Air pollution prediction with machine learning: a case study of Indian cities

1st Krishna B. Popat
Computer Science and Engineering
Nirma University
Ahmedabad, India
nilakrishna2004@gmail.com

2nd Dhruv S. Thakkar
Computer Science and Engineering
Nirma University
Ahmedabad, India
dhruvthakkar2511@gmail.com

3rd Jatin Varyani
Computer Science and Engineering
Nirma University
Ahmedabad, India
varyanijatin3@gmail.com

4th Namit K. Patel
Computer Science and Engineering
Nirma University
Ahmedabad, India
namit.k.patel@gmail.com

5th Tanmay Trivedi
Computer Science and Engineering
Nirma University
Ahmedabad, India
email address or ORCID

***Abstract*—By utilising the strength of Machine Learning (ML) and Deep learning (DL) models and in-depth examination of data, the present study looks into the field of air quality prediction. Predicting the Air Quality Index (AQI) has become crucial for effective governance of the environment and safety for society due to growing worries about the world's worsening air quality. The study uses a sophisticated ML model to anticipate AQI values based on many input factors that have been painstakingly trained and fine-tuned. The dataset used in this research paper contains historical AQI data collected from multiple stations in the cities of Ahmedabad, Mumbai, Kolkata and Delhi. The data was preprocessed, clean and normalized.**

I. Introduction

Among the biggest threats to the well-being of people is air pollution. The World Health Organisation(WHO) says that 7000000 people are in danger of health problems as a result of contaminants in the air. The effect of climate change is mostly caused by air pollution from factors like industrialization and car emissions, with the release of carbon dioxide (CO2) being one of the greatest drivers of the phenomena [1]. The air quality index can help people prepare in advance for safeguards. Measurement of air pollution has been more popular recently since it significantly affects both human well-being and the equilibrium of the environment. Industries in India contribute to 51% of all air pollution, making them one of the main contributors.Air pollution can harm children's health over time. The growth in air pollution in New Delhi is due to increased vehicle emissions, fossil fuel burning at power plants and local industry, and field burning by farmers in neighboring states [2]. A precise and up-to-date AQI index is needed to manage air pollution. India is one of the most polluted countries in the world, with rapid pollution growth causing environmental damage in several cities [3]. Nine out of 10 cities in the list of most polluted cities by WHO are Indian cities. Delhi is the most polluted city worldwide and it has the most amount of PM10 [4]. Besides all of that indoor air pollution has also become a major concern for the well-being of people [5]. Ninety per cent time the people who spend time indoors are exposed to various pollutants from refrigerators and various other electronic devices ultimately which can affect their work efficiency[6]. There are 0.2 billion individuals in India who use fuel for cooking, 47% of them use charcoal, 29.6% choose LPG, and 9.9% use dung from cow cakes. Kerosene is used by 2.4%, biogas by 0.9%, electricity by 0.2%, and various energy sources by 0.4% of people [6].

A. Why AQI

Using one combined index, a colour code, and general classifications of air quality levels (good, moderate, bad, etc.), the AQI was developed to assist in communicating to the public the seriousness of air pollution levels for numerous pollutants, the hazards they bring, and necessary precautions to take. Data on air quality is easier to convey to the general public thanks to AQI. Based on the present air quality situations, it allows users to make educated judgements regarding outdoor pursuits, such as working out or spending a lot of time outside [7]. AQI is used by authorities and environmental groups to develop and carry out air quality control plans. It is a crucial tool for determining how well air quality management methods are working and for implementing the required modifications to protect the public's well-being. When there are periods of bad air quality, such as when there is a lot of pollutants or a catastrophic fire, the AQI is very helpful in preparing emergency responses. It assists governments in effectively

allocating assets, distributing health alerts, and coordinating evacuation preparations [8].

B. How AQI is Calculated

The Air Quality Index (AQI), which includes several contaminants that may have an impact on the well-being of people, is an indicator of the total air quality in a particular area. The AQI is determined by a number of processes depending on quantities of important contaminants, such as particulate matter (PM2.5 and PM10), ground-level ozone (O3), sulphur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide (CO). Initially using predetermined boundaries corresponding to various intensity ranges, specific levels of pollutants are acquired and transformed into sub-indices. The most significant sub-index value for each of these contaminants determines the spot's total AQI. This method takes into consideration the many negative effects on well-being brought on by various contaminants. The AQI score is then separated into groups, each denoting a different degree of medical concern, such as "Good," "Moderate," "Unhealthy," and beyond. The ultimate AQI score provides information on potential hazards to health connected to the present atmosphere and acts as a transparent and understandable interface for the general public, decision-makers, and medical experts. In addition to quantifying pollutants, this thorough index equips citizens and government officials to make well-informed decisions, ranging from altering everyday routines to putting into place efficient air quality management measures. The AQI breakpoints and categories (per average timings) given by the Central Pollution Control Board are shown in Fig. 1.

AQI Category (Range)	PM10 (24hr)	PM2.5 (24hr)	NO2 (24hr)	O3 (8hr)	CO (8hr)	SO2 (24hr)	NH3 (24hr)	Pb (24hr)
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
Moderately polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Fig. 1. The AQI breakpoints and categories(as per average timings)

The Impact of AQI on health for various levels is also provided in Fig. 2.

PM2.5 & PM10 :

PM2.5 & PM10, and the Air Quality Index (AQI) are intrinsically linked since these contaminants are essential in assessing air quality. While PM10 comprises somewhat

AQI	Associated Health Impacts
Good (0-50)	Minimal impact
Satisfactory (51-100)	May cause minor breathing discomfort to sensitive people.
Moderately polluted (101–200)	May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Poor (201-300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease.
Very poor (301-400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.
Severe (401-500)	May cause respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

Fig. 2. The Impact of AQI on health for various levels

bigger particles up to 10 micrometres in diameter, PM2.5 consists of minuscule particles having a diameter of 2.5 micrometres or less. Both PM2.5 and PM10 play important roles in AQI estimates, considerably influencing the total review of air quality. Car exhaust, manufacturing operations, and environmental factors like dusty winds and wildfires are only a few of the causes of PM2.5 and PM10. Once these particulates are in the atmosphere, they can pose major health dangers since they can enter the lungs deeply and cause problems with the heart and lungs. Due to their exceptionally tiny dimensions and low weight, such particles frequently linger in the environment for a longer period of time than bigger, denser particles. They are now more likely to be ingested by people as a result. It is known that airborne emissions that contain particles smaller than 2.5 microns are more harmful to people's health than other pollutants. Particles like this can readily be inhaled into the trachea, where they can have detrimental effects on lung function and respiration[10]. Both PM 2.5 & PM 10 quantities are assessed in the overall picture of AQI, and their values are taken into account while calculating the index. Increased concentrations of such tiny particles levels result in an increase in AQI, which denotes worse air quality.

Ozone(O3) :

The existence of ozone appears as a key component in the assessment of the Air Quality Index (AQI), having a substantial influence on both human well-being and the surroundings. A notable airborne pollutant is atmospheric ozone (O3), which is produced when volatile organic compounds (VOC) and nitrogen oxides (NOx) combine chemically in the presence of sunshine. Ozone does not release directly as certain pollutants do; instead, it originates

as a result of additional reactions. Cities frequently experience higher levels of ozone due to the presence of industrial activity and substantial vehicular congestion. Given its immediate impact on health, ozone must be taken into account when calculating AQI. Ozone depletion is linked to breathing challenges, flare-ups of asthma, and various other illnesses related to breathing. Ozone also offers difficulties with the environment, damaging habitats and plants.

Carbon Monoxide :
With respect to its possible damage to both the health of humans and the ecosystem, carbon monoxide (CO) has a substantial impact on the Air Quality Index (AQI). CO is a colourless, odourless gas that is mostly produced when petroleum and coal are used inefficiently in automobiles, factories, and home heating systems. Several factors make incorporating it in AQI computations essential. Initially, carbon monoxide interacts with the capacity of blood to transport oxygen, reducing the amount of oxygen that reaches critical cells. This presents a major health danger, particularly in metropolitan locations with substantial vehicular traffic where levels of carbon monoxide might increase. The consequences of increased CO levels are especially harmful to people who have cardiac or respiratory disorders. Furthermore, CO is an antecedent to ozone layer production, another important AQI component. Ozone may affect the respiratory system and contribute to the creation of haze when it exists in excess.

Sulphur Dioxide :
AQI is greatly impacted by sulphur dioxide (SO2). As a key precursor to fine particulate matter (PM2.5), which is mostly released during the burning of fossil fuels, notably during industrial activities and the production of electricity, SO2 contributes to the development of suspended particles that can profoundly infiltrate the lungs. In terms of AQI, higher SO2 levels lead to higher index values, which indicate worse air quality. It is widely known that increased SO2 levels have detrimental impacts on breathing. brief exposure has been related to vision and respiratory system irritation, especially in people who already have illnesses like asthma.

