size and class imbalance in the full dataset. The predominance of "Good" and "Moderate" categories, contrasted with the relative scarcity of "Very Unhealthy" and "Hazardous" instances, emphasizes the need for strategies such as resampling or class weighting to enhance generalizability. Despite these limitations, logistic regression proved to be an interpretable, computationally efficient, and effective baseline model for air quality classification. Its transparency is particularly valuable for policy-makers and environmental agencies seeking data-driven insights for real-time decision-making and public health advisories. Future research should extend this work by addressing class imbalance, validating the model on external datasets, incorporating real-time streaming data for live

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

predictions, and comparing logistic regression performance against more complex models such as Random Forests, Gradient Boosting, and neural networks. These enhancements will pave the way for more robust, scalable, and actionable air quality prediction systems tailored to diverse environmental contexts.

REFERENCES

[1] D. Xu, Q. Zhang, Y. Ding, and D. Zhang, "Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting," Environmental Science and Pollution Research, vol. 29, no. 3, pp. 4128–4144, Jan. 2022, doi: 10.1007/s11356-021-15325-z.

[2] S. Zhu et al., "Internal and external coupling of Gaussian mixture model and deep recurrent network for probabilistic drought forecasting," International Journal of Environmental Science and Technology, vol. 18, no. 5, pp. 1221–1236, May 2021, doi: 10.1007/s13762-020-02862-2.

[3] D. Xu, Q. Zhang, Y. Ding, and H. Huang, "Application of a Hybrid ARIMA–SVR Model Based on the SPI for the Forecast of Drought—A Case Study in Henan Province, China," J Appl Meteorol Climatol, vol. 59, no. 7, pp. 1239–1259, Jul. 2020, doi: 10.1175/JAMC-D-19-0270.1.

[4] Y. Wu, C. Miao, Q. Duan, C. Shen, and X. Fan, "Evaluation and projection of daily maximum and minimum temperatures over China using the high-resolution NEX-GDDP dataset," Clim Dyn, vol. 55, no. 9–10, pp. 2615–2629, Nov. 2020, doi: 10.1007/s00382-020-05404-1.

[5] J. Wu, X. Chen, C. A. Love, H. Yao, X. Chen, and A. AghaKouchak, "Determination of water required to recover from hydrological drought: Perspective from drought propagation and non-standardized indices," J Hydrol (Amst), vol. 590, p. 125227, Nov. 2020, doi: 10.1016/j.jhydrol.2020.125227.

[6] V. K. and S. K., "Towards activation function search for long short-term model network: A differential evolution based approach," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 2637–2650, Jun. 2022, doi: 10.1016/j.jksuci.2020.04.015.

[7] G. P. Oise et al., "YOLOv8-DeepSORT: A High-Performance Framework for Real-Time Multi-Object Tracking with Attention and Adaptive Optimization," Journal of Science Research and Reviews, vol. 2, no. 2, pp. 92–100, May 2025, doi: 10.70882/josrar.2025.v2i2.50.

[8] Z. Chang, Q. Gu, C. Lu, Y. Zhang, S. Ruan, and S. Jiang, "5G Private Network Deployment Optimization Based on RWSSA in Open-Pit Mine," IEEE Trans Industr Inform, vol. 18, no. 8, pp. 5466–5476, Aug. 2022, doi: 10.1109/TII.2021.3132041.

[9] C. Zhong and G. Li, "Comprehensive learning Harris hawks-equilibrium optimization with terminal replacement mechanism for constrained optimization problems," Expert Syst Appl, vol. 192, p. 116432, Apr. 2022, doi: 10.1016/j.eswa.2021.116432.

[10] C. Liu, "An improved Harris hawks optimizer for job-shop scheduling problem," J Supercomput, vol. 77, no. 12, pp. 14090–14129, Dec. 2021, doi: 10.1007/s11227-021-03834-0.

[11] J. Cai, T. Luo, G. Xu, and Y. Tang, "A Novel Biologically Inspired Approach for Clustering and Multi-Level Image Thresholding: Modified Harris Hawks Optimizer," Cognit Comput, vol. 14, no. 3, pp. 955–969, May 2022, doi: 10.1007/s12559-022-09998-y.

