International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Hybrid Machine Learning Models for Fine-Grained Air Quality Forecasting

Raj Vikram Singh rajvikram[1], Abhay Pal Singh[2], Sandeep Kumar[3]

[1,2,3]Department of CSE School of Engineering and Technology, Sharda University, Greater Noida, India

ABSTRACT

Air pollution has become a major environmental issue that is causing many deaths each year and putting the environment and human health at serious risk. It causes the greenhouse effect, contributes to global warming, and increases the risk of lung cancer and other diseases that affect the respiratory system, including allergies. Setting and upholding strict air quality standards is essential to effectively combating air pollution. The air quality index (AQI) is a measurement used to determine the amount of pollutants in the atmosphere. By utilizing the capabilities of machine learning algorithms, precise forecasting of the fine-grained AQI is made feasible. To predict the AQI, a number of algorithms have been used, including logistic regression, decision tree regression, KNN, SVR, and linear regression. This project's main goal is to create models with machine learning algorithms and determine which model is best for AQI prediction.

Keywords: Air quality, Prediction, Algorithms, Random Forest, Linear Regression

1. INTRODUCTION

Since oxygen is necessary for all bodily cells to function, it is the most significant aspect of life. Air is an essential component on our planet, holding immense significance. While humans can survive without water for several days, our existence is limited to a mere three minutes without air. By circulating warm and humid air, it maintains a stable temperature on the Earth's surface. Furthermore, air plays a crucial role in influencing the water cycle. Our current breathing not only sustains our life but also plays a crucial role in how satisfied we can be with our lives. Low air quality has the potential to significantly impact fitness.

Among other respiratory conditions, contaminated air can lead to lung cancer, bronchitis, pneumonia, TB, and asthma. Approximately 7 million individuals worldwide are tragically lost each year due to air pollution, as per estimations. Furthermore, apart from its adverse effects on temperatures and sea levels, air pollution can also play a role in exacerbating global warming, which occurs when heat becomes trapped within the atmosphere. This can lead to infectious disease transmissions as well as temperature increases and increasing sea levels.

One can quantify the air's quality. The numerical indicator of the quality of the air is the air quality index, or AQI. It is a numerical value between 0 and 500 that represents the level of air pollution and the difficulty of staying fit. An AQI of fifty or lower, for instance, is regarded as precise air excellent, while an AQI of three hundred or more is classified as hazardous first-class. Based on potential health risks, the AQI is split into six categories. For ease of interpretation, every category has a different color. Green (zero-50), Yellow (51-100), Orange (100 and one-hundred fifty), and crimson are the six classes.

Since the AQI calculates vital information about the condition of the air, it is essential. When the AQI surpasses the maximum cost, we will take precautions to ensure that people are not harmed by pollutants. The AQI serves as warning system for the public about the quality of the air they breathe, educating them about which populations are most vulnerable to air pollution and how to minimize their exposure to it. The effects that breathing contaminated air can have on a person's health over a few hours to several days are the main focus of the AQI.

Air quality plays a crucial role in our surroundings, significantly affecting our overall health. When the air is polluted and contains high levels of harmful substances and particles, it can lead to various health issues including respiratory disorders, heart problems, and even premature death. To assess and communicate the condition of the air in a specific location, the AQI serves as a standardized measure. With the help of this indicator, people, organizations, and governments can reduce the negative effects of poor air quality by making educated decisions.

The significance of the AQI is emphasized by a number of crucial elements. Primarily, AQI plays a fundamental role in safeguarding public health. It acts as a crucial instrument in comprehending and conveying the potential health hazards linked to the air quality in a specific region. Furthermore, inadequate air quality can result in adverse impacts on the environment, such as the degradation of ecosystems, harm to vegetation, and the acidification of water bodies. Therefore, monitoring and forecasting AQI is imperative in order to protect the environment and preserve biodiversity.

High exposure to airborne pollutants and particulate matter can cause a variety of health issues, such as respiratory and cardiovascular problems, as well as premature death. To assess and report on the air quality in a specific area, AQI is used as a standard measurement. Accurate predictions of AQI are crucial for complying with the set standards that often specify limits on the concentration of specific pollutants. Hence, AQI prediction is not only essential for individual health but also for regulatory compliance.

Despite efforts to predict AQI, there are several obstacles that make it difficult. The inconsistency of air quality data poses as one of the main obstacles, which can fluctuate rapidly due to various factors such as weather conditions, industrial emissions, vehicular traffic, and natural disasters like wildfires. As a result, it is crucial to gather and analyse real-time data to ensure accurate predictions. Additionally, the complex and nonlinear interactions between air quality, weather, and pollutant concentrations pose another significant challenge that machine learning models must consider to produce dependable AQI forecasts.

Data quality poses a significant challenge within this domain. The reliability and accessibility of historical air quality and meteorological data can greatly differ across various locations. Inaccurate or insufficient data has the potential to impede the effectiveness of machine learning models. Additionally, proficient feature engineering plays a vital role. The careful selection of appropriate features (predictors) to incorporate into the model is of utmost importance. The air quality can be influenced by various factors, including temperature, humidity, wind speed, and geographical attributes. Hence, meticulous feature engineering is imperative to ensure model accuracy.

The quality of air can be influenced by various variables, including temperature, humidity, wind direction, and topography. Machine learning models require access to historical air quality data, meteorological data, and other relevant information in order to forecast the Air Quality Index (AQI). The data sources encompass government monitoring stations, satellite data, as well as crowd-sourced data obtained from sensors and mobile applications. Data preprocessing is the next critical step. Before feeding data into machine learning models, it must be pre-processed. This includes handling missing data, normalizing data,

and dealing with outliers. Preprocessing makes sure the data is ready for analysis and training in the right format.

Feature engineering follows data preprocessing. It entails choosing and developing significant features that can enhance the predictive capabilities of the model. Features might include past air quality readings, meteorological conditions, geographical features, and temporal patterns.

Model selection is another important decision. Numerous algorithms for AQI prediction are available thanks to machine learning. There are several models that are commonly utilized, such as random forests, decision trees, neural networks, support vector machines, and time series forecasting techniques like ARIMA and LSTM.

The subsequent stage involves training and validating the model. The selected model needs to be trained on previous data and then validated to assess its efficacy. The training data is used to improve the model, and the validation data is used to gauge how well the model performs on fresh data. Cross-validation techniques are often employed to prevent overfitting.

Hyperparameter tuning is crucial for optimizing model performance. It is crucial to adjust the machine learning model's hyperparameters. The Optimal hyperparameter configurations can be achieved using methods like grid search or random search. The last stage involves evaluating the model's performance using different metrics such as root mean square error (RMSE), mean absolute error (MAE), and correlation coefficients. An essential component in enhancing the model's resilience is cross-validation.

The aim of the AQI prediction models is to furnish prompt assessments of the air quality. These models rely on present and historical data, as well as weather predictions, to deliver the latest information. They play a vital role in enabling people to make swift choices, such as whether to venture outdoors or adopt preventive measures.

Prediction models, conversely, anticipate forthcoming AQI levels, frequently on a daily or hourly basis. These models utilize past data, meteorological predictions, and patterns to offer early warning of air quality circumstances. This data is significant for city planning, public health recommendations, and ecological administration.