Meteorological Factors :
Meteorological conditions exert a profound influence on the Air Quality Index (AQI), orchestrating a symphony of factors that intricately shape the presence and impact of pollutants in the atmosphere. Among these factors, wind speed emerges as a key player, dictating the dispersal or concentration of pollutants. Low wind speeds can lead to stagnation, trapping pollutants and elevating AQI levels, while higher speeds facilitate dispersion, potentially ushering in improved air quality. Temperature intricately weaves its effects into the AQI tapestry. Warmer temperatures foster the genesis of ground-level ozone, a pivotal AQI component, and heighten the volatility of select pollutants. Temperature inversions, marked by a warm air layer trapping cooler air near the ground, can compound pollution concentrations by impeding the vertical movement of pollutants. Humidity, a subtle yet influential actor, plays a dual role. High humidity contributes to particulate matter formation through processes like hygroscopic growth, while low humidity fosters dry conditions, potentially resulting in the suspension of dust and particulate matter in the air. Precipitation, the benevolent force in meteorology, emerges as a natural purifier for air quality. Rain orchestrates a cleansing act, removing pollutants through wet deposition and alleviating their concentration, thereby improving AQI levels. However, rain's impact varies, with certain pollutants experiencing resuspension during precipitation events.

II. Related Work

Air pollution analysis plays a crucial role in assessing pollution levels in different urban areas. Machine learning algorithms have gained significant prominence for predicting, forecasting, and controlling pollution levels. In a study by Shaban et al. (2016) [9], three machine learning algorithms, namely the Support Vector Machine (SVM), M5P Model Tree, and Artificial Neural Network (ANN), were explored for forecasting ground-level ozone, nitrogen dioxide, and sulfur dioxide levels. The research revealed that the M5P algorithm delivered the best forecasting performance when different features were incorporated into multivariate modelling.

Wang et al. (2020) used real-time sensor data from downtown Toronto, Canada to forecast PM2.5 and Black Carbon levels using Land Use Regression (LUR), Artificial Neural Network (ANN), and gradient boost machine learning algorithms [10]. LUR fared well with a smaller dataset, whereas ANN and the gradient boost technique improved with larger datasets.

Further study by Wood [11] (Wood, 2022) utilized ML and DL algorithms to forecast air quality. Support Vector Regression (SVR) predicted Dallas County CLAB best in 2019 and 2020.

Masmoudi et al. (2020) [12] proposed multi-target regression techniques to predict multiple air pollutant concentrations. Their method combined multi-target regression and the Random Forest paradigm, resulting in the Ensemble of Regressor Chains-guided Feature Ranking (ERCFR) framework.

Researchers have also used deep learning to predict air quality. LSTM, a long short-term memory model, predicts Air Quality Index in research by Li et al. (2017), Xayasouk et al. (2020), Zhao et al. (2019), and Liu et al. (2020) [13]. RNN, LSTM, and GRU were also used by Athira et al. (2018) to forecast air quality.

In a study by Chang et al. (2020) [14], the Aggregated Long Short-Term Memory (ALSTM) model was introduced, which integrated data from various sources to enhance prediction accuracy.

Notably, Nath et al. (2021) [15] compared deep learning methods with statistical methods, concluding that the Holt-Winters statistical model performed better in certain scenarios. Meanwhile, Bekkar et al. (2021) examined various deep learning algorithms, revealing that the hybrid CNN-LSTM

technique outperformed classical models on the UCI Machine Learning Repository.

In addition to deep learning, traditional time series analysis models such as Auto Regressive Integrated Moving Average (ARIMA) have been used for air pollution prediction (Bekkar, et al., 2021) [16]. The study suggested that LSTM is superior to ARIMA in forecasting pollution levels.

Lastly, a study by Liu and You (2022) [17] examined Beijing's Air Quality Index (AQI) from January 2019 to November 2021. They applied ARIMA and LSTM models for short-term AQI prediction.

Air pollution is a global problem that has been the subject of many research. Researchers created a model to forecast the Air Quality Index (AQI) for particular regions while taking pollutant levels into account in one study [18]. In a different study, [19], an accurate model for predicting Delhi's air quality using supervised algorithms and machine learning was presented. They made use of historical information from pollution control boards and websites.

Researchers created a system in [20] that uses a variety of algorithms to assess pollution in particular areas. This system analyses pollution patterns to forecast future levels of pollution and the AQI by predicting them based on basic characteristics. citebarthwal2018internet included building a mobile AQI station to track and gather information on different air contaminants. This made it possible to measure the AQI using statistical models for predicting in places where air quality data is scarce.

A system with two functions was proposed in another study [21]: gathering information on air contaminants and forecasting AQI levels for these pollutants. The Extreme Value Distribution (EVD) approach was introduced by [22] to predict the exceedance probability and return periods for extreme pollution. The authors of [23] presented a k-nearest Neighbour method that showed remarkable performance. A probability distribution was used in [24] to analyse and forecast extreme occurrences, and an Internet of Things (IoT) system was created to monitor air pollution in Delhi NCR.

The topic of air quality prediction has seen an increase in interest in recent years due to the pressing need to address environmental and public health challenges. Scholars have investigated diverse methodologies and tactics to precisely forecast air quality levels, including sophisticated machine learning algorithms and inventive strategies. Using Long Short-Term Memory (LSTM) networks is one such method, as seen in the research that was presented at the 2022 International Conference on Data Science and Information Technology [25]. Their study suggests using the CS-LSTM model as the foundation for a prediction tool for air quality.

Furthermore, a study on LSTM-based methods for Air Quality Index (AQI) prediction was presented at the 2019 IEEE Joint International Information Technology and Artificial Intelligence Conference [26]. A different study, given at the 2022 IEEE International Conference on Computer and Communications, presents a forecasting technique for AQI based on the Optimized GWO-LSTM Neural Network [27].

Additionally, studies have looked into how particular contaminants affect the environment and public health. An important contribution to environmental research is provided by an article in Current Opinion in Chemical Engineering that examines the effects of nitrogen oxides on the environment and suggests materials based on manganese for NOx abatement [28].

Studies have looked into the application of many machine learning algorithms concurrently. For example, a paper in Applied Sciences explores the use of machine learning methods for air quality index and air pollutant concentration prediction [29]. Similarly, demonstrating the breadth of machine learning applications, an article in Sensors offers a bionic electronic nose based on an MOS sensors array and machine learning algorithms for wine quality identification [30].

The application of diverse machine learning algorithms has been pivotal in these endeavors. Machine learning techniques, including the K-nearest neighbor method, have been instrumental in air quality index prediction, showcasing the breadth of methodologies utilized in this burgeoning field [28]. Additionally, the utilization of Random Forests, a versatile machine learning approach, has further enriched the spectrum of predictive modeling techniques, as detailed by Leo Breiman in the seminal paper published in Machine Learning in 2001 [31].

The landscape of air quality prediction extends further with the application of machine learning algorithms like Random Forests, as detailed by Leo Breiman in the seminal paper published in 2001 [31]. Additionally, diverse applications of machine learning, such as feature reduction for employee turnover prediction models, underscore the versatility and relevance of these techniques across various domains [32].

An important step has been taken in the direction of comprehending and reducing the negative effects of air pollution on society with the introduction of machine learning algorithms and environmental science principles into the multidisciplinary approach to air quality prediction. Through the utilization of sophisticated methodologies and interdisciplinary teamwork, scientists persist in creating precise and effective models for predicting air quality, thereby promoting a more salubrious environment for everybody.

System Model for AQI Prediction

In this section, we present a system model for predicting the Air Quality Index (AQI) of four different cities in India over the period from 2017 to 2023. The system consists of the following components:

1) **Sensors and Data Collection:**
   - We have a set of sensors denoted as $\{s_1, s_2, s_3, \ldots, s_n\} \in S$, where n represents the total number of sensors.
   - These sensors are responsible for measuring individual AQI components using satellite imaging techniques. The data collected by these sensors is represented as $S_i(t)$, where i is the sensor index, and t represents the time.

2) **Data Preprocessing:**
   - The data collected by the sensors is preprocessed to handle missing values, outliers, and noise. Let $D_i(t)$ represent the preprocessed AQI data from sensor s_i at time t.
3) **Machine Learning and Deep Learning Model (M):**
   - Our prediction model, denoted as M, uses time series analysis techniques to predict the future AQI pattern for each city.
   - The model takes historical AQI data from all sensors as input and generates predictions for the AQI of each pollutant component.
   - The model parameters, such as weights, biases, and hyperparameters, are optimized during training.
4) **Prediction Output:**
   - The output of the prediction model is represented as $\hat{AQI}_i(t)$, which is the predicted AQI value for city i at time t for a specific pollutant component.
5) **Overall AQI Prediction:**
   - To obtain the overall AQI prediction for each city, we combine the predictions of individual pollutant components using a suitable aggregation method. Let $AQI_i(t)$ represent the overall AQI prediction for city i at time t.

Mathematical equations to describe the components:

- **Data Collection:** $S_i(t)$: Measured AQI data from sensor s_i at time t.
- **Data Preprocessing:** $D_i(t)$: Preprocessed AQI data from sensor s_i at time t.
- **Machine Learning Model (M):** M: The machine learning model for AQI prediction, which takes $D_i(t)$ as input and outputs $\hat{AQI}_i(t)$.
- **Overall AQI Prediction:** $AQI_i(t)$: The overall predicted AQI for city i at time t.

By following this structure and providing the relevant mathematical equations, you can effectively convey the system model for predicting the AQI of the four different cities in India.

III. Proposed Framework

Developing high-quality services to address diverse healthcare needs stands as a pivotal objective. This section introduces a system architecture aimed at enhancing functionality, optimizing resource utilization, and refining tools within the medical domain. It comprises 2 distinct layers: the sensor layer and, the AI layer, These layers play vital roles in data collection, preprocessing, and prediction.

