IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSING THE EFFECT OF ATMOSPHERIC POLLUTION ON THE GLOBAL ENVIRONMENT USING MACHINE LEARNING TECHNIQUES

D P Singh, Professor Mathematics
Amity University Uttar Pradesh Greater Noida Campus

Abstract: This research investigates the use of machine learning methods to examine the effect of atmospheric pollution on the global environment. By utilizing sophisticated algorithms, we aim to uncover patterns, trends, and relationships within extensive datasets pertaining to air quality and environmental health. Our strategy involves employing predictive models to evaluate the long-term effects of different pollutants on climate change, ecosystems, and human health. The findings highlight the capability of machine learning to offer deeper understanding of the intricate dynamics of atmospheric pollution and to support policy-making and mitigation efforts. The rapid progress of machine learning algorithms has improved the capacity to explore the chemical properties of various pollutants, analyze chemical reactions and their influencing factors, and simulate different scenarios. When integrated with data from multiple fields, machine-learning models become a potent resource for analyzing atmospheric chemical processes and assessing air quality management, warranting increased focus in the future. Six different machine-learning models are utilized to forecast air quality. Their results are assessed using standard metrics.

The Random Forest model demonstrated the highest accuracy. This study emphasizes the capability of resource-constrained countries to predict air quality autonomously while they await larger datasets to enhance the accuracy of their predictions.

Keywords: - Machine Learning, Atmospheric Pollution, Global Environment, Air Quality, Predictive Models, Climate Change, Environmental Health, Data Analysis, Pollution Impact, Accuracy.

1.Introduction

In recent decades, atmospheric pollution has emerged as one of the most pressing environmental challenges, impacting ecosystems, human health, and the climate on a global scale. The complexity of atmospheric pollution lies in its multifaceted nature, where various pollutants interact with each other and with the environment, leading to a range of adverse effects. Traditional methods of studying these effects often fall short due to the sheer volume of data and the intricate patterns involved.

In the contemporary world, energy use and its consequences are inevitable in human activities. Human-induced air pollution includes emissions from industries, vehicles, aircraft, burning materials such as straw, coal, and kerosene, and aerosol cans. Every day, a range of harmful pollutants like CO, CO2, particulate matter (PM), NO2, SO2, O3, NH3, Pb, and others are released into the environment. These chemicals and particles, which make up air pollution, negatively affect the health of humans, animals, and plants. Human health dangers span from respiratory issues such as bronchitis to more serious illnesses like heart disease, pneumonia, and lung cancer. Furthermore, inadequate air quality worsens modern environmental concerns such as global warming, acid rain, decreased visibility, smog, aerosol creation, changes in climate, and early death. Scientists have acknowledged that air pollution also threatens historical monuments adversely (Rogers 2019).

Air pollution has a detrimental impact on both human health and ecosystem stability. It causes approximately 7 million premature deaths annually and results in a global economic loss of $2.9 trillion (IQAir, 2020). The number of premature deaths due to exposure to ambient PM2.5 and ozone is increasing worldwide (Chowdhury et al., 2020). To address these growing risks, the monitoring of pollutant emissions, transformation, and deposition through ground monitoring and remote sensing has become essential. This approach has led to a significant increase in available monitoring data (Cetin et al., 2018; Elsunousi et al., 2021; Hu et al., 2017; Kuerban et al., 2020; Li et al., 2023; Sevik et al., 2019; Wen et al., 2022; Xu et al., 2022a). Making effective use of this data can offer a strong scientific foundation for managing air pollution and informing policy decisions.

The atmosphere is a dynamic and open environmental system that holds various pollutants and involves complex chemical reactions. These reactions include intricate interactions among diverse components and controlling factors, making them difficult to interpret. Traditional statistical regression methods, such as parametric regression models, struggle to capture these nonlinear relationships, particularly with the increasing amount of data in atmospheric science. This limitation often results in poor prediction accuracy for nonlinear problems analyzed using large datasets (Feng et al., 2011).

These models are primarily created using example data for computer simulations in tasks involving classification and regression (Zhong et al., 2021). Machine learning models establish direct connections between data points, which helps mitigate the influence of outliers (Ucun Ozel et al., 2020) and achieve greater prediction accuracy and resilience with extensive datasets (Chen et al., 2022; Yuchi et al., 2019). They provide a more accurate fit to data with lower root mean square error, particularly in nonlinear situations (Chen et al., 2022; Feng et al., 2011; Zimmerman et al., 2018). Additionally, ML models can incorporate various data types, including integers and strings, for model construction. ML models are generally simpler, faster, and more cost-effective compared to numerical models. These benefits contribute to the increasing popularity of ML models in atmospheric science research (Liao et al., 2021; Zheng et al., 2021). By integrating multiple data fields, ML models play a critical role in diverse applications such as short-term forecasting (Yan et al., 2021), analysing pollutant chemical behaviours (Huang et al., 2021), and conducting impact assessments (Lv et al., 2023).

Moreover, some regression models are overly complex, involving numerous explanatory variables and strict assumptions about variable distributions. These complexities impede in-depth analysis of intricate data and extraction of valuable insights. Therefore, there is a critical need for more accessible and precise methods to facilitate effective data analysis.

Bibliometric analysis is a valuable method for assessing the research status of a specific field (Qin et al., 2022a; Zhang and Chen, 2020). It not only analyzes the development of research themes (Zhang et al., 2020b) but also reveals social networks by showing the interrelations between countries, institutions, and authors (van Eck and Waltman, 2010). To improve comprehension of machine learning (ML), we performed a bibliometric analysis of its uses and progression in worldwide air pollution studies. This provides additional insight into the most suitable future applications of ML in atmospheric science research.

Vehicle emissions, releases from power plants and factories, agricultural exhaust, and other sources contribute to the rise in greenhouse gases. These gases negatively impact climate conditions and subsequently affect plant growth (Fahad et al., 2021a). Emissions of inorganic carbons and greenhouse gases also influence interactions between plants and soil (Fahad et al., 2021b). Climate fluctuations not only impact humans and animals but also significantly affect agricultural factors and productivity (Sönmez et al., 2021), leading to economic losses. The Air Quality Index (AQI), a crucial parameter for assessing public health, directly correlates with the level of danger posed to human populations. As a result, the necessity to forecast AQI ahead of time has prompted scientists to observe and simulate air quality. The monitoring and prediction of AQI, particularly in urban regions, have grown more crucial and complex due to the rise in vehicular and industrial activities. Although studies on air quality often focus on developing countries, where pollutants like PM2.5 are found in significantly higher concentrations compared to developed nations (Rybarczyk and Zalakeviciute, 2021).

The research focuses on tackling data imbalance by applying a resampling method. It utilizes five well-known machine learning models alongside this approach, assessing their effectiveness through standard metrics widely accepted in the field (refer to Table 1). Relevant scholarly works such as Ayturan et al. (2020), Alade et al. (2019b), Al-Jamimi et al. (2019), and Al-Jamimi and Saleh (2019) are cited. Section 2 conducts a literature review and comparative analysis of existing studies on air quality prediction with machine learning. Section 3 outlines the dataset, its preprocessing steps, and the techniques used for feature selection. Section 4 deals with uncovering hidden patterns in the dataset through visualization. Section 5 outlines the experimental

setup, examines seasonal trends, presents empirical results, and engages in discussions. The concluding section provides a summary of the study's findings.
This is where machine learning (ML) techniques come into play. By leveraging vast amounts of data and sophisticated algorithms, machine learning offers a powerful tool to unravel the complexities of atmospheric pollution. This paper aims to explore how machine-learning techniques can be utilized to study the effects of atmospheric pollution on the global environment. We will delve into various ML models, their applications in pollution data analysis, and how these advanced techniques can lead to more accurate predictions and effective mitigation strategies.