Despite the considerable advancements achieved in the realm of AQI forecasting, there remain obstacles and constraints that necessitate attention. One major obstacle is the lack of high-quality data, which directly impacts the accuracy of AQI predictions. In certain areas, data may be scarce or of inferior quality, posing additional difficulties in making precise predictions.

The challenge of model generalization arises when machine learning models trained in one region fail to perform well in another region due to differences in local factors and pollutant sources. To overcome this challenge, localized models or transfer learning techniques can be employed.

Scalability is an important consideration as well. Developing machine learning models for AQI prediction necessitates a significant investment in computational resources for training and deployment. Expanding the scope to cover larger regions or multiple monitoring stations can present certain difficulties.

Moreover, the issue of model interpretability is gaining more attention. Deciphering machine learning models can be intricate, especially when it comes to deep neural networks. It is crucial to comprehend the mechanisms behind a model's specific predictions in order to establish trust and make well-informed choices.

AQI predictions have a broad and significant impact, despite the challenges and limitations they face. They play a crucial role in helping individuals make informed choices regarding outdoor activities and

health precautions. Additionally, these predictions empower health agencies to issue timely advisories, ensuring the protection of vulnerable populations.

AQI predictions can be utilized by cities in the field of urban planning to make informed decisions regarding the placement of schools, parks, and industrial zones.

Environmental management is another crucial application. Accurate AQI predictions help in identifying pollution sources and managing emissions effectively. This, in turn, aids in protecting ecosystems and preserving biodiversity.

Moreover, AQI predictions play a pivotal role in regulatory compliance

2. LITERATURE REVIEW

AUTHOR	ABSTRACT	RESEARCH GAP
R [1]	1. Unsupervised neural network techniques are employed alongside supervised algorithms like Decision Trees, Support Vector Machines, and K-Nearest Neighbor. When compared to these methods, neural networks demonstrate superior performance.	Anticipate the level of air pollution hourly.
R [2]	1. Fuzzy logic was utilized in the Study to predict PM2.5 and PM10 levels. 2. Outliers, which are undesired gases existing in the environment, are removed with the use of fuzzy logic.	Clusters are formed using fuzzy logic, which might result in predictions that are erroneous since they contain repetitive data.
R [3]	1. By utilizing recurrent neural networks, the issue of algorithm memorization power hindering hourly forecasts has been resolved in estimating the air pollution level at any given time.	Lacking in terms of operating without memory.
R [4]	1. They took into account neural networks with embedded sensors that provide precise air pollution levels.	It takes a long time to train and is unable to handle incomplete data sets, which makes it slow..
R [5]	1. The air quality index can be predicted with XG Boost, a method that minimizes the discrepancy between the actual value and the prediction values by constructing a robust classifier that leverages the shortcomings of previous weak classifiers.	One disadvantage is that it is dependent upon the preceding value and susceptible to the effects of outliers found in the atmosphere and unwanted pollutants.
R [6]	1. The author employs decision trees and multinomial logistic regression to predict the quantity of air quality pollutants. This method involves analyzing features and making predictions. 2. The actual values exhibit a significantly higher level of proximity to the predicted values in multinomial	Falling short of working with the entire dataset..

	logistic regression, thereby establishing its superiority in generating more precise outcomes.	
R [7]	1. The merging of the Rasberry Pi platform with machine learning has enabled accurate prediction of air pollution levels. 2. Multilayer perceptrons present a difficulty in both classifying discrete values and performing regression on continuous data.	The author failed to transmit the input to the activation function because they used a multilayer perceptron with backpropagation and discrete values. As a result, the output could only be limited to 0 or 1, signifying the degree of difference between the actual and predicted values. Even though the coefficient of determination (R2) showed improvement, incremental feeding may still be able to further improve the situation.
R [8]	1. Light trees and light gradient boosting are merged in their model. The suggested approach entails utilizing a sensor attached to a Raspberry PI to measure the PM2.5 level, storing the information gathered in the cloud, and predicting the level using a hybrid model. 2. The hybrid model has demonstrated superior performance in PM2.5 detection compared to Neural Network, Random Forest, Decision Tree, Lasso, Support Vector, Linear, and XGBoost regressions.	Requires more time even though images can be used to estimate PM2.5 Level more precisely.
R [9]	1. ELMs (Extreme Learning Machines) have been suggested as a reliable method for predicting air quality. 2. Sigmoidal function produces precision that is superior than sine and hard-limit activation. With the aid of 10 cross validations and additional hidden neurons, the ELM is utilized to compute the air pollution level.	Trouble handling large datasets.
R [10]	1. Considering the temperature, humidity, wind direction, and wind speed highlights the ozone layer. For this, a number of machine learning algorithms are used, including MLP, XG Boost, SVR, and DTR.	Lacking in their ability to use a few neural network algorithms.
R [11]	1. The study demonstrates the effectiveness of combining a multivariate data imputation method with	Continual improvement of forecasting accuracy and

	machine learning-driven forecasting to improve the precision of AQI forecasts. 2. The proposed approach improves the precision of predictions by addressing the absence of data and taking into account the interconnections among variables.	accessibility remains crucial in the ongoing battle to combat air pollution.
R [12]	1. This study's primary goal is to focus on the spatial forecast of AQI and present a novel hybrid model that combines boosted extreme learning machine, three-stage feature selection, and decomposition. 2. Through a three-step process, the model progressively extracts the spatiotemporal features. This procedure involves mutuality, binary grey wolf optimization (BGWO), and spatial correlation analysis. Furthermore, decomposition and ensemble techniques are utilized to enhance the accuracy and robustness of the model.	The proposed model can do more than just select the best features of air pollutants and monitoring stations.
R [13]	1. The research entails a preliminary randomized clinical trial in which an intervention is being examined to evaluate the influence of air quality (as indicated by the AQI) on childhood asthma. The trial is probably designed to evaluate whether enhancements in air quality yield beneficial outcomes on asthma symptoms and overall respiratory well-being in children.	Based on the results of the pilot trial, you could identify potential avenues for further research, such as larger-scale clinical trials or more comprehensive studies on the mechanisms linking air quality to asthma outcomes.
R [14]	1. This study set out to quantify the impact of heating emissions on air quality in a way that was both accurate and efficient. 2. In addition, the study sought to investigate the effects of heating emissions in 64 northern Chinese cities with different pollution levels.	Developing a new method, validating it, applying it to real-world scenarios, and deriving meaningful insights that can contribute to environmental management and policy decisions.
R [15]	1. An investigation was conducted to analyze the AQI, PM2.5 and NO2 levels in the troposphere 21 cities worldwide during three distinct lockdown phases: before, during, and after. The primary objective was to employ a straightforward before/after comparison method to observe the reduction in air pollution levels	Measurement of AQI using more data of before and after the lockdown.

