Environmental Systems Research

RESEARCH Open Access

AirNet: predictive machine learning model for air quality forecasting using web interface

Md. Mahbubur Rahman[1*], Md. Emran Hussain Nayeem[2], Md. Shorup Ahmed[2], Khadiza Akther Tanha[2], Md. Shahriar Alam Sakib[2], Khandaker Mohammad Mohi Uddin[3] and Hafiz Md. Hasan Babu[4]

Abstract

Air is one of the most significant elements of the environment. The increasing global air pollution crisis poses an unavoidable threat to human health, environmental sustainability, ecosystems, and the earth's climate. Air pollution has been referred to as a silent killer due to its insidious nature. Its indirect impact on human health further underscores its dangerous effects. Early detection of air quality can potentially save millions of lives globally. A unique and transformative approach can harness the power of machine learning to combat air pollution. This research presents a manual and web-based automatic prediction system that provides real-time alerts on air quality status and can help prevent premature deaths, chronic diseases, and other health problems. Air pollutants, including carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), and particulate matter (PM 2.5), are used in this study for feature analysis and extraction. The system utilizes publicly available data from 23,463 different cities worldwide. Data preprocessing was performed before feeding the data into the machine learning models for feature correlation and evaluation. The proposed research uses various machine learning models to predict air quality, including Random Forest (100%), Logistic Regression (79%), Decision Tree (100%), Support Vector Machine (93%), Linear SVC (98%), K-Nearest Neighbor (99%), and Multinomial Naïve Bayes (52%). A user-friendly Django-based web interface offers an accessible platform for users to monitor air quality in real-time, based on the two best-performing models: Random Forest and Decision Tree techniques.

Keywords Air pollutant feature extraction, Real-time air quality prediction, Django-based web interface, Machine learning in environmental science

Introduction

Air pollution remains a critical global environmental issue, with both short-term and long-term consequences for public health, ecosystems, and the climate. It greatly contributes to the occurrence of breathing problems, heart issues, stroke, and cancer, and it is also a major driver of climate change (Alahmad et al. 2023; Chandra et al. 2023). Efforts to combat air pollution continue through regulatory measures, technological innovations, and public awareness campaigns, with the goal of achieving cleaner and healthier air for all. Reducing air pollution not only saves lives but also ensures future generations will live in a healthier and more sustainable environment (Chandra et al. 2022). Based on the World Health Organization's (WHO) studies, air pollution, is a silent killer that produces the premature death of almost 7 million people each year, including 600.000 children and that number is expected to rise to 9 million by 2030 is deeply concerning (Tipton 2022). Children are

*Correspondence:
Md. Mahbubur Rahman
mmrmbstu@gmail.com
[1] Department of Computer Science and Engineering, Bangladesh University of Business and Technology (BUBT), Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh
[3] Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
[4] Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

especially vulnerable to the effects of air pollution, and they can suffer from a variety of health problems, including asthma, bronchitis, and pneumonia (Akhtar 2020; Likhon et al. 2024; Sram et al. 2013; Zhou et al. 2023a, b). The goal of this study is to provide an accurate technique for air quality forecasting using a suggested model as accurately as required.

Through incomplete combustion of carbon-containing fuels in sources like automobiles and industrial activities, carbon monoxide (CO) contributes to air pollution. Air quality and human well-being can be negatively impacted by high CO levels, which can also have an indirect effect on the development of other pollutants and have detrimental health impacts. Ground-level or tropospheric ozone, or O_3, contributes to air pollution even while it shields Earth from damaging UV rays in the stratosphere (Wang and Tang 2023). Pollutants from industry, automobiles, and other sources react with sunlight to produce ground-level ozone (Zhang et al. 2023). In addition to other health concerns, high ozone concentrations can cause respiratory disorders (Lu and Yao 2023). It is one of the main causes of pollution and lowers the quality of the air. One of the main causes of air pollution is nitrogen dioxide (NO_2), which is mostly released during automobile and industrial combustion (Park and Kim 2023). In addition to contributing to the development of other pollutants, it can cause respiratory issues. In order to lessen the effects of nitrogen dioxide on air quality and human health, it is essential to regulate emissions from sources such as cars and enforce air quality regulations (Kumar et al. 2024; Ravi et al. 2023). One major cause of air pollution is particulate matter, or PM2.5 (fine inhalable particles having a diameter of 2.5 μm or less). These microscopic particles can enter the respiratory system deeply and cause health hazards (Ding et al. 2023). They originate from many different places, including as industrial processes, natural sources, and vehicle emissions. Cardiovascular and respiratory issues are linked to higher PM2.5 exposure.

Traditional machine learning algorithm used for air pollution prediction depends on several aspects, including data availability, issue complexity, and required accuracy (Rahman et al. 2023a, b; Wu et al. 2023). Predicting the concentrations of crucial air pollutants, such as PM 2.5, O_3, CO, and NO_2, is the main objective of this study. So, we can actively participate in improving the effectiveness of a real-time air quality monitoring and alert system. We use some advanced machine learning techniques like random forest (RF), logistic regression (LR), decision tree (DT), multinomial naïve bayes (MNB), support vector machine (SVM), linear SVC, K-nearest neighbor (KNN) algorithms which are used to predict AQI and AQC based on several parameters. The generated models are valuable tools for identifying and understanding air contamination. This insights can be incredibly valuable for decision-makers. This information can help them to make informed decisions about how to mitigate the contamination and protect public health.

Kumar et al. constructed a model to predict air pollution. The main focus was on the air polluted cities in India. In the data preprocessing stage, outliers and missing data are handled by filling in missing values and transforming outlier data. The dataset consists of 12 features and 29,531 observations from 23 different Indian cities, collected between January 2015 and July 2020. Gaussian Naive Bayes, SVM, RF, XGBoost, KNN were used for classification. Among all classifiers, Gaussian Naive Bayes produced the highest accuracy, while the SVM model had the lowest accuracy (Kumar and Pande 2023, Sanjeev et al. 2021). developed an analytical and predictive model for predicting air quality using machine learning algorithms. In their air pollution dataset, there are features such as temperature, methane (CH_4), CO, non-methane hydrocarbons (NMHC), NO, NO_2, nitrogen oxides (NOx), O_3, PM10 and PM2, relative humidity (RH), and sulfur dioxide (SO_2). In the preprocessing stage, missing values are omitted, and new attributes are selected based on normalization procedures. For the air quality prediction process, RF, SVM, and artificial neural network (ANN) methods are utilized. The highest accuracy comes from RF, which is 97% (Sanjeev 2021). S. Abirami et al. proposed a deep learning-based architecture called "DL-Air" for air quality forecasting. The spatio-temporal attention augmented long short-term memory (STAA-LSTM) network is used for decoding data. In the processing stage, the absolute average deviation (AAD) and mean absolute error (MAE) are reduced by 37% and 30%, respectively. After applying root mean squared error (RMSE) and the correlation coefficient (R^2), the predictions are 8% more accurate in predicting the AQI category than the top baseline techniques (Abirami and Chitra 2023). Thomas Plocoste et al. developed a model to forecast PM10 concentrations in the Caribbean area. The dataset was not available due to ethical reasons. They measured PM10 averages from 2005 to 2012 in the city of Pointe-à-Pitre. They used six machine learning algorithms, with Gradient Boosting Regression (GBR) performing the best (Plocoste and Laventure 2023). (Xue-Bo Jin et al. 2023) produced a deep spatio-temporal graph neural network-based self-optimization model that revolutionizes air quality prediction. In their proposed model, spatio-temporal data on PM2.5 concentrations were collected from 16 districts in Beijing from January to April, 2017. They preprocessed the dataset to address missing values using the previous day's values. They evaluated two Bidirectional Graph Gated Recurrent Unit (BGGRU)

models, BGraphSAGE and BGraphGRU, and used Bayesian hyperparameter optimization to improve accuracy. The BGGRU achieved the highest accuracy of 99%. (Jin et al. 2023). Nishant Raj Kapoor et al. gathered real-time data and developed a method to assess indoor CO_2 levels. The proposed research utilized several features as input, such as the number of inhabitants, the area per person, the outside temperature, the wind speed, the relative humidity, and the air quality. ANN, SVM, DT, ensemble learning (EL), gaussian process regression (GPR), and LR are used to classify data. Besides, some models are optimized to increase prediction accuracy; they are GPR, EL, DT, and SVM. The highest root value is 98% (Kapoor et al. 2023). Chen et al. developed a system combining hyperspectral imaging (HSI) technology and deep learning methods for identifying air pollution. The visible-light HSI technology is used with a drone aerial camera to capture images. In the image classification stage, a 3D convolutional neural network and principal components analysis (PCA) technology are combined for a better outcome. The fusion of PCA and VGG-16 has the highest classification accuracy of 85.93% (Chen et al. 2021). (K. Basel et al. 2020) introduced an advanced predictive mechanism utilizing a Long Short-Term Memory (LSTM) architecture specifically designed for climate forecasting. The LSTM model effectively addresses the challenges inherent in long-term climate pattern detection by improving upon traditional Recurrent Neural Networks (RNNs), which often face difficulties in capturing dependencies over extended sequences. The architecture optimizes the loss function over multiple sequential time steps, enhancing its ability to identify complex temporal relationships in climate data. This deep learning model, integrated with convolutional layers, amplifies its feature extraction capabilities, enabling it to better differentiate between significant patterns and irrelevant noise within the data. The proposed system demonstrated remarkable accuracy in predicting global temperature trends, achieving high testing performance accuracy of 100% (El-Habil and Abu-Naser 2022), thus underscoring its potential for improving climate prediction models.

Table 1 serves as a comprehensive analysis of the various existing methodologies, offering a comparative perspective. After much study, many algorithms have been found to be better under various circumstances, such as the choice of pollutant and region selection in this case. It's essential to highlight the significance of utilizing meteorological data, including various major air pollutants like CO, O3, NO_2, and PM 2.5, which have proven to be pivotal in accurately predicting pollutant levels. The effectiveness of various algorithms becomes evident as they adapt to different conditions, such as the choice of pollutant and the specific area under study.

