Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.101. No 1
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INTELLIGENT DEEP LEARNING BASED AQI PREDICTION MODEL WITH POOLED FEATURES

SANTHANA LAKSHMI V[1] AND VIJAYA M S[2]

[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

ABSTRACT

Airborne pollution poses a significant threat to public health, leading to detrimental health effects. Despite global economicgrowth, ensuring access to clean air has become increasingly challenging worldwide. The contamination of air occurs as dust particlesand smoke, released by vehicles and industries, suspend into the atmosphere, exacerbating the challenge of providing clean air forpeople. Hence, it is imperative to predict the Air Quality Index (AQI) to safeguard the lives of people, especially considering the severe health effects caused by the inhalation of small particles. This paper outlines a deep learning methodology for constructing Air Quality Index (AQI) prediction models. The models utilize hourly meteorological data and pollutant information, aiming to fulfill the critical requirement for precise assessments of air quality. The aim of this paper is to formulate predictive models for AQI in Thiruvananthapuram, Kerala, employing deep learning algorithms, thereby addressing the escalating challenge of air pollution in theregion. Deep neural network architectures, such as Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BI-LSTM), and Gated Recurrent Unit (GRU), are implemented to construct the prediction model. When compared to other algorithms,GRU demonstrated promising outcomes. The findings of this research contribute not only to the advancement of AQI predictionmodels but also highlight the practical significance of employing deep learning techniques for accurate and timely air quality assessments. The outcomes have practical implications for public health and environmental management, providing a basis for informed decision-making in mitigating the adverse effects of air pollution.

Keywords: *Ammonia, CO, Pollution, Prediction Models, Meteorological Data*

1. INTRODUCTION

The Air Quality Index (AQI) is employed as a dailymetric to convey the state of air quality, offering a quantifiable measure of how air pollution affects human health. It serves as a crucial tool in representing and communicating the impact of air quality on well-being. Predicting AQI aims to aid individuals in comprehending how the air quality in their vicinity influences their health. The values of the seven pollutants $PM_{2.5}$, PM_{10}, Ammonia, Carbon Monoxide, Ozone, Sulphur Dioxide, Nitrogen Dioxide [1] are used to calculate Air Quality Index.

Particulate Matter denotes the solid particles and liquid droplets found in the atmosphere. Particulate Matter enters the lungs and stays in the tissues for a long time. This causes cancerand other respiratory diseases [2]. Sulfur dioxide is a toxic, colorless gas with a pungent smell. Sulphuric acid exposure can have devastating consequences for those who have asthma. Ground-level ozone is one of the most dangerous pollutants in the atmosphere. Ammonia is a hazardous gas that causes cardiovascular illness in people who breathe it for an extended period. Thus, the value of the air quality index depends upon the level of presence of these particles in the air.

Inadequate air quality not only correlates with the release of harmful pollutants but also presents a direct risk to respiratory health, making it a critical concern for communities and policymakers alike. Particularly vulnerable are children and individuals with pre-existing respiratory conditions, who face heightened health risks when exposed to compromised air quality. Thus, building an effective forecasting model for the airquality index is crucial so that individuals can protect themselves by avoiding exposure to outdoor pollutants.

The Central Pollution Control Board has set the National Ambient Air Quality Standard, defining the necessary air quality level and incorporating a specified safety margin to ensure the protection of public health. The AQI breakpoints and the pollutant breakpoints are provided in Table 1 & Table2.

Table 1. Threshold Levels of AQI

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

Table 2. Breakpoints for the pollutants

An effective forecast of AQI can be made using machine learning, which is a part of artificial intelligence that helps systems learn from data. Deep Learning methods are a subset of machine learning technique which includes a neural network with more layers that helps in decision making by learningthrough examples as humans do.

Numerous researchers actively engaged in predicting the Air Quality Index (AQI) value using a variety of algorithms, including both conventional Machine Learning and advanced Deep Learning techniques. In [3], the authors employed various machine learning models for Air Quality Index (AQI) prediction. The objective of their study was to construct predictive models for forecasting pollutant levels, including PM2.5, using publicly accessible data for New Delhi. The utilized algorithms encompassed Linear Regression, Lasso Regression, XG Boost, Random Forest Regression, and K- Nearest Neighbors Algorithm. To evaluate the performance of these models, the authors employed metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Conclusively, the authors found that the Random Forest Regressor yielded the most accurate results, showcasing a Mean Average Error of 24.75, Mean Squared Error of 1675.42, and Root Mean Squared Error of 40.93 in this AQI prediction task.

In the paper [4], the authors investigated and compared three air pollution prediction techniques—Linear Regression, Random Forest Regression, and Convolutional Neural Network (CNN). Emphasizing the importance of Root Mean Squared Error (RMSE) as an indicator of accuracy, the study revealed that the Random Forest algorithm outperformed others for city day data with an MSE of 936. Conversely, CNN demonstrated superior performance for city hour data, yielding the lowest MSE of 1834. The paper delved into data preprocessing, addressing missing values, and proceeded to compare various machine learning and deep learning models. The suggested model was lauded for its utility in visualizing air quality, showcasing its potential for forecasting pollutant levels.

In [5], the authors introduced a significant contribution to air quality management in Vietnam, employing the WRF model and machine learning algorithms, specifically highlighting the effectiveness of the Extra Trees Regression model in forecasting PM2.5 concentrations in Ho Chi Minh City. The study meticulously evaluated the model's performance, showcasing impressive statistical indicators such as RMSE = 7.68 µg m–3,MAE = 5.38 µg m–3, and an R-squared value of 0.68. The model's accuracy was further underscored by a 74% accuracy rate in the confusion matrix. Notably, the research emphasized the simplicity and stability of the predictive model, built on acomprehensive dataset spanning over three years and utilizing a single machine learning algorithm. The study's findings highlighted the model's capability to provide hourly PM2.5 concentration predictions for short to medium-term durations, addressing pollution concerns and offering valuable insights into health impacts for effective early warning systems and air qualitymanagement in major cities.

The research outlined in [6] provided a holistic methodology for predicting fine particulate matter (PM2.5) and nitrogen oxide (NOx) concentrations in Jaipur city, India, addressing the pressing concern of air pollution. Employing a machine learning-based multiple linear regression (MLR) model and utilizing two months of 2018 data, the authors selected input variables through the Pearson correlation coefficient method. The study highlighted the efficacy of the MLR model in forecasting PM2.5 and NOx concentrations across three locations in Jaipur, achieving commendable accuracy with R2values ranging from 0.59 to 0.68 for PM2.5 and 0.56 to 0.81 forNOx. The research not only contributed to the field of air qualityprediction but also emphasized the importance of machine learning methodologies in overcoming limitations posed by traditional approaches. Additionally, the paper underlined the potential for implementing such models in developing countrieslike India, where limited monitoring infrastructure exists. Overall, the study provided valuable insights for mitigating air pollution and managing air quality in urban environments.