	resulting from the implementation of lockdown measures. 2. The findings indicated that the variability of the frequency distribution for NO2 surpasses that of PM2.5, and the distribution appears to be less steep from 2020 compared to the baseline period of 2018-2019.	
R [16]	1. Data from the Environmental Pollution Agency (EPA) and the Centers for Disease Control (CDC) were used to analyze the relationship between various environmental pollutants and the spread of COVID-19 in the state. 2. The analysis of the pollutants included PM10, SO2, NO2, Pb, CO, VOC, and PM2.5. The study used statistical tests like the Spearman and Kendall correlation tests to determine the strength and direction of the relationship between these pollutants and COVID-19 cases. 3. The findings of the study suggest that there is a significant correlation between certain environmental pollutants, including COVID-19, PM10, PM2.5, SO2, NO2, and CO pandemic in California. This suggests that a higher number of COVID-19 cases is associated with elevated concentrations of these pollutants.	Falling short of working with the data after Covid-19
R [17]	1. Using both ground-based and satellite observations, the study looked at tropospheric NO2 levels, AQI, PM2.5, and other air quality indicators in several Indian cities. 2. Prominent cities like Delhi, Mumbai, Hyderabad, Kolkata, and Chennai saw a significant drop in PM2.5 and AQI levels during the lockdown period, according to the study. This suggests that fewer human activities and pollution sources contributed to the improvement in air quality.	They couldn't research in towns and rural areas during lockdown.
R [18]	1. The Central Pollution Control Board, an agency of the Indian government tasked with policing pollution, provided the data used in the study. The data includes the annual increase in three major pollutants: particulate matter with a diameter of 2.5 micrometers or less (PM 2.5), sulfur dioxide (SO2), and nitrogen oxides (NOx). 2. The study focuses on three major cities in India: Delhi, Bengaluru, and Chennai. These cities represent	Failing short of dataset as it covers only some of the cities of india

	different regions and urban environments, allowing for a diverse analysis of air pollution trends. 3. The study has identified multiple origins of contaminants that are contributing to the issue of air pollution. These origins comprise of diverse transportation methods (both small and large vehicles), power production from diesel, coal, and gas plants, industrial operations, construction, and household cooking.	
R [19]	1. The study indicates that a relationship exists between the Air Quality Index (AQI) and the propensity of individuals to migrate. According to the study, the number of searches pertaining to "emigration" is predicted to increase by 2.3% to 4.8% the following day if the AQI increases by 100 points on a given day. When the AQI level is above 200, which denotes days with "heavily polluted" and "severely polluted" air quality, the impact becomes more apparent. 2. The study found that the influence of air pollution on human health willingness to emigrate varied depending on the countries they were interested in relocating to and the particular urban areas they resided in. In simpler terms, the effect of air pollution on migration sentiment differed based on the destination countries and the geographical circumstances of the individuals' current location.	The research establishes a connection between air pollution and the inclination towards emigration, however, it does not explicitly establish the precise direction of this association or a cause-and-effect relationship.
R [20]	1. The inquiry has determined that there is no definitive response to this inquiry because of various factors, such as the absence of quantitative performance specifications from sensor manufacturers, a consensus on suggested end-use and performance objectives, and the capacity of users to establish requirements for their applications. 2. The analysis appears to have examined 17 significant projects that successfully reached the stage of practical implementation. It noted a transition in the monitoring of air quality, shifting from predominantly government-led efforts to initiatives driven by commercial entities and crowd-funding. This shift indicates a transformation in the paradigms of monitoring.	Sustainability, Long-Term Reliability
R [21]	1. This article presents an innovative deep learning model that utilizes LSTM networks to forecast	Failing short of data

	forthcoming air quality levels in the framework of a smart city. This indicates development of a unique methodology to tackle this particular prediction challenge.	
R [22]	1. The majority of reviewed studies rely heavily on meteorological data and source emissions data as predictors for their ANN models.	The paper acknowledges the opaque nature of ANNs and the difficulties it presents in understanding the rationale behind their predictions. Future research efforts could focus on developing methods to improve the interpretability of ANNs, allowing stakeholders to understand the key factors influencing predictions and increasing the transparency of the models.
R [23]	1. The performance of air pollution forecasting was evaluated by conducting experiments using SVR, GBTR, and LSTM machine learning methods. 2. In order to obtain the final result, three LSTMs were combined as Aggregated LSTM (ALSTM). The evaluation of performance was done using MAE, RMSE, and MAPE metrics.	PM 2.5 values are not accurate
R [24]	1. The experimental findings in the paper are derived from data collected from nine cities in China over a span of three years. DeepAir surpasses ten other methods used as a basis for comparison, demonstrating significant enhancements in accuracy. Specifically, it achieves a 2.4% improvement in short-term prediction, a 12.2% improvement in long-term prediction, and an impressive 63.2% improvement in predicting sudden changes compared to a previous online approach utilized in the AirPollutionPrediction system.	The exploration of how well the proposed DeepAir approach generalizes to cities in other developing countries with different urban layouts, sources of pollution, and meteorological patterns.
R [25]	1. The paper's principal goal is to address the difficulty in forecasting air pollutant levels, a crucial aspect in ensuring public health protection. The current approaches have faced criticism due to their limited	The article centers on 12 stations that monitor air quality. It would be beneficial to explore the

	capability in accurately capturing long-term relationships and neglecting spatial associations. 2. To improve the predictive accuracy, the model incorporates diverse supplementary information such as meteorological data and time stamp data. This integration is expected to enable the model to consider external factors that impact air pollution levels.	model's ability to accommodate a greater number of monitoring stations in a wider geographic region.
R [26]	1. 22 Indian cities saw improvements in their air quality as a result of COVID-19, with a notable drop in PM2.5 levels. 2. The correlation between different cities improved in 2020, especially in the north and east. This decrease in concentrations led to a four-fold decrease in total ER. Even if PM2.5 levels rise due to unfavorable weather conditions, they would still be within CPCB limits.	Shortage of dataset due to Covid-19 restrictions
R [27]	1. The concentrations of PM10 and PM2.5 dropped by roughly 50% prior to the lockdown. CO and NO2 concentrations dropped as well during the lockdown. 2. Air quality in transportation and industrial areas improved by almost 60%. The greatest improvement was seen in central and Eastern Delhi, with a 40% to 50% increase in the quality of the air on the lockdown's second and fourth days.	Sustainability and Long-Term Reliability
R [28]	1. The importance of air pollution and its detrimental effects on the environment and human health are emphasized in the opening section. It presents Particulate Matter (PM2.5), a class of air pollution consisting of particles smaller than 2.5 micrometers. Given its size and ease of entry into the respiratory system, this is concerning.	It doesn't go into detail about how the model's predictions can be interpreted.
R [29]	1. The model captures temporal patterns and variations in PM2.5 contamination by means of a Long Short-Term Memory (LSTM) network. Because LSTM excels at modeling sequential data, it can be used to monitor changes in air quality over time.	The model takes as inputs the day of the week, meteorological data, historical air quality data, and weather forecast data.
R [30]	1. The paper highlights that a significant portion (41%) of the publications focused on predicting air pollution for the following day, indicating the practical application of these predictions for short-term planning.	Long-Term Prediction and Trend Analysis
R [31]	1. The study aimed to create accurate air quality forecasting models for Beijing by incorporating spatiotemporal distribution characteristics.	Failing short of data