Air pollution prediction is the process of forecasting future air quality levels. It is important because air pollution has a significant impact on human health and the environment. Researchers use machine learning and deep learning algorithms, along with data sources such as historical air quality data, meteorological data, and traffic data to develop air pollution prediction models (Hameed et al. 2023; Kang et al. 2018; Méndez et al. 2023; Mitreska Jovanovska et al. 2023). These models are used for a variety of purposes, including public health warnings and climate change research. By applying several machine learning models and features, the primary goal of this research is to increase the accuracy of air pollution predictions. Our specific goals are to:

- Based on earlier studies, choose the features that are most crucial for AQI prediction.
- Employ seven classification techniques to predict AQI and AQC, with particular attention to the performance of RF and DT, which showed high accuracy.
- Develops a Django-based web interface that provides real-time alerts on air quality status.

Below are the parts that comprise the overall technique: "Methodology" and "Result and discussion" sections delineate the research methods and the comparative analysis, and finally, "Conclusions" sections reveals the conclusions.

Methodology

The proposed system aims to facilitate the detection of air quality in a more efficient and cost-effective manner, allowing for quick and accurate assessments. In addressing real-world challenges, our methodology focuses on the early detection of air quality issues through a highly accessible web interface. This system is crucial in tackling the escalating global problem of air pollution, as it enables real-time monitoring and prediction of air quality levels. By providing timely and accurate information, our system empowers individuals, communities, and governments to take proactive measures to mitigate the adverse effects of poor air quality. This early detection capability is crucial in preventing numerous health issues, reducing hospitalizations, and ultimately saving millions of lives worldwide. The practical application of our methodology ensures that it not only advances scientific understanding but also delivers tangible benefits in the real world, making a significant contribution to public health and environmental sustainability.

The initial phase of our methodology involved comprehensive data pre-processing. This step is critical as it ensures the quality and reliability of the data used in

Table 1 A comparison study of related works

References	Dataset	Methods	Strengths	Limitations	Accuracy
1. (Kumar and Pande 2023)	Data of Central Pollution Control Board, India	RF, KNN, GNB, SVM, XGBoost, and SMOTE	Outliers and missing data are handled, Gaussian naive bayes produced the highest accuracy	Transforming data sometimes produces wrong values, SVM model had the lowest accuracy	90%
2. (Sanjeev 2021)	Kaggle data	RF, SVM, and ANN	More than 11 features are used, Highest accuracy comes from RF	Normalized processes are not clearly demonstrating	97%
3. (Abirami and Chitra 2023)	UCI data	LSTM	A new methodology from LSTM, named STAA, is utilized,	Without RMSE and correlation coefficients, the system is under fitted,	37%
4. (Plocoste and Laventure 2023)	Data from city of Pointe-À-Pitter	GB Regression, SVR, KNN, RF Regression, and BRR	Gradient boosting regression performed the highest	Forecast PM10 concentration only instead of AQI	80%
5. (Jin et al. 2023)	PM2.5 concentration of 16 district in Beijing	BGraphSAGE, and BGraphGRU	The bidirectional graph gated recurrent unit was outstanding in terms of performance easement	The dataset to address missing values using the previous day's values	99%
6. (Kapoor et al. 2023)	Kaggle data	ANN, SVM, Dt, GPR, LR, and Ensemble Learning	Some models are optimized to increase prediction accuracy	Proposed architecture can only measure the CO level inside the home	98%
7. (Chen et al. 2021)	3D data captured by drone	PCA, and VGG-16	Hyperspectral imaging (HSI) technology used, Fusion of PCA and VGG-16 has the highest classification accuracy	Drone aerial camera images sometimes vary with features for visible light effects	86%
8. (El-Habil and Abu-Naser 2022)	Climate data records (netcdf format)	LSTM, and RNN	Leverages a deep convolutional based long short-term memory (LSTM) architecture	It's useful only for climate pattern detection-related problems	100%

subsequent stages. The pre-processing included the following steps: At first, we employed various techniques to address missing values in the dataset, ensuring that these gaps did not negatively impact the performance of the machine learning models. This might include methods such as the imputation or removal of records with missing data. Secondly, the correlation procedure helps to find the best-fitted categories like AQI value, CO AQI value, ozone AQI value, NO_2 AQI value, and PM2.5 AQI value. Those five parameters have tremendous importance and contribute to the prediction of air quality. This step helps in reducing the dimensionality of the dataset, improving model performance, and reducing computational complexity. Besides, data normalization is done using a label encoder, where categorical variables (good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, and hazardous) in the dataset are converted into numerical format. This transformation is necessary as most machine learning algorithms requires numerical input. The preprocessed data is then divided into two sections: 80% is allocated for training the models, and the remaining 20% is reserved for testing. This split ensures that the models are trained on a substantial portion of the data while still having a separate subset to evaluate their performance. We employed a diverse set of machine-learning algorithm to train the system, ensuring robust and accurate predictions. The algorithms, such as RF, LR, DT, SVM, linear SVC, KNN, and MNB.

To make our solution accessible and user-friendly, we have developed a web interface that provides real-time alerts on air quality status. This interface allows users to monitor air quality levels continuously and receive immediate notifications when air quality deteriorates. The primary focus of our research is on the early detection of poor air quality. Early detection is vital as it enables timely interventions and mitigative actions, reducing exposure to harmful pollutants and subsequently decreasing the risk of health issues associated with poor air quality. Figure 1 illustrates the structure for air pollution prediction utilized in this study. It depicts the flow from data preprocessing, through training and testing, to the final prediction and alert system. Algorithm 1 represents the overall working procedures of our proposed algorithm.

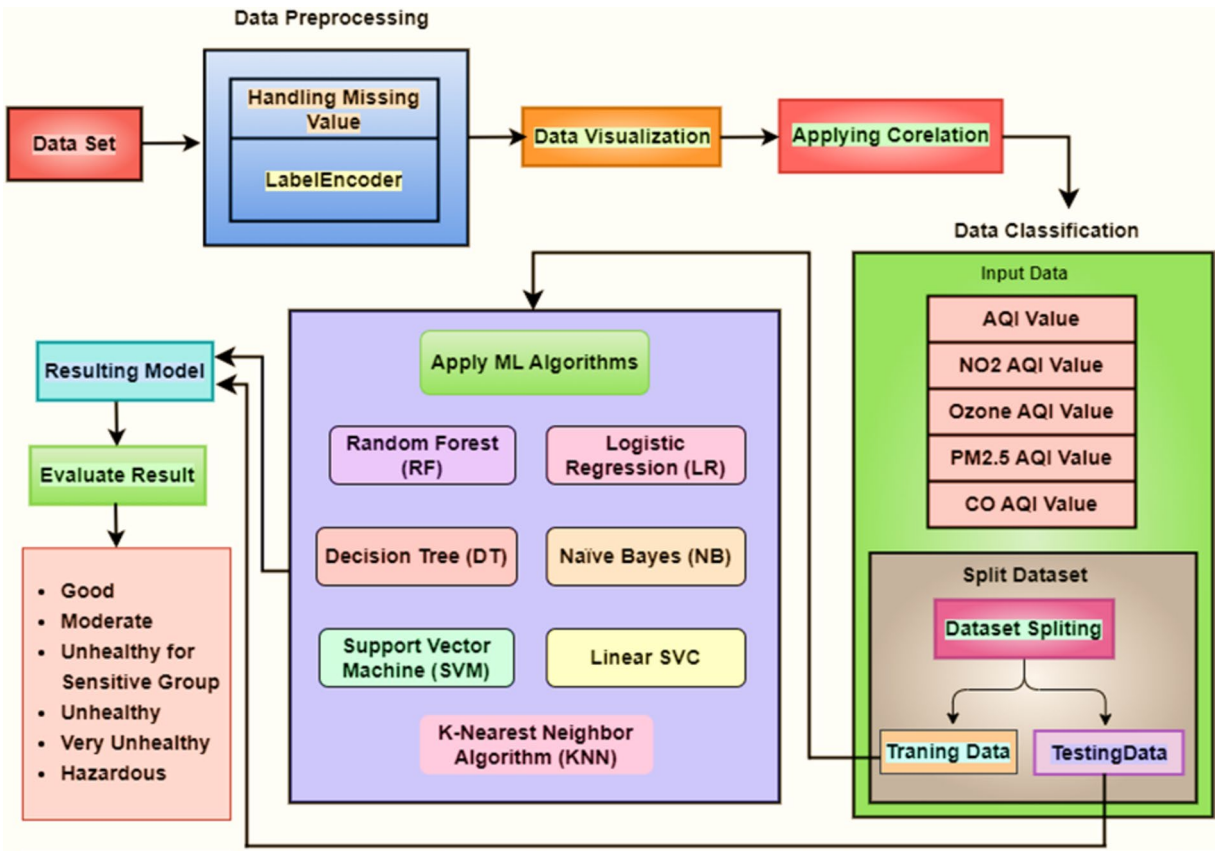


Fig. 1 Basic diagram of proposed 'AirNet' pollution prediction system

Algorithm 1 "AirNet" pollution prediction system

Input: AQI value, CO AQI value, Ozone AQI value, NO_2 AQI value, PM2.5 AQI value

Output: AQI categories **(**Good, Unhealthy for sensitive groups, Unhealthy, Very unhealthy, and Hazardous**)**

1. **Begin**
2. **data-preprocessing:**
3. dataset ←load dataset
4. **if** dataset.value is **equal** empty or missing
5. remove empty or missing_value
6. apply correlation
7. **feature-selection:**
8. select relevant features that contribute to predicting the AQI categories
9. convert AQI categories equal numerical labels using a label encoder
10. a ← dataset.drop [AQI categories]
11. b ← dataset.AQI categories
12. a1, a2, b1, b2 ←split_dataset of a and b
13. **model_training_testing:**
14. model ← train_model with a1 and b1
15. predict ← test_model with a2 and b2
16. evaluate the model performance using appropriate metrics (accuracy, precision, recall, F1-score)
17. compare the performance of all trained models using the evaluation metrics
18. identify the best-performing model based on the evaluation results
19. **End**

Table 2 Air quality index chart for various categories (Organization 2023)

AQI range	Air quality status	Application
0–50	Good	The air quality is at a safe level, and there is little to no harm expected from breathing in the air around you. It's a positive sign for our well-being
51–100	Moderate	The air quality is okay, but certain pollutants might be a bit worrisome for those who are extra sensitive to air pollution
101–150	Unhealthy for Sensitive Group	People in sensitive groups might feel some health effects, but the general public is unlikely to be impacted
151–200	Unhealthy	Everyone may begin to suffer some health impacts; however, those who are more vulnerable may have more several symptoms
201–300	Very Unhealthy	During emergency conditions, health warnings are issued, and there's a higher likelihood that the entire population may experience effects
301–500	Hazardous	Health alert more serious health effects may be experienced by everyone

Dataset used in experimental results

The study utilizes a dataset gathered from Kaggle [60], containing 23,463 instances and 12 features such as 'Country', 'City', 'AQI value', 'AQI category', 'CO AQI value', 'CO category', 'Ozone AQI value', 'Ozone AQI category', 'NO_2 AQI value', 'NO_2 AQI category', 'PM2.5 AQI value' and 'PM2.5 AQI category'. The 'AQI category' column is crucial, indicating 0 for Good, 1 for Hazardous, 2 for Moderate, 3 for Unhealthy, 4 for Unhealthy for Sensitive Groups, 5 for Very Unhealthy. Initially, there were 9936 instances of good, 9231 instances of moderate, 2227 instances of unhealthy, 1591 instances of unhealthy for sensitive groups, 287 very unhealthy and 191 instances of hazardous. This dataset forms a solid foundation for exploring patterns related to air pollution occurrence.