A. Sensor layer

The Sensor Layer serves as the cornerstone of our AQI prediction system, comprising an array of sensors denoted as $\{s_1, s_2, s_3, \ldots, s_n\} \in S$. These sensors are strategically distributed across the four Indian cities and utilize satellite imaging techniques to measure key components contributing to the AQI, including pollutants, particulate matter, and meteorological variables. Daily data collection is their primary function, but the data undergoes rigorous preprocessing to address missing values, outliers, and noise, ensuring data integrity. The preprocessed data, represented as $D_i(t)$ for each sensor s_i at time t, becomes the reliable input for our prediction model M. This Sensor Layer plays a pivotal role in delivering accurate AQI predictions by providing high-quality, refined data for further analysis.

B. AI Layer

1) Dataset description: The data consists of four .csv files which contain data from four major cities in India. These cities have different geology, and weather, and are located at large enough distances so that the results do not overlap. The four cities are Delhi, Mumbai, Ahmedabad and Kolkata. The dataset contains the AQI values of 6 pollutants, which are pm25, pm10, o3, no2, so2 and co, collected daily from the year 2017 to 2023.

2) Data preprocessing: The collected data is then stored in a comma-separated file(CSV). The first step is to convert the individual date, month and year into a proper time format. Then processing includes checking the NaN values. These values are filled with mean taken as a group of the last 6 months.
After the preprocessing, we assign the value of AQI as the highest value from the AQI of all the pollutants. Preprocessing also involves removing skewness from the data. Further, we visualize the data using different graphs:

1. Histogram: - The histogram illustrates the data distribution, showing how values are spread across different ranges.

2. Boxplot - The boxplot provides insights into the presence of outliers and summarizes the distribution's central tendency and variability.a

3. Q-Q Plot (Quantile-Quantile Plot): - The Q-Q plot assesses the data's adherence to a theoretical normal distribution by comparing the observed distribution to the expected quantiles of a normal distribution.

4. Autocorrelation Plot - The autocorrelation plot depicts how the variable correlates with its past values at various time lags, aiding in the identification of patterns and seasonality in time series data.

After the data analysis, the cleaned data is then passed onto the AI model. We have trained the data on multiple Machine Learning and Deep Learning models and used the best-performing model as our final output. Evaluation is also done on test data and is done on the basis of accuracy and mean-squared error of the predicted results. More detailed data analysis is explained in the results and discussion section.

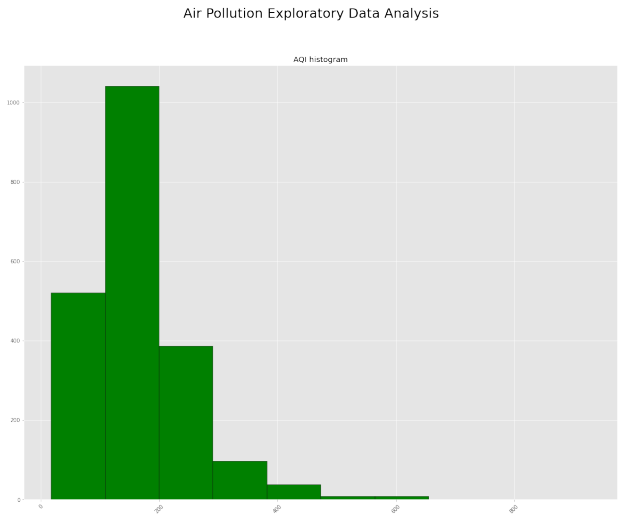

Fig. 3. Histogram

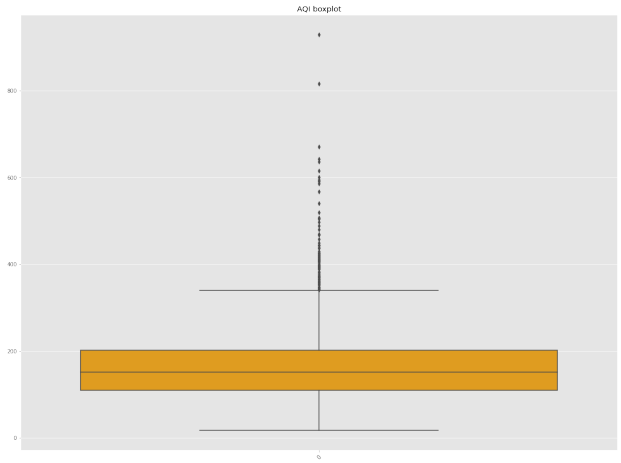

Fig. 4. Boxplot

IV. Results and Discussions

A. Experimental Setup

The proposed approach uses a time series analysis. The regression is performed on Jupyter Notebook a Python Integrated Development Environment(IDE). Libraries like Pandas, Numpy, Sklearn, Seaborn, Matplotlib, etc. Pandas have been used for preprocessing. Pandas have been used for data cleaning, and efficiently represent data in a structured manner. Numpy has been used for numerical calculations for eg. filling in the NaN values. The scikit-learn library is used for model selection and evaluation, regression and data preprocessing. Further, we have implemented Explainable Artificial Intelligence(XAI), which makes it easier to comprehend the models we have used. It also highlights the important features and their impact on the model's prediction for the selected data point.

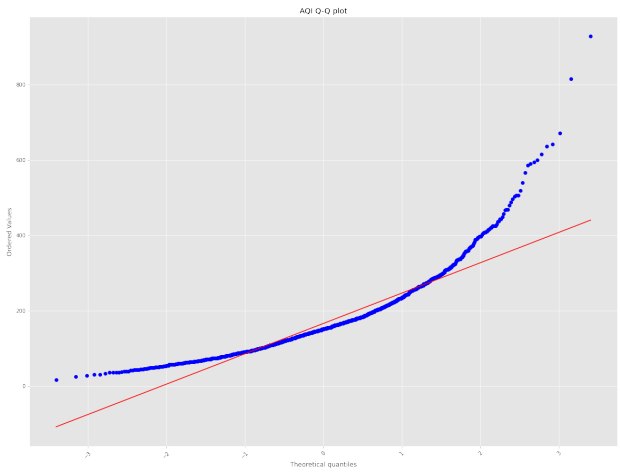

Fig. 5. AQI Q-Q plot

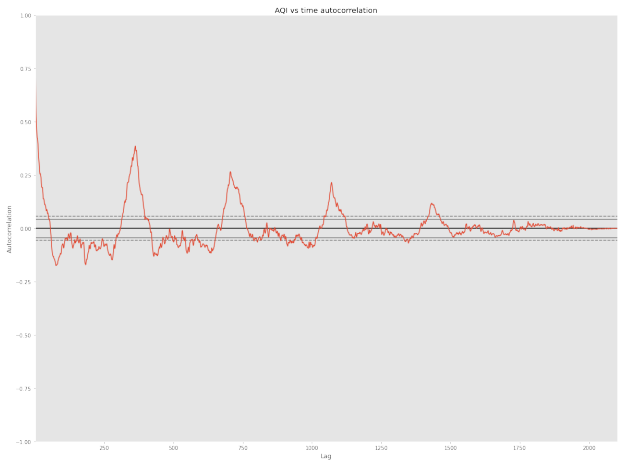

Fig. 6. Auto Correlation plot

B. Satellite Observations using ARCGIS

Fig 21 graphic depicts the spatial arrangement of carbon monoxide column density, providing insight into the complex characteristics of the atmosphere. The different shades and strengths indicate areas with different levels of CO concentrations, providing valuable information about how pollutants spread.

Fig. 22 depicts the temporal patterns, this figure displays the monthly dispersion of carbon monoxide concentrations. Every curve represents the flow of pollutants throughout the course of the year, providing a dynamic viewpoint that is essential for comprehending the seasonal impacts on air quality.

Fig. 23 graphic displays the spatial arrangement of nitrogen dioxide (NO2) concentration, specifically focusing on the distribution of NO2 column density. The visual spectrum is crucial for identifying and understanding places with high levels of concentration, which is vital for comprehending the intricacies of air quality.

Fig. 24 represents the monthly fluctuations of nitrogen dioxide levels, capturing the subtle changes throughout time.

