IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

Air Quality Prediction System using Machine Learning

K. Pazhanivel[1], U. Dinesh Kumar[2], K. Naveen[3], M. Niranjan[4]
Assistant Professor, Department of Computer Science Engineering[1]
Students, Department of Computer Science and Engineering [2,3,4]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, India

Abstract: *Environmental protection measures cannot now be effectively ensured due to the rapid industrialization of recent years. The main issue influencing the standard of living in the country now is the severity of environmental challenges. To comprehend the potential air pollution process beforehand, we must therefore develop a reasonably good air quality prediction model. To reduce air pollution, it is crucial to establish and implement the appropriate control measures, according to the model's forecast results. This study makes extensive use of data mining techniques like neural networks, mutual information theory, and intelligent optimization algorithms. We leverage the fundamental information from open monitoring locations' long-term predictions of air quality as our training and test sets. Secondly, the association between the various monitored pollutants is examined using the SOM neural network model for unsupervised grouping of pertinent pollutant data. A NSGA-II-optimized neural network is suggested as a solution to the issues of a vast amount of data and the lengthy computation time of the technique, paired with the findings of clustering. According to the experimental findings, contaminants can be predicted with an accuracy of more than 90%.*

Keywords: Air Quality

I. INTRODUCTION

Environmental issues including air pollution, water pollution, noise pollution, and others have become more prevalent in emerging nations like India due to the population growth and economic boom in urban areas.Human health is directly impacted by air pollution.;In our nation, there has been a rise in public knowledge of the same. Some of the long-term effects of air pollution include global warming, acid rain, and an increase in asthma cases. Predictive air quality forecasts can lessen the impact of maximum pollution on people and the environment. Therefore, one of the society's top priorities is to improve air quality predictions.Fossil fuel combustion, as well as the release of toxic gases and solid particles from moving vehicles and manufacturing processes, are the main causes of air pollution. Sulfur oxides, nitrogen dioxides, particulate matter, and carbon monoxide are examples of such substances. In order to live a healthy life, monitoring and examining air quality is currently a highly significant and crucial topic. Data mining techniques can be used to analyze air pollution so that appropriate measures can be taken to reduce it. Simple knowledge extraction from unstructured data collection is what data mining is all about. Massive datasets, which contain the most common patterns in a dataset, can also be explored via data mining. The purpose of the real data mining process is to extract information from a big pool of data and transform it into an understandable framework for further usage. Using data mining be put to use for classification, prediction, identification, and optimization. The process of extracting hidden predictive knowledge from sizable databases is known as data mining. It can be characterized as a logical procedure used to sift through a lot of data in an effort to identify relevant information. This method's primary objective is the discovery of unique information as well as previously unidentified patterns. Data mining and knowledge discovery are two different concepts. Using data mining techniques, it is possible to analyze, predict, and forecast pollutants related to air pollution. It is also possible to determine the cause of the air pollution. Optimization algorithms can be used to find the most optimal features for classification and prediction results that are effective. Programming that "learns" from its environment and makes appropriate adjustments is known as machine learning. A machine-learning algorithm improves over time by improvising, adapting to changes, and carrying out the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

assigned work more effectively. Hence, constructing prediction models for forecasting air .pollution can be done relatively well using machine learning approaches. Application of a machine learning methodology is problem-specific, so it is essential to use the best machine learning method possible based on both ecological and environmental aspects.

1.1 Air Quality

Our atmosphere is ~21% oxygen and ~78% nitrogen; the remaining 1% is considered "trace gases" and this includes everything else—from carbon dioxide to the noble gases like argon. Scientists and engineers study this 1%, as well as the many types of particles present in the atmosphere. When we think of air quality, we typically think of the air we are breathing and whether or not it is safe. However, air quality can refer to ambient outdoor conditions, indoor conditions, particular sources, good air quality vs. poor air quality, etc. Within the field of air quality, researchers specialized in many other specific topics. Due to the complexity of our atmosphere and the possibility of transport over long distances, we can also think of air quality in terms of scale—that is, pollution may cause local or global problems. For example, pollution in China can make its way over North America and add to existing pollutants there.

Poor air quality can negatively affect human and environmental health. In humans, poor air quality can lead to a multitude of problems that include respiratory and cardiovascular diseases. We tend to think first of asthma and respiratory problems, but some particles are so small that they can enter the blood stream through the lungs and cause inflammation leading to issues beyond our breathing. In plants, poor air quality can also cause disease that can result in crop loss. In addition to human and environmental health, many pollutants that we worry about are greenhouse gases and contribute to climate change. Finally, poor air quality can impact quality of life. Consider visibility issues in National Parks and odors near industrial areas of cities; in addition to potential health dangers, these air quality issues can make daily life unpleasant.

1.2 Forms of Pollution

Pollutants are typically thought of as either gas-phase molecules or particles (i.e., particulate matter, such as dust). Compounds that occur in the gas phase are known as gas-phase compounds (carbon dioxide is an example of a gas-phase pollutant). The size and makeup of particulate particles vary. An collection of either solid or liquid molecules makes up very tiny particulate matter. Pollen and dust are examples of bigger particulate matter. To help students comprehend how commonplace items like rubbing alcohol and burning wood produce different pollutants, refer them to the accompanying exercise Connecting Sources and Pollutants. The distinction between primary and secondary emissions is another crucial one in terms of air pollution. Direct emissions are the primary emissions. For instance, primary emissions come from things like smokestacks and tailpipes, whereas secondary emissions come from things like the atmosphere itself. One excellent example is ground-level or tropospheric ozone, which is produced by sunlight and a number of major pollutants. Because of the complexity of atmospheric chemistry, what we release may disperse, react and change into something else, or remain for a very long time, depending on the specific component and external factors.