Table 2 represents the air quality index categories. In the 'Good (0–50)' range, the air is exceptionally clean, perfect for outdoor activities, much like a refreshing day in nature's spa. Moving into the 'Moderate (51–100)' zone, the air quality is acceptable for most activities, similar to a peaceful day in the park. As we enter the 'Sensitive (101–150)' range, a few individuals might experience slight discomfort due to the air, so taking extra care is important, especially if you have respiratory sensitivities. When air quality falls into the 'Unhealthy (151–200)' zone, it's not the best time for outdoor adventures; it's wise to avoid strenuous activities for better health. In the 'Very unhealthy (201–300)' range, the situation becomes more serious, with the potential for severe health problems. It's advisable to stay indoors and take it easy. The 'Hazardous (301–500)' category represents an air quality emergency where everyone's health is at risk. During this period, it's crucial to stay inside to stay safe as the air quality emergency unfolds.

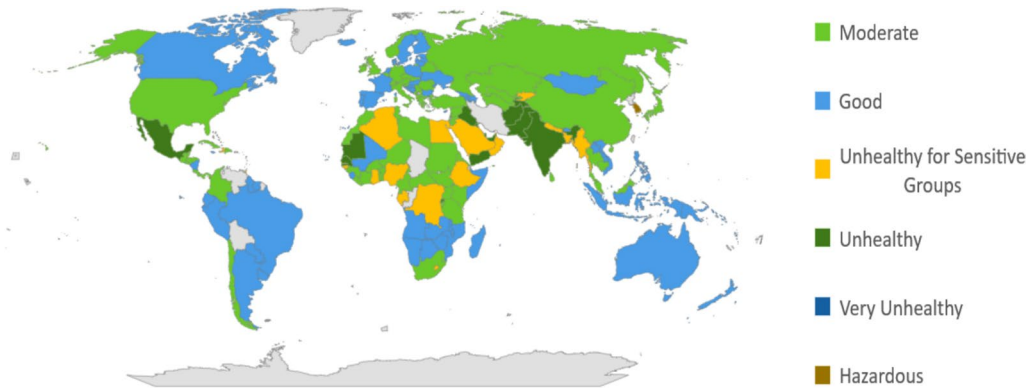


Fig. 2 World map color code distribution of AQI categories derived from dataset based on different cities in the world

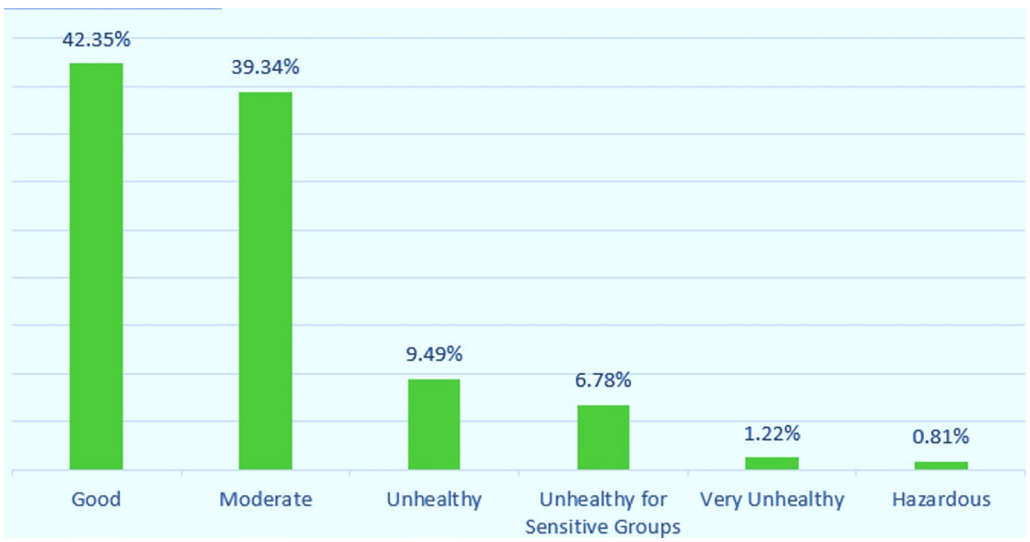


Fig. 3 Percentage of different AQI category in our dataset

In Fig. 2, we use a color code to represent the air quality index ranges. The color blue signifies 'good,' indicating air quality suitable for outdoor activities across the globe. Light green represents 'Moderate,' reflecting air quality acceptable for most daily endeavors. In the 'Sensitive' range, marked by yellow, caution is advised, especially for individuals with respiratory sensitivities. When air quality turns bottle green, it falls into the 'Unhealthy' category, prompting the avoidance of strenuous outdoor activities. Indigo designates 'Very unhealthy,' indicating a need for indoor refuge due to severe health concerns. Lastly, the color brown stands for 'Hazardous,' signifying an air quality emergency that requires everyone to stay indoors for safety. This color-coded system makes it easy to understand and assess air quality levels at a glance.

Figure 3 demonstrates the AQI values in category-wise of used dataset. Notably, 42.35% of the data falls into the 'Good' category, indicating favorable air quality. 'Moderate' air quality accounts for 39.34%, reflecting a balanced range suitable for most activities. Additionally, 'Unhealthy' air quality makes up 9.49% of the dataset, signaling areas of concern. 'Unhealthy for Sensitive Groups' represents 6.78%, highlighting a subset requiring extra attention. 'Very unhealthy' air quality is observed at 1.22%, and 'Hazardous' conditions account for 0.81%. This breakdown offers a detailed perspective on the prevalence of each AQI category, which is valuable for understanding the overall air quality landscape in our study.

Data pre-processing

Several preprocessing techniques were applied before the data were fed into the optimal variational machine learning system. The 'Global Air Pollution' dataset consists 23,463 data where 12 parameters are assigned. After completed all preprocessing technique the dataset consists 23,035 data, where all null and missing data are omitted. After addressing missing values, there are 9688 instances of good, 9087 instances of moderate, 2215 instances of unhealthy, 1568 instances of unhealthy for sensitive groups, 286 very unhealthy and 191 instances of hazardous data. To predicted air quality, we just need the AQI value of different air pollutants. As part of preprocessing, the best fitted features are encoded to normalize categorical data into numerical form. In order to simplify and improve user accessibility, the normalized dataset is shown in Table 3.

Table 3 Sample input and output values of normalized data

AQI Value	CO AQI Value	Ozone AQI Value	NO_2 AQI Value	PM2.5 AQI Value	AQI Category
66	1	39	2	66	2
51	1	36	0	51	0
66	1	39	2	66	2
51	1	36	0	51	0
51	1	36	0	51	0
34	1	34	0	20	3
51	1	36	0	51	0
51	1	36	0	51	0
66	1	39	2	66	2
66	1	39	2	66	2

The training and testing phases determine how well the dataset is classified. We split our whole dataset into two halves in order to improve the results: 80% of it was put aside for training, while the remaining 20% was set aside for testing.

Feature selection

In our research, we utilized a heatmap, as shown in Fig. 4, to serve as a comprehensive, color-coded representation of feature correlations within our dataset. The heatmap specifically illustrates the linear correlations between various variables, offering a visual and intuitive way to understand the relationships among the features. When it comes to selecting the most relevant features, the heatmap provides significant insights by highlighting which variables are strongly correlated. In our study, we employed correlation analysis to identify the features that were most appropriate for our model. The research focused on five key input features, and through Pearson's correlation coefficient, we determined the strength and direction of the relationships between these features. Pearson's coefficient, which ranges from − 1 to 1, measures the continuous association between two variables. A value of 1 or − 1 indicates a perfect positive or negative correlation, respectively, while a value of 0 indicates no correlation. This metric allowed us to quantitatively assess how closely the features are linked, both in terms of the strength and direction of their relationship.

The results from the air pollution dataset highlighted that certain features were highly dependent on the purity levels of the air.