2. Literature review:

Previous studies have investigated different kinds of models, including statistical, deterministic, physical, and Machine Learning (ML) models, for predicting AQI. Conventional methods based on probability and statistics are intricate and less effective. ML-based approaches, however, have shown higher reliability and consistency in predicting AQI. The advancement of technologies and sensors has facilitated precise and easy data collection. Achieving accurate and reliable predictions from extensive environmental data demands rigorous analysis, a task efficiently handled by ML algorithms.
This study presents a hybrid model that uses Artificial Neural Networks and Kriging to predict air pollution levels at various locations in Mumbai and Navi Mumbai. Historical data from the meteorological department and the Pollution Control Board are used for training the model. The model is implemented and evaluated using MATLAB for ANN and R for Kriging, with findings detailed in the subsequent sections (Suhasini V. Kottur , Dr. S. S. Mantha,2015).The system utilizes Linear Regression and Multilayer Perceptron (ANN) Protocol to forecast pollution levels for the next day. It helps predict future pollution by analyzing fundamental parameters and current pollution data, as well as forecasting future trends. Time Series Analysis was also used to identify upcoming data points and predict air pollution levels (Ruchi Raturi, Dr. J.R. Prasad,2018).
Bellinger et al. (2017) conducted an extensive review of ML and data mining in air pollution epidemiology. They noted significant research activity in Europe, China, and the USA, highlighting the widespread use of classifiers such as Decision Trees (DT), SVMs, K-means clustering, and the APRIORI algorithm. Rybarczyk and Zalakeviciute (2017) sought to develop a model linking traffic density and air pollution. They emphasized the cost-effectiveness of traffic data collection and the enhancement in accuracy when combining it with meteorological factors. Their hybrid model demonstrated superior performance, particularly with morning data showing the highest accuracy. Al-Jamimi et al. (2018) emphasized the importance of supervised ML algorithms in addressing environmental protection issues.
Castelli et al. (2020) aimed to forecast air quality in California by employing the Support Vector Regression (SVR) machine learning algorithm. Their approach focused on predicting pollutants and particulate levels, introducing a novel method for modeling hourly atmospheric pollution. Doreswamy et al. (2020) investigated machine learning models for forecasting air pollutant (PM) concentrations. They utilized six years of air quality data from Taiwan, applying existing models and demonstrating close alignment between predicted and actual values.
Liang et al. (2020) investigated the effectiveness of six machine learning classifiers in forecasting Taiwan's AQI using data spanning 11 years. They identified Adaptive Boosting (AdaBoost) and Stacking Ensemble as the most suitable methods for air quality prediction, noting that predictive accuracy varied across different geographical regions. Madan et al. (2020) conducted a comparative analysis of twenty literary works focusing on pollutant studies, machine learning algorithms utilized, and their respective performances. They observed that incorporating meteorological data such as humidity, wind speed, and temperature improved pollution level predictions. Neural Networks (NN) and boosting models demonstrated superior performance compared to other prominent machine learning approaches. Madhuri et al. (2020) highlighted wind speed, wind direction, humidity, and temperature as influential factors in air pollutant concentrations. They discovered through supervised machine learning methods that the Random Forest (RF) algorithm exhibited the fewest classification errors when predicting AQI.
Monisri et al. (2020) collected air pollution data from various sources and endeavoured to develop a mixed model for predicting air quality. The authors stated that their proposed model intends to assist residents of small towns in analyzing and forecasting air quality. Nahar et al. (2020) developed a model to predict AQI based on ML classifiers. The authors analyzed data gathered by Jordan's Ministry of Environment over 28 months to identify pollutant concentrations. Their model accurately pinpointed the most polluted areas. Patilet al. (2020) presented some literary works on various ML techniques for AQI modeling and forecasting. Most researchers utilized Artificial Neural Network (ANN), Linear Regression (LR), and Logistic Regression (LogR) models for predicting AQI, according to the authors' findings.

In a study by Gopalakrishnan in 2021, Google's Street View data and machine learning were combined to forecast air quality in different parts of Oakland, California. The emphasis was on regions with limited data, resulting in the creation of a web tool to predict air quality across various neighborhoods within the city.

Sanjeev (2021) analyzed a dataset containing pollutant concentrations and meteorological factors. Their study involved predicting air quality, with the Random Forest (RF) classifier identified as the most effective due to its reduced susceptibility to overfitting.

The authors discovered that the hybrid model outperformed others, achieving the highest accuracy specifically with morning data. It highlights a noticeable lack of research attention towards air quality analysis and prediction in Indian cities, despite the alarming statistic that nine out of the ten most polluted cities globally are in India (Deshpande, 2021). It endeavors to contribute to the field with innovative approaches to data visualization, utilizing correlation coefficient-based statistical outliers for analytical insights, and comparing the performance of five prominent ML models using standard metrics.

Authors used machine-learning algorithms to forecast air quality index (AQI) for specific areas. AQI serves as a standardized metric for assessing air quality, determined by concentrations of gases like SO2, NO2, CO2, RSPM, SPM, among others, monitored by agencies. The primary goal is to forecast air pollution levels in a city using available ground measurements(Aditya C R, et. al. 2018).The research also examined the advantages and disadvantages of each model (Gaganjot Kaur Kang, et. al., 2018). They developed a model employing gradient descent boosted multivariable regression, leveraging historical data from previous years to predict AQI for the upcoming year. Enhancing model efficiency, they applied cost estimation techniques for predictive problem solving. They assert that their model effectively predicts AQI for entire counties, states, or defined regions when provided with historical pollutant concentration data (Mrs. Ganga et al 2019).

The proposed system performs two key functions: firstly, it identifies PM2.5 levels using atmospheric data provided. Secondly, It predicts PM2.5 levels for specific dates using historical data. It employed several algorithms, including KNN, Random Forest, Decision Tree, AdaBoost, XGBoost and Support Vector Machines.

3. Data pre-processing and cleaning procedures:

We used data from Kaggle to examine the effects of air pollution on the global environment. (https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset)