1.3 Air Quality Monitoring

Scientists and engineers use the characteristics of various contaminants to quantify the issues such pollutants cause. A filter can be used to collect particles, which can then be weighed or examined. The behaviour of particles of various sizes can also be used to separate them. Additionally, gases have characteristics that we can use for measuring, like absorbance, for instance. By measuring which wavelengths of light are absorbed by a sample, we may calculate the amount of gas present. Certain gases absorb specific wavelengths of light. To preserve the health of people and the environment, laws must be developed and then enforced, which requires the ability to quantify air contaminants.

The main instrument employed throughout this unit is next-generation air quality monitoring technology. Low-cost equipment is now feasible because to developments in sensor technology. These technologies enable the collection of data with higher geographical and temporal precision, but they are not as reliable as more expensive conventional monitoring systems. Low-cost technology also make monitoring more accessible to communities, schools, and citizen scientists in underdeveloped nations. With the related exercise Understanding the Air through Data Analysis, students can further delve into the significance and procedures of air quality monitoring. Continue with the related activity Study

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

Design for Air Quality Research, where students practise project planning using a case study that contrasts conventional cook stoves with new and improved cook stoves intended for use in underdeveloped countries. The accompanying exercise, Presenting Your Project Findings with Professional Posters, can then be used by students to assist them present their final conclusions.

II. AIR POLLUTION

Globally, millions of people perish each year as a result of outdoor and indoor air pollution, according to the WHO's report on health [1]. The WHO has set guiding standards for lowering the high levels of contaminants. According to data from the WHO, 90% of people are exposed to levels of air pollution above the recommended upper limits. People of all ages experience health problems as a result of exposure to indoor air pollutants in addition to outdoor pollution. This includes everything from cancer to eye issues to respiratory illnesses. For both human health and the ecosystem, air pollution continues to pose a serious concern. A large number of fatalities occur each year as a result of the interaction between indoor and outdoor air pollution. This is a result of the rise in fatal illnesses brought on by air pollution, including heart disease and severe respiratory infections.

The issues brought on by air pollution are not exclusive to India. The highest attainable limits deemed suitable for human health are often exceeded by pollution levels in numerous locations. Greater metropolitan, heavily industrialised places like Delhi are experiencing a dire predicament.The national government and state governments are both making efforts to lower air pollution levels.

III. LITERATURE REVIEW

Predicting the concentration of air pollutants in a particular area over time is an important aspect of the field of research known as "air quality prediction." The management of air pollution and its detrimental consequences on both human health and the environment depends on accurate air quality forecasting. Air quality prediction using hybrid deep learning and time series analysis, by S. Zhang et al., is one of the most current articles on air quality prediction using machine learning. It was published in Atmospheric Pollution Research in 2022. The study suggests a unique method for forecasting air quality that combines deep learning and time series analysis. The authors present the idea of predicting air quality and stress the significance of good prediction for reducing the adverse effects of air pollution. The limits of conventional air quality prediction techniques are then discussed, and the necessity of a hybrid strategy that combines time series analysis and machine learning is highlighted. The proposed method is then discussed, which employs a deep neural network to extract characteristics from time series data on air quality before using a time series analysis technique to project future values. Using actual air quality data from Beijing, China, the authors validate their method and show how well it can accurately predict air quality. Overall, this study emphasizes the significance of air quality forecasting and offers a viable strategy based on time series analysis and machine learning methods. It also highlights the possibility of combining traditional air traffic control with machine learning. Increasing the precision and efficiency of air quality prediction techniques. Pollution prediction with the correlation of pollutants with other metrological variables for 5 cities of China was done in reference [14]. PM2.5 was the pollutant that has been used. RF was found to be the best method among the four methods that have been used.PM 2.5 values have been predicted by Qin et al. [15] using CNN and LSTM. In reference [19], also it is done with Aggregated LSTM model. NO2, SO2, CO, and O3 levels have been classified with the Decision tree and Naive Bayes algorithm [20].In reference [21],Support Vector Machines and neural networks were used to categorise the eight contaminants that affect air quality.

