EXPLORATORY DATA ANALYSIS ON AIR QUALITY DATA AND AQI FORECASTING MODEL USING DEEP LEARNING

Santhana Lakshmi V[1*], Vijaya M S[2]

Article History: Received: 20.04.2023 Revised: 30.04.2023 Accepted: 06.05.2023

Abstract:
Most of the people on earth are indeed exposed to high levels of air pollution. Air pollution is a significant contributor to a country's rising mortality and morbidity rates. The effective indicator that provides information about the level of air pollution to the people is the Air Quality Index (AQI). There are seven categories that make up the AQI value. It includes $PM^{2.5}$, PM^{10}, CO, SO2, NOX, ozone and NH_3. Each category has a different level of health concern. People when exposed to these pollutants for a long time are affected with respiratory diseases like asthma, emphysema etc. Therefore, developing a reliable AQI forecasting model is crucial to protect the people from the impact of outdoor air pollution. In this paper AQI prediction model is proposed using deep learning architectures such as Long Short Term Memory (LSTM), Bidirectional Long Short Term Memory (BILSTM) and Gated Recurrent Unit (GRU) by understanding the trends in time-series air quality data. It is found that a reliable AQI prediction model can be built using GRU algorithm.

Index Terms: Air Pollution, Deep Learning, Forecast, Machine Learning, Mortality.

[1*]Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India
[2]Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India

DOI: - 10.53555/ecb/2023.12.si5a.080

I.INTRODUCTION

Air Pollution is becoming a leading risk factor that impacts the health of human beings. Air is getting polluted when the pollutants in the air is communicated to the people in terms of a measure called Air Quality Index (AQI). It is a measure of the air quality in each area. The AQI is calculated based on levels of seven major air pollutants: ground-level ozone, particulate matter of two different sizes 2.5 and 10, carbon monoxide, sulfur dioxide, and nitrogen dioxide and ammonia. The AQI ranges from 0 to 500, with higher values indicating worse air quality. A value of 100 or below is considered to be good air quality, while values above 100 are considered to be increasingly unhealthy for certain groups of people.

A. Background and Motivation

Seven million people die from air pollution worldwide every year, according to estimates. According to WHO data, almost all the world's population (99%) breathes air that contains high levels of pollutants that exceeds WHO guideline limits [2]. Long-term exposure to these contaminants can cause lung, cardiac, and other health problems in people. The concentration of the pollutant and the duration of the exposure determine the health effects caused by the pollution.

The phrase "particulate matter," often known as "particle pollution" or PM refers to very minute solid and liquid droplets dispersed in the atmosphere. Nitrates, sulphates, organic compounds, metals, soil, or dust particles, etc. are just a few examples of the many different solid and liquid dust particles emitted by the industries, vehicles and burning fossil fuels suspend into the air. When the level of the pollutants increases, it causes a serious impact on human health [1].

The likelihood that particles will harm your health depends on their size. $PM^{2.5}$ and PM^{10} particles have diameters of 2.5 and 10, respectively. Due to their smaller size, these particles easily enter the mouth and induce lung diseases [3].

Ozone is composed of three oxygen atoms joined together. Two oxygen atoms joined together form the basic oxygen molecule O2. The additional third atom makes ozone an unstable, highly reactive gas. Nitrogen dioxide is a highly reactive gas formed by emissions from motor vehicles, industry, gas-heaters, and gas stove tops. Nitrogen dioxide is a respiratory irritant and has a variety of adverse health effects on the respiratory system. Sulphur dioxide is a highly reactive gas with a pungent irritating smell. It is formed by fossil fuel combustion at power plants and other industrial facilities. CO is an odourless, transparent, tasteless flammable gas. This is the most common air pollutant. The largest anthropogenic source of CO are vehicles and fossil fuels. Sulphur dioxide is a poisonous, transparent gas with a strong odour. Thus the level of these seven pollutants decides the level of AQI. Air Quality index is the measure used by the government to communicate to the people how polluted the air is. The Central Pollution Control board has defined the National Ambient Air Quality standard to indicate the level of air quality necessary with the given margin of safety to protect the health [4]. The breakpoints of the seven pollutants are provided in Table 1.

Table 1. Breakpoints for the pollutants

AQI Category	PM^{10}	$PM^{2.5}$	NO_2	O_3	CO	SO2
Good	0-50	0-30	0-40	0-50	0-1.0	0-40
Satisfactory	51-100	31-60	41-80	51-100	1.1-2.0	41-80
Moderate	101-250	61-90	81-80	101-168	2.1-10	81-380
Poor	251-350	91-120	181-280	169-208	10.1 – 17	381-800
Very Poor	351-430	121-250	281-400	209-748	17.1-34	801-1600
Severe	430+	250+	400+	748+	34+	1600+

It is very essential to build an AQI Forecasting model. Lot of research has been carried out to improve the accuracy of the prediction models.

B. Related Work

Hanin Alkabbani and et al. developed a forecasting model for predicting the value of Air Quality Index using machine learning algorithms such as SVM, SVR, Linear Regression and Random Forest. The time series data collected from Kuwait Environment Public Authority for AI-Jahra city for the period of 2 years was used for the analysis. They have compared the performance of two missing data imputation approaches such as linear and missForest method. Promising results were obtained while using the missForest method for imputing the missing data [5].

Yun-Chia Liang et al., used time series meteorological data and air quality data collected

from 3 monitoring stations in Taiwan for the period of 10 years from 2008 to 2018. They employed machine learning methods such as adaptive boosting (AdaBoost), Artificial Neural Network (ANN), Random Forest, stacking ensemble, and Support Vector Machine (SVM) for building AQI prediction models. They have also used stacking ensemble, AdaBoost, and Random Forest. They have found that the stacking ensemble method delivered consistently superior performance considering R Squared value and RMSE, whereas AdaBoost provided better results when considering MAE [6].