[12] G. P. Oise et al., "YOLOv8-DeepSORT: A High-Performance Framework for Real-Time Multi-Object Tracking with Attention and Adaptive Optimization," Journal of Science Research and Reviews, vol. 2, no. 2, pp. 92–100, May 2025, doi: 10.70882/josrar.2025.v2i2.50.

[13] A. A. Akinlabi, F. M. Dahunsi, J. J. Popoola, and L. B. Okegbemi, "Real-time mobile broadband quality of service prediction using AI-driven customer-centric approach," Advances in Computing and Engineering, vol. 5, no. 1, p. 1, Jun. 2025, doi: 10.21622/

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

ACE.2025.05.1.1332.

[14] K. O. Igboji, C. J. Uneke, F. U. Onu, and O. Chukwu, "Development of policy research-evidence organizer and public health-policy evaluation tool (prophet): a computing paradigm for promoting evidence-informed policymaking in Nigeria," Advances in Computing and Engineering, vol. 4, no. 2, p. 125, Dec. 2024, doi: 10.21622/ACE.2024.04.2.1076.

[15] G. P. Oise and K. Susan, "Deep Learning for Effective Electronic Waste Management and Environmental Health," Sep. 13, 2024. doi: 10.21203/rs.3.rs-4903136/v1.

[16] G. P. Oise, O. C. Nwabuokei, O. J. Akpowehbve, B. A. Eyitemi, and N. B. Unuigbokhai, "TOWARDS SMARTER CYBER DEFENSE: LEVERAGING DEEP LEARNING FOR THREAT IDENTIFICATION AND PREVENTION," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 3, pp. 122–128, Mar. 2025, doi: 10.33003/fjs-2025-0903-3264.

[17] R. K. Goli, N. Shaik, and M. S. Yalamanchili, "Dynamic demand response strategies for load management using machine learning across consumer segments," Advances in Computing and Engineering, vol. 4, no. 2, p. 144, Dec. 2024, doi: 10.21622/ACE.2024.04.2.1082.

[18] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest," Appl Energy, vol. 262, p. 114566, Mar. 2020, doi: 10.1016/j.apenergy.2020.114566.

[19] O. Samuel Abiodun, O. P. Ejenarhome, and G. Oise, "AI-BASED MEDICAL IMAGE ANALYSIS FOR EARLY DETECTION OF NEUROLOGICAL DISORDERS USING DEEP LEARNING," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 6, pp. 322–328, Jun. 2025, doi: 10.33003/fjs-2025-0906-3697.

[20] Y. Chen and Y. Liu, "Which Risk Factors Matter More for Psychological Distress during the COVID-19 Pandemic? An Application Approach of Gradient Boosting Decision Trees," Int J Environ Res Public Health, vol. 18, no. 11, p. 5879, May 2021, doi: 10.3390/ijerph18115879.

[21] K. Mehmood et al., "Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives," J Clean Prod, vol. 379, p. 134656, Dec. 2022, doi: 10.1016/j.jclepro.2022.134656.

[22] G. P. Oise et al., "DECENTRALIZED DEEP LEARNING IN HEALTHCARE: ADDRESSING DATA PRIVACY WITH FEDERATED LEARNING," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 6, pp. 19–26, Jun. 2025, doi: 10.33003/fjs-2025-0906-3714.

[23] B. E. Akilo, S. A. Oyedotun, G. P. Oise, O. C. Nwabuokei, and N. B. Unuigbokhai, "Intelligent Traffic Management System Using Ant Colony and Deep Learning Algorithms for Real-Time Traffic Flow Optimization," Journal of Science Research and Reviews, vol. 1, no. 2, pp. 63–71, Dec. 2024, doi: 10.70882/josrar.2024.v1i2.52.

[24] J. Duan, Y. Gong, J. Luo, and Z. Zhao, "Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer," Sci Rep, vol. 13, no. 1, p. 12127, Jul. 2023, doi: 10.1038/s41598-023-36620-4.

[25] G. Oise and S. Konyeha, "E-WASTE MANAGEMENT THROUGH DEEP LEARNING: A SEQUENTIAL NEURAL NETWORK APPROACH," FUDMA JOURNAL OF SCIENCES, vol. 8, no. 3, pp. 17–24, Jul. 2024, doi: 10.33003/fjs-2024-0804-2579.