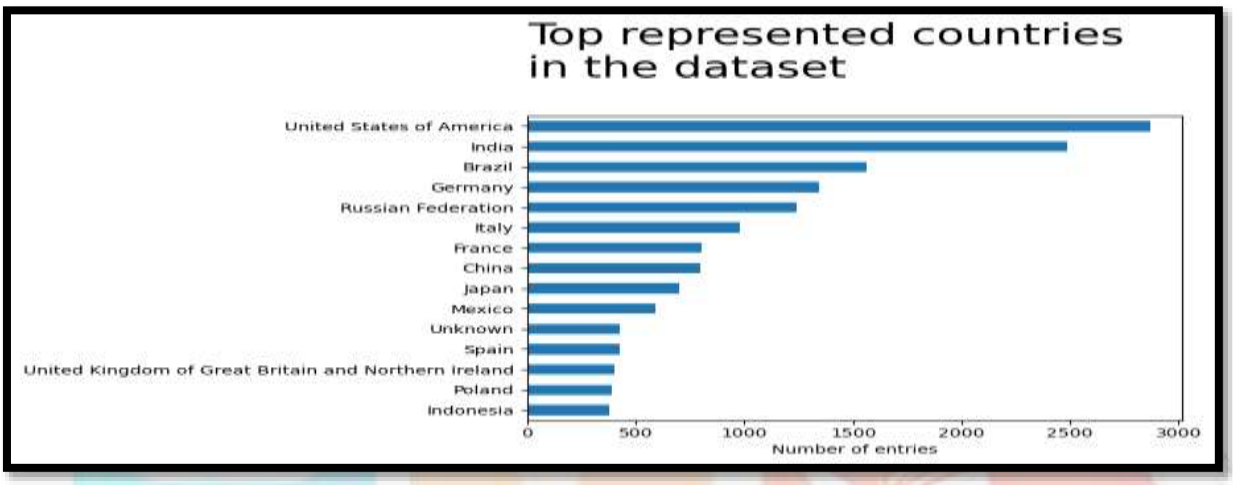
The quality of data stands as the primary and crucial requirement for effectively visualizing data and developing efficient machine learning models. Pre-processing steps are essential to diminish data noise, thereby enhancing the speed of processing and the ability of ML algorithms to generalize. Common data extraction and monitoring errors include outliers and missing data. Data pre-processing involves tasks like handling Non-values, eliminating or correcting outlier data, and more. According to a recent study by Dalberg Advisors and Industrial Development Corporation, air pollution in India results in annual economic losses estimated up to Rs 7 lakh crore ($95 billion) (Dalberg 2019).The main contributors to pollution are energy production, vehicle emissions, dust from soil and roads, waste incineration, power plants, and open burning of waste.