Samayan Bhattacharya and Sk Shahnawaz used Support Vector Regression for forecasting the pollutants and air quality index. The pollutant data are collected from the central pollution control board website and date features from the US embassy. They imputed the missing values, performed feature extraction and feature selection. Principal Component Analysis technique is used for feature selection. Forecasting model is developed using SVR-RBF. The performance of the model developed using the selected features are compared across the model developed with entire features. Better results were obtained while using entire features for developing the model [7].

Doreswamy, Harishkumar K S, and others looked into the amount of pollutants in the air. To create a forecasting model, they collected data from the Thaiwan air quality monitoring network. For analysis, the five-year period's worth of air quality data are used. For making predictions, methods like random forest, gradient boosting regression, decision tree regression, and MLP Regression are utilized. To compare the efficacy of the models, performance evaluation metrics such root mean squared error, mean absolute error, mean squared error, and coefficient of determination are used. They concluded that the gradient boosting regressor model is superior for predicting air pollution based on TAQMN data [8] after doing the study.

Most of the researchers have used machine learning for building an air quality prediction model. The accuracy of machine learning models depends heavily on the quality of the data used to train them. If the data is incomplete, outdated, or inaccurate, the model's predictions will likely be unreliable. Machine learning models are trained on specific datasets. Machine learning models can provide estimates of air quality, but there is always some level of uncertainty and variability in these predictions. This can be particularly problematic for public health applications, where even small errors in predicted air quality can have significant consequences.

In this paper deep-learning architectures have been adopted for modelling the prediction task. The prediction model is trained with seven pollutant features. The networks such as LSTM, Bidirectional LSTM and GRU are used for learning the patterns in the time series data. Deep learning algorithms are designed to capture complex and nonlinear relationships between variables. This makes them well-suited for forecasting tasks that involve multiple variables and where the relationships between those variables are not easily captured by traditional statistical models.

II. DATA COLLECTION

The time series data for the period of three years from 2017 to 2020 of Trivandrum city are collected from Central Pollution Control Room for Air Quality Management [9]. It is Kerala's biggest city. There are 957,730 people living in the city, and 1.68 million people live in the surrounding metropolitan region. The Thiruvananthapuram district is located between latitudes 8° 17′ and 8° 54′ in the north and 76° 41′ and 77° 17′ in the east. Kanyakumari, known as the "lands' end of India," lies 56 kilometres from Parasala, the world's most southern point [10]. The district extends for 78 kilometres [5] along the Arabian Sea's coast. Figure 1 provides the state's general layout.

Figure. 1. General layout of Trivandrum

The raw data includes seven pollutant features such as $PM^{2.5}$, PM^{10}, CO, SO_2, NOX, Ozone and NH_3. Raw data comprises 26,305 instances. $PM^{2.5}$ refers to tiny, inhalable particles with a diameter of 2.5 micrometres or less. Some of the sources of outdoor $PM^{2.5}$ pollution include emissions from the burning of wood, gasoline, or oil. PM^{10} is made up

of dust from landfills, farms, construction sites, wildfires, and the burning of fossil fuels. $PM^{2.5}$ travels down the airway and is deposited in the deepest part of the lungs. PM^{10} is deposited throughout the airways and reaches the lungs. Particles that land on the surface of the lungs can cause lung inflammation and tissue damage. Both $PM^{2.5}$ and PM^{10} exert harmful effects on health. Acute and chronic bronchitis, asthma, and higher hospital admissions for heart or lung conditions have all been linked to short-term exposure to $PM^{2.5}$. Most impacted are young children and elderly people who already have heart or lung conditions. The main effect of short-term exposure to PM^{10} is the deterioration of respiratory conditions. Premature death results from prolonged (months to years) exposure to these particles. Additionally, it slows down a child's lung development.

CO is a colourless, odourless gas that, if inhaled in excessive quantities, can be dangerous. Any time something burns, CO is released. Burning fossil fuels such as natural gas, gasoline, coal, and oil, wood smoke, and automobile and truck exhausts are all sources of carbon monoxide [11]. The amount of oxygen that can be carried in the bloodstream to vital organs like the heart and brain is decreased when breathing air with a high CO concentration. People with certain types of cardiac problems may be especially concerned when CO levels are high outside. These patients already have a diminished capacity for supplying their hearts with oxygenated blood when those conditions arise. When exercising or under more stress, they are especially susceptible to CO's effects. In these circumstances, brief exposure to increased CO may lead to decreased cardiac oxygenation and angina, or chest pain. The two groups most at danger are unborn children and those with cardiac problems. Headaches, fatigue, dizziness, and nausea are just a few of the "flu-like" symptoms that high amounts of carbon monoxide can produce.

Sulphur dioxide (SO2) is a colourless, highly toxic gas that is a common air pollutant. It is produced by the burning of fossil fuels, particularly coal and oil that contain sulphur compounds. Sulphur dioxide is a major contributor to acid rain and can also have negative effects on human health, including respiratory problems. In addition, sulphur dioxide can also react with other pollutants to form fine particulate matter, which is linked to a variety of health problems and can also damage crops and ecosystems [12].

Ammonia (NH_3) is a colourless gas that has a strong, pungent smell. It is a common air pollutant that is produced by a variety of sources, including agriculture (from the use of fertilizers and manure), industry (from the production of chemicals and textiles), and waste management (from the decomposition of organic waste). The primary effect of ammonia on human health is the irritation of eyes, nose, and respiratory system. High concentrations of ammonia can cause coughing, shortness of breath, and broncho constriction. Long-term exposure to low levels of ammonia can also damage the lungs, and individuals with pre-existing respiratory conditions may be more sensitive to the effects of ammonia [13]. Ammonia contributes to the formation of particulate matter and ground-level ozone, both of which can have negative effects on human health and the environment.