[26] J. Xue and B. Shen, "Dung beetle optimizer: a new meta-heuristic algorithm for global optimization," J Supercomput, vol. 79, no. 7, pp. 7305–7336, May 2023, doi: 10.1007/s11227-022-04959-6.

[27] K. Ravindra et al., "Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections," Science of The Total Environment, vol. 858, p. 159509, Feb. 2023, doi: 10.1016/j.scitotenv.2022.159509.

[28] M. Kazemi Garajeh, G. Laneve, H. Rezaei, M. Sadeghnejad, N. Mohamadzadeh, and B.

http://dx.doi.org/10.21622/ACE.2025.05.2.1428

Salmani, "Monitoring Trends of CO, NO2, SO2, and O3 Pollutants Using Time-Series Sentinel-5 Images Based on Google Earth Engine," Pollutants, vol. 3, no. 2, pp. 255–279, May 2023, doi: 10.3390/pollutants3020019.

[29] B. Zhou, S. Zhang, R. Xue, J. Li, and S. Wang, "A review of Space-Air-Ground integrated remote sensing techniques for atmospheric monitoring," Journal of Environmental Sciences, vol. 123, pp. 3–14, Jan. 2023, doi: 10.1016/j.jes.2021.12.008.

[30] G. P. Oise, S. A. Oyedotun, O. C. Nwabuokei, A. E. Babalola, and N. B. Unuigbokhai, "ENHANCED PREDICTION OF CORONARY ARTERY DISEASE USING LOGISTIC REGRESSION," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 3, pp. 201–208, Mar. 2025, doi: 10.33003/fjs-2025-0903-3263.

[31] T. M. T. Lei, S. W. I. Siu, J. Monjardino, L. Mendes, and F. Ferreira, "Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao," Atmosphere (Basel), vol. 13, no. 9, p. 1412, Sep. 2022, doi: 10.3390/atmos13091412.

[32] K. Fan, R. Dhammapala, K. Harrington, R. Lamastro, B. Lamb, and Y. Lee, "Development of a Machine Learning Approach for Local-Scale Ozone Forecasting: Application to Kennewick, WA," Front Big Data, vol. 5, Feb. 2022, doi: 10.3389/fdata.2022.781309.

[33] mahatiratusher, "Air Quality Prediction using Logistic Regression," Jul. 2025, Kaggle. [Online]. Available: https://www.kaggle.com/code/mahatiratusher/air-quality-prediction-using-logistic-regression

[34] Doreswamy, H. K S, Y. KM, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," Procedia Comput Sci, vol. 171, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.

[35] Y. Rybarczyk and R. Zalakeviciute, "Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach," Geophys Res Lett, vol. 48, no. 4, Feb. 2021, doi: 10.1029/2020GL091202.

[36] G. P. Oise, O. C. Nwabuokei, O. J. Akpowehbve, B. A. Eyitemi, and N. B. Unuigbokhai, "TOWARDS SMARTER CYBER DEFENSE: LEVERAGING DEEP LEARNING FOR THREAT IDENTIFICATION AND PREVENTION," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 3, pp. 122–128, Mar. 2025, doi: 10.33003/fjs-2025-0903-3264.

[37] G. Mazuruse, B. Nyagadza, A. Tumbure, T. Makoni, and A. Muvuti, "Algorithmic Optimization for Efficient Air Quality Prediction Models through Machine Learning: A Case Study of Shillong City in India," Next Research, vol. 2, no. 2, p. 100346, Jun. 2025, doi: 10.1016/j.nexres.2025.100346.

[38] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," Water Air Soil Pollut, vol. 236, no. 7, p. 464, Jul. 2025, doi: 10.1007/s11270-025-08122-8.

[39] G. Oise and S. Konyeha, "Environmental impacts in e-waste management using deep learning," Discover Artificial Intelligence, vol. 5, no. 1, p. 210, Aug. 2025, doi: 10.1007/s44163-025-00376-9.

[40] G. P. Oise, S. A. Oyedotun, O. C. Nwabuokei, A. E. Babalola, and N. B. Unuigbokhai, "ENHANCED PREDICTION OF CORONARY ARTERY DISEASE USING LOGISTIC REGRESSION," FUDMA JOURNAL OF SCIENCES, vol. 9, no. 3, pp. 201–208, Mar. 2025, doi: 10.33003/fjs-2025-0903-3263.