This dependency is crucial as it reflects how variations in one pollutant can significantly impact the overall air quality. Furthermore, the correlation between the AQI (Air Quality Index) category levels and other pollutants' functioning enabled us to better understand the environmental impact of specific air pollutants. By analyzing these relationships, we could draw meaningful conclusions about how certain pollutants influence air quality,

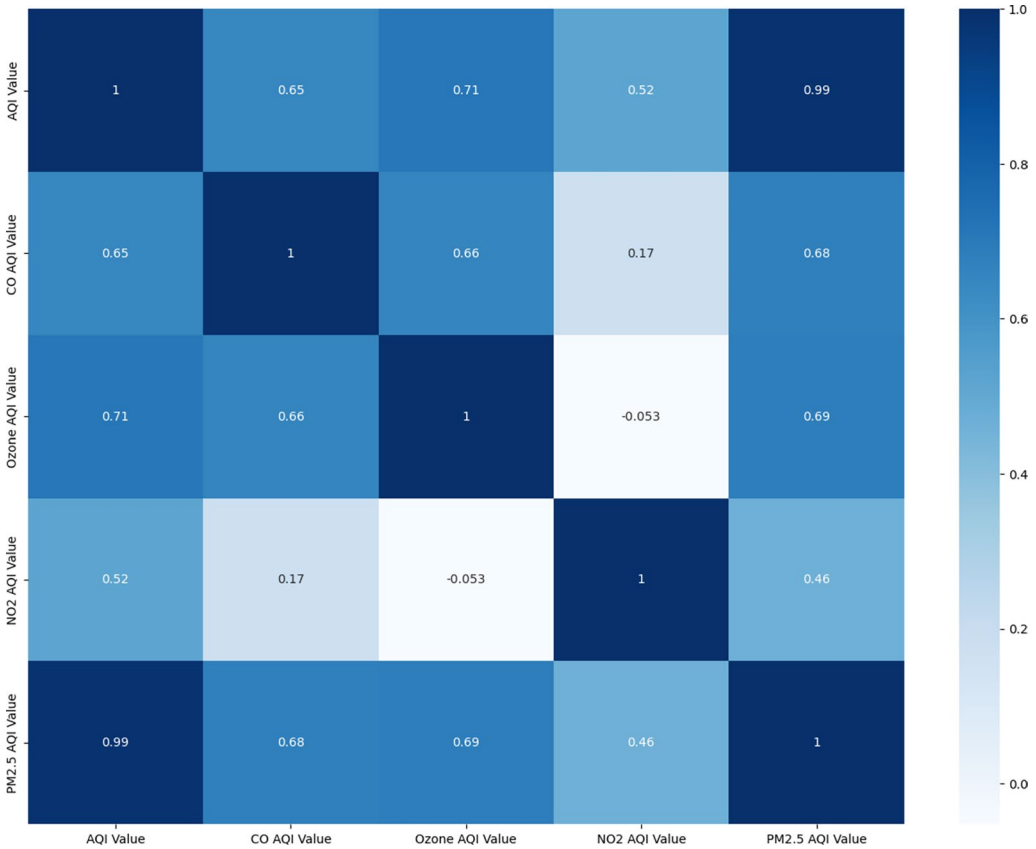


Fig. 4 Simple correlation plot of different air quality categories

which is pivotal in making accurate predictions and crafting effective mitigation strategies.

Data visualization

This useful tool offers a rapid overview of patterns and concentrations in our data, facilitating the identification of areas necessitating closer examination and in-depth analysis. It serves as a metaphorical traffic signal for our data, directing our attention to specific areas requiring focus. In our research, we used a pair plot that shown in Fig. 5. It visually explores the relationships and distributions between six different air quality categories. It's like joining the dots between different factors to spot patterns and trends. This tool is super useful because it gives us a sneak peek into what's happening with our data, making it easier to decide where to look more closely and explore further. It's like a visual guide for understanding the connections in our data.

In this study, a box plot was employed to visually represent the distribution of data, as illustrated in Fig. 6. A box plot, also known as a box-and-whisker plot, is an effective tool for summarizing the statistical distribution of data across different categories.

Each box plot in Fig. 6 depicts the distribution of AQI (Air Quality Index) values in relation to AQI categories. The box plot visualizes several key statistical metrics. The line within the box represents the median, which is the middle value of the data set. The box itself encompasses the interquartile range, which includes the middle 50% of the data. For Fig. 6a, this range shows where the central portion of ozone AQI values falls within the respective categories. Potential outliers are indicated by points outside the whiskers of the box plot. The whiskers extend to the smallest and largest values within 1.5 times the Interquartile Range (IQR) from the lower and upper quartiles. This range helps to identify values that deviate significantly from the rest of the data.

Figure 6a**,** which shows ozone AQI values across different AQI categories, the median values are consistently below 30%. This indicates that the central tendency of ozone AQI values is lower than 30% for the given categories. Figure 6b**,** This plot illustrates AQI values for a specific pollutant where the orange color denotes the AQI values for this category. Notably, there is no median value

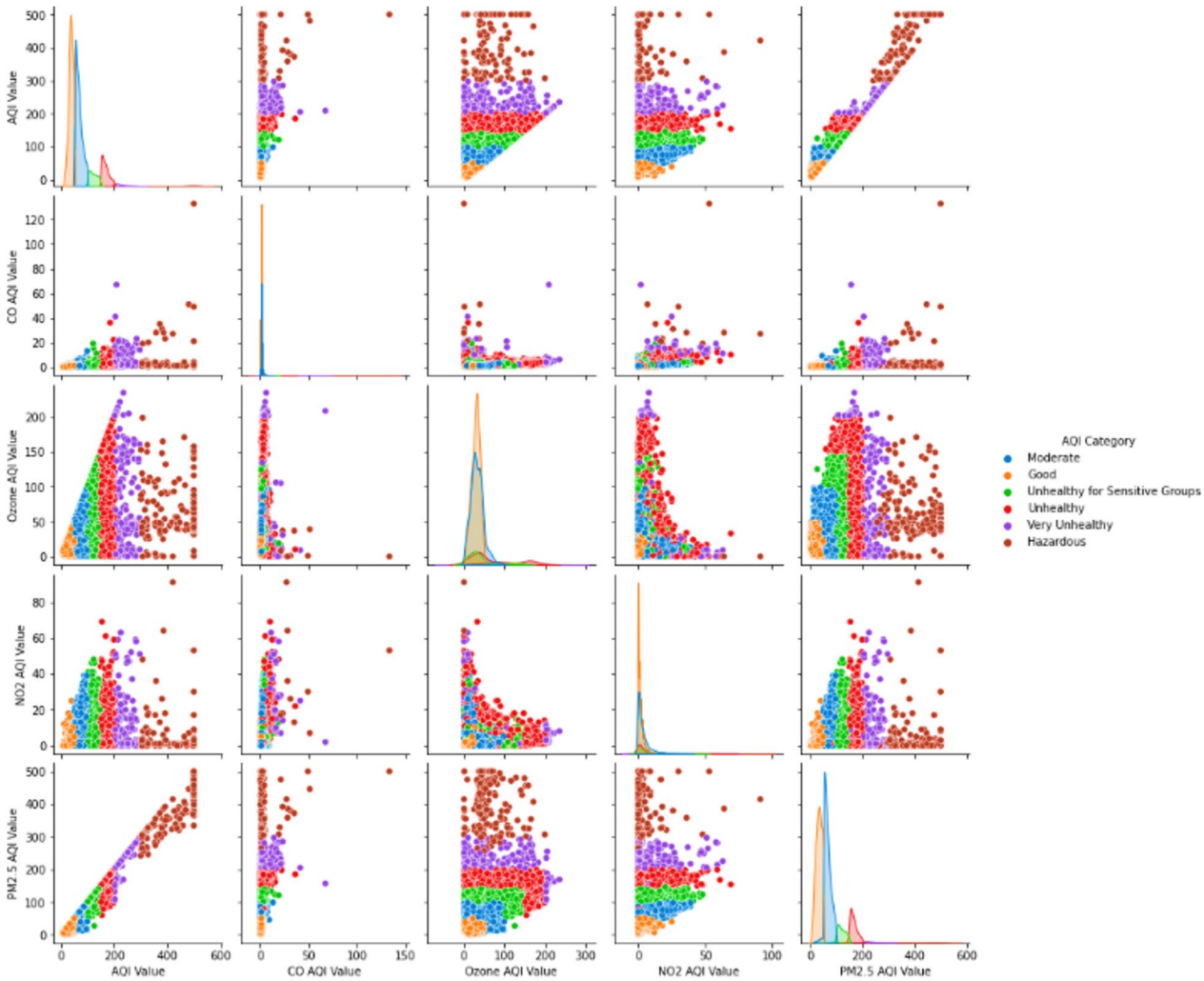
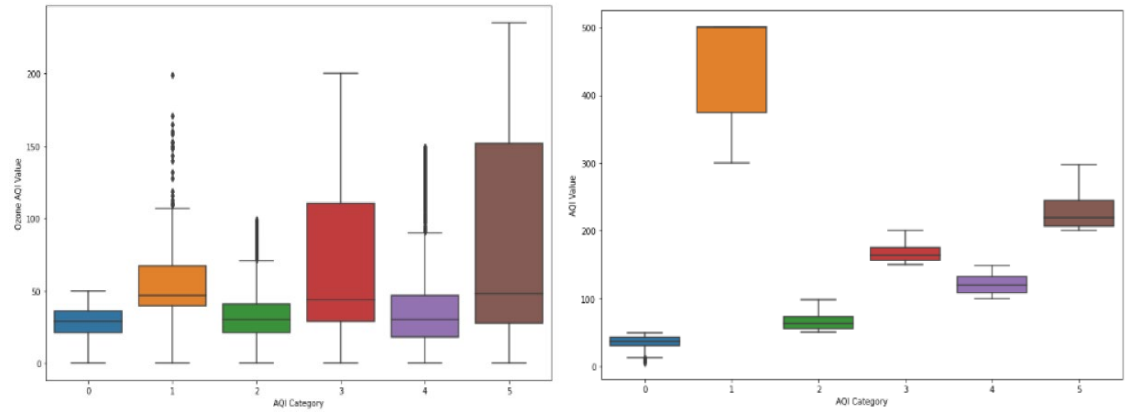


Fig. 5 Pair plot relationships different air quality categories

depicted for this category, suggesting that the data may not have a well-defined central tendency in this instance. Figure 6c**,** Displays the distribution of PM2.5 AQI values against AQI categories. Here, the median values are approximately 50%, and the IQR is positioned towards the lower quartiles for each category, indicating that PM2.5 values are more concentrated in the lower range of AQI categories. Figure 6d, e**,** Show the AQI values for NO_2 and CO, respectively. These figures reveal that the IQR values for both pollutants tend to fall within the upper quartiles of their categories, suggesting higher concentrations of NO_2 and CO in these ranges.