The literature survey provided a comprehensive overview of diverse studies conducted by researchers topredict the Air Quality Index (AQI) using various machine learning and deep learning algorithms. The details are summarized in Table 3. Despite the advancements in predicting AQI, there remains a need for a more robust and region-specific model that not only considers pollutant concentrations and meteorological factors but also incorporates the unique characteristics of the study area. Additionally, the current literature lacks a comprehensive comparison of deep learning models suchas LSTM and GRU with conventional machine learningalgorithms in the specific context of Thiruvananthapuram, Kerala. Thus, this study endeavors to fill this gap by formulating a problem that seeks to enhance the accuracy and applicability of AQI prediction models for the local environment.

Table 3. Performance Summary of Reviewed Papers

While considerable strides have been made in the field of air quality research, the persistence of air pollution as a pressing issue necessitates a closer examination of its enduring relevance. Despite advancements in technology and regulatory measures, the ubiquity and diverse sources of pollutants, coupled with evolving environmental challenges, contribute to the ongoing impact on air quality. This study aims to shed light on the continued significance of air quality concerns, emphasizingthe need for robust predictive models to address the dynamic nature of pollutants and their effects.

In this research work, we posit the hypothesis that implementing advanced deep learning architectures, specificallyLong Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BI-LSTM), and Gated Recurrent Unit (GRU),will lead to a significant improvement in the accuracy of Air Quality Index (AQI) predictions. Our hypothesis is based on the amalgamation of meteorological and pollutant features, coupledwith feature engineering, to capture nuanced relationships within the dataset. To test this hypothesis, we calculate AQI using pollutant features, adding it as the dependent attribute for developing robust AQI prediction models. The performance of these models is methodically assessed through metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-Squared value. This thorough evaluation ensures a comprehensive understanding of their predictive accuracy and reliability.

2. COLLECTION AND PREPARATION OF DATA

Air quality data is originated from the Central Control Room for Air Quality Management, a division of the Central Pollution Control Board in Delhi. The data is accessible throughan online portal [7]. Meteorological data spanning the years 2017 to 2020 for Thiruvananthapuram city in Kerala is procured from the Visual Crossing Website [8]. Comprising around 26,305 time series samples, this hourly dataset amalgamates meteorological parameters (Temperature, Relative Humidity, Dew, Sea Level Pressure, Cloud Cover, Visibility, Conditions, Icon, Solar Radiation, Barometric Pressure, Atmospheric Temperature, Rainfall, Feels-like, Wind Speed, and Wind Direction) with pollutant data (Ammonia, Particulate matter ofsize 2.5 and 10, CO, Ozone, SO2, and NOx) for in-depth analysis.

Relative Humidity is the percentage of water vapor present in the air out of the maximum amount [9]. Due to their small size and high polarity, water molecules can form strong bonds with various substances. The presence of water molecules suspended in the air significantly enhances the scattering of light by particles. Dew, in turn, is the condensed form of water vapor, appearing as droplets. The Dew Point signifies the temperature at which water vapor condenses to form droplets. The presence of dew in a surface layer reduces the pollutant concentration.

Barometric Pressure signifies the atmospheric pressure, while Sea Level Pressure reflects the thermal contrast between sea and land. Atmospheric pressure at any elevation, computed through a formula, is adjusted to a value approximating the pressure at sea level. Cloud Cover serves as an indicator of prevailing sky cloudiness.

Cloud Cover denotes the proportion of the sky typically obscured by clouds from a specific vantage point. Visibility gauges sky clarity, representing the maximum distance an object is discernible to the naked eye. The presence of aerosols in the air creates a white haze, impacting object identification in the distance. Solar radiation quantifies the sun's emitted energy [10]. Rainfall signifies the total water volume precipitated in an area. Wind Speed indicates the rate of wind flow from a particular direction. Notably, rainfall and wind speed exhibit an indirect relationship with the Air Quality Index: rainfall dissolves atmospheric pollutants, while wind speed disperses pollutant particles, contributing to their dissipation.

The acronym PM denotes particulate matter, encompassing both liquid droplets and solid air particles such as ashes, soot, and ash. Particle pollution encompasses inhalable particles of varying sizes, including small (PM10) and large (PM2.5) particles, with respective sizes of 10 and 2.5 micrometers. Exposure to these particles can cause respiratory issues, impacting the heart and lungs [11]. Carbon monoxide (CO), a colorless, odorless, and tasteless flammable gas, stands as the most prevalent airborne pollutant. It originates primarily from vehicles and the combustion of fossil fuels. Sulphur dioxide, a transparent gas with a pungent odor, is toxic. Its interaction with sulphuric acid produces sulphurous acid and sulphate particles. The main sources of sulphuric acidare humans and industrial waste.

Ozone gas (O3) consists of three oxygen atoms and is categorized into two types. The ozone found in the upper atmosphere protects humans from ultraviolet radiation, while ground-level ozone poses a significant threat as one of the most hazardous contaminants in the atmosphere [12].

Nitrogen oxide is produced through the reaction between nitrogen and oxygen, remaining inert at low temperatures but transforming into NOX during high temperatures. This NOX contributes to acid rain and smog. Transportation activities and fossil fuel combustion are the primary sources of nitrogen oxides. Ammonia, characterized by a pungent odor, combines with sulphates and nitrates to form PM2.5. Prolonged inhalation of this poisonous gas canlead to cardiovascular diseases [13].

Thusa total of 26,305 timeseries data with 22 features that includes 15 meteorological features and 7 pollutant features are prepared. To prepare the labelled instances, the values for the target variable AQI is calculated using the equation given below and added to the respective tuples.

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

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 features are summarized in Table 4 and sample meteorological data and air pollutant data is provided in Table5 and Table 6 respectively.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4. Meteorological and Pollutant features

Meteorological Features		Pollutant Features
Feelslike	Barometric Pressure	$PM_{2.5}$
Dew	Air Temperature	PM_{10}
Sea LevelPressure	Rainfall	Carbon Oxide
Cloudcover	Wind Speed	Sulphur Dioxide
Visibility	Wind Direction	Ozone
Temperature	Conditions	Nitrogen Oxide
RelativeHumidity	Icon	Ammonia
Solar Radiation		

Table 5. Sample Meteorological dataTable 6. Sample Pollutant data

2.1. Data Exploration

Exploratory data analysis involves utilizing statistical techniques and visualization methods to delve into the data,revealing latent patterns and trends. It serves as a vital preliminary step following data collection, wherein the data is observed, graphed and manipulated without making presuppositions. This process aids in evaluating data quality, performing data pre-processing, and selecting relevant features [14]. In our previous work the exploratory data analysis was performed which helped in identifying the most contributing features, outliers etc. The observations from the exploratory data analysis are explained below. Exploratory plots such as Heat Map, Boxplot, Pairplot, Barchart were used to get an overall understanding of the data and identify the correlation between meteorological features and pollutants. Heat Map gives a visual representation of the entire data. The objective of generating the heatmap is to understand the impact of all the values in the raw data in a single form [15]. The heat map generated for the data is given in Figure. 1.