R [32]	1. The article discusses the issue of air pollution in smart cities and emphasizes the significance of obtaining real-time pollution data to make informed decisions. 2. The study utilizes four sophisticated regression techniques, using support vector regression, random forests, decision trees, and linear regression to forecast air quality. Evaluating their suitability and efficacy in the context of smart cities is the goal.	Couldn't calculate linear and logistic regression
R [33]	1. The paper proposes a workable solution: the Transferred Bi-directional Long Short-Term Memory (TL-BLSTM) model. This model uses a bi-directional LSTM architecture and is based on deep learning. LSTM networks are a great option for predicting air quality because they are specifically made to capture long-term dependencies in time series data. The model's "transferred" element describes how knowledge acquired at lower temporal resolutions can be applied at higher ones through the use of transfer learning.	It is rely on an old dataset
R [34]	1. The storage and merging of data that measures airborne pollutants, public health, and environmental factors is on the rise. These large datasets hold immense potential, but they also pose a challenge to conventional epidemiological methods. 2. As a result, there is a growing interest in exploring alternative approaches to predict, identify patterns and gather data. The field of air pollution analysis is increasingly using machine learning and data mining algorithms to accomplish this.	Some Machine Learning Algorithms were incompatible with their work, which limited its effectiveness.
R [35]	1. The authors propose a new method to improve prediction accuracy by incorporating past meteorological data. 2. They also introduce a specific regularization method to ensure similarity between prediction models for consecutive hours. 3. This technique is compared to other standard regularization methods for multi-task learning.	Real-World Implementation and Deployment
R [36]	1. The article introduces a novel hybrid deep learning technique that merges 1D-CNNs and Bi-LSTM networks to address the complex problem of predicting PM2.5 air pollution levels. 2. The approach capitalizes on the strengths of each architecture to capture various aspects of the data's	Lacking in terms of operating without memory.

	characteristics, ultimately leading to accurate predictions.	
R [37]	1. The proposed paper presents a predictive model that employs a blend of neural networks to anticipate air quality for a period of up to 48 hours. 2. An artificial neural network (ANN), a convolution neural network (CNN), and a long short-term memory (LSTM) network are the neural networks used in this model.	Data imbalance, data generalization
R [38]	1. The proposed study introduces An LSTM model with deep multi-output (DM-LSTM) as a solution to these challenges. The primary objective of this model is to capture intricate spatio-temporal connections and minimize the accumulation and propagation of errors in forecasting air quality for multiple time steps ahead.	Feature engineering, selection, interpretability
R [39]	1. The observation that machine learning applications are mainly concentrated in the Eurasian and North American continents suggests potential disparities in research focus and resources allocation across different regions.	Couldn't include deep learning and hybrid models for testing and research
R [40]	1. This study uses the Beijing PM2.5 dataset from the UCI Machine Learning Repository to illustrate the efficacy of the CBGRU model. 2. According to the research, the CBGRU model surpasses conventional forecasting techniques by producing lower prediction errors and exhibiting superior overall prediction performance.	Model generalization, data variety, long term forecasting
R [41]	1. The research findings indicate that pollution levels and AQI values in Chinese urban areas have been consistently declining on a yearly basis, exhibiting a distinctive "U"-shaped pattern with monthly fluctuations. More specifically, pollutant levels tend to be higher during the winter season and decrease during spring, whereas they are lower during summer and rise again in the fall. However, this trend is reversed when it comes to ozone (O3), as its levels are higher during the summer months.	It takes a long time to train and is unable to handle incomplete data sets, which makes it slow..
R [42]	1. This model is intended to improve the overall cleanliness and appeal of shared vehicles, thus promoting their usage.	They couldn't work with big dataset

R [43]	1. Air pollution, resulting from various human activities, poses serious health risks, including diseases like cancer.	Data quantity, quality
R [44]	1. The study introduces a hybrid model for AQI forecasting that integrates multiple techniques.	Real time implementation and long term predictability
R [45]	1. Because of industrialization and human activity, air pollution is a global concern. The objective of this project is to compare the predictions of the Air Quality Index (AQI) and its factors using two machine learning algorithms: Random Forest Regression and SVM. 2. The Open Government Data Platform in India provides the data, which includes readings from various locations and contaminants like CO, NO2, PM10, PM2.5, and SO2.	Data quality and completeness, feature selection and engineering.
R [46]	1. The outcome of this research could provide insights into which machine learning method is most effective for predicting AQI in Ahmedabad. 2. This information could be used for better traffic routing by avoiding areas with high predicted pollution levels. 3. It could also help in identifying major pollutants and devising strategies to mitigate their impact on air quality.	Model uncertainty and real time predictions
R [47]	1. The objective of this research is to utilize the Random Forest machine learning technique to analyze time series data of particulate matter pollution in Ras Garib city, Egypt. 2. The study's ultimate objective is to create a susceptibility map for particulate matter pollution by forecasting and understanding air pollution in the city. The city's air pollution can be controlled and decreased with the help of this map.	Inclusion of additional variable, lack of dataset
R [48]	1. The research addresses the issue of data imbalance using a resampling technique. This involves adjusting the class distribution of the data to ensure accurate model training.	Short dataset in terms of years and quantity
R [49]	1. The investigation uses deep learning to predict air quality by analyzing a large dataset of environmental information. The research proposes a classification system for air quality using multi-output and LSTM neural networks.	Lacks in data learning models

	2. The study incorporates CHGS to optimize the deep learning model's parameters and improve accuracy. The deep learning approach contributes to sustainable urban development planning by implementing strategies to achieve target air quality values and reduce pollution.	
R [50]	1. The study emphasizes that machine learning has transformed the way prediction problems, particularly related to air pollution, are tackled. It's noted that machine learning offers advantages over conventional methods, which often rely on complex mathematical and statistical approaches that may be less accurate. 2. The field of air pollution epidemiology, which involves analyzing the effects of pollutants on both human health and the environment, seems to be the paper's main focus. Within this field, the paper explores machine learning as a possible replacement for traditional methods.	Lacks in accuracy, robustness, and generalization to different geographical areas and pollution patterns.