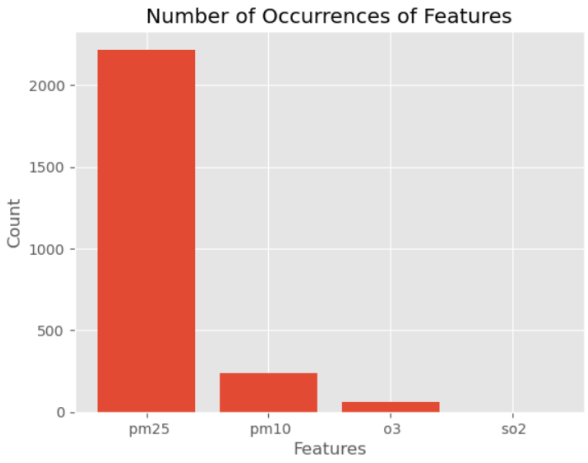

Fig. 7. Bar Graph for Number of Occurences of features

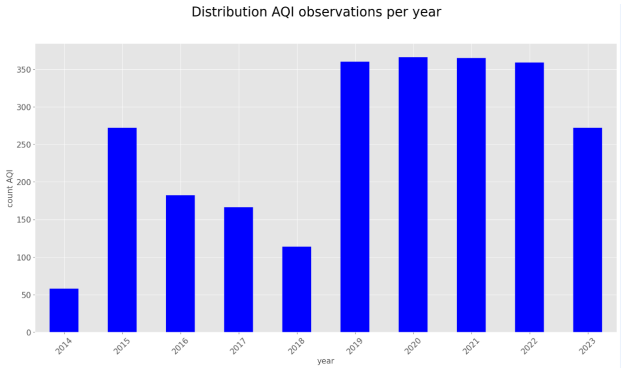

Fig. 8. Bar Graph for distribution of AQI Observations per year

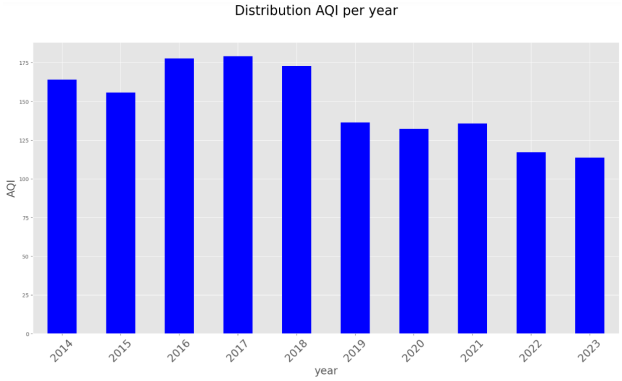

Fig. 9. Bar Graph for distribution of AQI per year

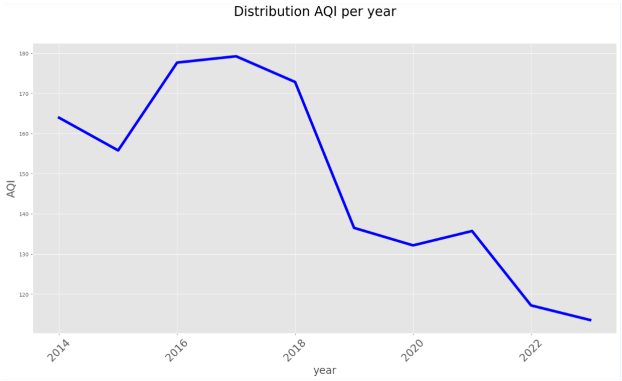

Fig. 10. Line plot for distribution of AQI per year

The graphic depiction reveals trends that lead to a full comprehension of NO2 fluctuations throughout the year.

Fig.25 visually explores the composition of the atmosphere by depicting the amount of sulphur dioxide present in the air. The utilisation of colour-coded representation facilitates the identification of regions with elevated levels of SO2, hence providing crucial data for air quality studies.

Fig 26 describes the monthly fluctuations of sulphur dioxide levels, with time being the main focus. The visual story effectively depicts the rhythmic patterns, facilitating a nuanced understanding of the evolution of SO2 throughout the year.

Fig.27 provides a visual representation of the distribution of ozone in various places, specifically highlighting the column density. The different hues serve as a visual guide to understanding the dispersion of this crucial atmospheric element.

Fig. 28 transforms time into a canvas, depicting the monthly variation of ozone levels. The many levels of intensity lead the spectator through the subtle changes, providing a significant understanding of the seasonal changes in atmospheric ozone.

In Fig. 29 the image provides a close-up view of particulate matter, specifically focusing on the geographical distribution of PM2.5. The different levels of intensity reveal areas with increased concentrations, providing a basis for comprehending the specific effects of fine particle pollution in certain locations.

Fig 30 image provides a close-up view of particulate matter, specifically focusing on the geographical distribution of PM2.5. The different levels of intensity reveal areas with increased concentrations, providing a basis for comprehending the specific effects of fine particle pollution in certain locations.

C. Preprocesssing Results

This section shows the model training and performances of different regression models, which are compared to find which one is the best model to use. Along with accuracy mean-squared error is also calculated and compared. Models that we have implemented are Linear Regression, Long short-term memory(LSTM), Artificial Neural Networks(ANN), and Support Vector Regressor(SVR). The fig *. shows a comparative analysis of the accuracies of all the machine learning and deep learning models. Fig. 1 is a bar plot which takes two arguments var and data. The x-axis represents the unique values of the var variable, and the y-axis represents the mean AQI for each group.

Fig. 2 shows a Boxplot that visualizes the distribution of the number of observations or counts of the Air Quality Index (AQI) based on a specific categorical variable in a dataset. This function takes two arguments: data, which is a Pandas DataFrame containing your dataset, and var, which represents

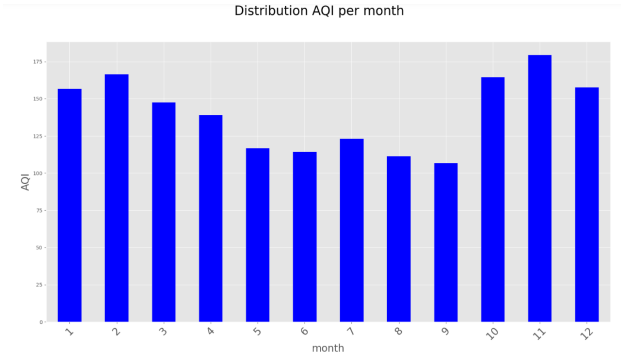

Fig. 11. Bar plot for distribution of AQI per Month

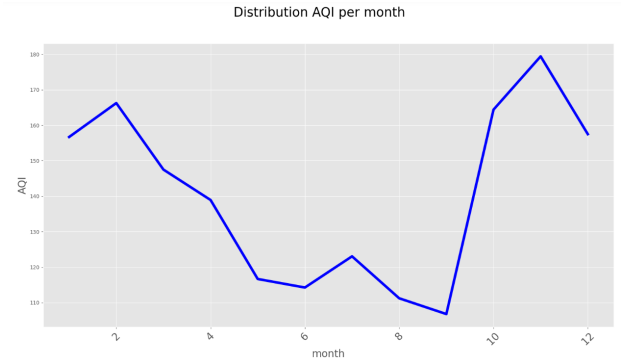

Fig. 12. Line plot for distribution of AQI per Month

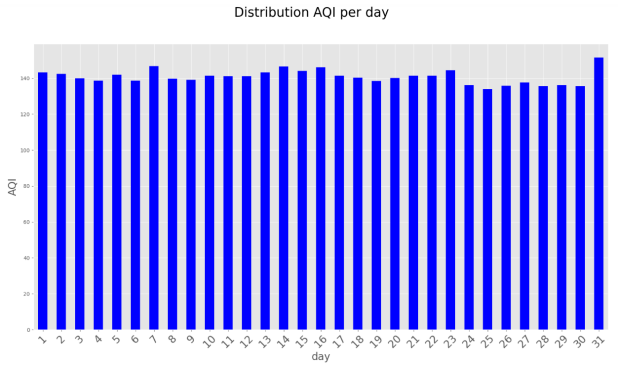

Fig. 13. Bar plot for distribution of AQI per Day

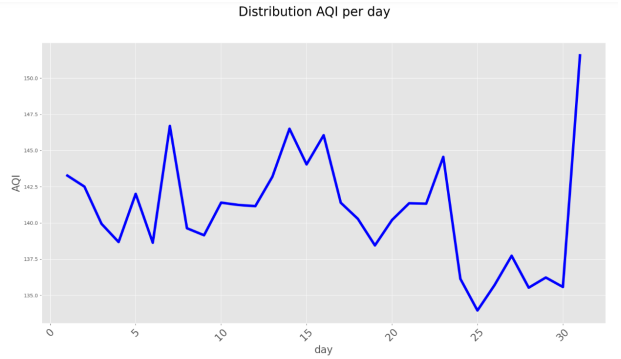

Fig. 14. Line plot for distribution of AQI per Day

the variable by which you want to group the data. It uses the group-by method to group the data by the values in the specified var variable and then counts the number of AQI observations within each group. The resulting counts are displayed as bars in the bar plot, with the x-axis representing the unique values of the var variable and the y-axis showing the count of AQI observations for each group.

Fig. 3 visualizes the distribution of the mean Air Quality Index (AQI) over different categories or values of a specific variable in a dataset.

Fig. 4 creates a time series line plot to visualize actual AQI values and their corresponding predictions over time.

To find out where pollution is found in Indian cities, we analyzed the dataset. The dataset's distribution of several pollution attributes was displayed in a bar graph plot shown in Fig. 5. The image provided a thorough grasp of the common pollutants by showing the frequency of several pollutants, including pm25, pm10, o3, no2, so2, and co. In addition, we extracted important temporal information like year, month, and day by processing the date column. This phase was essential for the model to identify any temporal patterns or dependencies in the data, which improved the models' overall predicted accuracy when used in the research.

We used a bespoke barplot method to better understand the distribution of Air Quality Index (AQI) observations shown in Fig. 6. The function made it easier to see how the distribution of AQI observations changed over a certain time. For example, we were able to determine the distribution of AQI observations throughout different years by utilizing the 'year' variable. The resulting bar plot gave a thorough picture of the distribution of air quality data across time by showing the number of AQI readings for each year. The AQI data from several years was made easier to detect by the clear visual depiction, which advanced our knowledge of the temporal dynamics of air pollution in Indian cities.