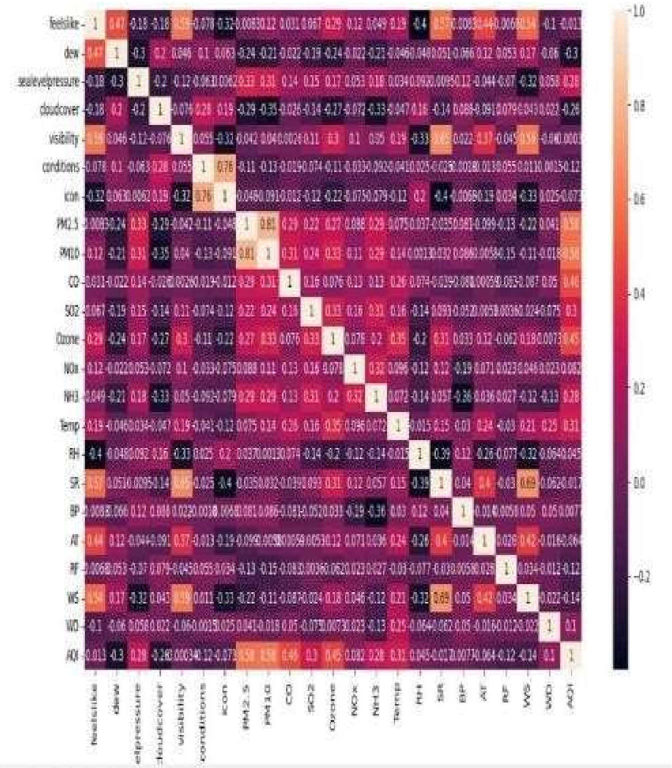

Figure. 1. Heat Map generated on the entire data

The heatmap analysis unveiled notable correlations within the dataset. Dew exhibited positive correlations with feels-like,cloud cover, atmospheric conditions, air temperature, and wind speed. Temperature showed positive correlations with feels-like, visibility, PM_{10}, sulfur dioxide, ozone, solar radiation, air temperature, wind speed, wind direction, and the Air Quality Index (AQI). Additionally, Sea Level Pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, Ozone, NOx, NH_3, temperature, and wind direction were positively correlated with AQI, while feels like, dew, cloud cover, atmospheric conditions, icon, rainfall, and wind speed exhibited negative correlations with AQI. However, no significant correlations were observed between cloud cover, barometric pressure, and AQI.

A histogram provides a visual representation of continuous data, organizing it into non-overlapping bins or ranges [16]. The primary purpose of a histogram is to illustrate the frequency of a feature's occurrence. Figure 2 displays the histogram generated from the raw data. The analysis indicates that, for the majority of days, humidity falls within the range of 65 to 70, temperature within 27 to 30, and dew value from 23 to 24. The observed minimum and maximum values for humidity were 50 and 100,respectively.

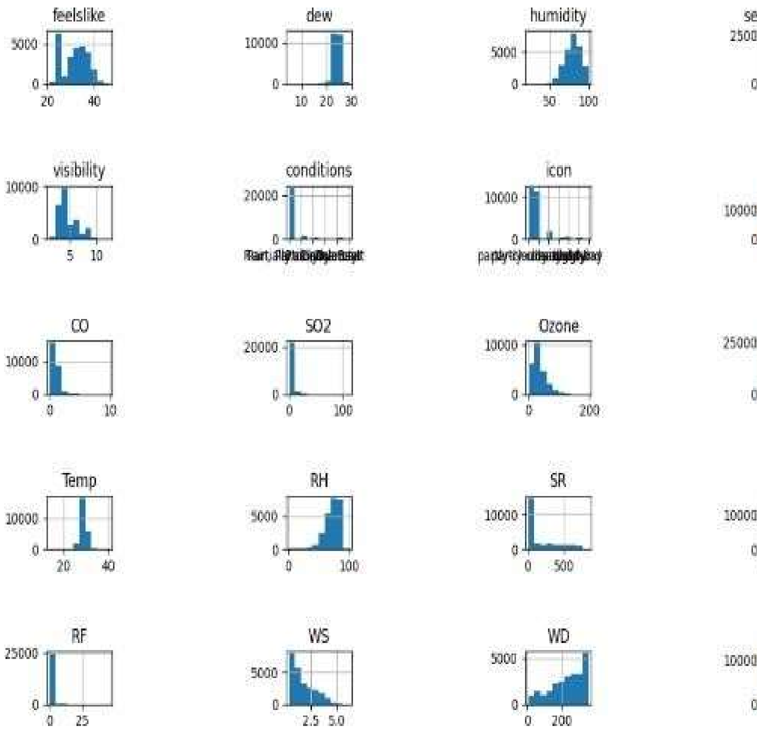

Figure. 2. Histogram generated on the entire data

Pair plots were used to identify pairwise relationships in the raw data. Boxplots provides a five number summary and was also used to identify the outliers present in the data. Pair plot generated for the raw data is given in figure 3. The pair plots generated depicts that the features sea level pressure, PM2.5, PM10, CO, NOx, NH3, SO2 and Ozone were positively correlated with Air Quality Index whereas feels like, dew, humidity, Wind Speed, Cloud Cover were negatively correlated with airquality index

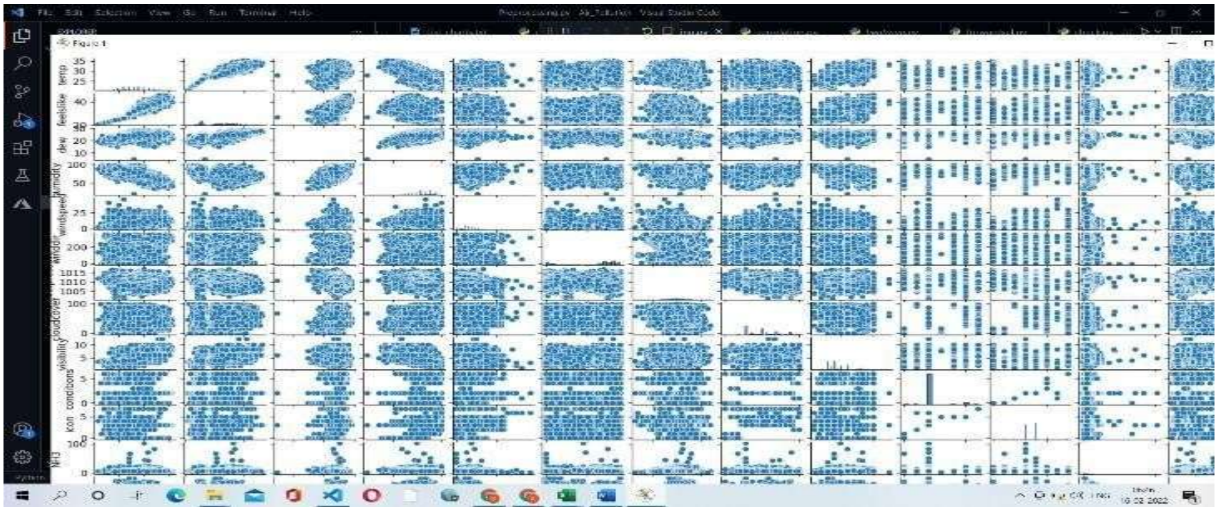

Figure. 3. Pair plot generated on the raw data

Correlation between each pair of features was identified to selectthe most contributing features and the results obtained are shownin Table 7. It has been identified that cloud cover and barometric pressure have no correlation with the prediction of AQI value.