Data pre-processing and cleaning are crucial stages in preparing data for analysis and modeling. These steps involve converting raw data into a well-structured format suitable for further analysis. Key procedures include addressing missing values, eliminating duplicates, standardizing features, encoding categorical variables, and potentially reducing dimensionality through feature selection. Effective pre-processing ensures data accuracy, completeness, and proper formatting, establishing a robust basis for reliable and insightful analytical results.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23463 entries, 0 to 23462
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Country                  23036 non-null  object
 1   City                     23462 non-null  object
 2   AQI Value                23463 non-null  int64
 3   AQI Category             23463 non-null  object
 4   CO AQI Value             23463 non-null  int64
 5   CO AQI Category          23463 non-null  object
 6   Ozone AQI Value          23463 non-null  int64
 7   Ozone AQI Category       23463 non-null  object
 8   NO2 AQI Value            23463 non-null  int64
 9   NO2 AQI Category         23463 non-null  object
 10  PM2.5 AQI Value          23463 non-null  int64
 11  PM2.5 AQI Category       23463 non-null  object
dtypes: int64(5), object(7)
memory usage: 2.1+ MB
```

	dtype	specimen	nunique	null_share
Country	object	Russian Federation	175	1.82%
City	object	Praskoveya	23462	0.00%
AQI Value	int64	51	347	0.00%
AQI Category	object	Moderate	6	0.00%
CO AQI Value	int64	1	34	0.00%
CO AQI Category	object	Good	3	0.00%
Ozone AQI Value	int64	36	213	0.00%
Ozone AQI Category	object	Good	5	0.00%
NO2 AQI Value	int64	0	59	0.00%
NO2 AQI Category	object	Good	2	0.00%
PM2.5 AQI Value	int64	51	383	0.00%
PM2.5 AQI Category	object	Moderate	6	0.00%

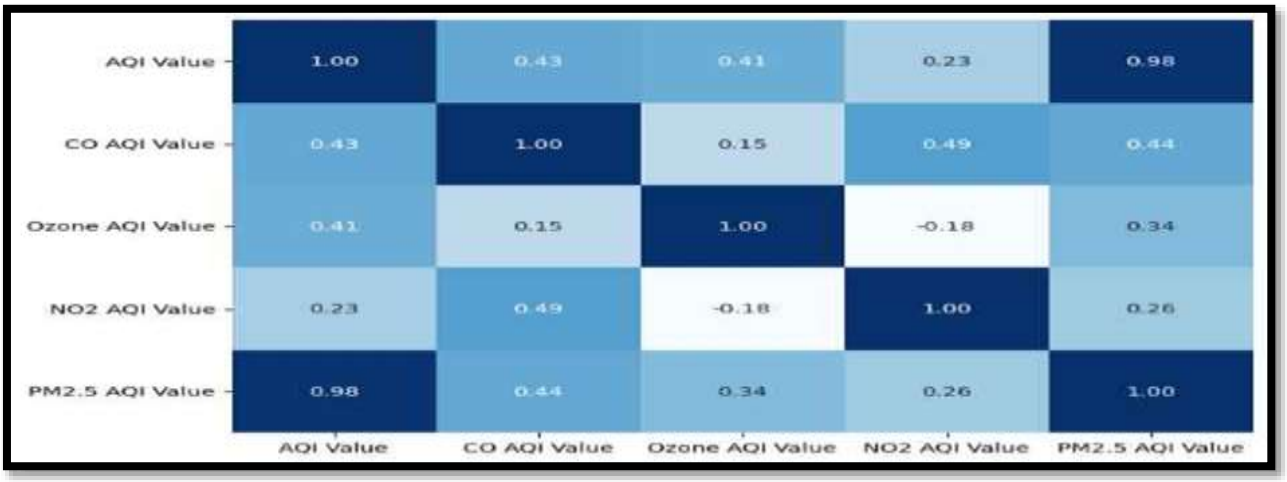
The dataset is meticulously prepared through pre-processing and cleaning procedures, followed by employing data visualization techniques to uncover deeper insights, hidden patterns, and trends.



The research also explores the correlation coefficient in conjunction with ML models, an approach that has been relatively underexplored in the existing literature (Alade et al., 2019a).
The table below displays the mean, standard deviation, minimum, lower quartile, median, upper quartile, and maximum values for AQI, CO AQI, Ozone AQI, NO2 AQI, and PM2.5 AQI:

	AQI Value	CO AQI Value	Ozone AQI Value	NO2 AQI Value	PM2.5 AQI Value
count	23463.000000	23463.000000	23463.000000	23463.000000	23463.000000
mean	72.010868	1.368367	35.193709	3.063334	68.519755
std	56.055220	1.832064	28.098723	5.254108	54.796443
min	6.000000	0.000000	0.000000	0.000000	0.000000
25%	39.000000	1.000000	21.000000	0.000000	35.000000
50%	55.000000	1.000000	31.000000	1.000000	54.000000
75%	79.000000	1.000000	40.000000	4.000000	79.000000
max	500.000000	133.000000	235.000000	91.000000	500.000000

The median values of the air quality index and the level of PM2.5 air pollution showed a highly significant positive correlation (98%):



The greater the former, the greater the latter. This suggests that poorer air quality typically involves these two features being closely linked. Carbon monoxide and ozone showed strong, though slightly weaker, positive correlations with the aforementioned factors when considering median values.

Nitrogen dioxide's behavior contrasts with others. It shows a moderate positive correlation with CO levels but a weak negative correlation with ozone median values.

The highest values generally showed moderate to strong positive correlations without significant exceptions. As before, the strongest correlation was observed between the air quality index and the level of PM2.5 air pollution. Nitrogen dioxide and ozone also exhibited a relatively weak, yet positive linear correlation.

4. Analysis Effect of Air Pollution on Global Environment:

Several researchers endeavoured to investigate the forecasting of air quality in global environment. Upon reviewing prior studies, it became clear that there was a notable void that could be addressed through the analysis and prediction of the Air Quality Index.

This study analyzes air pollution data encompasses observations from 175 countries, comprising 23463 cities instances across 12 features. In the study presents concise descriptive statistics of pollutants and the Air Quality Index (AQI) derived from this dataset. Key pollutants analyzed include PM2.5, NO2, CO, O3, among others. The research focuses on both analyzing these pollutants and predicting AQI levels.

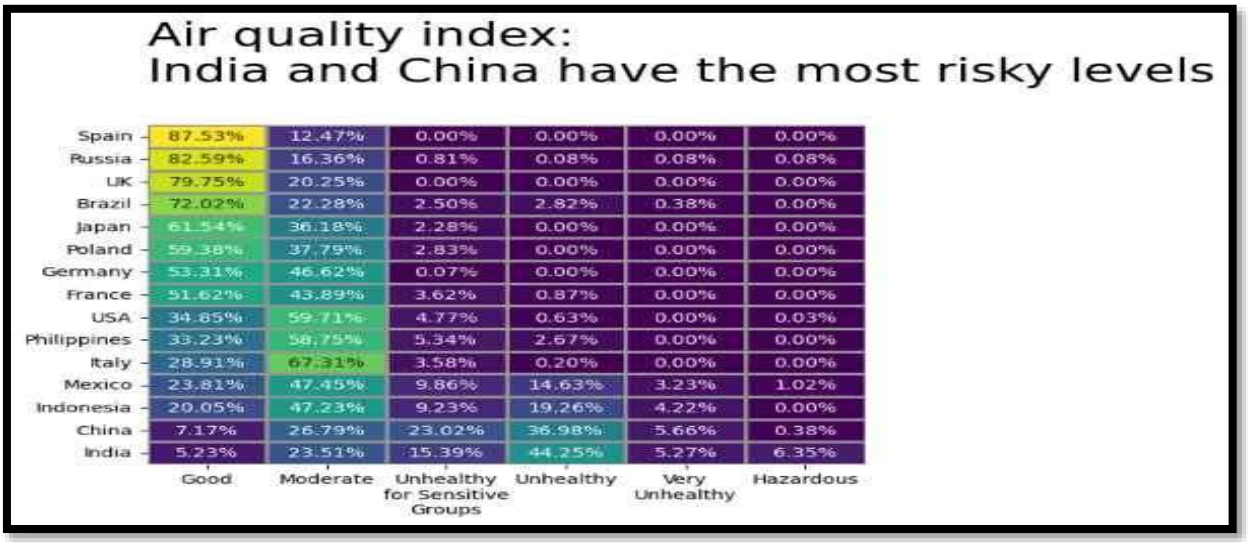
Air pollution significantly threatens the global environment, impacting climate, ecosystems, and human health. Through collective efforts, we can reduce the harmful effects of air pollution and safeguard the planet for future generations. An air quality index (AQI) informs the public about the current level of air pollution or predicts how polluted the air will be.

4.1Air quality index: India and China have notable percentages of locations classified as "Unhealthy for Sensitive Groups" or "Unhealthy." Specifically, in India, 15.4% of locations fall into the "Unhealthy for Sensitive Groups" category and 44.2% are "Unhealthy." In China, the corresponding percentages are 23% and 37%, respectively.

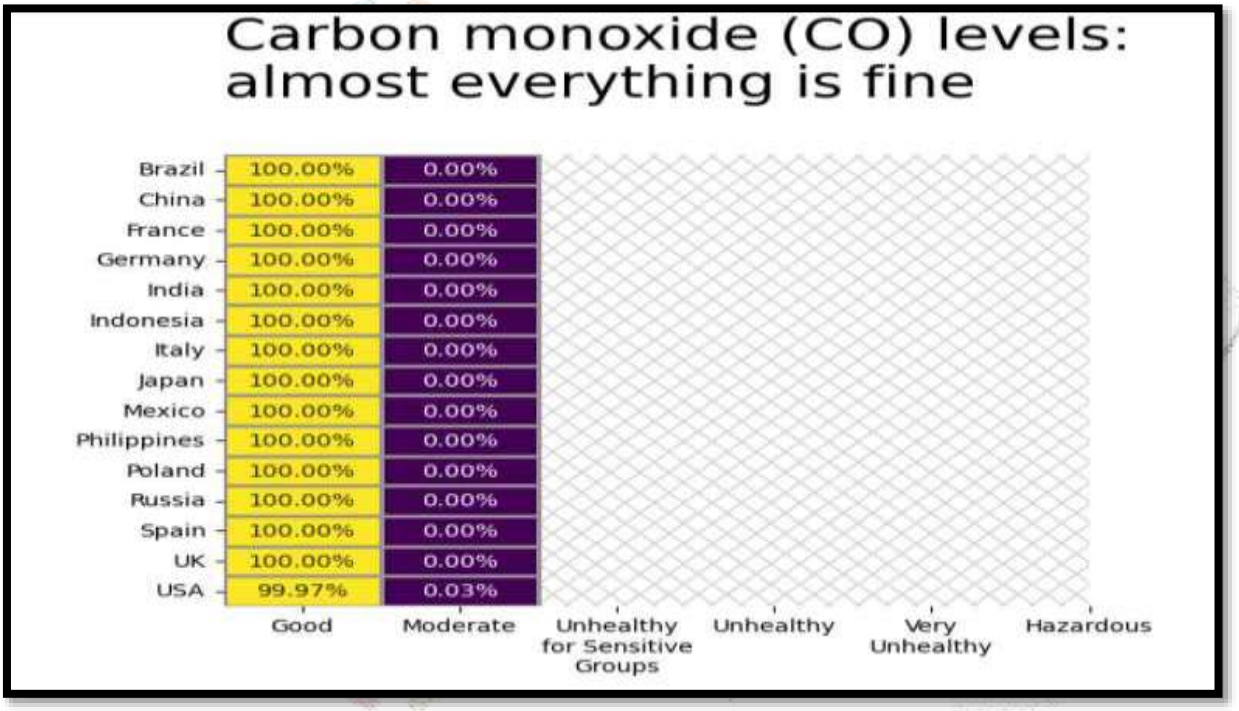
In India, a notable proportion of areas are classified under the worst category, 'Dangerous' (6.3%). Additionally, less than one-third of Indian locations are rated with a 'Good' or 'Moderate' air quality index, the rest fall below the threshold for healthy conditions.

India and China stand out from other countries with only about 5-7% of their locations rated as 'good.' In contrast, Indonesia and Mexico, which rank higher in air quality index, have more than 20-23% of their locations classified as 'good.'