Machine learning techniques can be used to categorise air quality into ranges from Excellent to Severe. Reference [20] employed a decision tree (J48) and the Naive Bayes method. 91% Accuracy was observed for the decision tree. A short data amount was used. The data set used was for US cities. Decision trees cannot act as good classifiers for time series. The challenges of forecasting the Air Quality Index (AQI) were addressed by Mahalingam et al. [21]. The paper aimed to minimize pollution. The prediction of AQ was performed using neural networks and support vector machines. The data set on air pollution was gathered from the CPCB in India. Data for Delhi city was considered. The results showed increased prediction accuracy of the proposed model over other models. A Support Vector Machine (SVM) based classifier was proposed in reference [4].The air quality can be classified as either good or dangerous using this classifier. The classifier did a good job of classifying the values of air quality, according to the authors. As potential

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

input for the classifier, the calculated values of Cumulative Indices were employed. Real data from three cities namely Kolkata, Delhi, and Bhopal were used to test the classifier. A predictive air quality map for the next 24 hours in Tehran was done efficiently [22]. They have used Apache Hadoop, Naive Bayes, and Logistic regression. They find Logistic Regression to be the best estimator. Logistic Regression can perform well for predicting classes. Here they use it for classification to produce the Predictive Air Quality Risk Map. So, both Naive Bayes, Logistic regression can be used to classify AQI. Air pollution trends in various cities in India were done by Sharma et al. [23]. Pollution data provided by the CPCB in India was considered. These data demonstrated the annual growth of pollutants like SO_2, NO_x, and PM2.5 over three years from 2015 to 2018. Data for three cities were used. The cities were Delhi, Bengaluru, and Chennai. Sources of pollution were analyzed. It was observed that NOx, SO_2, and PM 2.5 resulted from outdoor sources like various transport modes, power generation plants, industries, construction activities, and indoor sources like domestic cooking. Critically, highly, moderately, and low polluted were the four classifications used to group the areas under consideration. The findings assist in estimating city-wide pollution.

Sl. No.	Ref. No.	Topic discussed	Location / Datasets	Pollutants considered
1	De Leon et al. [6]	Effect of AQI on hospital admissions for respiratory diseases	The hospital admission data for 1987-88 and 1991-92 for London	Black smoke (BS), nitrogen dioxide (NO_2), Sulphur dioxide (SO_2), and ozone (O_3)
2	Agarwal et al. [9]	Effect of COVID-19 pandemic on AQI	6 cities each in India and China	PM2.5 and NO_2
3	Li et al. [12]	Generation of the prediction maps of pollutant	Chengdu	SO_2, NO_2, PM10
4	Ameer et al. [14]	Pollution prediction, Correlation of pollutants with other metrological variables	5 cities of China	PM2.5
5	Qin et al. [15]	Pollution prediction	Shanghai	PM2.5
6	Ghoneim et al. [16]	Ozone concentration prediction	Aarhus	Ozone
7	Sakarkar et al. [17]	Pollution prediction	Nagpur	SO_2, NO_2, RSPM, SPM, PM2.5
8	Liu et al. [18]	Forecasting AQI, Finding NO_X level	Beijing and an Italian city	NO_X
9	Chang et al. [19]	Prediction of pollutants	Data are given by Taiwan Environmental Protection Agency	PM2.5
10	Gore and Deshpande [20]	Classifying air quality	Data set US cities	NO_2, SO_2, CO, and O_3
11	Mahalingam et al. [21]	Classifying air quality	Delhi	8 pollutants including PM10 and Lead
12	Asgari et al. [22]	Predictive air quality map (classification)	Tehran	AQI, Wind direction, etc.

Table 1. Summary of machine learning methods for air quality forecasting

IV. PROBLEM STATEMENT

In metropolitan regions with more businesses, industries, and people living there, maintaining acceptable standards of air has become a major concern. With an increase in population come increases in transportation, power use, and fuel consumption. We are fully aware that a lot of trash has been dumped on the property. All types of life on earth are put in even greater danger because of the air's high level of contamination. This makes it vital to monitor and evaluate the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

air quality, and in response, the government should be alerted to take the appropriate measures. This study uses machine learning methods to effectively analyse all the significant efforts that have been made in this area.To develop a machine learning model to predict air quality based on various environmental parameters such as temperature, humidity, wind speed, and atmospheric pressure. The model should be able to accurately forecast air quality levels in a specific geographic location in advance, allowing people to take necessary precautions to mitigate the negative impacts of poor air quality. The goal is to create a tool that can help individuals, organizations, and policymakers make informed decisions and take appropriate actions to protect public health and the environment.

V. EXISTING CHALLENGES AND RELATED WORK

In your effort to estimate air quality, you could run into a number of difficulties. Some of the typical difficulties include:

- **Data accessibility:** Accurate and trustworthy air quality data must be available in order to train the machine learning model. Yet, the data could not be easily accessible or might only be available in select areas.
- **Data accuracy:** Accurate or incomplete data might cause a model to perform poorly, therefore data quality is also crucial. Before training the model, there may occasionally be outliers or abnormalities in the data that must be found and corrected
- **Feature engineering:**Itis the process of choosing and extracting pertinent characteristics from the data for the model. The process can take some time, and picking the best characteristics can have a big impact on how well the model performs.
- **Model selection:** Deciding which machine learning algorithm will be most effective for your task of predicting air quality can be difficult. There are several models to pick from, and each has advantages and disadvantages.
- **Interpretability:** Particularly in disciplines like public health and environmental science, it is crucial to comprehend how the machine learning model generates its predictions. The model's interpretability can be used to pinpoint the root causes of air pollution and create efficient remedies.
- **Generalization:** To make accurate predictions in real-world circumstances, the trained model must be able to generalise well to data that has not yet been observed. Poor model performance and incorrect predictions can result from either overfitting or underfitting.
- **Deployment:** After creating an accurate model, it can be difficult to implement it in a real-world setting. It include managing input data, integrating the model into current systems, and resolving potential problems such data drift over time.