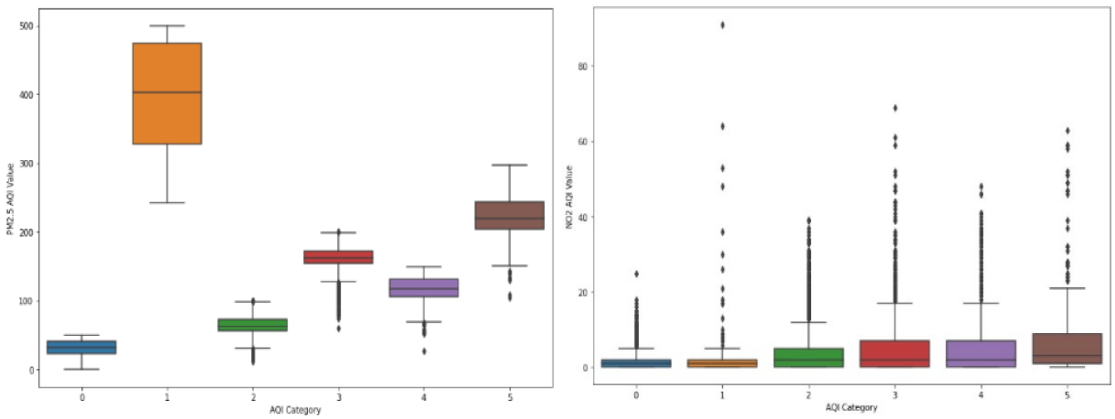
Hyperparameter tuning

Hyperparameters are important features that enhances the classification result. GridSearchCV and Mutual Information sometimes used in the realm of numerical data (Rahman et al. 2024). It leverages the dependency between performance matrix like accuracy, precision, recall and specificity and hyperparameters (Progga et al. 2023). To get the optimal values Mutual Information are applied to get more insightful outcome. In this process the value of 'k' mainly provides the most promising and top-rated parameters values among all features. Two base function named 'mutual_info_regression' and 'mutual_info_classif' are more promising in term of numerical values. In proposed system we have 12 features, among them 7 features are not numerical values like Country name, City name, pollutants category (Good, Healthy, Unhealthy…). That's why our proposed system not concern with the string values and without those 7 features we utilized rest of the 5 numerical features for further classification.



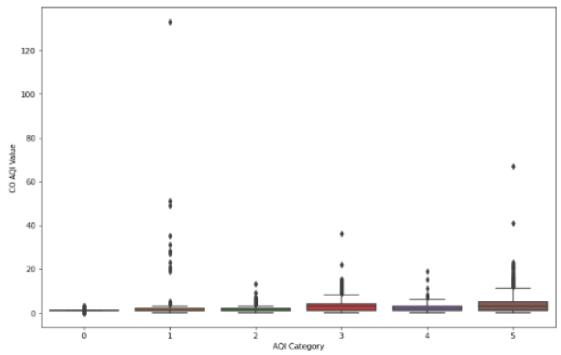
(a) Ozone AQI value vs AQI category

(b) AQI value vs AQI category



(c) PM2.5 AQI value vs AQI category

(d) NO_2 AQI value vs AQI category



(e) CO AQI value vs AQI category

Fig. 6 Box plot representation of different AQI value vs. AQI category:(**a**) Ozone AQI value vs. AQI category, (**b**) AQI value vs. AQI category, (**c**) PM2.5 AQI value vs. AQI category, (**d**) NO_2 AQI value vs. AQI category, (**e**) CO AQI value vs. AQI category.

Machine learning models

Utilizing data patterns to provide precise projections in many fields, machine learning models strengthen prediction systems and improve decision-making and productivity (Nur-A-Alam et al. 2024; Rahman et al. 2022; Ripa et al. 2024; Uddin et al. 2023). The performance is assessed by seven classifiers that are covered in this section: random forest, multinomial naïve bayes, linear svc, decision tree, random forest, and logistic regression. Subsequently, among all these classifiers, the model that

performs the best is assessed with the greatest accuracy. The computational complexity for the machine learning algorithms like RF, LR, SVM, Linear SVC, MNB is O(n). Some algorithms like DT and KNN is $O(n^2)$ (Ali et al. 2023; Hassan et al. 2022; Kearns 1990).

Random forest

Random forest is a powerful method that uses a group of decision trees to make accurate predictions for various tasks. To reduce overfitting, it creates an ensemble of decision trees, each of which is trained on a random data subset with feature subsets (Zhou et al. 2023a, b). Through bootstrapping, multiple tree models are built and their predictions combined, resulting in a robust and noise-resistant final output. This approach handles complex data patterns effectively while maintaining strong generalization performance, making RF a favored choice in machine learning.

Logistic regression

Logistic regression is a key player in supervised machine learning, particularly suited for classification tasks. Unlike traditional regression, it transforms outcomes into binary decisions, effectively drawing a decision boundary that separates data points into distinct categories (Geerts and De Weerdt 2023). It achieves this using a logistic function that strikes a balance between minimizing false positives and false negatives, making it a valuable tool for classifying data. The formula is

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1X)}} \tag{1}$$

Decision tree

Within the realm of supervised learning algorithms, the decision tree stands out as a prominent tool. Decision trees are crafted by leveraging inputs with high entropy (Wan et al. 2023). These trees take on a hierarchical structure, commencing with a root node and branching out to leaf nodes, with each leaf node corresponding to a distinct class label, while the internal nodes represent attributes (Roy et al. 2020). In essence, within an entropy-driven framework, decision trees excel at data refinement, sifting out extraneous samples to focus on the most informative aspects, which are encapsulated in the root node. In this case, p_{jk} is the ratio of the i_{th} class to the total number of models, and k is the response variable.

$$\text{Entropy } = \sum_{k=1}^{n} p_{jk} \log 2\, p_{jk} \tag{2}$$

Support vector machine

A supervised machine learning approach that may be applied to regression and classification problems is called support vector machines. The way SVMs operate is by identifying a hyperplane in the data that divides the various classes of data points. The distance between the hyperplane and the nearest data points is called the margin between the two classes, and it is maximized when the hyperplane is selected in this manner (Ghosh et al. 2019; Suthaharan and Suthaharan 2016).

$$w^t x + b = 0 \tag{3}$$

Here w is the hyperplane's weight vector, x denotes the input data vector, and b denotes the hyperplane's bias term.

Linear SVC

The Linear Support Vector Machine is a formidable classification technique known for its simplicity and effectiveness. It identifies a hyperplane that maximizes the margin between two classes, effectively separating them. By focusing on critical data points, or support vectors, it constructs a decision boundary that minimizes classification errors. Linear SVM is particularly valuable when dealing with high-dimensional data, offering a powerful tool for binary classification tasks, thanks to its ability to find an optimal balance between accuracy and generalization (Suthaharan and Suthaharan 2016).

$$\text{f(x)} = \text{w} \cdot \text{ x} + \text{b} \tag{4}$$

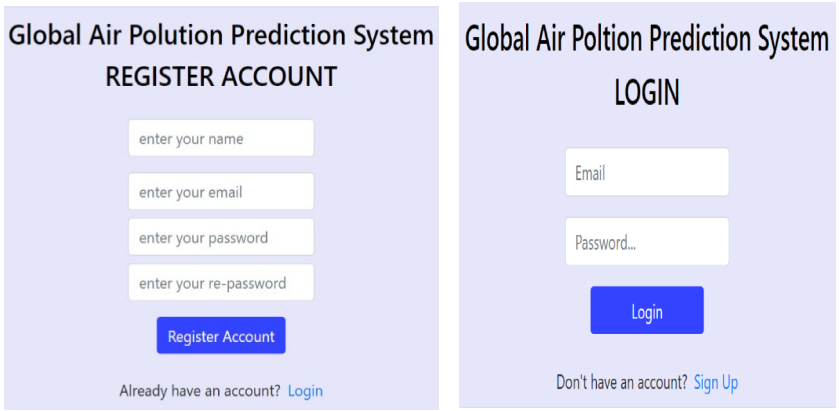
Here, x is the input vector, w is the weight vector that specifies the hyperplane's orientation, b is the bias factor (also referred to as the threshold) that moves the hyperplane away from the origin, and f(x) is the decision function for the given input vector.

Multinomial naive bayes

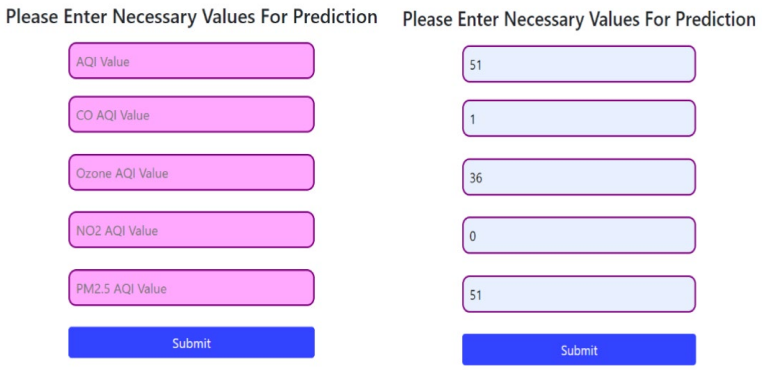
With reference to Bayes' theorem, Multinomial Naive One probabilistic machine learning technique is called Bayes. It operates under the premise that an object's characteristics are independent of one another. This implies that the existence of one characteristic has no bearing on the existence of any other feature (Jiang et al. 2016). From P(c), P(x), and P(x|c), one can calculate the posterior probability P(c|x) using the Bayes theorem. Have a look at the following equation.

$$p_{(c|x)} = \frac{p(x|c)p(c)}{p(x)} \tag{5}$$

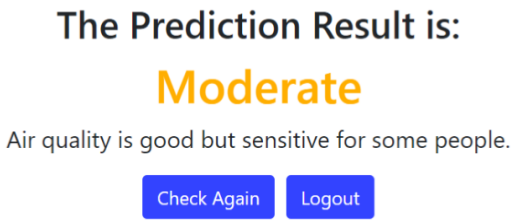
$$p_{(c|x)} = p_{(x_1|c)} * p_{(x_2|c)} * p_{(x_3|c)} * \ldots\ldots p_{(x_n|c)} * p(c) \tag{6}$$



(a) Registration and Login form.



(b) Attributes of AQI of proposed web interface.



(c) Prediction result of proposed web interface.

Fig. 7 Web interface for our proposed system: (**a**) Registration and login form, (**b**) Attributes of AQI in the proposed web interface, (**c**) Prediction result in the proposed web interface.