Utilizing a customized barplot function, we examined the Air Quality Index (AQI) distribution with respect to a certain variable. We were able to identify any temporal trends or patterns in the air quality readings by using the 'year' variable to gain insight into the average AQI for each year. The resulting bar plot made it possible to see the average AQI for each year clearly and to have a thorough knowledge of how the quality of the air has changed over time. This graphic depiction was helpful in emphasizing any significant fluctuations in air pollution levels, which added to the thorough examination of the dynamics of air quality in Indian cities. Similarly in Fig. 9, the same is shown for the given data per month and Fig. 11 shows the same per day.

A custom line plot function visualized the average Air Quality Index (AQI) over time for a specified variable. Using the 'year' variable, we might show average AQI values as a continuous trendline, revealing air quality trends over time. The line plot shown in Fig. 8 illustrates AQI value fluctuations over time, helping explain long-term air pollution changes.

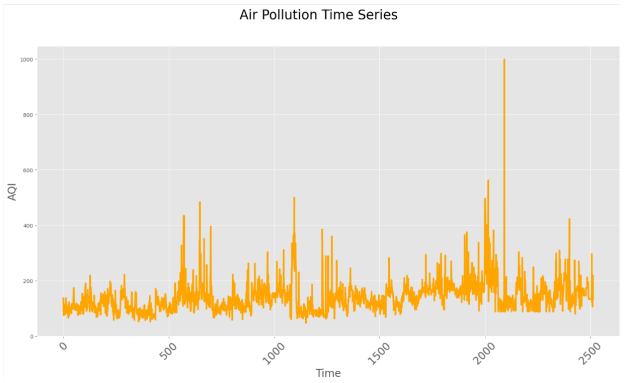

Fig. 15. Time Series plot Air Pollution

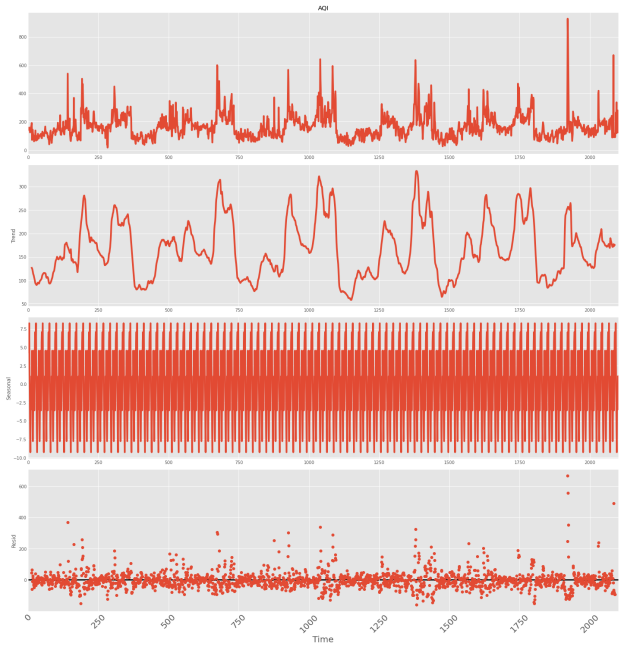

Fig. 16. Additive Model to Decompose Time Series

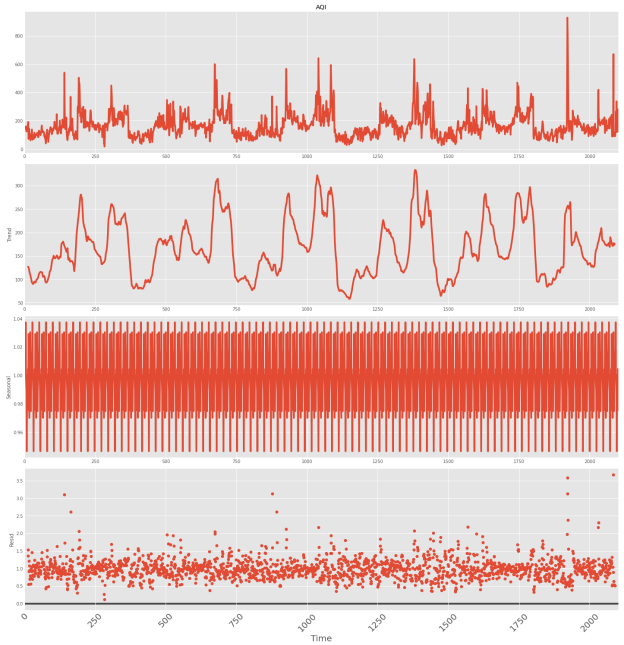

Fig. 17. Multiplicative Model to Decompose Time Series

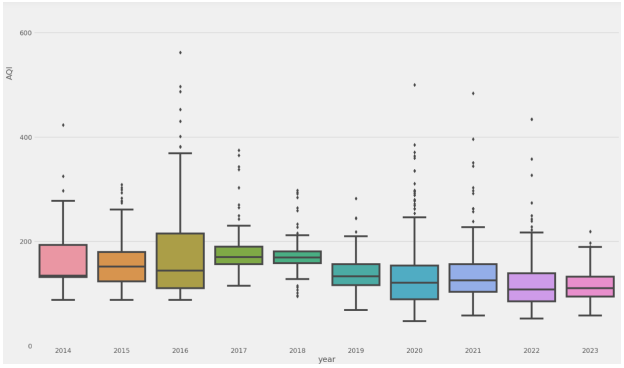

Fig. 18. Box Plot of AQI values over years

This visualization identified air quality dynamics patterns and shifts, helping analyze air pollution trends in Indian cities. Similarly in Fig. 10, the same is shown for the given data per month and Fig. 12 shows the same per day.

Using the custom time series plot function tsplot, the Air Quality Index (AQI) temporal dynamics could be seen across time. This tool showed AQI changes over time, helping identify dataset patterns, trends, and seasonality. By setting 'period' to 'all,' the function presented the whole AQI time series, revealing the dataset's temporal pattern. This visualization revealed long-term trends and patterns, helping analyze air pollution variances in Indian cities over time. This is shown in Fig.13.

The stats models library's seasonal decompose function was used to study Air Quality Index (AQI) time series data seasonal patterns and trends shown in Fig. 14 . The function used a 24-hour additive model to decompose the AQI time series into trend, seasonal, and residual components also a 24-hour multiplicative model was used to decompose the AQI time series into trend, seasonal, and residual components shown in Fig. 15. The visualization revealed seasonal variations and fluctuations in the data, revealing any repeating patterns or periodic fluctuations that affect Indian city air pollution dynamics. The AQI time series' systematic changes and seasonal trends were revealed by this study, helping to understand the variables affecting air pollution levels over time.

We have also analyzed Air Quality Index (AQI) data over time using a box plot shown in Fig. 16. The box plot displays the distribution of AQI values across years, revealing the dataset's central tendency, dispersion, and outliers. Each plot box shows the interquartile range (IQR) of AQI readings for a single year, with a line indicating the median. The map also shows any notable data deviations or anomalies, which may help explain air quality trends over time. The box plot helps identify air quality changes over time, helping

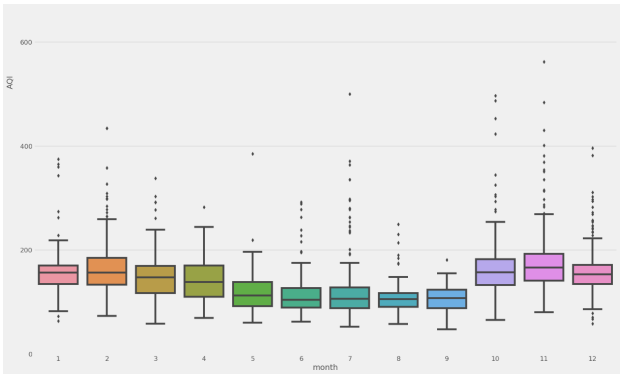

Fig. 19. Box Plot of AQI values over months

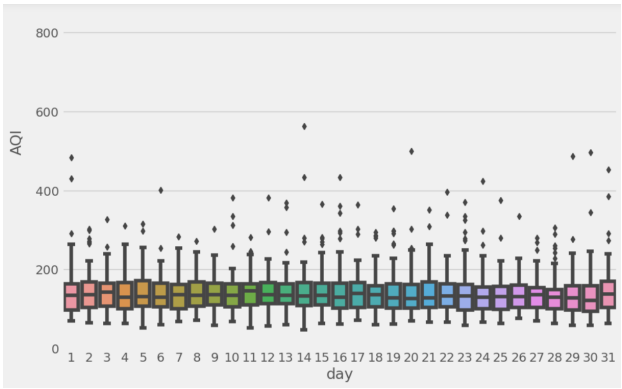

Fig. 20. Box Plot of AQI values over days

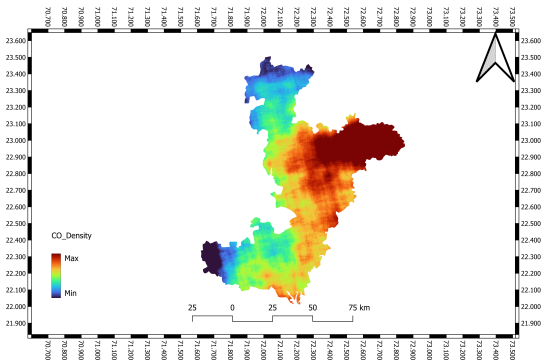

Fig. 21. Carbon Monoxide Column Density

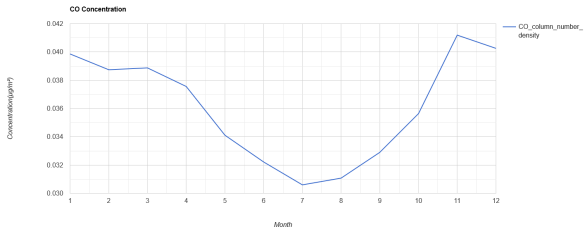

Fig. 22. Carbon Monoxide Monthly Distribution

environmental policymakers and stakeholders comprehend air pollution dynamics. The same Boxplots are plotted that display the distribution of AQI values across months and days which is shown in Fig. 17 and 18 respectively.