VI. RELATED WORKS

There have been many works related to air quality prediction. Here are some of them:

1. "Deep Learning-Based Air Quality Prediction Models" by Zhengyang Song, Yujie Sun, Yuying Hao, and Qingquan Li. This paper proposes a deep learning-based model for predicting air quality using historical data.
2. "Air Quality Prediction Using Machine Learning Algorithms" by Piyush Gupta and Aman Agrawal. This paper compares different machine learning algorithms for air quality prediction and evaluates their performance.
3. "Forecasting Air Quality with Machine Learning and Internet of Things" by Wei Yang, Wei Guo, Zhiqiang Zhang, and Wei Zhang. This paper proposes an air quality forecasting system that combines machine learning algorithms with Internet of Things (IoT) technology.
4. "Real-time Air Quality Prediction Using Hybrid Machine Learning Models" by Jia Liu, Weifeng Liu, and Xiangxu Meng. This paper presents a hybrid machine learning model for real-time air quality prediction that combines support vector regression and artificial neural networks.
5. "Air Quality Prediction Using Ensemble Learning Techniques" by Raj Kumar Gupta, Ravi Prakash Gupta, and Sanjay Kumar. This paper proposes an air quality prediction model based on ensemble learning techniques, including bagging, boosting, and random forests.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

6. "Short-term Air Quality Prediction Using Recurrent Neural Networks" by Pengcheng Wang, Lei Meng, Yuhong Wang, and Mengdi Zhang. This paper proposes a recurrent neural network-based model for short-term air quality prediction.

6.1 Health effects of air pollution

De Leon et al. [6] demonstrated the impact that air pollution has on hospital admissions. Using Poisson regression, this was done in the context of respiratory disease. Pollution elements included ozone, NO2, SO2, and black smoke (BS) (O3). The analysis examined hospital admission data for the City of London for the years 1987–1988 and 1991–1992. Ammasi Krishnan et al. evaluated the pollution caused by fireworks during the Diwali festival season. [7]. In India, Diwali is observed between October and November. Around Diwali in 2017, there was an upsurge in particle pollution in Chennai. People's health problems during the festival season increase as a result of this rise in pollution levels.The most common health problem identified was eye irritation.Krishnan et al. [8] monitored the amount of pollutants, including sulphur dioxide (SO2), in Kodungaiyur, Chennai, India. In 2017, over the course of six months in the spring and summer, the study was conducted. To determine the effects of pollutants, particularly PM2.5, on humans, a survey on the illnesses observed in the Kodungaiyur area was conducted. The ailments included in the survey include respiratory and ocular allergies. The findings showed a connection between exposure to pollution and an increase in illnesses.

6.2 Impact of COVID-19 pandemic on air quality

There were few studies that examined how the COVID-19 epidemic affected air quality metrics. Agarwal et al. [9] looked at six cities in China and six cities in India. As anticipated, stringent lockdown procedures and a decrease in significant human activities would have resulted in an improvement in air quality. They looked at the three-month improvement in air quality. The outcomes showed that the decrease in NO2 AQI was instantaneous. The PM2.5 AQI was steadily declining. Measures to lower pollutant concentrations may be implemented in the future with the help of an analysis of how these have been carried out.The goal of the study in reference [10] was to ascertain the effects of the COVID-19 pandemic's shutdown on the air quality in India's four largest cities from January to May 2020. Delhi, Mumbai, Kolkata, and Chennai were the locations. Analysing the pollutant levels of gases like nitrogen dioxide and particulate matter was done to achieve this (NO2). The data showed that the studied regions had significantly improved air quality.

VII. METHODOLOGY

Air quality prediction is an essential task that can be achieved by using machine learning (ML) algorithms. The following are the steps that can be taken to develop an ML-based air quality prediction model:

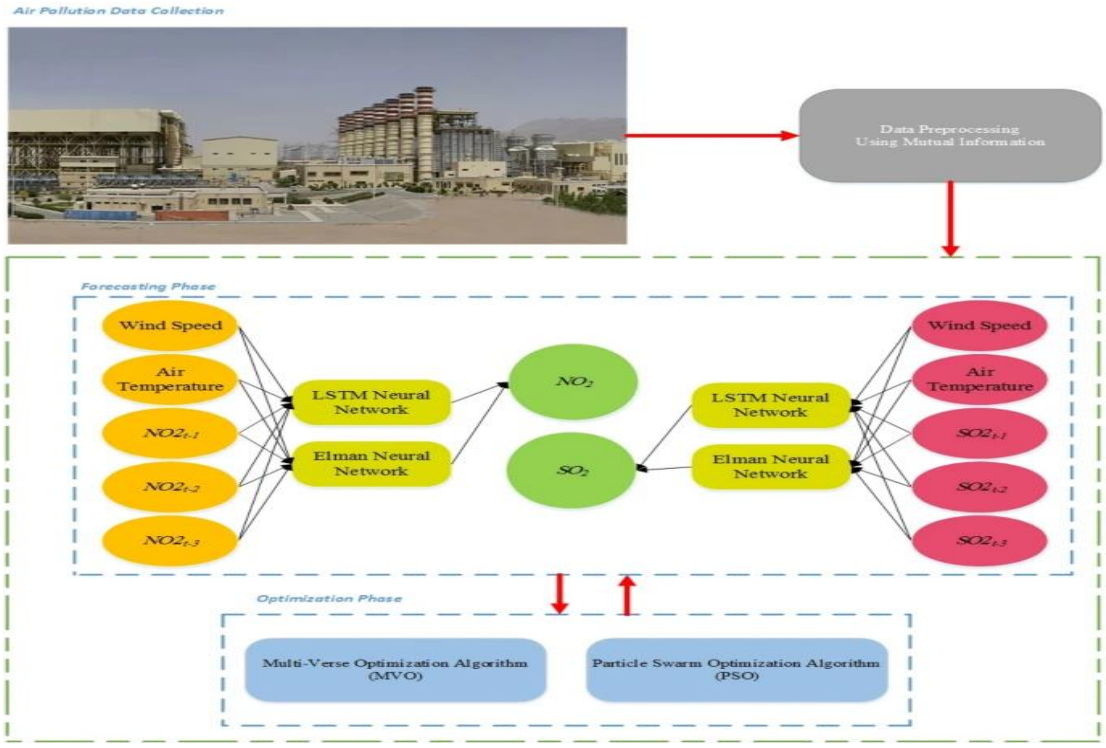
- **Data Collection:** The first step is to collect the data required for building the air quality prediction model. This includes data on various air pollutants, weather conditions, geographical location, and other factors that impact air quality. There are several sources of data available, such as government monitoring stations, satellite imagery, and IoT devices.
- **Data Preprocessing:** Preprocessing is the following stage after data collection.This includes cleaning the data, handling missing values, and transforming the data into a format that can be used by ML algorithms.
- **Data Cleaning:** This entails deleting any extraneous information, including duplicate records, irrelevant columns, and null or missing values.
- **Data Integration:** This step involves combining data from multiple sources, if necessary.
- **Data Transformation:** This involves converting the data into a suitable format for analysis, such as scaling, normalization, and feature engineering.
- **Data Reduction:** This step involves reducing the amount of data by sampling, feature selection, or principal component analysis (PCA).
- **Data Discretization:** This involves converting continuous data into discrete data by binning or bucketing.
- **Data Encoding:** This involves converting categorical data into numerical data, such as one-hot encoding.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

- **Data Splitting:** This involves splitting the data into training, validation, and testing sets for machine learning algorithms
- **Feature Engineering**: Feature engineering is the process of selecting the most relevant features from the dataset. This involves analyzing the data to identify the variables that have the most significant impact on air quality. Feature engineering can also involve transforming the data to improve its quality.
- **Model Selection**: The next step is to select the appropriate ML algorithm for building the air quality prediction model. This depends on the nature of the data, the type of problem, and the desired accuracy of the model. Common ML algorithms used for air quality prediction include linear regression, decision trees, neural networks, and support vector machines.
- **Model Training:** Once the ML algorithm is selected, the next step is to train the model. This involves feeding the preprocessed data into the ML algorithm and adjusting the model's parameters to improve its performance. The model is trained on a portion of the data, and the remaining data is used for testing.
- **Model Evaluation:** The model's performance is evaluated using various metrics such as root mean square error (RMSE), mean absolute error (MAE), and R-squared. These metrics indicate how accurately the model predicts air quality based on the input data.
- **Model Deployment:** Once the model is trained and evaluated, it can be deployed to make predictions on new data. The model can be integrated into an application or system that provides real-time air quality predictions. In summary, the methodology for air quality prediction based on ML involves data collection, data preprocessing, data cleaning, data integration, data transformation, data reduction, data discretization, data encoding, data splitting, feature engineering, model selection, model training, model evaluation, and model deployment. By following these steps, it is possible to build an accurate air quality prediction model that can be used to improve public health and environmental conditions.

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.