Since these pollutants create adverse impacts on the health of the people and environment, federal and state governments take necessary steps to ensure that emissions remain below a certain level. It is essential to build an effective forecasting model which can help the government and people to take precautionary steps when the pollution level is high.

A. Calculating AQI

The air quality index is used to understand the quality of the air around us. The AQI is calculated based on the seven pollutants $PM^{2.5}$, PM^{10}, CO, SO2, ozone and NH_3. It can also be calculated when one or 2 pollutant values are not available. At least the value of 3 pollutants is essential. Among that one pollutant should include either $PM^{2.5}$ or PM^{10}. The sub index must be calculated for each pollutant which can then be used to identify the value of AQI. The lowest sub index value will be considered as the AQI value [14]. The formula for calculating air quality index is given below.

$$Ip = [IHi – ILo / BPHi – BPLo] (Cp – BPLo) + ILo \quad (1)$$

Where,

Ip = index of pollutant p

Cp = truncated concentration of pollutant p

BPHi = concentration breakpoint i.e. greater than or equal to Cp

BPLo = concentration breakpoint i.e. less than or equal to Cp

IHi = AQI value corresponding to BPHi

ILo = AQI value corresponding to BPLo

The calculated AQI value should be below the levels defined by the Central Pollution Control Board. The breakpoints of the AQI value is given in the table 2.

Table 2. Breakpoints for AQI

Index	AQI Category
0 – 100	Good
101 – 200	Moderate
201 – 300	Poor
301 – 400	Very Poor
401 – 500	Severe

Colour codes were also used to communicate how healthy the air is to breathe on that day. The colours go
from green to yellow to orange to red to purple to maroon, each colour specifies that the air is less clean than the colour before. Green is the best air quality. When the AQI is green, the air is clean.

III.AIR QUALITY INDEX PREDICTION MODEL

Air quality prediction is a key tool for improving public health and protecting the environment. Air quality prediction is essential as it allows individuals and organizations to take proactive measures to protect their health and the environment. The main objective of this paper is to build an effective AQI forecasting model using the seven pollutants. The process of developing a time series prediction model involves various phases such as understanding the data through exploratory data analytics, preparing the dataset by selecting the most contributing features and eliminating the unwanted features, building an AQI prediction model and finally evaluating its performance.. Exploratory data analysis (EDA) is performed to understand the data and gain insights from it. EDA is followed by pre-processing the data. The pre-processed data is then used for building the forecasting model. The architecture of the Air Quality Prediction model is provided in the figure 2.

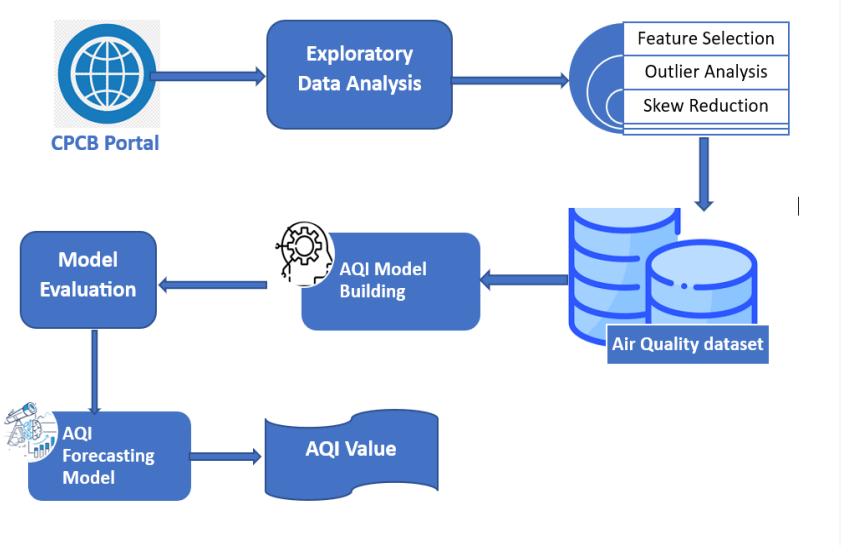

Figure. 2. Architecture of developing AQI forecasting model

A. Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach for analysing the data in which the main goal is to understand the underlying structure and relationships within the data. The main objective is to identify patterns, outliers, and other interesting features in the data. Exploratory data analysis is essential for several reasons. It helps to develop an understanding of the data and to identify outliers. It helps to identify which features should be selected for further analysis and modelling [15].

Any time series data can be decomposed into four components. Trend, cyclic, seasonal, and irregular. Trend represents the overall direction or pattern of the data over time. A cyclic pattern in time series data refers to a repeating pattern or cycle within the data. This pattern can occur over different time periods, such as daily, weekly, monthly, or even yearly. Seasonality represents repeating patterns within the data whereas irregular pattern denotes random variations in the data [16]. Identifying these pat0terns helps to make more accurate predictions about future values. Air quality data is decomposed, and the observed pattern is provided in figure 3. Cyclic pattern is not identified. The data do not have either an upward trend or downward trend. Season pattern is identified in the data.

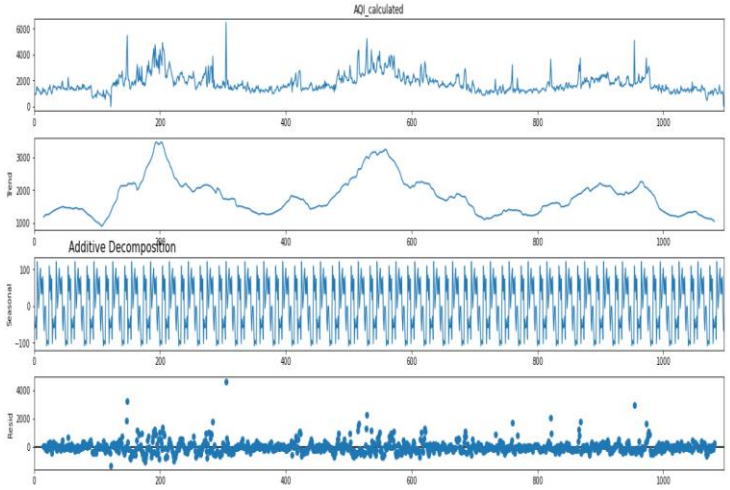

Figure. 3. Decomposition of Pollutant data

In this work, various charts like Heat map, Bar chart, Box plot etc., are used for analysing air pollutant data.