We have plotted a time series plot of the average Air Quality Index (AQI) throughout a period. The code initially aggregates AQI values by date to build a data frame with the mean AQI for each date. The time series plot shown in Fig. 19 shows average AQI swings and trends. The figure shows air quality patterns, seasonal fluctuations, and long-term trends, revealing air pollution's temporal dynamics. The time series' visual depiction helps identify recurring trends and abnormalities in air quality changes, which can help develop successful environmental policies and actions. The plot also helps stakeholders understand temporal fluctuations in air quality data and design informed air pollution management and mitigation strategies.

D. Methodology

For training the dataset we are using four models, which can be used for time series analysis, such as LSTM, LR, ANN and SVR. We will be discussing more about the architecture of these algorithms in this section.

1) Long Short-Term Memory (LSTM): Long Short-Term Memory, abbreviated as LSTM [33], is a specialized recurrent neural network (RNN) architecture designed to address the challenges of capturing and predicting sequential data, making it particularly suited for time series analysis. At its core, an LSTM unit operates through intricate mathematical computations, facilitating the selective learning and storage of temporal dependencies. The key mathematical components of an LSTM unit involve gate mechanisms that regulate the flow of information, including the input gate (i_t), the forget gate (f_t), and the output gate (o_t). These gates control the update of cell state (C_t) and the hidden state (h_t), ensuring that relevant information is retained and irrelevant information is discarded.

2) Linear Regression (LR): Linear Regression, a fundamental and interpretable machine learning model, is rooted in elementary mathematical principles. It follows the concept of fitting a linear equation to the data to model the relationship between a dependent variable (Y) and one or more independent variables ($X_1, X_2, \ldots, X_n$). The core mathematical formulation of linear regression can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

In this equation, Y represents the dependent variable, $X_1, X_2, \ldots, X_n$ denote the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients to be estimated, and ϵ is the error term. The primary objective of linear regression is to determine the optimal coefficients that minimize the sum of squared residuals, resulting in the best linear fit to the given data.

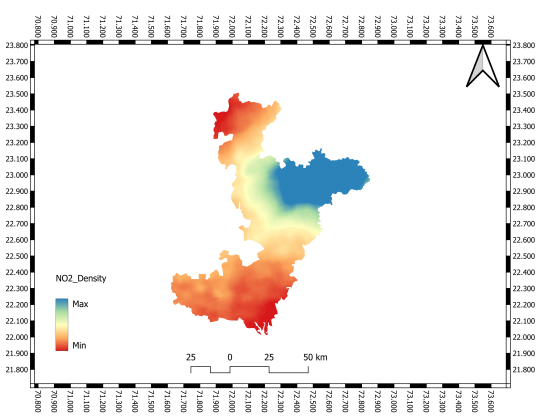

Fig. 23. Nitrogen Dioxide Column Density

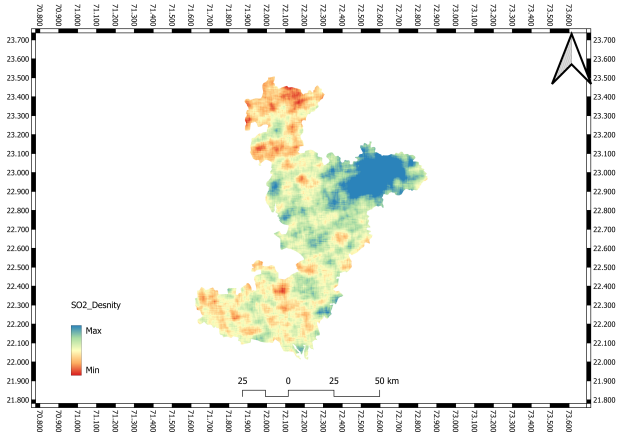

Fig. 25. Sulphur Dioxide Column Density

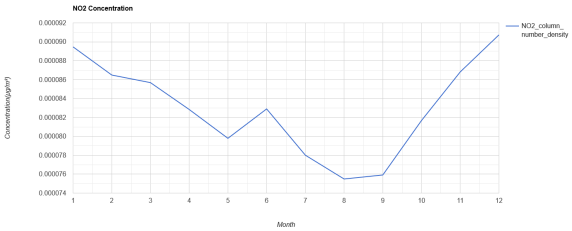

Fig. 24. Nitrogen Oxide Monthly Distribution

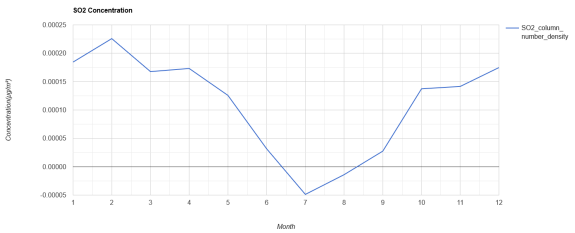

Fig. 26. Sulphur Dioxide Monthly Distribution

3) Artificial Neural Network (ANN): Artificial Neural Networks, a cornerstone of deep learning, consist of interconnected layers, including input, hidden, and output layers. The mathematical foundation of an artificial neural network involves feedforward propagation, characterized by linear combinations of inputs with corresponding weights, followed by activation functions introducing non-linearity. A simplified mathematical representation of an ANN is as follows:

$$y = f(W_2 \cdot f(W_1 \cdot X + b_1) + b_2)$$

Here, X denotes the input, W_1 and W_2 represent weight matrices, b_1 and b_2 are bias vectors, and $f(\cdot)$ signifies activation functions applied element-wise. Through the process of training, the neural network adapts the weights and biases to make predictions.

4) Support Vector Regression (SVR): Support Vector Regression (SVR) is an extension of support vector machines tailored for regression tasks. It is grounded in the concept of identifying the optimal hyperplane that best fits the data while minimizing prediction errors. Mathematically, SVR aims to find a function ($f(x)$) satisfying the conditions:

$$y - \epsilon \leq f(x) \leq y + \epsilon$$

In this formulation, y stands for the true value, $f(x)$ represents the predicted value, and ϵ denotes a user-defined margin of tolerance for prediction errors. The primary objective is to locate the hyperplane that maximizes the margin while ensuring that the majority of data points fall within this margin. otion

E. AI-based results

Our research delves into the LSTM(Long Short-Term Memory) model utilized for forecasting an Indian city's AQI (Air Quality Index) under the AI layer. The two 50-unit LSTM layers that make up the LSTM architecture are a primary focus. The first layer returns a sequence of outputs, while the second layer combines these outputs into a single prediction. The final AQI forecast is then generated using two dense layers, one with 25 units and the other with 1. We apply the Adam optimizer to the model compilation process, using the mean squared error (MSE) as the loss function—a popular option for regression tasks such as AQI prediction. To successfully capture temporal patterns in time series data, we trained our model for 15 epochs with a small batch size of 1.

For our model, preparing the data is crucial. In order to guarantee a pertinent input sequence for forecasting, we scale the input data using the scaled data variable and choose the latest 60 data points from the scaled data for the test data. The trained model is utilised to generate predictions, which are first scaled. To get the original AQI values, the predictions undergo inverse transformation. We utilise the Root Mean Square Error (RMSE) and the R-squared ($R\hat{2}$) score to evaluate the performance of the LSTM model, which is very important.