VIII. SYSTEM ARCHITECTURE



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

- **Data Collection**: Data collection involves gathering relevant data from various sources, including air pollutant concentrations, weather conditions, topography, and population density. This data can be collected using sensors installed in different locations or obtained from public sources. The quality and quantity of data collected play a crucial role in the accuracy of the prediction models.
- **Data Preprocessing**: Data preprocessing involves cleaning and preparing the collected data for use in machine learning models. This step includes removing any outliers, filling in missing values, data normalisation, and data transformation. The goal of data preprocessing is to ensure that the data is accurate, consistent, and ready for use in the prediction models.
- **Feature Engineering:** Feature engineering involves selecting and extracting relevant features from the preprocessed data that will be used as input variables in the machine learning models. The process includes selecting the most relevant variables and transforming them into features that can help the models predict the air quality levels accurately.
- **Elman Neural Network:**For time-series prediction issues, the Elman Neural Network is a kind of recurrent neural network (RNN). In order to store previous inputs and outputs, it features an extra layer of neurons that acts as a memory. The Elman Neural Network can learn and recall the time dependencies in the data, which makes it particularly valuable for forecasting changes in air quality over time.
- **LSTM Neural Network:** Another RNN variety that excels at solving time-series prediction issues is the Long Short-Term Memory (LSTM) Neural Network. It can learn long-term dependencies in the data and has a more intricate design than the Elman Neural Network. While predicting air quality, the LSTM Neural Network is utilised because long-term relationships in the data must be captured.
- **Multi verse optimization algorithm:**A metaheuristic optimisation approach called the Multi-Verse Optimisation Algorithm (MVO) is used to improve the hyperparameters of machine learning models. MVO simulates numerous universes, each of which is a potential solution to the optimisation issue. The optimal collection of hyperparameters to optimise the performance of the prediction models are then found by the algorithm using these solutions to explore the solution space.
- **Particle Swarm Optimization Algorithm:**Another metaheuristic optimization approach that is frequently used to tune the hyperparameters of machine learning models is the Particle Swarm Optimisation (PSO) Algorithm. Each particle in the simulation represents a potential answer to the optimisation issue and moves around the solution space while the algorithm runs. The programme then searches for the best collection of hyperparameters to optimise the performance of the prediction models using the collective behaviour of the particles.
- **Model Deployment:** Once the models are developed and tested, they are deployed in a production environment, where they can be used to make predictions about air quality levels. The models are integrated into the system architecture and configured to receive input data and provide output predictions.
- **Monitoring and Evaluation:** The performance of the air quality prediction system is continuously monitored and evaluated to ensure that it is working as expected. This step involves tracking the accuracy of the models, identifying any issues or errors, and making necessary adjustments to improve the performance of the system.
- **Advantages of Random Forest:** Tasks involving classification and regression can be completed using Random Forest. Large datasets with a high dimensionality can be handled by it. In addition to preventing overfitting, it improves the model's accuracy.

Implementation Steps are given below:

Random forest is a powerful machine learning algorithm used for classification, regression, and other types of data analysis tasks. It is an ensemble learning method that combines multiple decision trees to improve the accuracy of the predictions.

Here is a step-by-step description of how random forest works:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

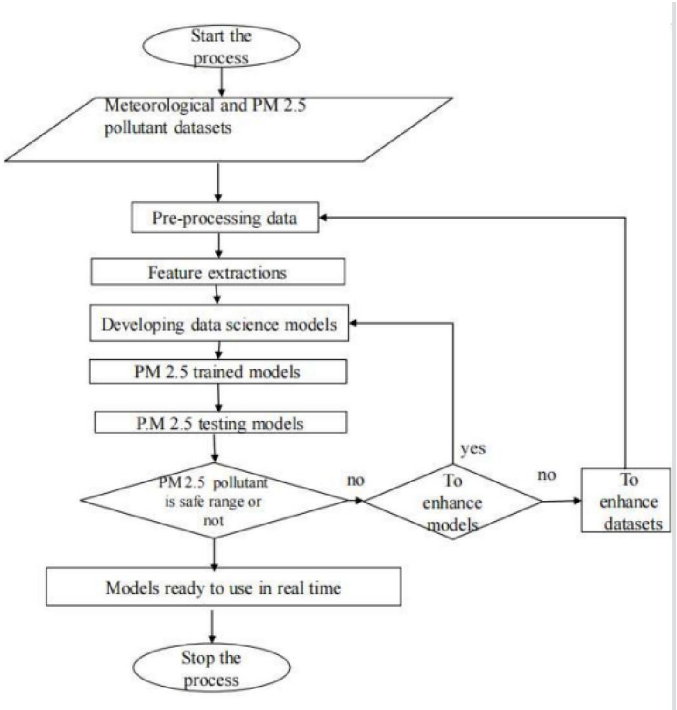
- **Data Preparation**: The first step in implementing a random forest is to prepare the data. This involves cleaning the data, handling missing values, encoding categorical variables, and splitting the data into training and testing sets.
- **Random Sampling:** In random forest, multiple decision trees are built using a random sample of the data with replacement. This is known as bootstrap aggregating or bagging.
- **Decision Tree Building**: For each sample of the data, a decision tree is built using a random subset of the features. This helps to reduce overfitting and improve the generalization of the model.
- **Tree Optimization:** The decision tree building process involves optimizing various parameters such as the number of nodes, the depth of the tree, and the splitting criteria. This helps to improve the accuracy of the individual decision trees.
- **Prediction**: Once the decision trees are built, the random forest algorithm combines the predictions of all the trees to make a final prediction. In classification problems, the class with the highest number of votes is selected, while in regression problems, the mean of the predictions is used.
- **Model Evaluation:** The final step is to evaluate the performance of the random forest model. This is done by comparing the predicted values with the actual values in the test set using metrics such as accuracy, precision, recall, F1-score, and mean squared error.

The main advantages of using a random forest algorithm are:

1. It is highly accurate and can handle complex data sets with high dimensionality.
2. It is robust to outliers and missing data.
3. It can handle both categorical and numerical data.
4. It can handle both classification and regression problems.
5. It can provide estimates of feature importance.

Overall, random forest is an effective and flexible machine learning technique that can be applied to a variety of data analysis applications. Its ability to combine multiple decision trees and handle complex data sets make it a popular choice for many real-world applications.

IX. FLOW CHART



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

Performance Assessment of Air Quality Forecasting:

When the prediction is used, the following parameters are used for the performance assessment of Air Quality Forecasting as to how accurate the predicted values are: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), etc. [11].

MAE (Mean absolute error) is the average of the absolute errors. Absolute error is given by the difference between the actual and forecasted values.

$$MAE=\sum Ni=1|Yforecast -Xactual |/N$$

MSE (Mean Squared Error) is the measure of the average of the squares of the errors. The error is given by the difference between the estimated values and the actual value.