A heat map is a graphical representation of data that uses a color-coding system to indicate different levels of a particular variable. It is a way to visualize a two-dimensional matrix of data, where the individual values are represented by colours. Typically, the colour scale is represented by a colour bar or a legend, and the data values are represented by different shades of colour, where darker shades indicate higher values and lighter shades indicate lower values [17]. Heat Map generated using the raw data is provided in figure 4. From the figure it is observed that the attributes such as $PM^{2.5}$, PM^{10},CO and NH_3 have strong correlation with air quality index.

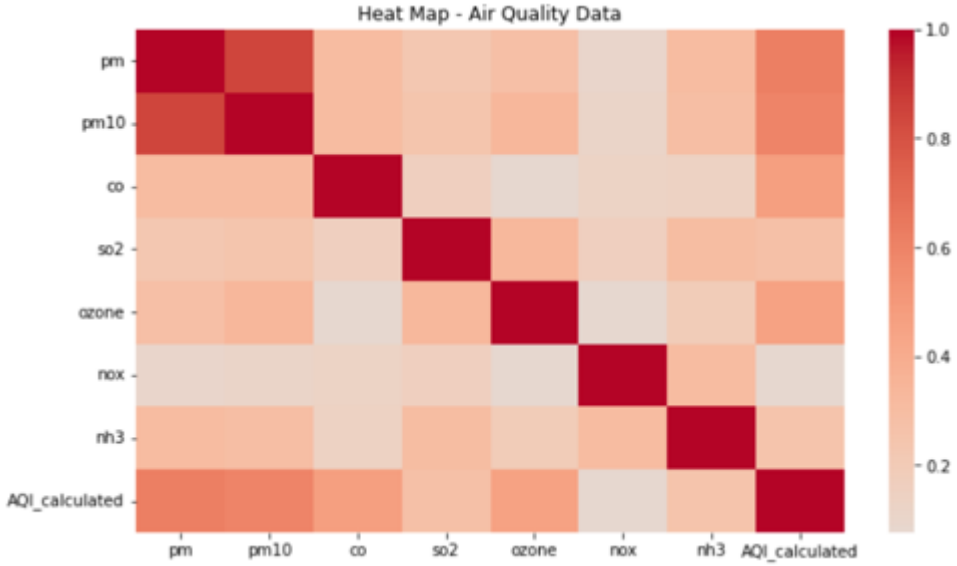

Figure. 4. Heatmap

A histogram is a graphical representation of the distribution of a dataset. It is a way to visualize the frequency of different values in a dataset. It is a bar graph that shows the frequency of different ranges of values in a dataset. The x-axis represents the range of values, and the y-axis represents the frequency of those values. In time series analysis, histograms can be used to visualize the distribution of the data over time [18]. They can be used to identify patterns and trends in the data, such as whether the data is skewed or symmetric, and the presence of any outliers. Histogram generated for all the attributes including the target attribute is provided in the figure 5. From the chart it was understood that for most of the days value of particulate matter falls within the range of 0 to 50. Value of sulphur dioxide falls within the range 0 to 10. AQI_calculated is the target attribute. It falls within the range 10 to 200. Most of the days the values fall within the range 50 to 60. The entire data is not normally distributed. Necessary steps need to be taken to normalize the data.

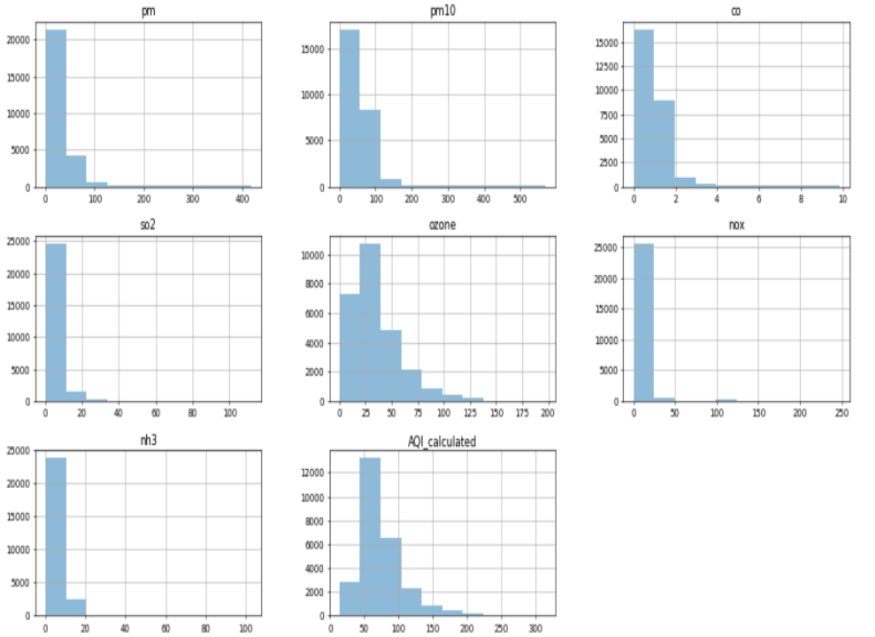

Figure. 5. Histogram

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It is a way to visualize the five-number summary of a dataset, which consists of the minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value. Boxplot generated on the raw data is given in the figure 6. The chart clearly depicts that outliers are available in almost all the attributes other than carbon monoxide.