Spain boasts the best air quality index, ranging from 'Good' to 'Moderate'. Unlike other countries, Spain has no locations listed under higher-risk categories.



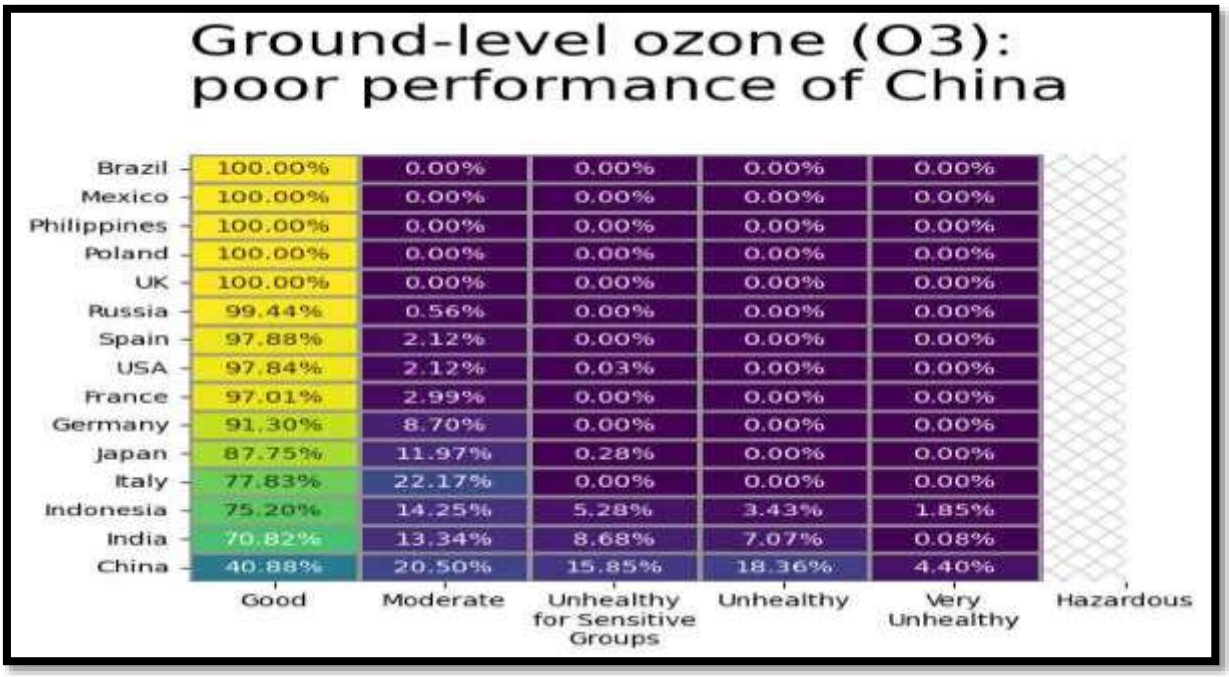
4.2CO levels: No country's locations are categorized as high-risk. It's likely that only the 'Good' to 'Moderate' scale is used for assessments. All the countries exhibit 'Good' performance, with only the USA having a small proportion of 'Moderate' locations.



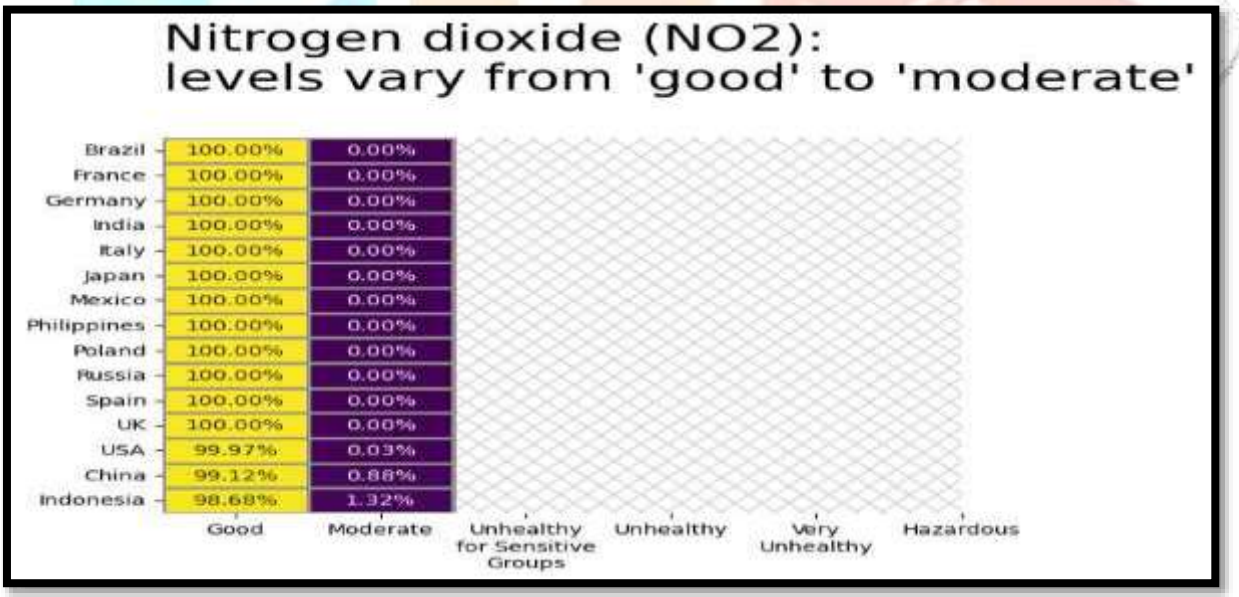
4.3Ozone Level: There is a lack of data on hazardous locations. China has shown particularly poor conditions, with less than 40% of areas classified as 'Unhealthy for Sensitive Groups' to 'Very Unhealthy'. However, more than 61% of locations were categorized as 'Good' to 'Normal', suggesting the situation is less severe compared to India's air quality index.

In the context of India, more than 70% of the locations are deemed to have favourable conditions within this category.

Five countries stand out as top performers in their category for ground-level ozone, with all achieving a 'Good' rating of 100%: Brazil, Mexico, the Philippines, Poland, and the UK.

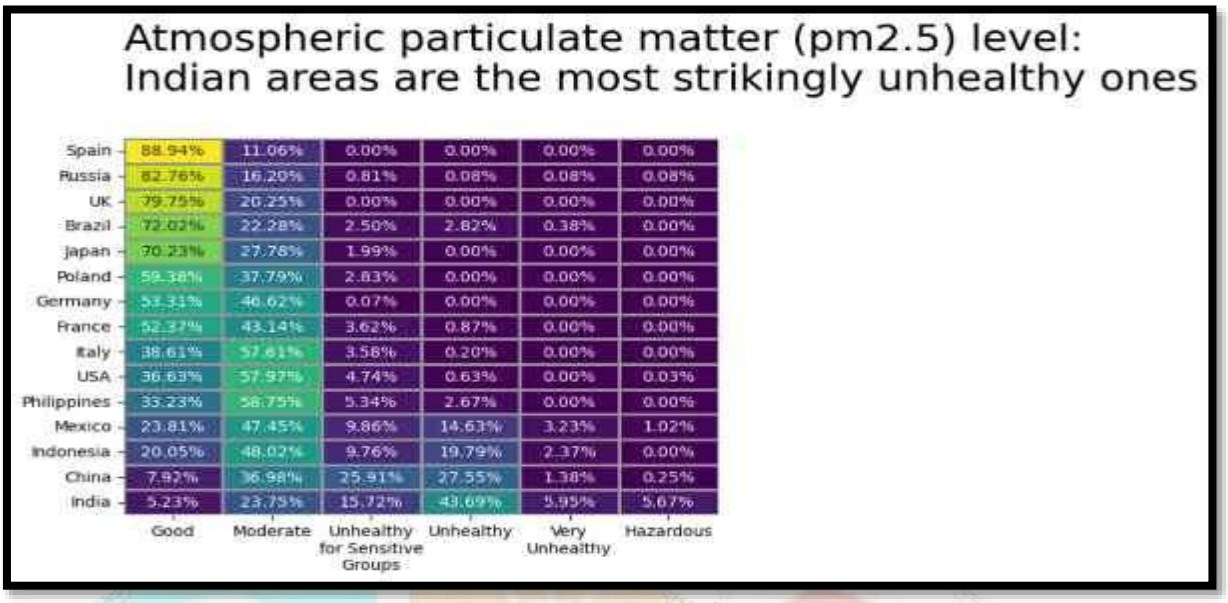


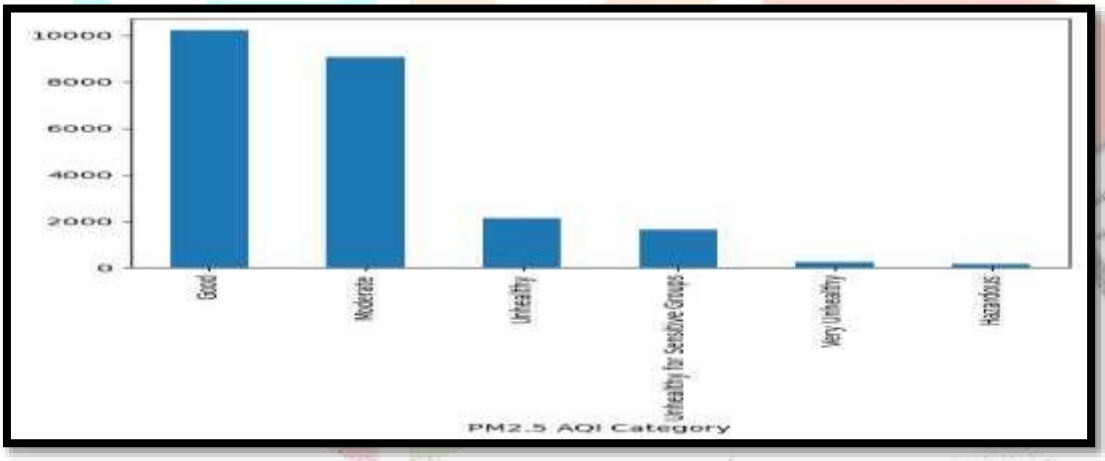
4.4 Nitrogen Dioxide Level: Similar to carbon monoxide, a two-part scale ranging from 'Good' to 'Moderate' is used. Indonesia shows relatively worse conditions with 1.3% of areas labeled as 'Moderate', while China has 0.9% in the same category. A very small proportion of 'Moderate' areas exists in the USA, amounting to just 0.3%.



4.5 PM2.5 Level: Like the air quality index, the most severe conditions are observed in India, China, Indonesia, and Mexico. The disparity in air quality index remains consistent between India and China, where 'Good' areas constitute approximately 5-8%, and between Indonesia and Mexico, where it ranges from 20-24%

Again, fewer than 30% of locations in India are categorized as 'Good' to 'Moderate', while approximately 44% are classified as 'Unhealthy'.





The Air Quality Index (AQI) is a numeric system used to indicate the extent of air pollution in a particular region at a specific moment. It relies on measuring concentrations of various pollutants such as particulate matter (PM2.5), ground-level ozone (O3), nitrogen dioxide (NO2), and carbon monoxide (CO). PM2.5 and ozone depletion are primary factors contributing to deteriorating air quality, while levels of CO and NO2 are currently managed effectively.

City	Country	AQI Value	CO AQI Value	PM2.5 AQI Value	NO2 AQI Value
Puranpur	India	500	2	481	1
Padampur	India	500	1	441	0
Yazman	Pakistan	500	1	428	0
Rohtak	India	500	1	467	1
Sasni	India	500	1	433	1
Bisalpur	India	500	1	480	1
Ganganagar	India	500	1	424	0
Sardulgarh	India	500	1	500	1
Gajraula	India	500	1	500	1
Kasganj	India	500	1	476	3

India and Pakistan host some of the most polluted cities on the planet, but their neighboring countries fare relatively better in terms of air quality.

5.ML Models: We employed six different machine learning algorithms to assess the effects of air pollution on the global environment.

5.1 K-Nearest Neighbors (KNN): KNN is a fundamental yet crucial classification technique in machine learning, falling under supervised learning. It is significantly relevant in fields such as pattern recognition, data mining, and intrusion detection. The K-NN algorithm is known for its simplicity and adaptability and is widely used in machine learning. Unlike many other methods, it doesn't assume specific data distribution, making it suitable for various datasets. Its versatility includes handling both numerical and categorical data, making it practical for classification and regression tasks. As a non-parametric method, K-NN predicts outcomes by evaluating the similarity between data points and is robust against outliers. It functions by identifying the K nearest neighbors to a data point using a chosen distance measure, like Euclidean distance (D P Singh, 2024a, D P Singh, 2024b).

5.2 Decision Trees: A decision tree is a highly effective tool in supervised learning, used for both classification and regression tasks. It is structured as a tree-shaped flowchart where each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a class label. The tree is built by iteratively splitting the training data into subsets based on attribute values, stopping according to criteria such as maximum tree depth or minimum samples required for node splitting. In classification, the tree uses input features to determine outcomes, with leaf nodes representing the final decision and internal nodes containing dataset characteristics and decision-making rules. The input feature with the highest information gain is selected to predict the output, with information gain calculated at each node for every attribute in the tree (D P Singh, 2024a, D P Singh, 2024b).

5.3 Random Forest: Random Forest, a widely used ensemble learning method involving decision trees, creates a 'forest' of multiple trees. These trees are usually trained with the 'bagging' technique, which merges multiple models to improve the overall result. Random Forest boosts the performance of Decision Trees by reducing variance, achieved by growing more trees and introducing more randomness into the model. Rather than always choosing the most significant feature for splitting nodes, it selects the best feature from a random subset of features, leading to a more robust model.Random Forest Regression is a machine learning ensemble method capable of managing both regression and classification tasks by utilizing multiple decision trees and applying Bootstrap and Aggregation, commonly known as bagging. Rather than relying on a single decision tree, this approach combines the outputs of several trees to produce the final result. In Random Forest, numerous decision trees act as the core learning models. The process includes randomly selecting rows and features from the dataset to generate sample datasets for each tree, a technique called Bootstrap sampling (D P Singh, 2024a, D P Singh, 2024b).