III METHODOLOGY

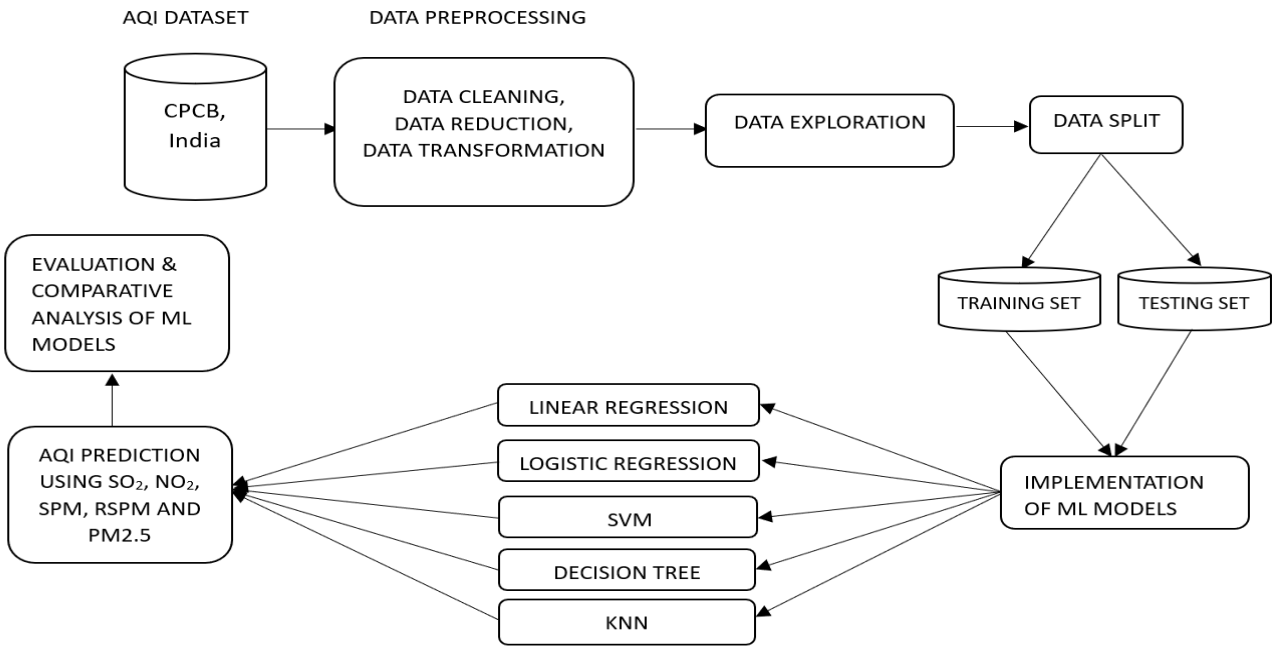

Figure 1. Flowchart of Air Quality Prediction Model

A. DATA

The official website of the Indian government, known as (CPCB) Central Pollution Control Board, has made available a comprehensive dataset on air quality and AQI for different towns in India. This dataset is quite large, consisting of 435,743 rows and thirteen columns, and has a file size of 61 MB. It includes

various variables like CO, PM2.5, ammonia, NO2, PM10, SO2, NO, and nitric x-oxide, among other substances.

B. DATA PREPROCESSING

First off, there are missing, inconsistent, and noisy data in the statistics set. Preprocessing of data is essential to eliminate irrelevant information and extract valuable insights. The application of statistical techniques enables the transformation of the data into a format that is appropriate for analysis and utilization. Statistics preprocessing has been concerned with the subsequent steps.: Data Cleaning, Data Reduction and Data Transformation.

Data Cleaning: It refers to the procedure for eliminating undesired information from a set of records, including erroneous data, redundant statistics, and unformatted facts. By eliminating these inaccuracies, we can enhance the accuracy of the results. The following actions were done even as the records were being cleaned:

- **Remove Duplicates:** The likelihood of duplicate entries in statistics is high when data is gathered from various sources. The results will be unclear if there are duplicates. It would be better if we eliminated those duplicates.
- **Remove Irrelevant Data:** Performing evaluation on irrelevant statistics hampers the efficiency of the process as it does not yield any benefits. To illustrate, if our analysis solely requires attention to particulate matter, we must exclude various additives that are highly unrelated to our investigation in order to optimize time utilization.
- **Handling Missing Values:** The management of absent records can be approached by either deleting the complete collection of data or by adding the necessary values to the gaps. In case of absentee subjects, an estimated cost can be determined. If the missing values significantly impact the overall statistics, the tuple containing those values will be excluded.
- **Clear Formatting**: Data that has undergone extensive formatting cannot be processed by machine learning models, making it challenging to work with if our information is structured in a particular way.
- **Convert Data Types:** Textual representations of numbers are sometimes mistakenly inputted. Consequently, the numerical data takes the form of strings. Since they are in string format, it becomes impossible to carry out any mathematical calculations on them. Therefore, it becomes necessary to convert the string representations of the statistical figures in order to perform mathematical operations on them.

Data Transformation: The process of enhancing the visual appeal of data enables better decision-making based on data. This involves modifying the values, structure, or presentation of information. Techniques such as normalization, feature selection, discretization, and concept hierarchy generation are all integral to this process.

Data Reduction: Analysis gets harder when one is working with a lot of statistics. We use an information reduction technique to address this. Data reduction aims to reduce the costs associated with record evaluation and storage while boosting garage efficiency. In order to discount records, a number of steps were followed, including records cube aggregation, dimensionality reduction, numerosity discount, and attribute subset selection.

C. TRAINING ANDF TESTING THE MODEL

1. LINEAR REGRESSION

In machine learning, one method for determining the correlation between a dependent and independent variable is called linear regression. The independent variable, also known as the predictor variable, is utilized to forecast the dependent variable, which is referred to as the response variable. The most effective form of the regression function is linear regression, which represents a linear equation involving variables. The linear form of the regression characteristic simplifies the translation of parameters.

Hypothesis function for linear regression:

$$y = \theta_1 + \theta_2 x \qquad \ldots(i)$$

Given:

x: Input Training Data (Univariate – One Input Variable (Parameter))

y: Labels To Data (Supervised Learning)

θ1: Intercept

θ2: Coefficient Of x

Cost Function (J): In order to minimize the gap between the actual prediction and the error, the model seeks to estimate the y-values through a smooth linear regression. Therefore, it's crucial to adjust 1 and 2 to determine the optimal value that reduces the discrepancy between the expected y-value and the actual y-cost (y) (pred).

$$J = \frac{1}{N}\sum_{i=1} (pred_i - y_i)^2 \qquad \ldots(ii)$$

- The cost function (J) of linear regression is the root mean squared error (RMSE) between the true value (y) and the predicted value (pred).

2. LOGISTIC REGRESSION

One common type of regression technique is logistic regression. Regression techniques use it to make precise predictions. reduction trends. The average shrinkage of the parent material toward the center is called shrinkage. The lasso method works with sparse, basic patterns (like lesser models).

Calculate residual sum of squares:-

$$RSS = \sum^{n}_{i=1} (y_i - f(x_i))^2 \qquad \ldots(iii)$$

Where:

y_i = The i^{th} Value Of The Variable To Be Predicted

$f(x_i)$ = Predicted Value Of y_i

n = Upper Limit Of Summation

- Executes L1 Regularization, meaning it assigns a penalty equal to the absolute cost of the coefficients' importance.

  RSS + α * (Total Of The Absolute Value Of Coefficients) = Minimization Objective …(iv)
- Similar to the ridge, α (alpha) in this instance offers a trade-off between balancing RSS and coefficient importance. Like the ridge, there is a range of values for α. Let's quickly restate it here:
  1. The coefficients in α = 0 are the same as in simple linear regression.
  2. α = ∞: Every coefficient is zero (no logic change)

3. $0 < \alpha < \infty$: Coefficients Of Simple Linear Regression Between 0 And That

3. DECISION TREE REGRESSOR

The Decision Tree Regressor utilizes recursive binary splitting as a technique to produce predictions. This algorithm starts at the root node of the decision tree and works its way down to a leaf node. Next, to ascertain the target value for a particular instance, the value of the leaf node is utilized as the prediction. Typically, the mean of the target values within the leaf node serves as the prediction for a Decision Tree Regressor.