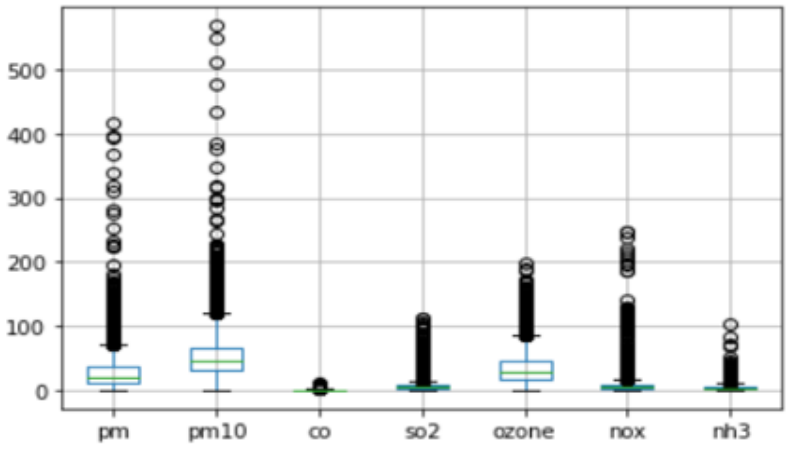

Figure. 6. Boxplot

A. Data Pre-Processing and Preparation Of Dataset

Pre-processing refers to the various techniques used to prepare data for analysis. Exploratory data analysis is performed on the raw data. The insights gained from the analysis helps to identify the pre-processing tasks that need to be performed to form the dataset for building the air quality prediction model. From the analysis it was understood that the distribution of the target attribute AQI_calculated is left skewed. Skewed data is problematic when building predictive models as it affects the accuracy and performance of the model [19]. Transformations, such as log or square root, can be applied to the data to make it more normally distributed, making it easier to model. In this paper square root transformation technique is applied to normalize the values.

The scale of the data varies drastically for all the attributes. The entire data must be normalized. Normalization is a technique used in data mining to transform variables to a common scale. This allows for comparison of variables that may have originally been measured on different scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling. Min-max normalization scales the

data so that it falls between a specified range, such as 0 and 1. Z-score normalization standardizes the data by transforming it so that it has a mean of 0 and a standard deviation of 1. Decimal scaling normalizes the data by moving the decimal point to a specified number of places to the left or right. In this work, Min Max normalization technique is employed to normalize the dataset.

All the features do not contribute equally for building the prediction model. Feature selection needs to be performed. There are several methods for feature selection in data mining, which aim to identify a subset of the most relevant features for a given task. Some of the most common methods include filter methods, wrapper methods, embedded methods, and hybrid methods. In this research work Select K best method is used to identify the features that contribute more for the prediction of air quality index [20]. The SelectKBest method, also known as k-best feature selection, is a filter method for feature selection. The basic idea behind this method is to select a fixed number (k) of features that have the highest scores based on a statistical test. It was identified that the features $PM^{2.5}$, pm^{10}, carbon monoxide, sulphur-di-oxide, ozone and ammonia play a major role in building the prediction model. The dataset is formed after performing all the pre-processing tasks.

Finally a dataset with 26305 instances and 7 features has been developed. The air pollutant instances are tagged with AQI value to facilitate supervised learning. Among the 7 features, six features $PM^{2.5}$, PM^{10}, carbon monoxide, sulphur-di-oxide, ozone and ammonia are independent attributes. The Air Quality Index (AQI) is the dependent attribute for prediction task.

B. Methods and Algorithms

The AQI pollutant data is modelled as time series data and the AQI prediction problem is formulated as a regression task. Machine learning method of data analysis is used here for regression and model building. Deep learning approach is employed to build AQI prediction model by learning the representations in time series pollutant data. It involves training artificial neural networks, which are composed of layers of interconnected nodes, on a vast amount of data. These networks can learn to recognize patterns and features in the data, which can then be used for any prediction tasks and decision making. In this work deep learning architectures such as LSTM, BILSTM and GRU have been used for building an Air Quality Index prediction model.

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) architecture specifically designed to handle sequential data with long-term dependencies. LSTM networks have a memory cell that can store information for an extended period, along with gates that control the flow of information into and out of the cell. This allows LSTMs to selectively remember or forget information, making them well-suited for tasks such as language modelling, speech recognition, and time series forecasting. LSTMs can be stacked on top of each other to create deeper networks, which can improve performance on some tasks [21].

A Bi-directional LSTM (BILSTM) is a type of LSTM network that processes the input sequence in both forward and backward directions. This allows the network to consider context from both past and future time steps, which can be useful for tasks such as natural language processing and speech recognition. In a BILSTM, two LSTM networks are trained on the input sequence, one processing the input in the forward direction and the other processing it in the backward direction. The final output is obtained by concatenating the hidden states of the two networks at each time step [22]. This allows the network to extract features from both past and future context, which can improve the performance of the model.

GRU (Gated Recurrent Unit) is another type of Recurrent Neural Network (RNN) architecture, like LSTM, but with a simpler structure. GRUs have two gates, a reset gate and an update gate, that control the flow of information into and out of the hidden state. The reset gate determines how much of the previous hidden state to forget, while the update gate determines how much of the new input to incorporate into the hidden state. GRUs are computationally less expensive than LSTMs, as they have fewer parameters to train, and are therefore faster to train and use. But sometimes LSTM is better in some specific cases where the long-term dependencies need to be captured. GRUs are often used in natural language processing, speech recognition, and other sequential data tasks [23].

Various hyperparameters such as learning rate,batch size,number of epochs, optimizers, dropouts, number of hidden layers and neurons are defined for LSTM, BILSTM, GRU networks while building the model in order to improve the efficiency of prediction.