KNN

K-Nearest Neighbors (KNN) stands as a straightforward yet effective instance-based learning method in predictive modeling. It operates by finding the k-nearest data points in the training set, determined through distance metrics like Euclidean or Manhattan distance. This local neighborhood approach allows k-NN to make predictions for classification or regression tasks, drawing upon the class labels or values of nearby points (Sabry 2023). Its simplicity, adaptability, and reliance on data proximity make KNN a versatile tool widely employed in machine learning.

$$d(x,y) = \sqrt{\sum_{i}^{k} (x_i - y_i)^2} \tag{7}$$

where, d (x, y) is the Euclidean distance between two points, x and y, x_i and y_i are the coordinates of the two points in the ith dimension.

Web interface

The proposed method provides a global air pollution prediction system through a web interface, as illustrated in

Fig. 7. This system allows users to assess air quality by first creating a user ID. The figure is divided into three subfigures for detailed explanation:

Figure 7a**,** This subfigure displays the parameters of the login and registration form. Users must enter their credentials to access the system. Figure 7b**,** Shows the interface for predicting air quality, which includes five input parameters. Users enter data based on specific characteristics to evaluate air quality. Figure 7c**,** Presents the output of the prediction system. The result is categorized into one of six air quality levels: 'Moderate,' 'Good,' 'Unhealthy for Sensitive Groups,' 'Unhealthy,' 'Very Unhealthy,' and 'Hazardous.' The system uses machine learning classification methods to predict air quality, providing accuracy metrics and algorithmic outcomes for each method.

Experimental setup and discussion

In this section, we delineate the experimental setup, results, and discussions pertaining to the proposed method. The classification outcomes for various models, including support vector machine, random forest, k-nearest neighbor, decision tree, multinomial naïve bayes, and logistic regression, are presented here in. We conducted individual assessments for each model, examining both the confusion matrix and plot diagram. This systematic approach aimed to discern and select the most suitable model tailored to our dataset (Albahri et al. 2023).

Environmental setup

The proposed framework for air quality forecasting was developed utilizing TensorFlow and Keras, leveraging their capabilities in melding Python programming with neural networks (Sarkar et al. 2018; Vasilev et al. 2019). Training and evaluation of the model were executed within a Jupiter Notebook environment, running on 2.80 GHz Intel 11th Generation Core i7 CPU and supported by 16 GB of RAM, this PC configuration runs 64-bit Windows 11, providing the computational power necessary for the intricate tasks involved in training and assessing the predictive model for air quality.

Result and discussion

A crucial tool for evaluating a machine learning model's efficacy is a confusion matrix. It applied to a given dataset for classification tasks (Rahman et al. 2023a, b). It plays a pivotal role in quantifying the model's performance by summarizing how well it predicts categorical labels for input instances (Heydarian et al. 2022; Rahman 2022). In this paper when we try to measures such as Precision, Recall, False positive rate, True negative rate, F1-Score then we try confusion matrix. The matrix provides a comprehensive breakdown of four key metrics: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics offer a clear picture of the model's ability to correctly identify and classify instances of positive and negative outcomes, aiding in the evaluation and refinement of the model's classification accuracy. Equation (8, 9, 10), and (11, 12, 13) represents the computing formula of these performance metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{F1} - \text{score} = 2\left(\frac{Sensitivity * Precision}{Sensitivity + Precision}\right) \tag{12}$$

$$\text{Classification Error} = \frac{FP + FN}{TP + FN + TN + FP} \tag{13}$$

Result analysis

The global air pollution dataset's confusion matrix, as seen in Fig. 8. Seven pre-trained models were used to the dataset in order to perform classification. Figure 8a, b show that 9214 AQI categories in total have all been accurately identified using the Random Forest and Decision Tree algorithms. Figure 8a, b demonstrate that 100% classification accuracy has been achieved by both the Random Forest and Decision Tree designs. From Fig. 8c–g, it is clear that the Logistic Regression, Multinomial Naïve Bayes, SVM, Linear SVC, K-Nearest Neighbor failed to classified 931, 2187, 291, 50 and 23 AQI Category.

The performance evaluation of several machine learning models reveled varying levels of effectiveness based on precision, F1 score, recall, accuracy and classification error. Both RF and DT models achieved perfect score across all metrics, indicating flawless performance. The K-Nearest Neighbors (KNN) model followed closely, with near-perfect precision, F1 score, recall, and accuracy, and a minimal classification error, demonstrating its robustness. Linear SVC also performed exceptionally well, with high scores across all metrics and a low classification error. The Support Vector Machine (SVM) showed

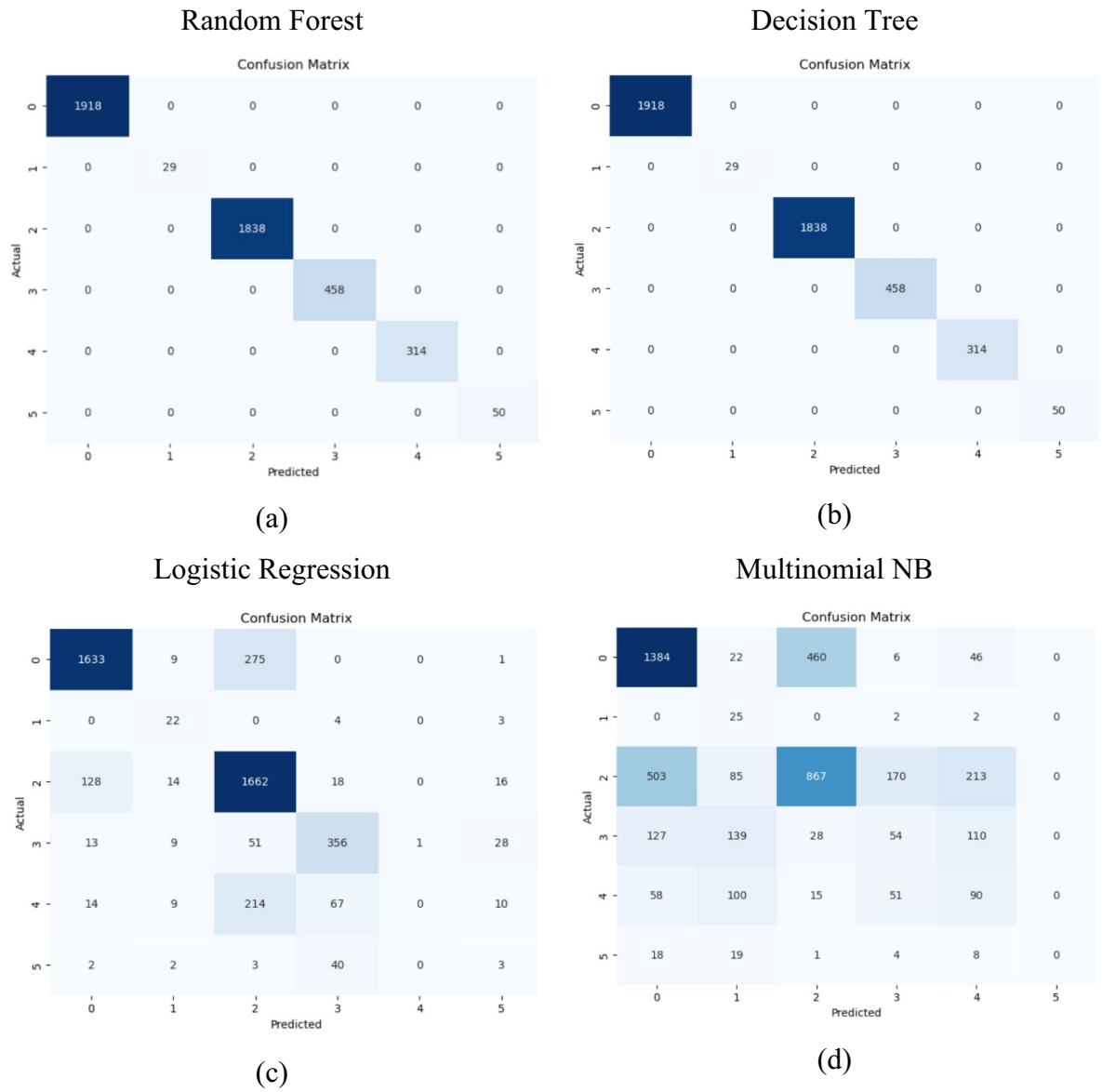


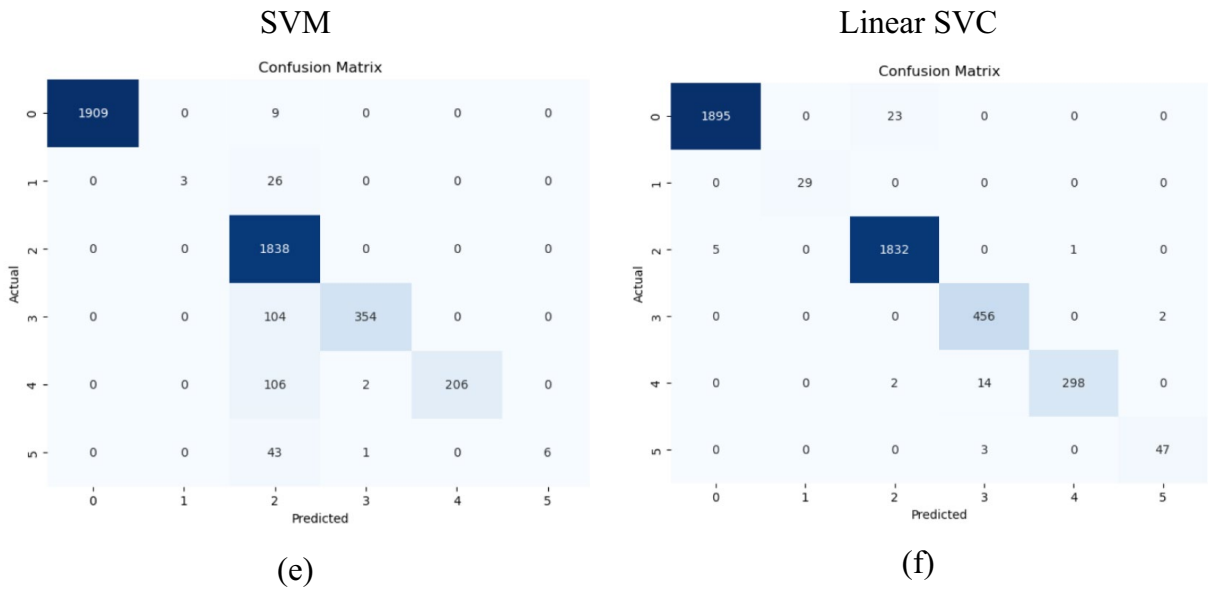
Fig. 8 Confusion matrix different machine learning models: (**a**) Decision Tree, (**b**) Logistic Regression, (**c**) Multinomial Naive Bayes, (**d**) Support Vector Machine (SVM), (**e**) Linear Support Vector Classifier (SVC), (**f**) K-Nearest Neighbor.

strong performance with slightly lower scores and a higher classification error than the top models. Logistic Regression had moderate performance, with uniform scores across metrics, but was notably less competitive. Lastly, the Multinomial Naive Bayes (MNB) model had the lowest performance, with lower scores and a significantly higher classification error, indicating it may not be suitable for this dataset. Table 4 shows the precision, F1 Score, Recall Score, accuracy, and classification error of different algorithms.