RMSE (Root Mean Squared Error) is the error rate given by the square root of MSE.

The equation is shown below:

$$RMSE=\sum Ni=1(Yforecast -Xactual)2/N----------\sqrt{}(3)$$

When classification is used, the following parameters are used for the Performance Assessment of Air Quality Forecasting: Accuracy, Precision, and Recall.

Accuracy: Accuracy is the ratio of the number of observations predicted correctly to the total number of observations.

$$Accuracy =(TP+TN)(TP+FP+FN+TN)$$

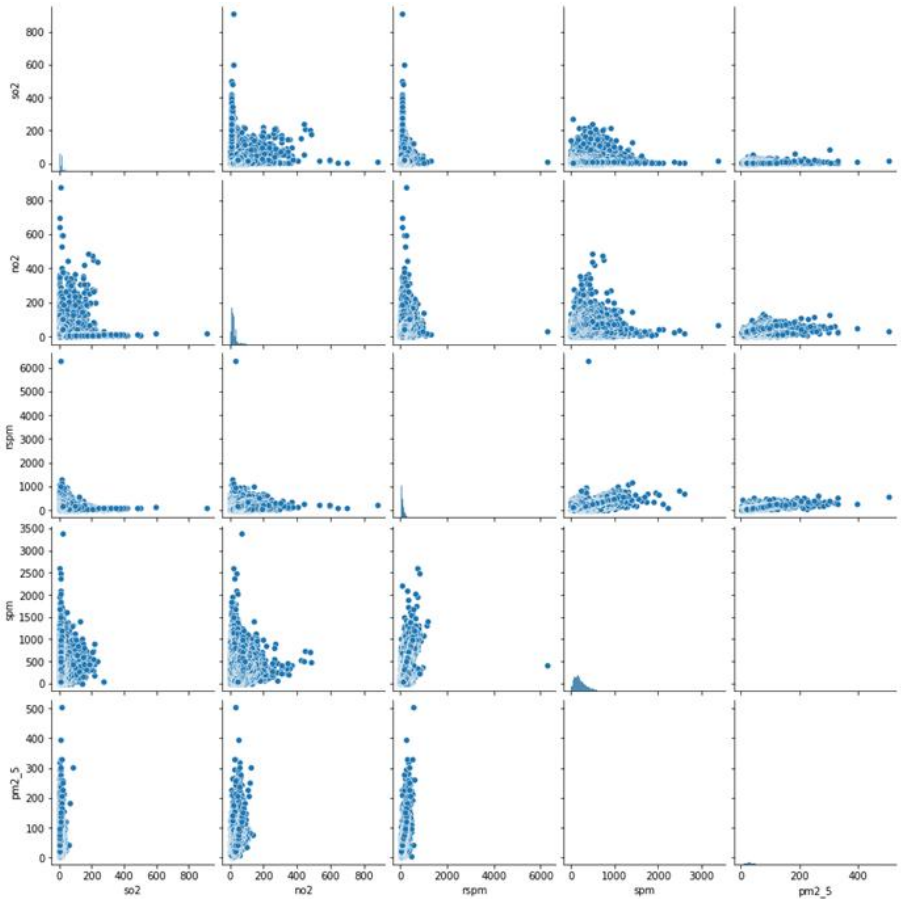
Precision (P): Precision is the ratio of the number of positive observations predicted correctly to the total number of the predicted positive observations.

$$Precision =TP(TP+FP) \quad (5)$$

Recall (R): Recall is the attribute of the model that sums up the model's ability to predict all the positive samples.

$$Recall =TP(TP+FN) \quad (6)$$

X. DATA VISUALIZATION



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

XI. CONCLUSION

An AQI forecast is required to stop air pollution. To perform this task more accurately, a few machine learning approaches can be used. We talked about the application of machine learning techniques to the classification or prediction of pollution levels. Some techniques are more suited to the prediction, categorization, or accuracy of AQI. The following conclusions were drawn from a comparison of the research methods used in the numerous research publications presented here; SVM, SVM with Naive Bayes, Logistic Regression AQI can be categorised using decision trees, each with their own benefits and drawbacks. Decision trees are one example that can be utilised, although they are not effective time series data classifiers. For class prediction, logistic regression can be effective. Forecasting the Air Quality values can be done using DT, RF, MLP, GB regression, and deep neural networks. About the data set, each author has used data sets for a certain city or region. The project could be expanded to a wider geographic area. Similar to this, different studies differed in the quantity and kind of contaminants they considered. Future work could involve comparing the findings of a study of the various approaches utilising a bigger data set for all main pollutants, all available ML methods, in order to determine the strategy that provides the highest level of accuracy.

REFERENCES

[1]. World Health Organization Health Topic on Air Pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1, accessed on 25 April 2021.

[2]. Report on National Air Quality Index. (2015). Central Pollution Control Board, Ministry of Environment, Forests and Climate Change, Government of India. https://app.cpcbccr.com/ccr_docs/FINAL-REPORT_AQI_.pdf, accessed on 25 April 2021.