Hyperparameter tuning is an essential step in deep learning, which involves finding the optimal set of hyperparameters for a given neural network architecture to achieve the best possible performance on a particular task. Hyperparameters are parameters that cannot be learned from the data and must be set by the user. Hyperparameters include learning rate, batch size, number of epochs, optimizers, dropouts, number of hidden layers, and number of neurons in each layer. These hyperparameters can significantly affect the performance of the model, so it is important to choose the right values [24]. In this paper the hyperparameters such as epochs, batch size, and optimizers are defined for the deep learning architectures LSTM, BILSTM and GRU.

Air quality index forecasting model is developed by training the tagged time series data with AQI as the output variable under predefined hyperparameters settings. The performance of the models developed using deep learning architectures such as LSTM, BILSTM and GRU are evaluated based on the metrics such as R Squared value, root mean squared value and mean absolute error.

IV.EXPERIMENTS & RESULTS

In our previous work, the experiments were carried out by training the time series air quality dataset that consists of only samples with pollutant parameters. Prediction models were built by employing machine learning algorithms such as linear regression, support vector regression and decision tree regression. AQI Prediction model built using the decision tree algorithm generates low RMSE value 0.7493 whereas high RMSE value 0.8166 is obtained for the model built using linear regression. Mean absolute error of 0.6170 is produced by the SVR-AQI model. DT-AQI model produced low MAE value of 0.5806 whereas a high MAE value 0.6433 is produced for LR-AQI. Mean squared error is low in DT-AQI model whereas high MSE value 0.6669 is obtained for LR-AQI model. High R2 score 0.6246 is acquired for DT-AQI model whereas low R2 score 0.5541 is obtained for LR-AQI model.

So while considering all the measures, DT-AQI model produces lowest MAE, MSE, and RMSE values, which means it is the most accurate model among the three. Moreover, the R2 score for the DT-AQI model is the highest, indicating that this model explains the highest proportion of variance in the AQI values. Conversely, the LR-AQI model has the highest MAE, MSE, and RMSE values, indicating that it is the least accurate model among the three. The prediction results of the machine learning models discussed above are provided in Table 3. The same is illustrated in figure 7.

Table 3. Performance of Machine Learning based AQI Models

Model	MAE	MSE	RMSE	R2
LR-AQI	0.64	0.66	0.81	0.55
SVR-AQI	0.61	0.62	0.79	0.58
DT-AQI	**0.58**	**0.56**	**0.74**	**0.62**

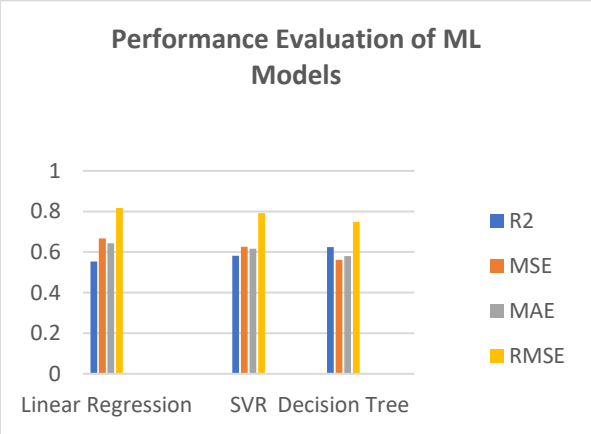

Figure. 7. Comparative performance analysis of Machine Learning models.

In this work air quality index prediction models are built using deep learning architectures such as LSTM, BILSTM and GRU. Experiment is carried out by modifying the hyperparameters. The hyperparameters such as optimizers, batch sizes, dropouts are experimented. Optimizers are used to adjust the parameters of a model, such as weights and biases, to minimize the error or loss function of the model. Optimizers such as Adam, Adagrad and RMSprop are experimented and the results are given in table 4.

When the adam optimizer is combined with deep learning architectures like the LSTM, BILSTM, and GRU, the BILSTM-AQI model yields a low MAE value of 0.3355 while the GRU-AQI model yields a high MAE of 0.3509. When taking root mean squared value into account, BILSTM AQI model produces a low value of 0.4557 whereas GRU-AQI model produces a high value of 0.4834. With the BILSTM-AQI model, a high R2 score of 0.8286 was obtained, while the R2 score for the GRU-AQI model was low at 0.8071.

The Adagrad optimizer is used in the experiment along with the LSTM, BILSTM, and GRU architectures. Using the BILSTM-AQI model, low RMSE value 0.7759 and MAE value 0.6412 are attained. In the BILSTM-AQI, high R2 value of 0.5032 is obtained. In the GRU-AQI model, a low R2 score of 0.3701 and high RMSE value of 8736 are obtained.

When RMSprop optimizer is used , the error rate is high in the LSTM AQI model whereas a high R2 score is obtained for GRU-AQI model. Low RMSE value 0.4701 and MAE value of 0.3787 are obtained for GRU-AQI whereas high RMSE value of 0.6253 and MAE value 0.5192 are obtained for LSTM AQI. The R2 score obtained for GRU-AQI is 0.8176 and low R2 score 0.6773 is obtained for LSTM AQI. While comparing the performance of the three optimizers adam, adagrad and rmsprop, adam optimizer outperforms the other two optimizers. Low RMSE value of 0.4557 and MAE value 0.3355 and high R2 score 0.8286 are obtained while using adam optimizer. The results acquired for the various optimizers are provided in table 4.

Table 4. Performance of the Models using optimizers

Model	Adam Optimizer		
	RMSE	R^2	MAE
LSTM-AQI	0.4687	0.8186	0.3476
BILSTM-AQI	0.4557	0.8286	0.3355
GRU-AQI	0.4834	0.8071	0.3509

Model	Adagrad Optimizer		
	RMSE	R^2	MAE
LSTM-AQI	0.7866	0.4894	0.6501
BILSTM-AQI	0.7759	0.5032	0.6412
GRU-AQI	0.8736	0.3701	0.7280

Model	Rmsprop Optimizer		
	RMSE	R^2	MAE
LSTM-AQI	0.6253	0.6773	0.5192
BILSTM-AQI	0.5419	0.7576	0.3969
GRU-AQI	0.4701	0.8176	0.3787

The performance of the deep learning models experimented with various optimizers are illustrated in the figure 8.