Table 7. Correlation between the attributes andAQI

S.No	Attribute	Correlation Value	Result
1	Feelslike	- 0.012688724606251351	NegativelyCorrelated
2	Dew	- 0.3015968201961364	NegativelyCorrelated
3	Sea LevelPressure	0.28390300278636227	PositivelyCorrelated
4	Cloudcover	0.10126312865660123	Positively Correlated
5	Visibility	0.000343232025569354	No correlation
6	PM2.5	0.5756957872056067	Positively Correlated
7	PM10	0.5810464548312541	Positively Correlated
8	Carbon Oxide	0.45764413603029847	Positively Correlated
9	Sulphur Dioxide	0.30074500670790966	Positively Correlated
10	Ozone	0.44706794048480963	Positively Correlated
11	Nitrogen Oxide	0.18179587794498741	Positively Correlated
12	Ammonia	0.28033571115301464	Positively Correlated
13	Temperatur e	0.3119078283575685	Positively Correlated
14	Relative Humidity	0.45415418707998985	Positively Correlated
15	Solar Radiation	- 0.017490877066920537	Negatively Correlated
16	Barometric Pressure	0.0076884989859714635	No Correlation
17	Air Temperatur e	- 0.06382006724829173	Negatively Correlated
18	Rainfall	- 0.1207312355884903	Negatively Correlated
19	Wind Speed	- 0.1365778017395007	Negatively Correlated
20	Wind Direction	0.10193251948936194	Positively Correlated
22	Conditions	- 0.11623631138814557	Negatively Correlated
23	Icon	- 0.0733477949698372	Negatively Correlated

The analysis highlighted positive correlations between Sea Level Pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, Ozone, NOx, NH_3, temperature, and wind direction with AQI. Conversely, feels-like, dew, cloud cover, atmospheric conditions, icon, rainfall, and wind speed showed negative correlations with AQI. No correlations were found between cloud cover, barometric pressure, and AQI. Furthermore, features such as relative humidity, wind speed, sea level pressure, and all pollutant data were identified to have outliers, necessitating preprocessing. Tasks such as filling missing values, outlier removal, and normalization were identified through Exploratory Data Analysis (EDA) for attributes like NH3 andsea level pressure, ensuring effective preprocessing.

2.2. Preprocessing

For any data science project, data preprocessing and Exploratory Data Analysis (EDA) stand as crucial tasks. As a result of exploratory data analysis, it is understood that there is no correlation between Visibility, Barometric Pressure and Air Quality Index. So, these two features are not used for further processing. Among the other features, two features' conditions and icon are non-numerical variables. Most machine learning and deep learning algorithms cannot handle categorical data. The best practice is to convert them to numerical data and apply themin the algorithms. One of the methods to convert categorical data to numerical data is one hot encoding which encodes the categorical data to binary vectors. In this paper Label Encoder function from scikit-learn library is used here to convert them intonumerical variables.

Handling missing data is the essential step to be performed during pre-processing since the presence of missing values produces biased results. The techniques to fill the missing values include replacing with median value, replacing with mode value. Ignoring the entire record and Interpolation[17]. In this paper Interpolation technique is used to fill the missing values. Interpolation is the process of filling the missing values with the neighboring values.

Building an efficient model requires the dataset with no outliers. In this paper outliers are detected and replaced using an interpolation technique which falls under the category of clustering-based methods. Scaling refers to the handling of highly variable magnitudes, values, or units, which is another crucial pre-processing activity. When feature scaling is absent, machine learning algorithms tend to favor larger values over smaller ones, irrespective of the measurement unit. Standardization and Min-Max scaling represent approaches to address this issue.

Min-Max scaling is a method that adjusts a feature orobservation value, transforming it to a range between 0 and 1 within its distribution [18]. A feature value is rescaled using thevery effective standardization procedure so that its distribution has a mean value of 0 and a variance of 1. In this paper Min Max scaling is used for normalizing the values of all the attributes.

3. AIR QUALITY PREDICTION MODEL -DESIGN & METHODOLOGY

Predicting air quality is crucial as it serves as an early warning system to protect lives. This paper aims to explore the correlation between meteorological data and air pollutant features and leverage them to construct a predictive model for the AirQuality Index. The efficacy of the prediction model hinges on the quality of the data utilized. Therefore, collecting relevant data, conducting thorough pre-processing, and selecting pertinent features are pivotal in developing an efficient prediction model. The system architecture is illustrated in Figure 4.

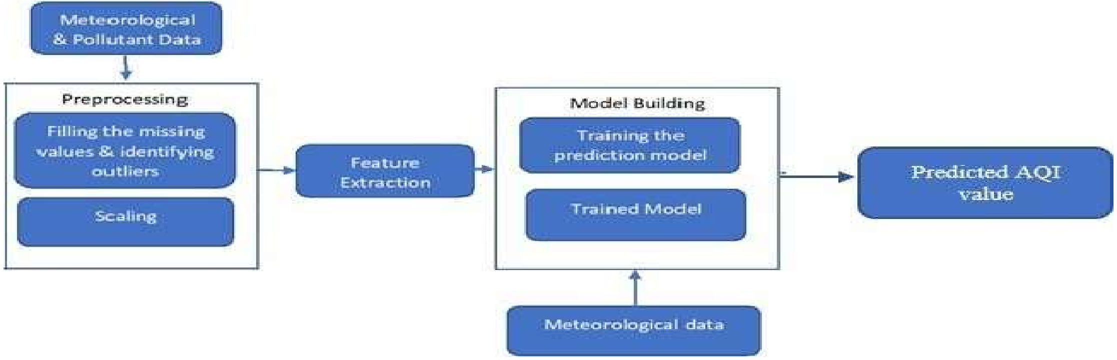
Meteorological & Pollutant Data
Preprocessing
Filling the missing values & identifying outliers
Scaling
Feature Extraction
Model Building
Training the prediction model
Trained Model
Predicted AQI value
Meteorological data

Figure 4. System Design Framework

3.1 Data Modelling and Training Dataset

The air quality time series data collected for Trivandrum area, has been analyzed and preprocessed as described in section 2. Feature selection methods are categorized into Filter methods, wrapper methods, embedded methods, and hybrid methods. In this paper one of the filter techniques called correlation coefficient is used to identify the attributes that do not contribute to the prediction. One ofthe wrapper methods called select k best is used to identify the most contributing features [19]. As a result of identifying correlation coefficients, the features such as Visibility and Barometric Pressure are removed as there is no correlation between them and AQI. As a result of the select k best method, the best 15 most contributing attributes are identified.