The prediction made by a Decision Tree Regressor at a leaf node is determined by calculating the training data's mean of the target values samples that have reached that specific leaf node during the tree-building process. Mathematically, you can represent it as follows:

$$\hat{y}(x) = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \text{... (iv)}$$

Where,

- $y^{(x)}$ = predicted value for a new instance x.
- N = number of training data samples that have reached the specific leaf node.
- y_i = target value of each of those training data samples.

The Decision Tree Regressor essentially averages the target values of the data points that end up in the same leaf node, and this average becomes the predicted value for any new data point that follows the same path in the tree.

4. K-NEAREST NEIGHBOURS

It is an algorithm for supervised machine learning that is simple and efficient for solving classification and regression problems. It utilizes the similarity between a data point and its K nearest neighbors in the training dataset to make predictions. The formula for predicting a target value in the context of regression using KNN is as follows:

$$\hat{y}(x) = \frac{1}{K}\sum_{i=1}^{K} y_i \qquad \text{... (v)}$$

Where,

5. $y^{(x)}$ = predicted value for a new instance x.
6. K = number of nearest neighbors to consider.
7. y_i = target value of the i^{th} of the nearest neighbors datapoint.

5. RANDOM FOREST CLASSIFIER

A versatile machine learning ensemble algorithm that combines the predictions of several decision trees is called a Random Forest. It's widely used for classification and regression tasks. The process involves bootstrapping the training data (randomly selecting subsets with replacement), constructing individual decision trees, and aggregating their predictions. For classification, it tallies the votes from each tree to determine the final class prediction, while for regression, it computes the average of the individual tree predictions. By introducing randomness through feature selection and sampling, Random Forest mitigates overfitting, improves prediction accuracy, and provides a reliable tool for various machine learning applications.

D. PERFORMANCE METRICS

Machine learning models' accuracy and efficacy are primarily assessed using performance metrics. These metrics help measure the error or deviation between model predictions and actual results. Some of the performance measures used in AQI measurement are:

Mean Absolute Error (MAE): The MAE is obtained by dividing the mathematical difference between the actual and predicted values. Another way to define it is by counting the number of errors in paired observations that represent the same phenomenon. This metric measures the extent to which the predicted outcome differs from the actual outcome. Below is a mathematical representation of MAE.:

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |y_i - x_i| \quad \ldots(vi)$$

where,

yi = Prediction

xi = True Value

N = Total Number Of Data Points

R-squared (R^2) :- This indicator calculates the ratio of variables (air quality) that can be explained by individual variables (weather data, emissions data, and so on). A higher R2 indicates greater precision..

$$R^2 = 1 - \frac{\text{Sum Squared Regression(SSR)}}{\text{Total Sum Of Squares(SST)}}$$

$$R^2 = 1 - \frac{\sum(y_i - y_i)^2}{\sum(y_i - y)^2} \quad \ldots(vii)$$

Root Mean Square Error (RMSE): The statistical metric known as root mean square error, or RMSE, calculates the average of the squared difference between the actual value and the predicted value produced by the model. It is essentially the mean square error (MSE) squared. Its implementation methodology may bear similarities to MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_i - x_i)^2} \quad \ldots(viii)$$

Where,

x_i = True Value

N = Total Number Of Data Points

Performance metrics are evaluated in order to verify the machine learning models. The system learning version performs better when the r-squared, RMSE, and MAE are reduced.

E. IMPLEMENTATION

This section includes implementation codes. To implement In order to predict the air quality index (AQI) using machine learning, we have considered variables like PM2.5, PM5, PM10, SO, NO, CO, ammonia and NO2.

Function to calculate so2 individual pollutant index(si)

```
def cal_SOi(so2):
    si=0
    if (so2<=40):
     si= so2*(50/40)
    elif (so2>40 and so2<=80):
     si= 50+(so2-40)*(50/40)
    elif (so2>80 and so2<=380):
     si= 100+(so2-80)*(100/300)
    elif (so2>380 and so2<=800):
     si= 200+(so2-380)*(100/420)
    elif (so2>800 and so2<=1600):
     si= 300+(so2-800)*(100/800)
    elif (so2>1600):
     si= 400+(so2-1600)*(100/800)
    return si
df['SOi']=df['so2'].apply(cal_SOi)
data= df[['so2','SOi']]
data.head()
# calculating the individual pollutant index for so2(sulphur dioxide)
```

Figure 2. Calculating SO2

Function to calculate no2 individual pollutant index(ni)

```
def cal_Noi(no2):
    ni=0
    if(no2<=40):
     ni= no2*50/40
    elif(no2>40 and no2<=80):
     ni= 50+(no2-40)*(50/40)
    elif(no2>80 and no2<=180):
     ni= 100+(no2-80)*(100/100)
    elif(no2>180 and no2<=280):
     ni= 200+(no2-180)*(100/100)
    elif(no2>280 and no2<=400):
     ni= 300+(no2-280)*(100/120)
    else:
     ni= 400+(no2-400)*(100/120)
    return ni
df['Noi']=df['no2'].apply(cal_Noi)
data= df[['no2','Noi']]
data.head()
# calculating the individual pollutant index for no2(nitrogen dioxide)
```

Figure 3. Calculating NO2

Function to calculate rspm individual pollutant index(rpi)

```
def cal_RSPMI(rspm):
    rpi=0
    if(rpi<=30):
     rpi=rpi*50/30
    elif(rpi>30 and rpi<=60):
     rpi=50+(rpi-30)*50/30
    elif(rpi>60 and rpi<=90):
     rpi=100+(rpi-60)*100/30
    elif(rpi>90 and rpi<=120):
     rpi=200+(rpi-90)*100/30
    elif(rpi>120 and rpi<=250):
     rpi=300+(rpi-120)*(100/130)
    else:
     rpi=400+(rpi-250)*(100/130)
    return rpi
df['Rpi']=df['rspm'].apply(cal_RSPMI)
data= df[['rspm','Rpi']]
data.head()
# calculating the individual pollutant index for rspm(respirable suspended particualte matter concentration)
```

Figure 4. Calculating RSPM

Function to calculate spm individual pollutant index(spi)

```
def cal_SPMi(spm):
    spi=0
    if(spm<=50):
     spi=spm*50/50
    elif(spm>50 and spm<=100):
     spi=50+(spm-50)*(50/50)
    elif(spm>100 and spm<=250):
     spi= 100+(spm-100)*(100/150)
    elif(spm>250 and spm<=350):
     spi=200+(spm-250)*(100/100)
    elif(spm>350 and spm<=430):
     spi=300+(spm-350)*(100/80)
    else:
     spi=400+(spm-430)*(100/430)
    return spi

df['SPMi']=df['spm'].apply(cal_SPMi)
data= df[['spm','SPMi']]
data.head()
# calculating the individual pollutant index for spm(suspended particulate matter)
```

Figure 5. Calculating SPM

Figure 6. Calculating all factors together

F. RESULTS AND DISCUSSION

The analysis of the experiment involved the utilization of established Machine Learning algorithms and the preprocessed dataset. To create distinct versions for the comparable air quality data, the following algorithms were employed: logistic regression, k-nearest neighbors, decision tree classifier, random forest classifier, and decision tree regressor. The experiment involved using the attributes of using PM 2.5, SO2, and NO2 as input to get the desired result, which is the AQI attribute.