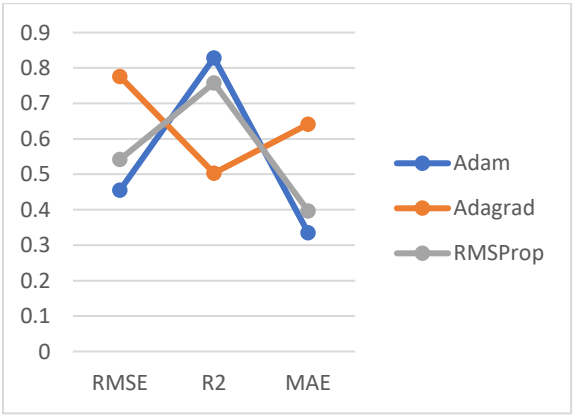

Figure. 8. Comparative performance analysis of deep learning models with various optimizers

The experiment is carried out by changing the epoch sizes from 50 to 150. The performance of the models observed are provided in table 5. Mean absolute error is high when the epoch is 50 whereas it is reduced to 0.2977 when the epoch is increased to 100. The value of mean absolute error was further increased when the epoch size was increased to 150. Similar to mean absolute error, root mean squared error was more when the epoch is 50 and 100 and low RMSE 0.4278 was obtained for the epoch 100. Root mean squared error is less when the epoch is 100. Similarly high R Squared value is obtained when the epoch is 100. High R Squared value 0,8489 was obtained when the size of the epoch is 100 whereas low value is obtained for the other epochs 50 and 100.

Table 5. Performance of the LSTM Model for Different Epochs

	LSTM		
Epoch	50	100	150
MAE	0.3476	**0.2977**	0.3211
RMSE	0.4687	**0.4278**	0.4437
R2	0.8186	**0.8489**	0.8375

The performance of the model evaluated using the metrics for BILSTM algorithm are provided in table 6. High R Squared value 0.8457 is obtained for the epoch 100 and low R Squared value 0.8157 is obtained when the epoch is 150. Similarly high root mean squared error is obtained for the epoch is 150 and low value 0.4323 is obtained for the epoch 100.

Table 6. Performance of the BILSTM Model

	BILSTM		
Epochs	50	100	150
MAE	0.3355	**0.3164**	0.3702
RMSE	0.4557	**0.4323**	0.4725
R2	0.8286	**0.8457**	0.8157

The performance of the model evaluated using the metrics for GRU algorithm are provided in table 7. High R Squared value 0.8658 is obtained for the epoch 100 and low R Squared value 0.8071 is obtained when the epoch is 50. Similarly high root mean squared error 0.4834 is obtained for the epoch is 50 and low value 0.4031 is obtained for the epoch 100. From the experiment it is observed that a better AQI prediction model can be built using GRU algorithm.

Table 7. Performance of the GRU Model

	GRU		
Epochs	50	100	150
MAE	0.3509	**0.2918**	0.3702
RMSE	0.4834	**0.4031**	0.4725
R2	0.8071	**0.8658**	0.8157

The performance of the LSTM-AQI, BILSTM-AQI, and GRU-AQI prediction models built using deep learning algorithms such as LSTM, BILSTM and GRU are compared with the performance of traditional machine learning based AQI prediction models and the comparative results are analysed. The performance of these models are compared based on the metrics mean absolute error, mean squared error, root mean squared error and R squared value. Comparing the models, it was understood that the LSTM-AQI, BILSTM-AQI, and GRU-AQI models have the lowest MAE, MSE, and RMSE values. This indicates that these models are more accurate in predicting AQI values.

The GRU-AQI model has the lowest MAE as 0.2918 whereas linear regression produces the high value of 0.6422. Mean Squared Error is 0.1624 for GRU-AQI model and high MSE of 0.6669 is produced by linear regression model. In terms of R Squared Value, GRU-AQI produced a high value of 0.8658 whereas low value 0.5541 is obtained in the linear regression model indicating that GRU-AQI has the best fit for the data. However, it's important to note that while the R2 score is a useful metric, it is not the only measure of a good model, and other metrics such as MAE and RMSE should also be considered. By considering all the metrics, the LSTM-AQI, BILSTM-AQI, and GRU-AQI models perform better than the LR-AQI, SVR-AQI, and DT-AQI models in predicting AQI values, with the GRU-AQI model being the most accurate having the best fit for the data. The comparative results observed are shown in table 8.

Table 8. Performance of the Machine Learning & Deep Learning models

Model	**MAE**	**MSE**	**RMSE**	**R2 Score**
LR-AQI	0.6433	0.6669	0.8166	0.5541
SVR-AQI	0.6170	0.6257	0.7910	0.5816
DT-AQI	0.5806	0.5615	0.7493	0.6246
LSTM-AQI	0.2977	0.1830	0.4278	0.8489
BILSTM-AQI	0.3164	0.1868	0.4323	0.8457
GRU-AQI	**0.2918**	**0.1624**	**0.4031**	**0.8658**

The comparative performances of the deep learning models and machine learning models experimented are illustrated in the figure 9. The AQI prediction model created using GRU offers greater accuracy when analysing the performance of the models created using machine learning and deep learning algorithms. GRU-AQI model provides 86% accuracy with minimum RMSE of 0.4031.