Finally, a dataset with 26,305 instances and 16 attributes has been developed. Since AQI prediction process is modelled as regression task, the attributes such as Temperature, Dew, Atmospheric Temperature, Cloud cover, conditions, icon, Rainfall, Wind Speed, $PM_{2.5}$,PM_{10}, CO, SO_2, Ozone, NOx and NH_3 are considered as independent variables and the AQI is a target variable.

To create a prediction, model the dataset is split for training and testing. 80% of the records are used for the training and 20% are used for testing the model. The algorithms such as LSTM, Bidirectional LSTM and GRU are used to build a forecasting model.

3.2 Methodology

Deep Learning, a subset of Artificial Intelligence and a variant of machine learning, employs densely layered architectures reminiscent of human decision-making[20].With the ability to handle intricately interconnected and diverse unstructured data, deep learning equips machines to tackle complex problems. As deep learning algorithms accumulate knowledge, their performance improves, and they acquire expertise through learning from examples. The potency of deep learning is harnessed with considerable processing power and extensive information, making it adaptable to various data types. Notable deep learning algorithms encompass Convolutional Neural Network, Long Short-Term Memory, Recurrent Neural Network, Generative Adversarial Network, Radial Basis Function, Multilayer Perceptron, Deep Belief network, and Autoencoders. This paper explores advancements in Recurrent Neural Network, such as LSTM, Bidirectional LSTM, and Gated Recurrent Unit, for constructing AQI prediction models.

3.2.1Long Short-Term Memory (LSTM)

Long Short-Term Memory Network is a type of RNN suitable for time series prediction. The LSTM cells could addlong term memory which makes the forecasting model powerful and provides accurate results. If a long trend is available in the data, LSTM is the best choice of algorithm for prediction. Even Though GRU is faster than LSTM, LSTM provides better accuracy than GRU as it includes 3 gates whereas GRU has only 2 gates. It can also handle the vanishing gradient problem faced by RNN. RNN keeps track of previousdata and employs it when processing new input. Addressing the challenge of vanishing gradients, traditional Recurrent Neural Networks (RNNs) exhibit a drawback in recalling long-term dependencies. This limitation is specifically overcome by Long Short- Term Memory (LSTM) networks. An LSTM consists of three distinct components, each serving a unique purpose. The initial segment determines the necessity of retaining or discarding information from the previous timestamp. The subsequent segment focuses on assimilating new insights from the input data. Finally, the third segment conveys updated information from the current timestamp to the succeeding one [21]. These components are categorized as follows, Forget Gate, responsible for the first part, Input Gate, constituting the second part and Output Gate, representingthe final element.

3.2.2. Bidirectional Long Short-Term Memory

The concept of incorporating sequence information in both directions—backward (future to past) and forward—is encapsulated in the term "Bidirectional Long Short-Term Memory" (Bidirectional LSTM), where "bidirectional" pertains to the flow of information from both past to future.Distinct from the conventional LSTM, a Bidirectional LSTM (BI-LSTM) facilitates dual input streams that traverse in both directions. While the standard LSTM enables input to flow in one direction—either backward or forward— BI-LSTMs accommodate bidirectional input, capturing insights from both past and future contexts [22]. This augmentation equips the network with enriched informational resources, enhancing its access to contextual understanding.

3.2.3 Gated Recurrent Unit (GRU)

Gated Recurrent Unit is one of the most advanced RNN that are widely used for time series forecasting. GRU has simple architecture. Reset and update gates are two ofits gates. The gates are used to regulate the information flow. The reset gate is responsible for short term memory and update gate for long term memory. GRU works by creating candidate hidden states by obtaining values such as input, hidden state from previous timestamp and output of reset gate applied to tanh function. The value of the reset gate determines how much information from the previous information to be considered. The candidate state and update gate are then used to build hidden states [23]. The output of the update gate is very critical as it controls both the historical information and new information which comes from the candidate state. Due to the simple architecture GRU is faster to train when compared to LSTM and BILSTM.

3.3. Hyperparameter Tuning

The pre-processed data undergoes application to deep learning algorithms for constructing an effective AQI prediction model. Enhancing the performance of deep learning models involves hyperparameter tuning. Hyperparameters such as the number of hidden layers, activation function, epochs, batch size, and dropout rate are adjusted to establish an optimized model. Positioned between the input and output layers, hidden layers play a pivotal role in model architecture [24]. The decision regarding the quantity of hidden layers hinges on the intricacy of the problem at hand.

Optimizers serve as strategies to minimize loss by adjusting weights or learning rates. Diverse optimizers encompass gradient descent, stochastic gradient descent, momentum-based gradient descent, RMSProp, Adagrad, Adam, and others. In this context, the Adam optimizer is leveraged for optimization. Adam optimizers exhibit efficacy due to the integration of momentum gradient descent's history preservation and RMSProp's adaptive learning rate. The concept of epochs signifies the number of times the network processes the complete training dataset. Training with numerousepochs could lead to overfitting, this is mitigated through an early stopping technique. Early stopping automatically halts training after a certain number of epochs when performance stabilizes. Additionally, to counteract overfitting, dropout techniques are employed. Drop out entails the random exclusion of neurons during training, temporarily removing their contribution to downstream neuron activation during the forward pass, and bypassing weight updates during the backward pass. The quantity of training samples used in a single iteration is termed the "batch size" in machine learning..

As AQI prediction task is formulated as a regression problem, the pattern of independent variables is self learnt through representation learning from the above training dataset to model the target variable using LSTM, BIRNN, GRU. Thus independent AQI prediction models are built. Model performance assessment employs metrics including Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R Square. Root Mean Squared Error serves as a primary evaluation metric. It gauges the dissimilarity between predicted and observed values, quantifying the average error through squaring and subsequent calculation of the square root [25]. This approach lends substantial weight to significant errors due to the prior squaring of errors before averaging. TheMean Absolute Error, on the other hand, represents the mean of absolute prediction errors across all instances within the test set. A pivotal evaluation metric, the R2 score (pronounced as R squared), stands as the coefficient of determination. Operating by contrasting expected and observed values, a higher R2 score signifies better model performance. The effectiveness of the model is validated through these metrics. Model performance is deemed favorable when Root Mean Squared Error and Mean Absolute Error exhibit low values, while the R2 score demonstrates a high value.

4. EXPERIMENT AND RESULTS

In the deep learning-based approach to predict the air quality index the dataset with meteorological features and pollutant features are used for building the model. The dataset includes 26,305 instances with 16 features. Among which 80 percent of the records are used for training and 20 percent of the records are used for testing. Different optimizers, including Adam, Adagrad, and Gradient Descent, were evaluated, and dropout rates ranging from 10 to 20 were tested. Batch sizesbetween 20 and 64 were utilized during experimentation, and the impact of various epoch sizes, such as 100, 200, and 300, was observed. The LSTM, BILSTM, and GRU algorithms were fine-tuned with hyperparameters such as hidden layers, optimizer, epoch, batch size, and dropout to optimize the deep network. The experiments were executed with the parameter configurations outlined in Table 8, leading to the developmentof the prediction model.