A. LINEAR REGRESSION

Train Model Accuracy: 0.9849533579250526

Test Model Accuracy: 0.9847286394495923

Kappa Score: 0.982377382981496

B. LOGISTIC REGRESSION

Train Model Accuracy: 0.7776012426913104

Test Model Accuracy: 0.7619365216071491

Kappa Score: 0.764377382981496

C. RANDOM FOREST CLASSIFIER

Train Model Accuracy: 0.998153774486465

Test Model Accuracy: 0.987909441913

Kappa Score: 0.9851205476925056

D. K-NEAREST NEIGHBOURS

Train Model Accuracy: 0.998153774486465

Test Model Accuracy: 0.9967105949441913

Kappa Score: 0.9951205476925056

E. DECISION TREE

Train Model Accuracy: 1.0

Test Model Accuracy: 0.9998400500712821

Kappa Score: 0.9997627702215676

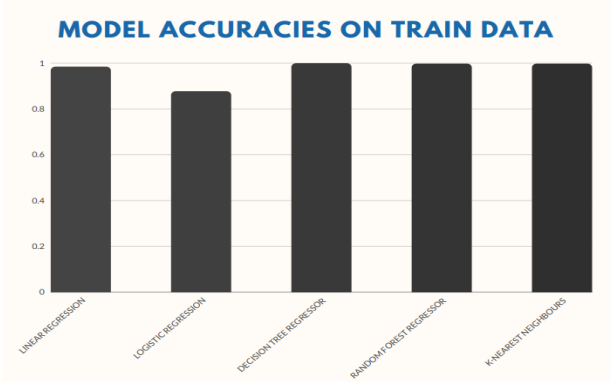

Figure 7. Graph of model accuracy on train data

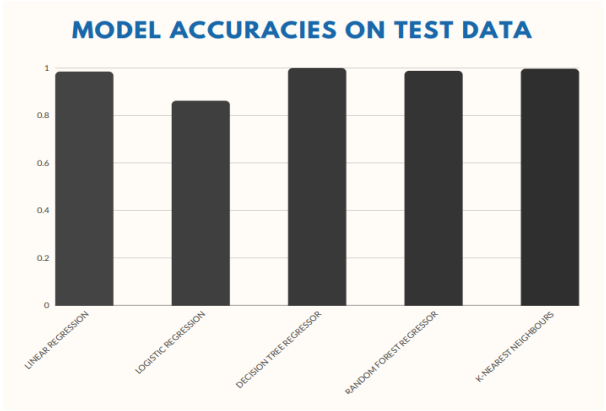

Figure 8. Graph of model accuracy on test data

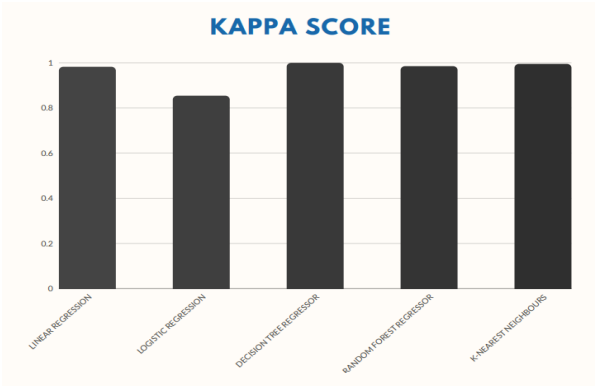

Figure 9. Graph of Kappa Score

The above values shows that the model using the KNN and Decision tree regression has a larger Train and test accuracy when compared to all other algorithms. This shows how accurately the fine grained AQI is predicted by Decision tree regression and KNN versions. However, Linear regression AND random forest regression demonstrate that those models accurately predict the AQI in this scenario due to their accuracy. After that, by looking at performance metrics for each fashion, we will conclude that the random forest regression model and decision Tree regression version do a good job of forecasting AQI for the information that is provided.

The task of predicting air quality is essential for maintaining public health and environmental monitoring. Support vector regression, one of the machine learning algorithms, offers an effective method for modeling the relationship between air quality and An efficient technique for simulating the connection between air quality and fine-grained air quality prediction. Cross-validation, high-level benchmarks, feature selection, and

performance measure selection all need to be done carefully when training and testing machine learning models. Furthermore, testing is a crucial step in choosing the best model for a particular weather operation.

G. COMPARISION AND ANALYSIS

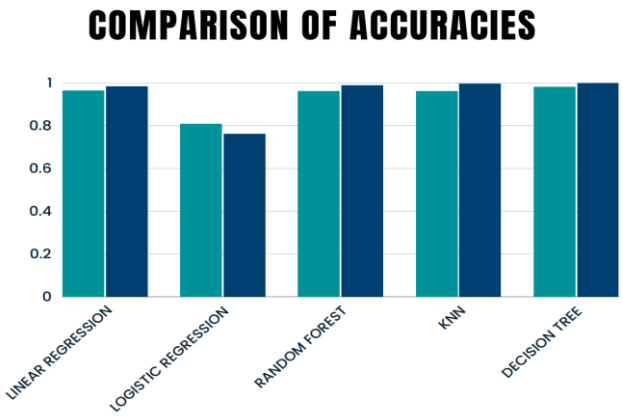

Figure 10. Graph of Comparison of Accuracies

The below table shows the comparison of author's model accuracy from different authors.

Table.2 Comparison of Accuracies

MODEL USED FOR PREDICTION	DIFFERENT PAPERS FOR COMPARISION	ACCURACY OF THE PAPERS	DATASET USED FOR PREDICTION VALUES	ACCURACY OF PREDICTED MODEL
LINEAR REGRESSION	R[12]	0.9640876	DATASET OF KUWAIT (2021)	0.9847286
LOGISTIC REGRESSION	R[21]	0.8087209	DATASET OF CHINA (2014)	0.76193659
RANDOM FOREST	R[9]	0.96183549	DATASET OF EUROPE (2016)	0.9879865
KNN	R[38]	0.9619876	DATASET OF CHINA (2013)	0.99671059
DECISION TREE	R[43]	0.982385472	DATASET OF CHINA (2011)	0.99984005

H. FUTURE WORK

Regarding future research, more studies are required to increase the precision and generalizability of climate forecasting models, particularly when it comes to the forecasting of uncommon events. New data and technologies, like satellite imagery and Internet of Things (IoT) sensors, should also be incorporated into the modeling process. All things considered, machine learning has the power to completely change how we track and regulate air quality, creating safer and healthier communities.