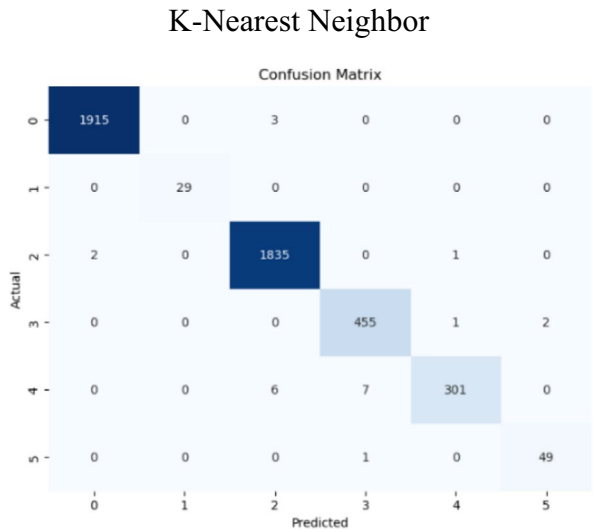
The accuracy of the provided data using various methods is shown in Fig. 9. Among all approaches, Random Forest and Decision Tree offer the best results, accuracy (100%) and lowest classification errors for same dataset. Whereas two classifiers MNB and LR has less than 80% accuracy.

Discussion with existing works

Due to the presence of several contaminants that can negatively impact the respiratory, cardiovascular, and other physiological systems (Rahman 2022; Shadid et al. 2019). Accurate air quality detection is essential for safeguarding the environment and public health in its early



(e) (f)



(g)

Fig. 8 continued

Table 4 Performance metrics result of different classifiers

Model	Precision	F1 Score	Recall Score	Accuracy	Classification Error
Random Forest	1.00	1.00	1.00	1.00	0.00
Decision Tree	1.00	1.00	1.00	1.00	0.00
Logistic Regression	0.79	0.79	0.79	0.79	0.21
Multinomial NB	0.52	0.52	0.52	0.52	0.48
SVM	0.93	0.93	0.93	0.93	0.07
Linear SVC	0.98	0.98	0.98	0.98	0.02
KNN	0.99	0.99	0.99	0.99	0.01

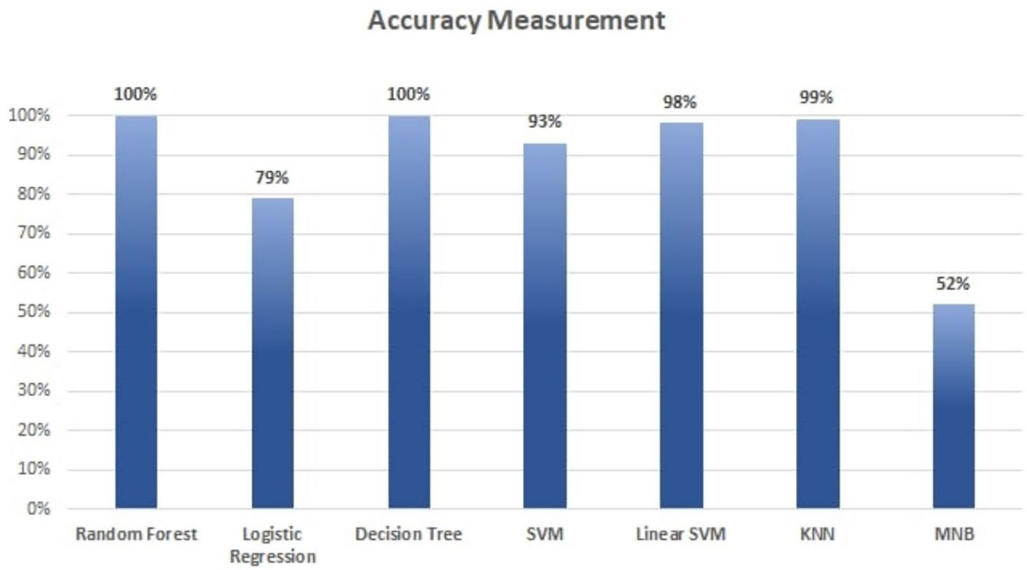


Fig. 9 Comparative results of different classifier

stages (Edo et al. 2024). The analysis shows a range of accuracies across different research, from 37 to 100%. By utilizing a comprehensive feature set and advanced machine learning methods, previous work has significantly improved prediction accuracy. Basel et al. (2020) achieved 100% accuracy but with a limited set of features, focusing only on PM2.5, air pressure, humidity, temperature, and seasonal data. In contrast, our proposed method incorporates a broader set of air pollutants, such as AQI value, CO AQI value, Ozone AQI value, NO2 AQI value, and PM2.5 AQI value, leading to a similarly high accuracy of 100%. This highlights the effectiveness of expanding the feature set and applying a diverse array of machine learning models, such as Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Linear SVC, K-Nearest Neighbors (KNN), and Multinomial Naive Bayes (MNB).

Previous works, such as Sanjeev, (2021), used temperature, methane (CH4), CO, non-methane hydrocarbons (NMHC), and PM10/PM2.5 alongside

Table 5 Related works based on air quality prediction

Reference	Features	Methodology	Accuracy
K. Kumar et al. 2023	PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, NO_x, NO	KNN, GNB, SVM, RF, XGBoost	90%
Sanjeev, 2021	Temperature, CH_4, CO, NMHC, NO, NO_2, NOx, O_3, PM10, PM2.5, RH, SO_2	RF, SVM, and ANN	97%
Xue-Bo Jin et al. 2023	PM2.5	CNN, GRU, LSTM	99%
K. Nishant Raj Kapoor et al (2023)	Air Pollution, PM2.5, Air Pressure, Humidity, Temperature, Season	Neural Network, LSTM, GRU, LASSO	98%
Chen. C.-W et al. (2021)	PM2.5, PM10, SO_2, CO	PCA, VGG-16	85%
K. Basel Y et al. 2020	PM2.5, Air pressure, Humidity, Temperature, Season	LSTM, Unsupervised Deep Learning Network Model, RNN	100%
AlThuwaynee, O.F. et al (2021)	CO, SO_2, NO_2, O and PM	DT, Gradient Boosted Tree, RF	82%
Avila et al (2023)	O_3, PM2.5, PM10, SO_2, CO, And NO_2	ARIMA, ARIMAX and RNN Models	75%
Abirami, S. et al. 2023	Encoder, STAA-LSTM Network	AAD, RMSE, MAE, and R2	37%
Kapoor, N.R., et al. 2023	Number of Occupants, Area Per Person, Outdoor Temperature, Outer Wind Speed, Relative Humidity, and Air Quality	ANN, SVM, DT, GPR, LR, EL, Optimized GPR, Optimized EL, Optimized DT and Optimized SVM	98%
Proposed Method	AQI value, CO AQI value, Ozone AQI value, NO_2 AQI value, and PM2.5 AQI value	RF, LR, DT, SVM, Linear SVC, KNN and MNB	100%

models like RF, SVM, and artificial neural networks (ANN), achieving an accuracy of 97%. Similarly, Xue-Bo Jin et al. (2023) utilized PM2.5 data with deep learning models like convolutional neural networks (CNN), gated recurrent units (GRU), and long short-term memory (LSTM), achieving 99% accuracy. Our method, which achieves 100% accuracy, represents a further step forward by applying a broader spectrum of AQI values and using both traditional and advanced machine learning algorithms.

The comparison of our results with existing works is presented in Table 5, showing that our method provides competitive performance.

Conclusions

In this work, we focus on the early detection of air quality, crucial for safeguarding public health and the environment, with the potential to save millions of lives globally. Our research introduces an advanced air quality prediction system that integrates various machine learning techniques, including K-Nearest Neighbor, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and Linear SVC, achieving a remarkable 100% classification accuracy with both Random Forest and Decision Tree models. This work contributes uniquely by combining conventional and advanced methods and offering a user-friendly web interface for real-time monitoring and alerts. Future research directions include exploring additional deep learning models, incorporating real-time data analytics, and expanding the system's scalability to cover more geographic regions. The model's limitations include its reliance on historical data, challenges in real-time data integration, and the need for further validation across diverse conditions.

Author contributions

Md. Mahbubur Rahman(MMR) and Md. Emran Hussain Nayeem(MEHN) designed the study. MMR, MEHN, and Md. Shorup Ahmed(MSA) wrote the manuscript; Khadiza Akther Tanha (KAT) collected and preprosseed data. Md. Shahriar Alam Sakib(MSAS) designed the web interface; Hafiz Md. Hasan Babu(HMHB), and Khandaker Mohammad Mohi Uddin(KMMU) edited the manuscript; MMR, and MEHN performed the analyses. MEHN, and MSA generated all figures and tables. All of the authors have read and approved the final version of the paper.

Funding

None.

Availability of data and materials

All the materials can be retrieved from https://github.com/shimulmbstu/AirNet. Training and testing data of proposed research is collected from Kaggle and cited as [37] in main manuscript. Reference: Al Muzaddid, H. (Year). Global Air Pollution Dataset. Kaggle. https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset.

Declarations

Informed consents

On behalf of all authors, the corresponding author states that informed consent was obtained from all participants involved in the study.

Human and animal rights

On behalf of all authors, the corresponding author affirms that human and animal rights were upheld in the study.

Competing interests

The authors declare that they have no competing interests.