5.4 XGBoost: XGBoost, which stands for Extreme Gradient Boosting, is an advanced machine learning technique that enhances the traditional gradient boosting method. It includes regularization, which improves its performance and speed over standard gradient boosting. Additionally, XGBoost excels with datasets that have a combination of numerical and categorical variables.

Evaluation of performance involved metrics such as R-squared value, root mean squared error (RMSE), and magnitude relative error (MRE):

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (y_i - \check{y}_i)^2} \quad \text{and} \quad MRE = \frac{|y_i - \check{y}_i|}{|y_i|}$$

n represents the total count of observations within the dataset. Each observation, denoted as y_i, corresponds to the actual value for the i-th entry, while $\check{y}_i$ stands for the predicted value for the same i-th observation (D P Singh, 2024a, D P Singh, 2024b).

5.5 AdaBoost: AdaBoost is a boosting method that builds a strong model by progressively adding weak learners. The alpha parameter in AdaBoost is inversely related to the weak learner's error. This differs from XGBoost's Gradient Boosting, where the alpha parameter is determined based on the errors of the weak learners (D P Singh, 2024a, D P Singh, 2024b).

5.6 Support Vector Machines (SVM): The Support Vector Machine (SVM) is a powerful machine learning method used for both linear and nonlinear classification, regression, and outlier detection. It is applicable in diverse areas like text and image classification, spam filtering, handwriting and face recognition, gene expression analysis, and anomaly detection. SVMs are particularly effective at handling high-dimensional data and capturing nonlinear relationships, making them highly adaptable and efficient for various tasks. Their strength comes from their ability to find the optimal separating hyperplane that distinguishes between different classes in the target feature (D P Singh, 2024a, D P Singh, 2024b).

6. Result Discussion:

The distribution of major types of PM2.5 varied significantly across different global regions. North America, Central Africa, West Asia, South Asia, and East Asia showed notably high levels of PM2.5. The potential health impacts from prolonged exposure are likely to vary among these regions and need careful evaluation. A meta-analysis approach was used to assess the global effects of these diverse PM2.5 types. We have demonstrated in paragraph (4. Analysis Effect of Air Pollution on Global Environment) discussing the impact of air pollution on the global environment.

Six machine learning models including KNN, Decision Tree, Random Forest, XGBoost, AdaBoost, and Support Vector Machine were employed to predict air quality. Performance evaluation was conducted using standard metrics, revealing that the Random Forest model achieved the highest level of accuracy.

Best Regression model: The Random Forest model is the top performer with the highest Mean Cross Val Score (0.995934), the lowest MSE (0.011197), the highest R^2 Score (0.993914), and Mean Residuals nearly zero (0.002355).

Conversely, the SVR (Support Vector Regression) model is the least effective, with the lowest Mean Cross Val Score (0.849935), the highest MSE (0.268666), the lowest R2 Score (0.853981), and Mean Residuals far from zero (-0.007298).

Table Regression evaluation results:

	Mean Cross Val Score	MSE	R2 Score	Mean Residuals
KNeighbors	0.978200	0.036493	0.980166	0.007465
Decision Tree	0.995550	0.020833	0.988677	0.003472
Random Forest	0.995934	0.011197	0.993914	0.002355
XGBoost	0.982792	0.037276	0.979741	0.003241
AdaBoost	0.956112	0.080916	0.956023	-0.022816
SVR	0.849935	0.268666	0.853981	-0.007298

Best Classifier Model: The Random Forest Classifier model excels across all metrics (Accuracy: 0.999349, Precision: 0.999355, Recall: 0.999349, F1 Score: 0.999350, Mean Cross Validation Score: 0.999219), demonstrating its consistent accuracy in classifying instances correctly. Its high mean cross-validation score also highlights its robustness and ability to generalize well to new data.

In contrast, the AdaBoost Classifier model has the lowest values across all metrics (Accuracy: 0.558377, Precision: 0.357617, Recall: 0.558377, F1 Score: 0.419883, Mean Cross Validation Score: 0.492881), indicating significant difficulties in accurately classifying instances. The low precision suggests a high number of false positives, and the low mean cross-validation score shows poor generalizability to new data.

AdaBoost has significantly lower values compared to other models like Logistic Regression, KNeighbors, Decision Tree, Random Forest, XGBoost, and SVC.

Table Classification Evaluation results:

	Accuracy	Precision	Recall	F1 Score	Mean Cross Val Score
KNeighbors	0.990885	0.990916	0.990885	0.990868	0.991014
Decision Tree	0.998698	0.998776	0.998698	0.998704	0.999045
Random Forest	0.999349	0.999355	0.999349	0.999350	0.999219
XGBoost	0.995877	0.995840	0.995877	0.995797	0.993532
AdaBoost	0.558377	0.357617	0.558377	0.419883	0.492881
SVC	0.977865	0.978212	0.977865	0.977846	0.979858

The two tables demonstrate the exceptional effectiveness of the Random Forest algorithm, and it can be used for better prediction impact of air pollution on global environment.

7.Conclusion