Table 8. List Of Hyperparameters And Optimum Value

Hyperparameter	Values
Optimizer	Adam
Batch size	64
Dropout	10
Learning rate	0.01

The performance metrics such as Root Mean Square value, Mean Absolute error and R-Squared value observed for evaluating the performance of the AQI prediction models for various epochs using LSTM algorithm are provided in Table 9. From the prediction result it was observed that, as the number of epochs increases, the performance of the model also increases and becomes stable at the epoch 300 for LSTM. Mean Absolute Error was more when the epoch is 100 and got reduced when the epoch is 300. Mean absolute error was 0.5157when the epoch was 100 and reduced to 0.4182 when the epoch was 300 for LSTM. The high Root Mean Square error 0.7722 was observed at the epoch 100 whereas it is reduced to 0.6279 at the epoch 300. R-Squared value was 0.4651 for LSTM at the epoch 100 and increased to 0.6685 at the epoch 300.

Table 9. – Performance Evaluation Of LSTMModel For Various Epochs

	LSTM		
Epochs	100	200	300
MAE	0.5157	0.4772	0.4182
RMSE	0.7722	0.6708	0.6279
R2	0.4651	0.5292	0.6685

The difference between the expected value and predicted value captured when using LSTM for the epoch 300 are provided in Figure 5.

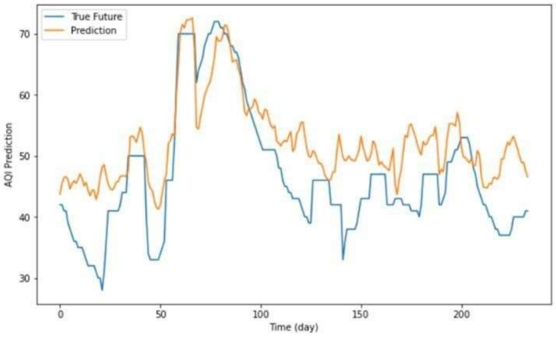

Figure. 5. Difference Between Expected Value And Observed Value For LSTM

The performance of the BILSTM model evaluated using the metrics RMSE, MAE and R2 value are provided in Table10.Mean absolute error is 0.7714 at the epoch 100 and gradually reduced and reached 0.5287 at the epoch 200. Root mean squared error is less at the epoch 200 whereas high RMSE is observed during the epoch 100.R squared value was 0.2108 at the epoch 100 whereas it is increased to 0.6000 at the epoch 200.

Table 10. – Performance Evaluation Results Of BI - LSTM ModelFor Various Epochs.

	BI-LSTM		
epochs	100	200	300
MAE	0.7714	0.5287	0.7846
RMSE	0.9405	0.7987	0.8203
R2	0.2108	0.6000	0.5901

The difference between the expected value and predicted value captured when using BILSTM for the epoch 200 are providedin Figure 6.

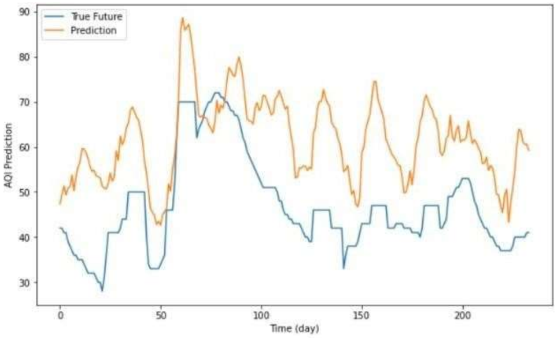

Figure. 6. Difference Between Expected Value And ObservedValue For BI -LSTM

Performance of the generated GRU model is evaluated and the results observed at the various epochs are provided in Table 11. Mean absolute error was 0.7662 for epoch 100 and got reduced to 0.2136 when the epoch size was increased to 300. Similarly high RMSE value 0.8529 was obtained at the epoch 100 and low error value 0.3168 was observed at the epoch 300. R2 value was 0.4539 at the epoch 100 and increased to 0.8566 at the epoch 300.

Table 11. – Performance Evaluation Results OfGRU Algorithm For Various Epochs.

	GRU		
Epochs	100	200	300
MAE	0.7662	0.6634	0.2136
RMSE	0.8529	0.7139	0.3168
R2	0.4539	0.6690	0.8566

The difference between the expected value and predicted value captured when using GRU for epoch 300 are provided inFigure 7.

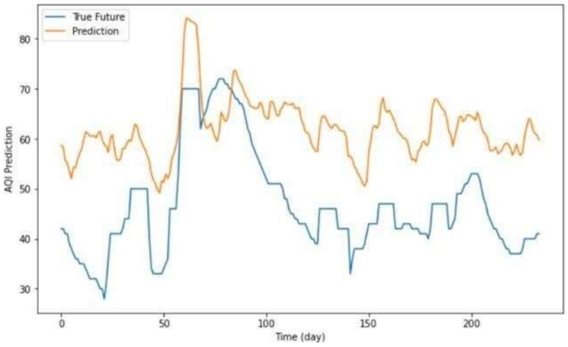

Figure. 7. Difference Between Expected Value And Observed Value For GRU

After comparing the performance of the models built using LSTM, BI-LSTM and GRU, it was understood that high R2 value was obtained for the prediction model built using GRU. Root Mean Squared error and Mean absolute error was also low in GRU when compared to LSTM and BILSTM.

The performance of the AQI prediction models developed using deep learning algorithms such as LSTM, BILSTM and GRU observed at the epoch 200 are compared with prediction models developed using traditional machine learning algorithms. The machine learning algorithms such as Linear Regression, Support Vector Regression, Decision Tree Regression and XGBoost algorithms were used for building the AQI prediction model. The same performance evaluation metrics are used for comparison. The values observed for the performance metrics of deep models and traditional machine learning based AQI models are provided in Table 12.

Table 12. Comparative Performance ResultsOf AQI Prediction Models

When considering the error rate XGBoost algorithm isbetter than the other machine learning algorithms. The high RootMean squared error 0.9462 is observed in linear regression whereas the error rate is reduced to 0.6937 in XGBoost algorithm. The R Square value0.4896 was observed for support vector regression whereas it got improved to 0.5936 with XGBoost algorithm. The comparison of deep learning models vsmachine learning based AQI prediction models is illustrated in Figure. 8.

Figure. 8. Comparative Performance Analysis Of AQIPrediction Models

By comparing the performance of deep learning algorithms with machine learning algorithms, it was clearly understood that the lower error rate was observed for GRU whencompared to all the other algorithms. Similarly high R Squared value was obtained for GRU. Thus, an accurate AQI prediction model can be built using the GRU algorithm.