Received: 20 June 2024 Accepted: 27 September 2024
Published online: 09 October 2024

References

Abirami S, Chitra P (2023) Probabilistic air quality forecasting using deep learning spatial–temporal neural network. GeoInformatica 27(2):199–235

Akhtar J (2020) Non-small cell lung cancer classification from histopathological images using feature fusion and deep CNN. Int J Eng Adv Technol. 9:2249–8958

Alahmad B, Khraishah H, Althalji K, Borchert W, Al-Mulla F, Koutrakis P (2023) Connections between air pollution, climate change, and cardiovascular health. Canad J Cardiol. 39:1182–1190

Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Informat Fusion 96:156–191

Ali YA, Awwad EM, Al-Razgan M, Maarouf A (2023) Hyperparameter search for machine learning algorithms for optimizing the computational complexity. Processes 11(2):349

AlThuwaynee OF, Kim SW, Najemaden MA, Aydda A, Balogun AL, Fayyadh MM,Park HJ (2021). Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off inextreme-gradient boosting (XGB) and random forest (RF) algorithms. Environ Sci Pollut Res. 28:43544–43566. https://doi.org/10.1007/s11356-021-14615-5

Avila ML, Alonso AM, Peña D (2023). Modelling multiple seasonalities with ARIMA:Forecasting Madrid NO2 hourly pollution levels.

Chandra R, Agarwal S, Singh N (2022) Semantic sensor network ontology based decision support system for forest fire management. Eco Inform 72:101821

Chandra R, Tiwari S, Agarwal S, Singh N (2023) Semantic web-based diagnosis and treatment of vector-borne diseases using SWRL rules. Knowl-Based Syst 274:110645

Chen C-W, Tseng Y-S, Mukundan A, Wang H-C (2021) Air pollution: sensitive detection of PM2.5 and PM10 concentration using hyperspectral imaging. Appl Sci 11(10):4543

Ding J, Li J, Qi J, Fu L (2023) Characterization of dental dust particles and their pathogenicity to respiratory system: a narrative review. Clin Oral Investigat 27:1–15

Edo GI, Itoje-akpokiniovo LO, Obasohan P, Ikpekoro VO, Samuel PO, Jikah AN, Nosu LC, Ekokotu HA, Ugbune U, Oghroro EEA (2024) Impact of environmental pollution from human activities on water, air quality and climate change. Ecol Front. https://doi.org/10.1016/j.ecofro.2024.02.014

El-Habil BY, Abu-Naser SS (2022) Global climate prediction using deep learning. J Theor Appl Inf Technol 100(24):4824–4838

Geerts M, De Weerdt J (2023) An evolutionary geospatial regression tree. Proceedings of the 2nd International Workshop on Spatio-Temporal Reasoning and Learning (STRL 2023) co-located with the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023),

Ghosh S, Dasgupta A, Swetapadma A (2019) A study on support vector machine based linear and non-linear pattern classification. 2019 International Conference on Intelligent Sustainable Systems (ICISS),

Hameed S, Islam A, Ahmad K, Belhaouari SB, Qadir J, Al-Fuqaha A (2023) Deep learning based multimodal urban air quality prediction and traffic analytics. Sci Rep 13(1):22181

Hassan SU, Ahamed J, Ahmad K (2022) Analytics of machine learning-based algorithms for text classification. Sustain Operat Comput 3:238–248
Heydarian M, Doyle TE, Samavi R (2022) MLCM: Multi-label confusion matrix. IEEE Access 10:19083–19095
Jiang L, Wang S, Li C, Zhang L (2016) Structure extended multinomial naive Bayes. Inf Sci 329:346–356
Jin X-B, Wang Z-Y, Kong J-L, Bai Y-T, Su T-L, Ma H-J, Chakrabarti P (2023) Deep spatio-temporal graph network with self-optimization for air quality prediction. Entropy 25(2):247
Kang GK, Gao JZ, Chiao S, Lu S, Xie G (2018) Air quality prediction: big data and machine learning approaches. Int J Environ Sci Dev 9(1):8–16
Kapoor NR, Kumar A, Kumar A, Arora HC (2023) Prediction of indoor air quality using artificial intelligence. Machine intelligence, big data analytics, and iot in image processing: practical applications, 447–469.
Kearns MJ (1990) The computational complexity of machine learning. MIT press.
Kumar K, Pande B (2023) Air pollution prediction with machine learning: a case study of Indian cities. Int J Environ Sci Technol 20(5):5333–5348
Kumar SS, Chandra R, Agarwal S (2024) Rule based complex event processing for an air quality monitoring system in smart city. Sustain Cities Soc 112:105609
Likhon MNH, Rana SU, Akter S, Ahmed MS, Tanha KA, Rahman MM, Nayeem MEH (2024) SkinMultiNet: advancements in skin cancer prediction using deep learning with web interface. Biomed Mater Devices 25:1–17
Lu J, Yao L (2023) Observational evidence for detrimental impact of inhaled ozone on human respiratory system. BMC Public Health 23(1):929
Méndez M, Merayo MG, Núñez M (2023) Machine learning algorithms to forecast air quality: a survey. Artif Intell Rev 56:1–36
Mitreska Jovanovska E, Batz V, Lameski P, Zdravevski E, Herzog MA, Trajkovik V (2023) Methods for urban air pollution measurement and forecasting: challenges, opportunities, and solutions. Atmosphere 14(9):1441
Nur-A-Alam M, Uddin KMM, Manu M, Rahman MM, Nasir MK (2024) An automatic system to detect colorectal polyp using hybrid fused method from colonoscopy images. Intell Syst Applicat 22:200342
Organization WH (2023) WHO ambient air quality database, 2022 update: status report. World Health Organization.
Park Y-K, Kim B-S (2023) Catalytic removal of nitrogen oxides (NO, NO2, N2O) from ammonia-fueled combustion exhaust: a review of applicable technologies. Chem Eng J 461:141958
Plocoste T, Laventure S (2023) Forecasting PM 10 concentrations in the caribbean area using machine learning models. Atmosphere 14(1):134
Progga NI, Jahan F, Uddin M, Shafkat A, Azim MA, Islam MK (2023) Emotion detection using deep learning approach. 2023 International conference on information and communication technology for sustainable development (ICICT4SD),
Rahman MM (2022) A web-based heart disease prediction system using machine learning algorithms. Network Biology 12(2):64
Rahman MM, Khan MSI, Babu HMH (2022) BreastMultiNet: a multi-scale feature fusion method using deep neural network to detect breast cancer. Array 16:100256
Rahman MM, Basar MA, Shinti TS, Khan MSI, Babu HMH, Uddin KMM (2023a) A deep CNN approach to detect and classify local fruits through a web interface. Smart Agric Technol 5:100321
Rahman MM, Nasir MK, Nur-A-Alam M, Khan MSI (2023b) Proposing a hybrid technique of feature fusion and convolutional neural network for melanoma skin cancer detection. J Pathol Informat 14:100341
Rahman MM, Islam A, Islam F, Zaman M, Islam MR, Sakib MSA, Babu HMH (2024) Empowering early detection: a web-based machine learning approach for PCOS prediction. Informat Med Unlock 47:101500
Ravi SS, Osipov S, Turner JW (2023) Impact of modern vehicular technologies and emission regulations on improving global air quality. Atmosphere 14(7):1164
Ripa R, Uddin KMM, Alam MJ, Rahman MM (2024) Hepatitis C prediction using machine learning and deep learning-based hybrid approach with biomarker and clinical data. Biomedical Materials & Devices 45:1–18
Roy D, Panda P, Roy K (2020) Tree-CNN: a hierarchical deep convolutional neural network for incremental learning. Neural Netw 121:148–160
Sabry F (2023) K Nearest Neighbor algorithm: fundamentals and applications (Vol. 28). One billion knowledgeable.
Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int J Eng Res Technol 10(3):533–538
Sarkar D, Bali R, Ghosh T (2018) Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras. Packt Publishing Ltd.
Shadid SH, Shafkat A, Yasmeen MF, Sibli SA, Rafi MR (2019) Prediction of heart disease using data mining techniques: A case study. International Journal of Information and Decision Sciences, 41(1):11–11
Sram RJ, Binkova B, Dostal M, Merkerova-Dostalova M, Libalova H, Milcova A, Rossner P Jr, Rossnerova A, Schmuczerova J, Svecova V (2013) Health impact of air pollution to children. Int J Hyg Environ Health 216(5):533–540
Suthaharan S, Suthaharan S (2016) Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207–235.
Tipton J (2022) Leaving the city: health and happiness in the other America. John Hunt Publishing.
Uddin KM, Rahman N, Rahman MM, Dey SK (2023) Artificial intelligence based domotics using multimodal security. I.J. Intelligent Systems and Applications, 15(3):44–55. https://doi.org/10.5815/ijisa.2023.03.04
Vasilev I, Slater D, Spacagna G, Roelants P, Zocca V (2019) Python deep learning: exploring deep learning techniques and neural network architectures with Pytorch, Keras, and TensorFlow. Packt Publishing Ltd.
Wan X, Li Z, Yu W, Wang A, Ke X, Guo H, Su J, Li L, Gui Q, Zhao S (2023) Machine learning paves the way for high entropy compounds exploration: challenges, progress, and outlook. Adv Mater. https://doi.org/10.1002/adma.202305192
Wang J, Tang D (2023) Air pollution, environmental protection tax and well-being. Int J Environ Res Public Health 20(3):2599
Wu C-L, Song R-F, Zhu X-H, Peng Z-R, Fu Q-Y, Pan J (2023) A hybrid deep learning model for regional O3 and NO2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network. Environ Pollut 320:121075
Zhang S, Zhang Z, Li Y, Du X, Qu L, Tang W, Xu J, Meng F (2023) Formation processes and source contributions of ground-level ozone in urban and suburban Beijing using the WRF-CMAQ modelling system. J Environ Sci 127:753–766
Zhou X, Guo M, Li Z, Yu X, Huang G, Li Z, Zhang X, Liu L (2023a) Associations between air pollutant and pneumonia and asthma requiring hospitalization among children aged under 5 years in Ningbo, 2015–2017. Front Public Health 10:1017105
Zhou Z, Qiu C, Zhang Y (2023b) A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models. Sci Rep 13(1):22420

Publisher's Note