The integration of machine learning techniques into the study of atmospheric pollution marks a significant advancement in environmental science. By harnessing the capabilities of ML, researchers can process and analyze vast datasets with unprecedented accuracy and speed, unveiling patterns and correlations that were previously obscured. The insights gained from these analyses are crucial for developing effective policies and strategies to mitigate the adverse effects of pollution on the global environment. As this field continues to evolve, it is imperative to address the challenges of data quality, algorithmic transparency, and interdisciplinary collaboration. Nevertheless, the potential of machine learning to revolutionize our understanding and management of atmospheric pollution is immense, promising a future where data-driven decisions contribute to a cleaner and healthier planet.

Key findings from the implementation of machine learning in this field include improved accuracy in predicting pollutant dispersion, enhanced understanding of the correlation between pollution levels and health outcomes, and more effective identification of pollution sources. Machine learning models have facilitated the creation of real-time monitoring systems and predictive tools that enable policymakers and researchers to respond more swiftly and effectively to environmental threats.

References:

[1].van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538.

[2].Feng, Y., Zhang, W., Sun, D., Zhang, L., 2011. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. Atmos. Environ. 45, 1979–1985.

[3].Suhasini V. Kottur , Dr. S. S. Mantha,2015. “An Integrated Model Using Artificial Neural Network and Kriging For Forecasting Air Pollutants Using Meteorological Data”. International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online) :
2278-1021 ISSN (Print) : 2319-5940 Vol. 4, Issue 1, January 2015

[4].Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944.

[5].Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. https://doi.org/10. 1186/s12889-017-4914-3

[6].Rybarczyk Y, Zalakeviciute R (2017) Regression models to predict air pollution from affordable data collections. In: H. Farhadi (Ed.), Machine learning advanced techniques and emerging applications pp 15–48. IntechOpen. https://doi.org/10.5772/intechopen.71848

[7].Cetin, M., Onac, A.K., Sevik, H., Sen, B., 2018. Temporal and regional change of some air pollution parameters in Bursa. Air Qual. Atmos. Health 12, 311–316.

[8].Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.

[9] Al-Jamimi HA, Al-Azani S, Saleh TA (2018) Supervised machine learning techniques in the desulfurization of oil products for environmental protection: a review. Process Saf Environ Prot 120:57–71. https://doi.org/10.1016/j.psep.2018.08.021

[10].Sweileh WM, Al-Jabi SW, Zyoud SH, Sawalha AF (2018) Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidiscipline Respiratory Med. https://doi.org/10.1186/s40248-018-0128-5

[11].Zhu D, Cai C, Yang T, Zhou X (2018) A machine learning approach for air quality prediction: model regularization and optimization. Big Data and Cognitive Comput. https://doi.org/10.3390/bdcc2010005

[12]. Ruchi Raturi, Dr. J.R. Prasad .“Recognition of Future Air Quality Index Using Artificial Neural Network”. International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[13]. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine Learning Models”. International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

[14].Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

[15].Alade IO, Rahman MAA, Saleh TA (2019a) Predicting the specifc heat capacity of alumina/ethylene glycol nanofuids using support vector regression model optimized with Bayesian algorithm. Sol Energy 183:74–82. https://doi.org/10.1016/j.solen er.2019.02.060

[16].Alade IO, Rahman MAA, Saleh TA (2019b) Modeling and prediction of the specific heat capacity of Al2 O3/water nanofuids using hybrid genetic algorithm/support vector regression model. Nano-Struct Nano-Objects 17:103–111. https://doi.org/10. 1016/j.nanoso.2018.12.001

[17].Al-Jamimi HA, Saleh TA (2019) Transparent predictive modelling of catalytic hydrodesulfurization using an interval type-2 fuzzy logic. J Clean Prod 231:1079–1088. https://doi.org/10.1016/j. jclepro.2019.05.224

[18].Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J Comput Appl Technol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006

[19].Dalberg (2019) Air pollution and its impact on business: the silent pandemic. https://www.cleanairfund.org/wp-content/uploads/ 2021/04/01042021_Business-Cost-of-Air-Pollution_LongForm-Report.pdf

[20].Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G (2019) A machine learning model for air quality prediction for smart cities. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET). IEEE 452–457. https://doi.org/10.1109/WiSPNET45539.2019. 9032734

[21].Soundari AG, Jeslin JG, Akshaya AC (2019) Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 14(11):181–186

[23]. Mrs. A. Gnana SoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C., 2019.“Indian Air Quality Prediction And Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue)

[24].Rogers CD (2019) Pollution's impact on historical monuments pollution's impact on historical monuments. SCIENCING. https://scien cing.com/about-6372037-pollution-s-impact-historical-monuments.html

[25].Chowdhury, S., Pozzer, A., Dey, S., Klinginuellei, K., Lelieveld, J., 2020. Changing risk factors that contribute to premature mortality from ambient air pollution between 2000 and 2015. Environ. Res. Lett. 15, 074010.

[26].Kuerban, M., Waili, Y., Fan, F., Liu, Y., Qin, W., Dore, A.J., Peng, J., Xu, W., Zhang, F., 2020. Spatio-temporal patterns of air pollution in China from 2015 to 2018 and implications for health risks. Environ. Pollut. 258, 113659.

[27].Ayturan YA, Ayturan ZC, Altun HO, Kongoli C, Tuncez FD, Dursun S, Ozturk A (2020) Short-term prediction of PM2.5 pollution with deep learning methods. Global NEST J 22(1):126–131