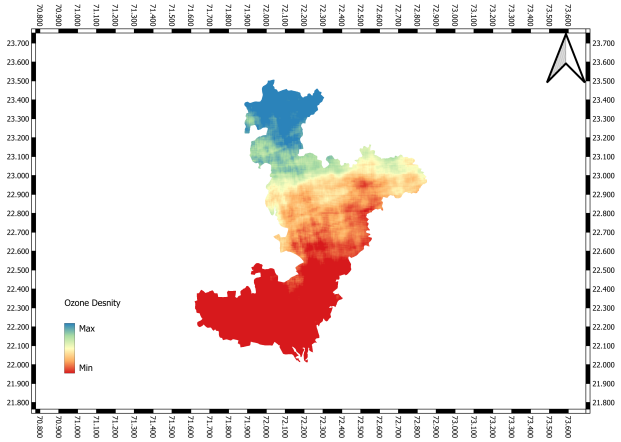

Fig. 27. Ozone Column Density

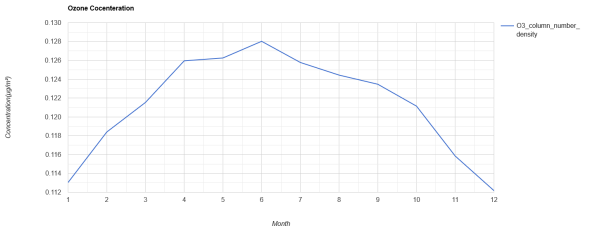

Fig. 28. Ozone Monthly Distribution

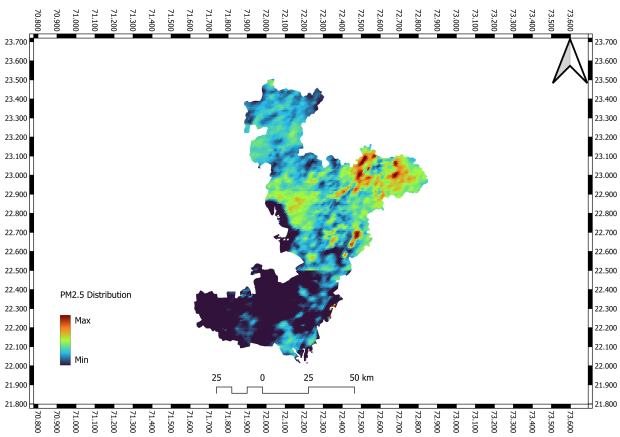

Fig. 29. PM2.5 Mean Value

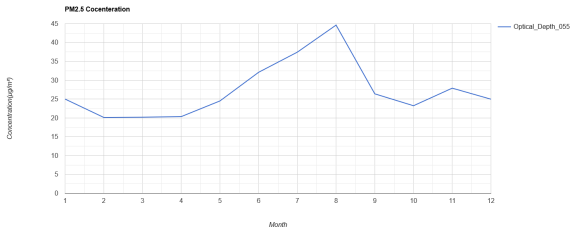

Fig. 30. PM2.5 Monthly Distribution

$R\hat{2}$ provides a thorough evaluation of the model's efficacy by measuring the percentage of variance in the AQI values that the model explains, whereas RMSE evaluates the average prediction error in the original AQI units.

The graph offers a visual depiction of our LSTM (Long Short-Term Memory) model's AQI (Air Quality Index) prediction process. It displays three different lines to depict the time evolution of AQI values: the real training data on one line, the validation dataset on another, and the model's AQI predictions on a third line. The model's capacity to forecast AQI values and capture complex temporal patterns is clearly communicated through the graph, which also allows for the useful comparison of predicted values with actual observations for validation. A wider audience may now more easily understand the intricate task of air quality prediction thanks to this graphic portrayal, which also increases the transparency of our study findings.

We used a reliable and traditional machine learning technique called linear regression in our research study that aims to estimate an Indian city's Air Quality Index (AQI). Without regard to the other models that are employed, this model is an essential part of our research. To make sure the dataset is clean and prepared for analysis, we start with extensive data pretreatment. This includes importing the data, dealing with null values, and converting the date column—a critical stage in temporal analysis—into a datetime format.

Our Linear Regression approach's primary function is to forecast the air quality index (AQI) using specific pollutant characteristics, such as PM2.5, PM10, O3, NO2, SO2, and CO. These contaminants are important to our research because they have a significant effect on air quality. The training and testing sets of the dataset are carefully separated, allowing us to evaluate the predictive performance of the model. Feature scaling is utilised to improve the convergence and efficacy of the model by guaranteeing that every variable has a uniform scale. Standard regression measures like R-squared (R^2) and Root Mean Squared Error (RMSE) are used to assess the accuracy of the model after it has been trained. Our thorough grasp of AQI dynamics is enhanced by the results of the Linear Regression model, which also provides insightful information on trends and patterns in air quality. The model has a really good performance and gives an accuracy of 0.953638 and RSME given by the model is 0.953638.

Our ANN model, specifically designed for regression analysis, had an input layer with 12 units and an activation function called Rectified Linear Unit (ReLU). A hidden layer with eight units and an additional ReLU activation function came next. The output layer had a single unit and was made for accurate regression analysis. The ANN was constructed with the mean squared error (MSE) loss function and the Adam optimizer, providing a strong foundation for model training in terms of model setup.

In order to improve the readability of our study and the assessment of our ANN model's predictive power, we painstak-

ingly created a graphic depiction of the AQI predictions. The temporal history of AQI values was successfully highlighted in Fig. 31, which made it possible to evaluate the predicted accuracy of the model rigorously. Three different lines were crucial to this portrayal. Direct comparisons between the model's predictions and the observed values were made easier by the train line, which graphically represented the actual AQI values taken from the training dataset. The Validation line provided a strong validation dataset by representing AQI values obtained during a separate validation period that was not subjected to the model during training.

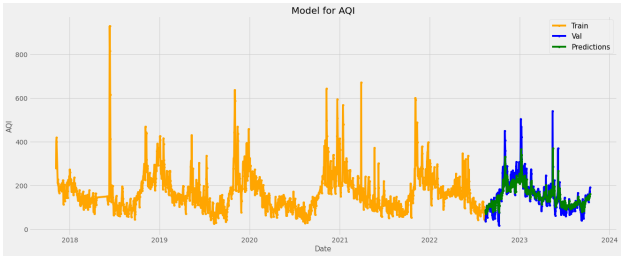

Fig. 31. ANN Predicions

Out of four models that have been used, ANN is getting the highest accuracy of 0.999689 out of 1 and RMSE is 1.559937. The last model is SVR which gives an accuracy of 0.593439 and the RSME is 56.474519.

For all our models we have employed two key evaluation metrics to assess the performance of our machine learning models—specifically, the R-squared (R^2) score and the Root Mean Square Error (RMSE). The R^2 score (Fig. 32), often referred to as the coefficient of determination, is a metric that quantifies the proportion of the variance in the dependent variable (in our case, the AQI) that is predictable from the independent variables. It ranges from 0 to 1, where a higher R^2 score indicates a better fit of the model to the data. The formula for calculating R^2 is:

$$R^2 = 1 - \frac{SSR}{SST}$$

Here, SSR represents the sum of the squared residuals, which measures the difference between the predicted values and the actual values, while SST stands for the total sum of squares, indicating the total variance in the dependent variable.

On the other hand, RMSE (Fig. 33), or Root Mean Square Error, serves as a measure of the model's predictive accuracy by quantifying the average magnitude of the errors between predicted and observed values. It is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y_i represents the actual AQI values, $\hat{y}_i$ denotes the predicted AQI values, and n is the total number of data points.

Our graphs, which portray these metrics, play a pivotal role in illustrating the effectiveness of our models in capturing

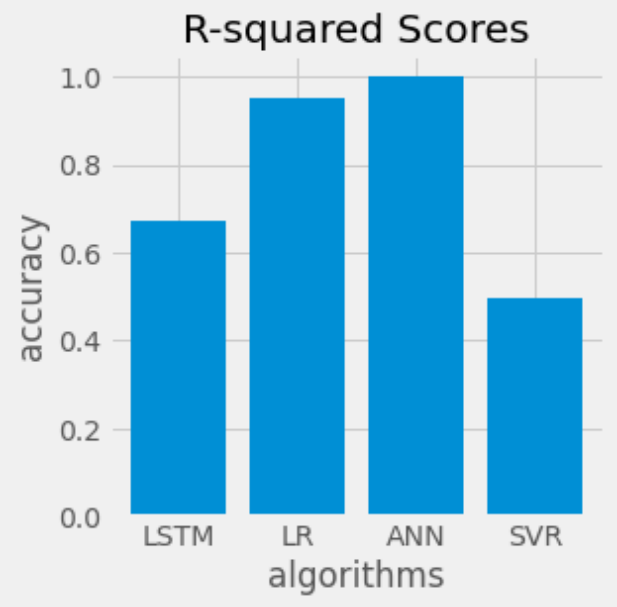

Fig. 32. R2 score comparision

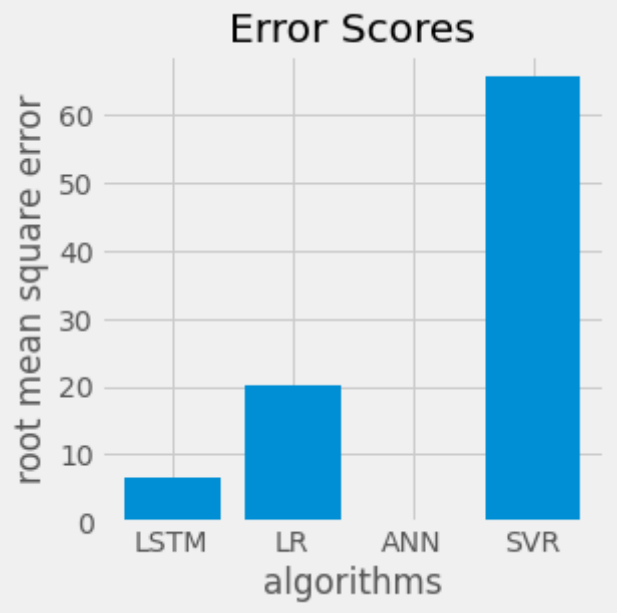

Fig. 33. RMSE comparision

We can infer that the Artificial Neural Network is the best model, not only because it has the highest accuracy but also because it has the lowest RMSE value.

An essential component of our research, the loss curve Fig. 34 shows how our Artificial Neural Network (ANN) model is trained to predict the Air Quality Index (AQI). It displays the Mean Squared Error (MSE) loss of the model, which measures

the variation between the actual and predicted AQI values throughout the training epochs. The model's performance on the training dataset is represented by the training loss, which decreases as the model gains knowledge from the data. Simultaneously, overfitting is gauged by the validation loss, which is determined on a different dataset. The significance of the loss curve in maximizing the performance and generalization abilities of our AQI prediction model is shown by the fact that a convergence of training and validation losses indicates efficient learning, but a sizable gap may indicate overfitting.