When comparing the performance of Air Quality Index (AQI)models constructed through deep learning architectures such as LSTM, BILSTM, and GRU with the outcomes detailed in Table 3 from the reviewed papers, significant enhancements in AQI prediction are evident. The traditional Machine Learning Algorithms, including Linear Regression, Lasso Regression, XGBoost, Random Forest, and K Nearest Neighbors, yielded higher errors with a MAE of 24.74, MSE of 1675.42, and RMSE of 40.93. In contrast, the GRU-based AQI prediction model demonstrated substantially lower errors, achieving a minimum MAE of 0.2136 and RMSE of 0.3168. The Extra Trees Regression model, used for PM2.5 prediction, exhibited an RMSE of 7.68, MAE of 5.38, and R-squared value of 0.68. Comparatively, the GRU-based AQI prediction model outperformed with an accuracy of 0.8566. The superiority of the GRU architecture was further evident in predicting PM2.5 and NOx, where it achieved an impressive accuracy of 0.8566, surpassing the R-squared values of 0.68 and 0.81 reported in the literature. These findings underscore the remarkable efficacy of deep learning models, particularly GRU, in enhancing the precision and accuracy of AQI predictions.

From this research work, the following observations are made. The experimental results demonstrate that machine learning algorithms can also be used for time series AQI forecasting. But when comparing the performance with deeplearning algorithms, the prediction accuracy is more for deeplearning based AQI prediction models as it can handle multiple input variables with noisy complex dependencies. One of the advantages of deep learning networks is their capacity to extractpatterns from input data that spans relatively extended sequences. Through feature selection best contributing features are identified which helped the deep learning architectures to identify the trendpresent in the data. The ability to fine-tune the parameters totheir ideal values improved forecast accuracy and decreased error rate. The quality of AQI prediction is improved by including Meteorological features with pollutant data since they have a greater impact on determining the air quality. Thus, the enhanced air quality prediction model with meteorological andpollutant time series data has proven to be an effective tool inpredicting the air quality in different locations.

5. CONCLUSION

This work models the prediction of the Air QualityIndex (AQI) as a task centered around time series forecasting. It showcases the application of deep learning methodologies to forecast the AQI value, employing sophisticated neural network structures like LSTM, BILSTM, and GRU. The study incorporates a time series dataset comprising 8 meteorological attributes and 7 pollutant features, amalgamating them into a unified representation of air quality data. The initial step involvedExploratory Data Analysis (EDA), which provided insights intodata distribution and the significance of individual parameters inpredicting the air quality index. Through various preprocessing techniques, the air quality dataset was refined to ensure its suitability for analysis. Subsequently, deep learning-based models for AQI prediction were devised, employing LSTM, BILSTM, and GRU architectures, followed by meticulous performance evaluation. The performance of the models is evaluated using Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and R Squared value. When comparing the performance of all the algorithms, the model builtusing GRU showed superior accuracy. Notably, the study raises pertinent questions regarding the applicability of predictive models in regions with limited historical data, suggesting avenues for future research to explore techniques like data augmentation and transfer learning to enhance the accuracy insuch scenarios.

REFERENCES

[1]. Prana Air. What is Air Quality Index (AQI) and ItsCalculation. [Online]. Available: https://www.pranaair.com/blog/what-is-air-quality-index- aqi-and-its-calculation

[2]. Yadav IC, Devi NL. Encyclopedia of EnvironmentalHealth. 2019.

[3]. Sajjan A, Begam MF, Dubey A. Predicting Air QualityIndex using most suitable ML model. Int J Adv Res ComputCommun Eng. 2023 Oct;12(10). doi: 10.17148/IJARCCE.2023.121015.

[4]. Banara S, Singh T, Thenuia Y, Nandanwar H, Chauhan A. Air Pollution Forecasting using Machine Learning and DeepLearning Techniques. J Xi'an Shiyou Univ, Nat Sci Ed.2018;18(5):42-46.

[5]. Minh VTT, Tin TT, Hien TT. $PM_{2.5}$ Forecast System byUsing Machine Learning and WRF Model, A Case Study:Ho Chi Minh City, Vietnam. Aerosol Air Qual Res. 2021;21. https://doi.org/10.4209/aaqr.210108

[6]. Goyal S, Sharma R. Prediction of the concentrations of $PM_{2.5}$and NOx using machine learning-based models. Mater Today Proc. 2023. https://doi.org/10.1088/1755- 1315/446/3/032113.

[7]. Central Pollution Control Board (CPCB). [Online]. Available: https://app.cpcbccr.com/ccr/#/caaqm-

dashboard-all/caaqm-landing/data.

[8]. Visual Crossing. [Online]. Available:https://www.visualcrossing.com/weather-data.

[9]. National Weather Service. Humidity. [Online]. Available:https://www.weather.gov/lmk/humidity.

[10]. Gomez I, et al. Evaluating the influence of air pollution on solar radiation observations over the coastal region of Alicante (Southeastern Spain). J Environ Sci. 2023;pp.633-643.

[11]. World Health Organization (WHO). Ambient (Outdoor) Air Quality and Health. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[12]. Environmental Protection Agency (EPA). Ground-level Ozone Basics. [Online]. Available: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics.

[13]. Breeze Technologies. Major Air Pollutants: Their Impact and Sources. [Online]. Available: https://www.breeze- technologies.de/blog/major-air-pollutants-their-impact-and-sources/

[14]. Analytics Vidhya. Exploratory Data Analysis using DataVisualization Techniques. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/explorator y-data-analysis-using-data-visualization-techniques.

[15]. Nair JP, Vijaya MS. Exploratory Data Analysis of Bhavani River Water Quality Index Data. In: Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems. Springer, Singapore. 2023. https://doi.org/10.1007/978-981-19-3951-8_74.

[16]. ASQ. Histogram. Available from: https://asq.org/quality-resources/histogram

[17]. Missing Values in Time Series. [Online]. Available: https://www.section.io/engineering-education/missing- values-in-time-series.

[18]. Machine Learning Mastery. StandardScaler and MinMaxScaler Transforms in Python. [Online]. Available:https://machinelearningmastery.com/standardscaler-and- minmaxscaler-transforms-in-python/.

[19]. Nair R, Bhagat A. Feature Selection Method To Improve The Accuracy of Classification Algorithm. Int J InnovativeTechnol Explor Eng (IJITEE). 2019;8:124-127.

[20]. TechTarget. Deep Learning. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network.

[21]. Towards Data Science. LSTM Networks: A DetailedExplanation. [Online]. Available: https://towardsdatascience.com/lstm-networks-a-detailed- exlanation-8fae6aefc7f9.

[22]. Ali W, Yang Y, et al. Aspect-Level Sentiment Analysis Based on Bidirectional-GRU in SIoT. IEEE Access. 2021.https://doi.org/10.1109/ACCESS.2021.3078114.