[28].Ucun Ozel, H., Gemici, B.T., Gemici, E., Ozel, H.B., Cetin, M., Sevik, H., 2020. Application of artificial neural networks to predict the heavy metal contamination in the Bartin River. Environ. Sci. Pollut. Res. Int. 27, 42495–42512.

[29].Zhang, Y., Chen, Y., 2020. Research trends and areas of focus on the Chinese Loess Plateau: A bibliometric analysis during 1991–2018. Catena 194, 104798.

[30].Zhang, Y., Pu, S., Lv, X., Gao, Y., Ge, L., 2020b. Global trends and prospects in microplastics research: a bibliometric analysis. J. Hazard. Mater. 400, 123110.

[31].Zhang, Y., Vu, T.V., Sun, J., He, J., Shen, X., Lin, W., Zhang, X., Zhong, J., Gao, W., Wang, Y., Fu, T.M., Ma, Y., Li, W., Shi, Z., 2020c. Significant changes in chemistry of fine particles in wintertime Beijing from 2007 to 2017: Impact of clean air actions. Environ. Sci. Technol. 54, 1344–1352.

[32].Castelli M, Clemente FM, Popoviˇc A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/ 2020/8049504

[33].Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[34].Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning based prediction of air quality. Appl Sci 10(9151):1–17. https:// doi.org/10.3390/app10249151

[35].Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ ICACCCN51052.2020.9362912
[36].Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123
[37].Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J Adv Sci Technol 29(5):6934–6943
[38].Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840
[39].Patil RM, Dinde HT, Powar SK (2020) A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms 5(8):1148–1152
[40].Elsunousi, A.A.M., Sevik, H., Cetin, M., Ozel, H.B., Ozel, H.U., 2021. Periodical and regional change of particulate matter and CO2 concentration in Misurata. Environ. Monit. Assess. 193, 707.
[41].Deshpande T (2021) India Has 9 of World's 10 most-polluted cities, but few air quality monitors. India spend. https://www.india spend.com/pollution/india-has-9-of-worlds-10-most-pollutedcities-but-few-air-quality-monitors-792521

[42].Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013
[43].Fahad, S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021b) Sustainable soil and land management and climate change (1st ed.). CRC Press. https://doi.org/10.1201/9781003108894
[44].Gopalakrishnan V (2021) Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascie nce.com/hyperlocal-air-quality-prediction-using-machine-learn ing-ed3a661b9a71
[45].Gurjar BR (2021) Air pollution in India: major issues and challenges. Energy future 9(2):12–27.
[46].Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. Geophysics Res Lett. https://doi.org/10.1029/2020GL091202 5348 International Journal of Environmental Science and Technology (2023) 20:5333–5348 1 3
[47].Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538
[48].Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan V (2021) Climate change and plants: biodiversity, growth and interactions (S. Fahad, Ed.) (1st ed.). CRC Press. https://doi.org/10.1201/9781003108931
[49].Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S., Duan, C., 2021. Statistical approaches for forecasting primary air pollutants: a review. Atmosphere 12 (6), 686.
[50].Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., Li, F., 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst. Appl. 169, 114513.
[51].Qin, F., Li, J., Zhang, C., Zeng, G., Huang, D., Tan, X., Qin, D., Tan, H., 2022a. Biochar in the 21st century: a data-driven visualization of collaboration, frontier identification, and future trend. Sci. Total. Environ. 818, 151774.
[52].Qin, X., Zhou, S., Li, H., Wang, G., Chen, C., Liu, C., Wang, X., Huo, J., Lin, Y., Chen, J., Fu, Q., Duan, Y., Huang, K., Deng, C., 2022b. Enhanced natural releases of mercury in response to the reduction in anthropogenic emissions during the COVID-19 lockdown by explainable machine learning. Atmos. Chem. Phys. 22, 15851–15865.
[53].Wen, Z., Wang, R., Li, Q., Liu, J., Ma, X., Xu, W., Tang, A., Collett Jr., J.L., Li, H., Liu, X., 2022. Spatiotemporal variations of nitrogen and phosphorus deposition across China. Sci. Total. Environ. 830, 154740.

[54].Xu, W., Zhao, Y., Wen, Z., Chang, Y., Pan, Y., Sun, Y., Ma, X., Sha, Z., Li, Z., Kang, J., Liu, L., Tang, A., Wang, K., Zhang, Y., Guo, Y., Zhang, L., Sheng, L., Zhang, X., Gu, B., Song, Y., Van Damme, M., Clarisse, L., Coheur, P.F., Collett Jr., J.L., Goulding, K., Zhang, F., He, K., Liu, X., 2022a. Increasing importance of ammonia emission abatement in PM2.5 pollution control. Sci. Bull. 67, 1745–1749.
[55].Li, Y., Hong, T., Gu, Y., Li, Z., Huang, T., Lee, H.F., Heo, Y., Yim, S.H.L., 2023. Assessing the spatiotemporal characteristics, factor importance, and health impacts of air pollution in seoul by integrating machine learning into Land-use Regression modeling at high spatiotemporal resolutions. Environ. Sci. Technol. 57 (3), 1225–1236.

[56].Lv, Y., Tian, H., Luo, L., Liu, S., Bai, X., Zhao, H., Zhang, K., Lin, S., Zhao, S., Guo, Z., Xiao, Y., Yang, J., 2023. Understanding and revealing the intrinsic impacts of the COVID-19 lockdown on air quality and public health in North China using machine learning. Sci. Total. Environ. 857, 159339.

[57].Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S.B., Janes, C. R., Lanphear, B.P., McCandless, L.C., Takaro, T.K., Venners, S.A., Webster, G.M., Y. Li et al. Ecotoxicology and Environmental Safety 257 (2023) 114911 10

[58]. D P Singh, 2024a, An Extensive Examination of Machine Learning Methods for Identifying Diabetes, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055 Vol. 45 No.2,

[59]. D P Singh, 2024b, An Extensive Analysis of Machine Learning Models to Predict the Breast Cancer Recurrence, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055 Vol. 45 No. 2