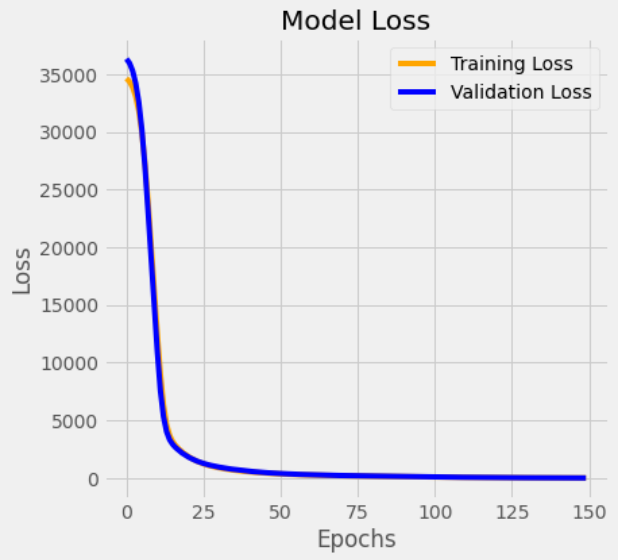

Fig. 34. Loss curve

V. Conclusion

In this study, we conducted a comprehensive analysis of air quality data for the city of Pusa, Delhi, India, with the aim of predicting the Air Quality Index (AQI). Four different machine learning and deep learning models, namely Long Short-Term Memory (LSTM), Linear Regression (LR), Artificial Neural Network (ANN), and Support Vector Regression (SVR), were employed for AQI prediction.

The Linear Regression model, based on a simple linear relationship between pollutant features and AQI, demonstrated its effectiveness in providing quick and interpretable predictions. The Artificial Neural Network, a more complex model, exhibited the capability to model complex non-linear relationships within the data. Additionally, the Support Vector Regression model, which is based on the concept of support vectors, showed competitive performance in AQI prediction, but the best performance was give by ANN, because ultimately, the choice of the best model depends on specific requirements, such as the need for interpretability, computational resources, and the particular characteristics of the data. This study offers valuable insights into the application of machine learning and deep learning models for AQI prediction, facilitating more informed decision-making regarding air quality management and environmental policies.

Future research could explore model enhancements, feature engineering, and the integration of meteorological and geographical data to further improve AQI predictions. Additionally, the deployment of these models in real-time monitoring systems can contribute to proactive air quality management and public health awareness.

VI. abs

References

[1] F. C. Moore, "Climate change and air pollution: Exploring the synergies and potential for mitigation in industrializing countries," *Sustainability*, vol. 1, no. 1, pp. 43–54, 2009.

[2] V. Kanawade, A. Srivastava, K. Ram, E. Asmi, V. Vakkari, V. Soni, V. Varaprasad, and C. Sarangi, "What caused severe air pollution episode of november 2016 in new delhi?," *Atmospheric Environment*, vol. 222, p. 117125, 2020.

[3] R. Sharma, R. Kumar, D. K. Sharma, L. H. Son, I. Priyadarshini, B. T. Pham, D. Tien Bui, and S. Rai, "Inferring air pollution from air quality index by different geographical areas: case study in india," *Air Quality, Atmosphere & Health*, vol. 12, pp. 1347–1357, 2019.

[4] M. S. Hammer, A. van Donkelaar, C. Li, A. Lyapustin, A. M. Sayer, N. C. Hsu, R. C. Levy, M. J. Garay, O. V. Kalashnikova, R. A. Kahn, M. Brauer, J. S. Apte, D. K. Henze, L. Zhang, Q. Zhang, B. Ford, J. R. Pierce, and R. V. Martin, "Global estimates and long-term trends of fine particulate matter concentrations (1998–2018)," *Environmental Science & Technology*, vol. 54, no. 13, pp. 7879–7890, 2020. PMID: 32491847.

[5] G. R. Ana, A. S. Alli, D. C. Uhiara, and D. G. Shendell, "Indoor air quality and reported health symptoms among hair dressers in salons in ibadan, nigeria," *Journal of Chemical Health & Safety*, vol. 26, no. 1, pp. 23–30, 2019.

[6] A. Kankaria, B. Nongkynrih, and S. K. Gupta, "Indoor air pollution in india: Implications on health and its control," *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, vol. 39, no. 4, p. 203, 2014.

[7] M. R. Haider, M. M. Rahman, F. Islam, and M. M. Khan, "Association of low birthweight and indoor air pollution: biomass fuel use in bangladesh," *Journal of Health and Pollution*, vol. 6, no. 11, pp. 18–25, 2016.

[8] P. Chhikara, R. Tekchandani, N. Kumar, M. Guizani, and M. M. Hassan, "Federated learning and autonomous uavs for hazardous zone detection and aqi prediction in iot environment," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15456–15467, 2021.

[9] K. B. Shaban, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," *IEEE Sensors Journal*, vol. 16, no. 8, pp. 2598–2606, 2016.

[10] A. Wang, J. Xu, R. Tu, M. Saleh, and M. Hatzopoulou, "Potential of machine learning for prediction of traffic related air pollution," *Transportation Research Part D: Transport and Environment*, vol. 88, p. 102599, 2020.

[11] D. A. Wood, "Local integrated air quality predictions from meteorology (2015 to 2020) with machine and deep learning assisted by data mining," *Sustainability Analytics and Modeling*, vol. 2, p. 100002, 2022.

[12] S. Masmoudi, H. Elghazel, D. Taieb, O. Yazar, and A. Kallel, "A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection," *Science of the Total Environment*, vol. 715, p. 136991, 2020.

[13] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," *Environmental pollution*, vol. 231, pp. 997–1004, 2017.

[14] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin, "An lstm-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, 2020.

[15] P. Nath, P. Saha, A. I. Middya, and S. Roy, "Long-term time-series pollution forecast using statistical and deep learning methods," *Neural Computing and Applications*, pp. 1–20, 2021.

[16] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi, "Air-pollution prediction in smart city, deep learning approach," *Journal of big Data*, vol. 8, no. 1, pp. 1–21, 2021.

[17] T. Liu and S. You, "Analysis and forecast of beijing's air quality index based on arima model and neural network model," *Atmosphere*, vol. 13, no. 4, p. 512, 2022.

[18] V. Singh, "Indian air quality prediction and analysis using machine learning," 2021.

[19] C. Aditya, C. R. Deshmukh, D. Nayana, and P. G. Vidyavastu, "Detection and prediction of air pollution using machine learning models," *International journal of engineering trends and technology (IJETT)*, vol. 59, no. 4, pp. 204–207, 2018.

[20] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Science of the Total Environment*, vol. 426, pp. 244–255, 2012.

[21] K. D. Hutchison, S. Smith, and S. J. Faruqui, "Correlating modis aerosol optical thickness data with ground-based pm2. 5 observations across texas for use in a real-time air quality prediction system," *Atmospheric Environment*, vol. 39, no. 37, pp. 7190–7203, 2005.

[22] A. Barthwal, D. Acharya, and D. Lohani, "Iot system based forecasting and modeling exceedance probability and return period of air quality using extreme value distribution," in *2019 IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, IEEE, 2019.

[23] P. D. Rosero-Montalvo, J. A. Caraguay-Procel, E. D. Jaramillo, J. M. Michilena-Calderón, A. C. Umaquinga-Criollo, M. Mediavilla-Valverde, M. A. Ruiz, L. A. Beltrán, and D. H. Peluffo, "Air quality monitoring intelligent system using machine learning techniques," in *2018 International Conference on Information Systems and Computer Science (INCISCOS)*, pp. 75–80, IEEE, 2018.

[24] A. Barthwal and D. Acharya, "Extreme value analysis of urban air quality using internet of things.," *International Journal of Next-Generation Computing*, vol. 10, no. 1, 2019.

[25] Y. Zhongjie, W. Shengwei, and W. Ze, "Air quality prediction method based on the cs-lstm," in *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pp. 1–5, IEEE, 2022.

[26] Y. Jiao, Z. Wang, and Y. Zhang, "Prediction of air quality index based on lstm," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pp. 17–20, IEEE, 2019.

[27] H. Wu, Y. Zhang, Q. Zhang, K. Duan, Y. Lin, and S. Du, "Prediction of air quality index (aqi) based on optimized gwo-lstm neural network," in *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pp. 1445–1449, IEEE, 2022.

[28] B. Leo, "Random forests," *Machine learning*, vol. 45, pp. 5–23, 2001.

[29] T. Boningari and P. G. Smirniotis, "Impact of nitrogen oxides on the environment and human health: Mn-based materials for the nox abatement," *Current Opinion in Chemical Engineering*, vol. 13, pp. 133–141, 2016.

[30] E. G. Dragomir, "Air quality index prediction using k-nearest neighbor technique," *Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII*, vol. 1, no. 2010, pp. 103–108, 2010.

[31] H. Liu, Q. Li, B. Yan, L. Zhang, and Y. Gu, "Bionic electronic nose based on mos sensors array and machine learning algorithms used for wine properties detection," *Sensors*, vol. 19, no. 1, p. 45, 2018.

[32] M. M. Alam, K. Mohiuddin, M. K. Islam, M. Hassan, M. A.-U. Hoque, and S. M. Allayear, "A machine learning approach to analyze and reduce features to a significant number for employee's turn over prediction model," in *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2*, pp. 142–159, Springer, 2019.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.