[23]. Zhang A, et al, Dive into Deep Learning,2022

[24]. DeepAI Hidden Layer, Machine Learning. [Online] Available : https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning.

[25]. Analytics India Magazine. A Guide to Different EvaluationMetrics for Time Series Forecasting Models. [Online].Available: https://analyticsindiamag.com/a-guide-to- different-evaluation-metrics-for-time-series-forecasting- models/

Table 2. Breakpoints For The Pollutants

AQI Category	PM$_{10}$ 24-hr	PM$_{2.5}$ 24-hr	NO$_2$ 24-hr	O$_3$ 8 hr	CO 8 hr (mg/m3	SO$_2$ 24-hr
Good	0-50	0-30	0-40	0-50	0-1.0	0-40
Satisfactory	51-100	31-6	41-80	51-100	1.1-2.0	41-80
Moderate	101-250	61-9	81-80	101-168	2.1-10	81-380
Poor	251-350	91-12	181-280	169-208	10.1 – 17	381-800
Very Poor	351-430	121-25	281-400	209-748	17.1-34	801-1600
Severe	430+	250+	400+	748+*	34+	1600+

Table 3. Performance Summary Of Reviewed Papers

	Objective	Algorithm	Performance
[3]	Predict AQI value	Linear Regression, LassoRegression, XG Boost, Random ForestRegression, and K-Nearest Neighbors Algorithm.	MAE - 24.74 MSE - 1675.42 RMSE – 40.93
[4]	Predict AQI	Linear Regression, Random Forest Regression, and Convolutional NeuralNetwork (CNN)	MSE – 1834 RMSE – 42.82

[5]	Forecast PM2.5 value	Extra Trees Regression	RMSE - 7.68 µg m–3
			,
			MAE = 5.38 µg m–3 ,
			R-Squared = 0.68

[6]	Forecast PM2.5 and NOX	Multiple regressior linear (MLR)	R-Squared – 0.68 forPM2.5 R-Squared – 0.81 forNOX

Table 5. Sample Meteorological Data

Datetime	Dew	Cc	Vis	Pm	SR	bp	Ws
2017-07-01T00:00:00	23	50	3	12	85	22	6.5
2017-07-01T01:00:00	23	50	3	14	85	22	1.25
2017-07-01T02:00:00	23.7	44.3	3.6	11	86	22	1.75
2017-07-01T03:00:00	22	36.4	3	11.5	86	22	26.25
2017-07-01T04:00:00	22	54.5	2	12.25	86.5	22	10
2017-07-01T05:00:00	23.1	75.1	3.6	10.25	87.75	22.25	19.25
2017-07-01T06:00:00	22	50	3	12	88	33.5	43.25
2017-07-01T07:00:00	22	36.4	3	16.75	85.75	136.5	1

Table 6. Sample Pollutant Data

Datetime	PM	PM_{10}	CO	SO_2	ozone	NOX
2017-07-01T00:00:00	12	44	0.56	3.45	10.12	2.77
2017-07-01T01:00:00	14	38.25	0.56	3.9	12.52	2.73
2017-07-01T02:00:00	11	36.75	0.55	4.38	15.9	3.62
2017-07-01T03:00:00	11.5	30.25	0.56	4.25	15.87	2.53
2017-07-01T04:00:00	12.25	31.5	0.28	4.37	15.93	2.73
2017-07-01T05:00:00	10.25	31.75	0.34	4.07	12.08	4.12
2017-07-01T06:00:00	12	29.5	0.41	3.2	12.2	3.7
2017-07-01T07:00:00	16.75	36.75	0.45	2.98	12.08	4.02
2017-07-01T08:00:00	21.75	47	0.58	2.8	16.63	3.85
2017-07-01T09:00:00	19	54	0.56	2.67	13.3	4.35

Table 12. Comparative Performance ResultsOf AQI Prediction Models

	LSTM	BI-LSTM	GRU	LR	SVR	DTR	XGBoost
MAE	0.4182	0.5287	0.2136	0.8462	0.7048	0.6639	0.5849
RMSE	0.6279	0.7987	0.3168	0.9462	0.80	0.835	0.693
R2	0.685	0.6000	0.8566	0.5278	0.4896	0.5813	0.5936

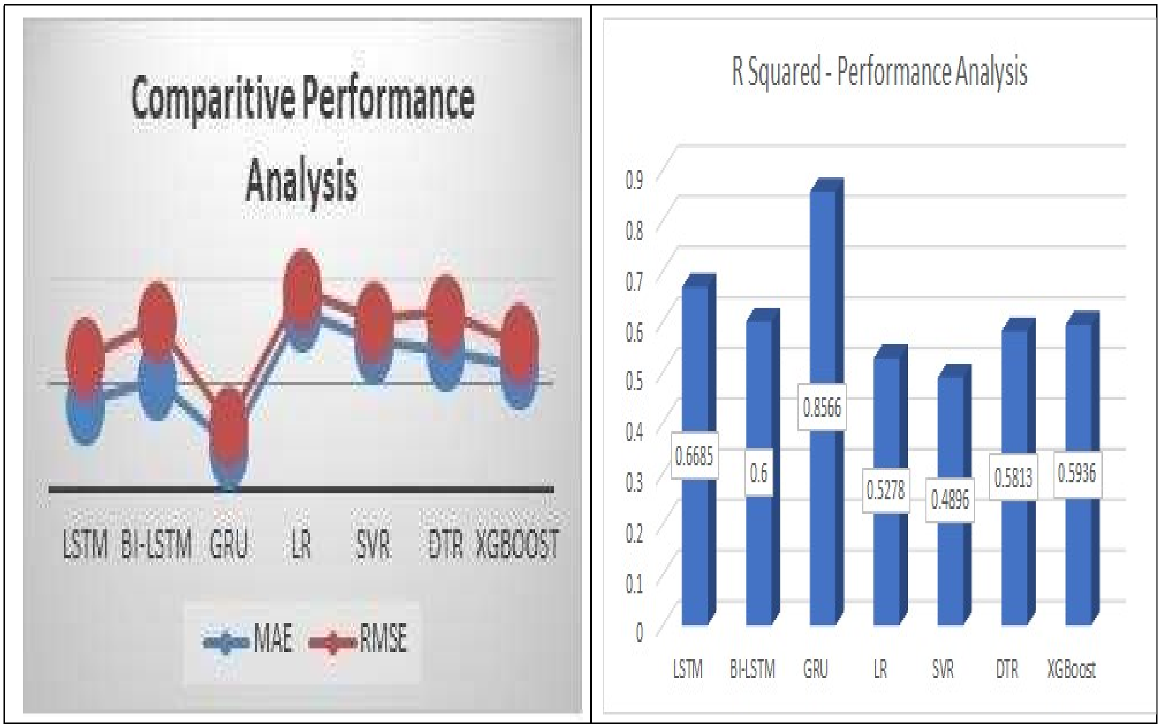

Figure. 8. Comparative Performance Analysis Of Aqi Prediction Models