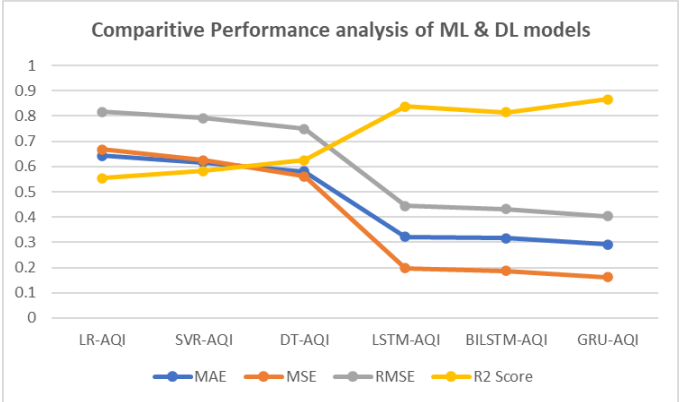

Figure. 9. Comparative performance analysis of deep learning and machine learning models.

The research work demonstrates that pollutants play a major role in building the air quality prediction models. Among all the pollutants particulate matter contributes more for the prediction as high correlation is obtained between particulate matter and AQI. The pre-processing tasks such as filling the missing values, removing outliers and feature selection helped in improving the performance of the forecasting model. This ensures that the model is trained on high-quality data, which leads to more accurate results. Normalizing the data helped to speed up the training process and improve the accuracy of the model.

When compared to machine learning algorithms, deep learning algorithms give better results in predicting AQI. Deep Learning algorithms can acquire hierarchical representations of the data and manage large amounts of data. When dealing with intricate relationships between the input features and the target variable, DL algorithms can achieve greater accuracy. By adjusting the hyperparameters, the performance of the deep learning models can be further enhanced. Setting the proper learning rate correctly leads to faster convergence and better accuracy. Finding the ideal sample size facilitates generalisation and prevents overfitting. The ability of the model to understand complex relationships in the data is improved by adding more layers. Thus optimizing the hyperparameters significantly improved the accuracy of AQI prediction models. Thus an accurate AQI prediction model can be built using GRU algorithm.

V. CONCLUSION

This work models the AQI prediction as a time series prediction task and demonstrates the machine learning and deep learning approach for forecasting the value of air quality index. Deep neural network architectures such as LSTM, BILSTM and GRU are used. Air quality data of 7 pollutant features were used in this study. EDA was applied to the air quality dataset to understand the distribution of data and the importance of each parameter in predicting air quality index. Various pre-processing tasks were employed to prepare the quality dataset. Machine Learning algorithms such as linear regression, support vector regression and decision tree regression were employed in building the prediction models. The deep learning based AQI prediction models have also been developed using LSTM, BILSTM, GRU and their performances were analysed. As a scope of future work, the model efficiency can further be improved by exploring additional information about air quality data and deep learning architectures.

REFERENCES

1. National Institute of Environmental Health Science: Air Pollution and your Health, https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm
2. World Health Organization : Air Pollution, https://www.who.int/health-topics/air-pollution#tab=tab_1
3. Karuna Singh, Dhananjay Tripathi, Particulate Matter and Human Health, Environmental Health, IntechOpen
4. Banega swasth India: Air Pollution: What Is Air Quality Index, How Is It Measured And Its Health Impact, https://swachhindia.ndtv.com/air-pollution-what-is-air-quality-index-how-is-it-measured-and-its-health-impact-40387/
5. Hanin Alkabbani ,Ashraf Ramadan , Qinqin Zhu & Ali Elkamel.(2022). An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach", Atmosphere, https://doi.org/10.3390/atmos13071144.
6. Yun-Chia Liang, Angela Chen, Josue Rodolfo Cuevas Juarez. (2020). Machine Learning-Based Prediction of Air Quality", Applied Sciences, 10(24):9151, https://doi.org/ 10.3390/app10249151
7. Samayan Bhattacharya, Sk Shahnawaz. (2021). Using Machine Learning to Predict Air Quality Index in New Delhi, https://doi.org/10.48550/arXiv.2112.05753
8. Gad, I. M., Alharbi, M. A., Mohammad, A. M., Alshehri, A. H., & Alqahtani, N. A. (2019). Forecasting Air Pollution Particulate Matter (PM 2.5) Using Machine Learning Regression Models. Procedia Computer Science.
9. Central Control Room for Air Quality Management Delhi, NCR: Average Report Criteria, https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data.
10. Wikipedia:Thiruvananthapuram, https://en.wikipedia.org/wiki/Thiruvananthapuram
11. EPA United States Environmental Protection Agency: Basic Information about Carbon Monoxide (CO) Outdoor Air Pollution, https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution
12. Wexler, P. (2014). Encyclopedia of Toxicology (3rd ed.). Academic Press.
13. National Library of Medicine : Ammonia,

https://pubchem.ncbi.nlm.nih.gov/compound/Ammonia
14. Pranaair : What is Air Quality Index (AQI) & how is it calculated?, https://www.pranaair.com/blog/what-is-air-quality-index-aqi-and-its-calculation/
15. Nair, J.P., Vijaya, M.S. (2023). Exploratory Data Analysis of Bhavani River Water Quality Index Data. Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems. https://doi.org/10.1007/978-981-19-3951-8_74.
16. Unacademy : Components of Time Series, https://unacademy.com/content/ca-foundation/study-material/statistics/components-of-time-series/
17. Tukey, J. W. (1977). Exploratory data analysis (Vol. 2). Addison-Wesley Publishing Company.
18. Towardsdatascience : Visualizing your Exploratory Data Analysis, https://towardsdatascience.com/visualizing-your-exploratory-data-analysis-d2d6c2e3b30e
19. Watthanacheewakul, L. (2021). Transformations for Left Skewed Data. In Proceedings of the World Congress on Engineering.
20. Aman Gupta. (2023, February 15). Types of Feature Selection Methods in ML. Feature Selection Techniques in Machine Learning, https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
21. Machine Learning Mastery : How to Develop LSTM Models for Time Series Forecasting, https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/
22. Bhanumathi S, Dr. S N Chandrashekara. (2021). Deep learning based BiLSTM architecture for lung cancer classification.12, 492-503.
23. Ali Jaber Almalki. (2020). Forecasting method based upon GRU-based deep learning model.