[3]. Sowlat, M.H., Gharibi, H., Yunesian, M., Mahmoudi, M. T., Lotfi, S. (2011). A novel, fuzzy-based air quality index (FAQI) for air quality assessment. Atmospheric Environment, 45(12): 2050-2059. https://doi.org/10.1016/j.atmosenv.2011.01.060

[4]. Saxena, A., Shekhawat, S. (2017). Ambient air quality classification by grey wolf optimizer based support vector machine. Journal of Environmental and Public Health, 2017: 3131083. https://doi.org/10.1155/2017/3131083

[5]. Acharjya, D.P., Ahmed, K. (2016). A survey on big data analytics: Challenges, open research issues and tools. International Journal of Advanced Computer Science and Applications, 7(2): 511-518. https://dx.doi.org/10.14569/IJACSA.2016.070267

[6]. De Leon, A.P., Anderson, H.R., Bland, J.M., Strachan, D.P., Bower, J. (1996). Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. Journal of Epidemiology & Community Health, 50(S1): s63-s70. https://doi.org/10.1136/jech.50.suppl_1.s63

[7]. Ammasi Krishnan, M., Devaraj, T., Velayutham, K., Perumal, V., Subramanian, S. (2020). Statistical evaluation of PM2.5 and dissemination of PM2.5, SO2 and NO2 during Diwali at Chennai, India. Natural Hazards, 103(3): 3847-3861. https://doi.org/10.1007/s11069-020-04149-8

[8]. Krishnan, M.A., Jawahar, K., Perumal, V., Devaraj, T., Thanarasu, A., Kubendran, D., Sivanesan, S. (2019). Effects of ambient air pollution on respiratory and eye illness in population living in Kodungaiyur, Chennai. Atmospheric Environment, 203: 166-171. https://doi.org/10.1016/j.atmosenv.2019.02.013

[9]. Agarwal, A., Kaushik, A., Kumar, S., Mishra, R.K. (2020). Comparative study on air quality status in Indian and Chinese cities before and during the COVID-19 lockdown period. Air Quality, Atmosphere & Health, 13(10): 1167-1178. https://doi.org/10.1007/s11869-020-00881-z

[10]. Pant, G., Garlapati, D., Gaur, A., Hossain, K., Singh, S. V., Gupta, A.K. (2020). Air quality assessment among populous sites of major metropolitan cities in India during COVID-19 pandemic confinement. Environmental Science and Pollution Research, 27(35): 44629-44636. https://doi.org/10.1007/s11356-020-11061-y

[11]. Senthil, K.P. (2019). Improved prediction of wind speed using machine learning. EAI Endorsed Transactions on Energy Web, 6(23). https://doi.org/10.4108/eai.13-7-2018.157033

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 4, April 2023
Impact Factor: 7.301

[12]. Li, S., Song, S., Fei, X. (2011). Spatial characteristics of air pollution in the main city area of Chengdu, China. In 2011 19th International Conference on Geoinformatics, Shanghai, China, pp. 1-4. https://doi.org/10.1109/GeoInformatics.2011.5981082

[13]. Chang, Y.S., Lin, K.M., Tsai, Y.T., Zeng, Y.R., Hung, C.X. (2018). Big data platform for air quality analysis and prediction. In 2018 27th Wireless and Optical Communication Conference (WOCC), Hualien, Taiwan, pp. 1-3. https://doi.org/10.1109/WOCC.2018.8372743

[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. https://doi.org/10.1109/ACCESS.2019.2925082

[15]. Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., Zhang, B. (2019). A novel combined prediction scheme based on CNN and LSTM for urban PM 2.5 concentration. IEEE Access, 7: 20050-20059. https://doi.org/10.1109/ACCESS.2019.2897028

[16]. Ghoneim, O.A., Manjunatha, B.R. (2017). Forecasting of ozone concentration in the smart city using deep learning. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, pp. 1320-1326. https://doi.org/10.1109/ICACCI.2017.8126024

[17]. Sakarkar, G., Pillai, S., Rao, C.V., Peshkar, A., Malewar, S. (2020). Comparative study of ambient air quality prediction system using machine learning to predict air quality in smart city. Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India, pp. 175-182. https://doi.org/10.1007/978-981-15-3020-3_16

[18]. Liu, H., Li, Q., Yu, D., Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences, 9(19): 4069. https://doi.org/10.3390/app9194069

[19]. 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. Atmos Pollut Res, 11(8): 1451-1463. https://doi.org/10.1016/j.apr.2020.05.015

[20]. Gore, R.W., Deshpande, D.S. (2017). An approach for classification of health risks based on air quality levels. In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, pp. 58-61. https://doi.org/10.1109/ICISIM.2017.8122148

[21]. 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), Chennai, India, pp. 452-457. https://doi.org/10.1109/WiSPNET45539.2019.9032734

[22]. Asgari, M., Farnaghi, M., Ghaemi, Z. (2017). Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster. In ICCBDC 2017: Proceedings of the 2017 International Conference on Cloud and Big Data, London United Kingdom, pp. 89-93. https://doi.org/10.1145/3141128.3141131

[23]. Sharma, R., Kumar, R., Sharma, D.K., et al. (2019). Inferring air pollution from air quality index by different geographical areas: Case study in India. Air Quality, Atmosphere & Health, 12(11): 1347-1357. https://doi.org/10.1007/s11869-019-00749-x