REFRENCES

1. Alkabbani, H., Ramadan, A., Zhu, Q., & Elkamel, A. (2022). An improved air quality index machine learning-based forecasting with multivariate data imputation approach. *Atmosphere*, *13*(7), 1144.
2. Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE access*, *7*, 76690-76698.
3. Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12), 2570.
4. Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., & Wang, Y. S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of cleaner production*, *209*, 134-145.
5. Soh, P. W., Chang, J. W., & Huang, J. W. (2018). Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *Ieee Access*, *6*, 38186-38199.
6. L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: an overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, p. 780, 2018.
7. A. C. Kemp, B. P. Horton, J. P. Donnelly, M. E. Mann, M. Vermeer, and S. Rahmstorf, "Climate related sea-level variations over the past two millennia," *Proceedings of the National Academy of Sciences*, vol. 108, no. 27, pp. 11017–11022, 2011.
8. J. Wang, H. Jiang, Q. Zhou, J. Wu, and S. Qin, "China's natural gas production and consumption analysis based on the multicycle Hubbert model and rolling Grey model," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 1149–1167, 2016.
9. X. Su, X. Yan, and C.-L. Tsai, "Linear regression," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 4, no. 3, pp. 275–294, 2012.
10. Du, S., Li, T., Yang, Y., & Horng, S. J. (2019). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2412-2424.
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 computing*, *2*(1), 5
12. Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health*, *17*, 1-19.
13. Ma, J., Cheng, J. C., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment*, *214*, 116885.
14. Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, *7*, 128325-128338.
15. 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 Systems with Applications*, *169*, 114513.
16. S. Ameer, M. Ali Shah, A. Khan et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.
17. A. Plaia and M. Ruggieri, "Air quality indices: a review," Reviews in Environmental Science and Bio/Technology, vol. 10, no. 2, pp. 165–179, 2021
18. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: model regularization and optimization," *Big Data Cognitive Computing*, vol. 2, no. 1, pp. 5–15, 2018.

19. D. Ramesh, "Enhancements of artificial intelligence and machine learning," *International Journal of Advanced Science and Technology*, vol. 28, no. 17, pp. 16–23, 2019.
20. M. Somvanshi, P. Chavan, S. Tambade, and S. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in 2016 international conference on computing communication control and automation (ICCUBEA). IEEE, 2016, pp. 1–7.
21. 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.
22. Nidhi Sharma , ShwetaTaneja , VaishaliSagar , Arshita Bhatt, "Forecasting air pollution load in Delhi using data analysis tools", ScienceDirect, 132 (2018) 1077–1088.
23. A. Kumar and P. Goyal, "Forecasting of air quality in delhi using principal component regression technique," Atmospheric Pollution Research, vol. 2, no. 4, pp. 436–444, 2021.
24. YusefOmidiKhaniabadi, GholamrezaGoudarzi, Seyed Mohammad Daryanoosh, Alessandro Borgini, Andrea Tittarelli, Alessandra De Marco, "Exposure to PM10, NO2, and O3 and impacts on human health", Environ SciPollut Res, 2016.
25. R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian, "MONITORING OF AMBIENT AIR QUALITY IN SALEM CITY, TAMIL NADU", International Journal of Current Research, ISSN: 0975-833X, Vol. 4, Issue, 03, pp.275-280, March, 2019
26. Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Applied Sciences, 10(7), 2401.
27. Zhao, J., Deng, F., Cai, Y., & Chen, J. (2019). Long short-term memory-Fully connected (LSTM-FC) neural network for PM2. 5 concentration prediction. *Chemosphere*, *220*, 486-492.
28. Huang, C. J., & Kuo, P. H. (2018). A deep CNN-LSTM model for particulate matter (PM2. 5) forecasting in smart cities. Sensors, 18(7), 2220.
29. Mahato, S., Pal, S., & Ghosh, K. G. (2020). Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. *Science of the total environment*, *730*, 139086.
30. Sharma, S., Zhang, M., Gao, J., Zhang, H., & Kota, S. H. (2020). Effect of restricted emissions during COVID-19 on air quality in India. *Science of the total environment*, *728*, 138878.
31. Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, *231*, 997-1004.
32. Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018, July). Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 965-973).
33. Chang, Y. S., Chiao, H. T., Abimannan, S., Huang, Y. P., Tsai, Y. T., & Lin, K. M. (2020). An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, *11*(8), 1451-1463.
34. Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, *119*, 285-304.
35. Kök, İ., Şimşek, M. U., & Özdemir, S. (2017, December). A deep learning model for air quality prediction in smart cities. In *2017 IEEE international conference on big data (big data)* (pp. 1983-1990). IEEE.

36. Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., ... & Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?. *Environment international*, *116*, 286-299.
37. Qin, Y., & Zhu, H. (2018). Run away? Air pollution and emigration interests in China. *Journal of Population Economics*, *31*(1), 235-266.
38. Qin, Y., & Zhu, H. (2018). Run away? Air pollution and emigration interests in China. *Journal of Population Economics*, *31*(1), 235-266.
39. Singh, R. P., & Chauhan, A. (2020). Impact of lockdown on air quality in India during COVID-19 pandemic. *Air Quality, Atmosphere & Health*, *13*, 921-928.
40. Bashir, M. F., Jiang, B., Komal, B., Bashir, M. A., Farooq, T. H., Iqbal, N., & Bashir, M. (2020). Correlation between environmental pollution indicators and COVID-19 pandemic: a brief study in Californian context. *Environmental research*, *187*, 109652.
41. Benchrif, A., Wheida, A., Tahri, M., Shubbar, R. M., & Biswas, B. (2021). Air quality during three covid-19 lockdown phases: AQI, PM2. 5 and NO2 assessment in cities with more than 1 million inhabitants. *Sustainable Cities and Society*, *74*, 103170.
42. Zhan, D., Kwan, M. P., Zhang, W., Yu, X., Meng, B., & Liu, Q. (2018). The driving factors of air quality index in China. *Journal of Cleaner Production*, *197*, 1342-1351.
43. Liu, H., & Chen, C. (2020). Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China. *Journal of Cleaner Production*, *265*, 121777.
44. Rosser, F. J., Rothenberger, S. D., Han, Y. Y., Forno, E., & Celedón, J. C. (2023). Air Quality Index and Childhood Asthma: A Pilot Randomized Clinical Trial Intervention. *American Journal of Preventive Medicine*, *64*(6), 893-897.
45. Li, H., You, S., Zhang, H., Zheng, W., Lee, W. L., Ye, T., & Zou, L. (2018). Analyzing the impact of heating emissions on air quality index based on principal component regression. *Journal of cleaner production*, *171*, 1577-1592.
46. Güçlü, Y. S., Dabanlı, İ., Şişman, E., & Şen, Z. (2019). Air quality (AQ) identification by innovative trend diagram and AQ index combinations in Istanbul megacity. *Atmospheric Pollution Research*, *10*(1), 88-96.
47. Xu, K., Cui, K., Young, L. H., Wang, Y. F., Hsieh, Y. K., Wan, S., & Zhang, J. (2020). Air quality index, indicatory air pollutants and impact of COVID-19 event on the air quality near central China. *Aerosol and Air Quality Research*, *20*(6), 1204-1221.
48. Tan, X., Han, L., Zhang, X., Zhou, W., Li, W., & Qian, Y. (2021). A review of current air quality indexes and improvements under the multi-contaminant air pollution exposure. *Journal of environmental management*, *279*, 111681.
49. Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., & Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. *Environmental pollution*, *231*, 1232-1244.
50. Xue, J., Xu, Y., Zhao, L., Wang, C., Rasool, Z., Ni, M., ... & Li, D. (2019). Air pollution option pricing model based on AQI. *Atmospheric Pollution Research*, *10*(3), 665-674.
51. Cheng, W. L., Chen, Y. S., Zhang, J., Lyons, T. J., Pai, J. L., & Chang, S. H. (2007). Comparison of the revised air quality index with the PSI and AQI indices. *Science of the Total Environment*, *382*(2-3), 191-198.