https://doi.org/10.48047/AFJBS.6.12.2024.303-315

African Journal of Biological Sciences
Journal homepage: http://www.afjbs.com
Research Paper Open Access

Enriching Air Quality Index prediction through hybrid neural network

Ms.Shubhangi P.Phadtare[1][0009-0007-1200-1261],
Mrs.Varsha A.Jujare[2][0009-0009-4596-0711]
Dr.Amit J.Chinchawade[3][0000-0002-3815-3163]

[1]Sharad Institute of Technology college of Engineering,Yadrav,Kolhapur,Maharashtra 416115.
phadtareshubhangi782@gmail.com

[2]Sharad Institute of Technology college of Engineering,Yadrav, Kolhapur,Maharashtra 416115.
varshajujare@sitcoe.org.in

[3]Sharad Institute of Technology college of Engineering,Yadrav,Kolhapur,Maharashtra 416115.
amitchinchawade@sitcoe.org.in

Article History

Volume 6 Issue 12, 2024
Received: 25 May 2024
Accepted : 25 June 2024
doi:
10.48047/AFJBS.6.12.2024.303-315

***Abstract* - The health, comfort, and well-being of both humans and animals are directly affected by the air quality index (AQI), making its monitoring essential for multiple reasons. By predicting toxic gases and AQI matter quickly, allowing for immediate actions to enhance air quality. Some machine learning models exist that operate in solo mode, predicting the AQI index without yielding a satisfactory result. Therefore, it is necessary to hybridize neural networks to accurately predict the air quality prediction index, enabling the concerned departments to take appropriate action. As a result, the proposed model considers not only air quality index data, but also weather data, which is a blend of an artificial neural network and a bidirectional Convolution Neural network (CNN)-Long short term neural network(LSTM) model. The genetic algorithm catalyzes this hybrid model to accurately predict the air quality index. The rigorous evaluation of the implemented model for the parameters RMSE and accuracy yields good results of 2.987 and 98.98, respectively, which is again far better than the other solo models that predict the air quality index.**

Keywords: *Air Quality Index Prediction (AQI), Artificial Neural network(ANN), Long Short term memory(LSTM), Hybrid Neural network,* Bidirectional Convolution Neural network (CNN), *Genetic algorithm.*

I INTRODUCTION

There are three aspects that affect air quality: input àir contaminants from both nearby and faraway places, the air's dispersive qualities as a function of weather conditions, and the cleaning effects of natural disasters like rain, thunderstorms, and cyclones. We have some say over emissions from nearby sources of air pollution, but we have no say over more distant ones, such as dust from deserts, forest fires, volcanoes, dust storms, etc. The air's dispersive qualities are also very varied, depending on factors including inversion, wind speed, and direction. Same goes for where we are—in a valley, for example, or where the water drains away. Any location on Earth might experience unpredictable air quality due to a conglomeration of uncontrollable natural phenomena such as rain, cyclones, thunderstorms, dust storms, etc. Things are obviously different for Antarctica. Pollution is occurring on a worldwide scale, with no identifiable origin in this case.

The air quality index is used to measure the quality of the air for day-to-day analysis. The range of the air quality index can vary from 0 to 500. Air pollution levels and their respective health hazards are generally measured with the Air Quality Index (AQI) proportions. If the air quality index value is less than 100, it is considered satisfactory; on the other hand, any air quality index value that crosses more than 100 is considered unsatisfactory. If the value continues to rise, it is another alarming sign of the devastating effects on health, particularly on the lungs. The major gases that are considered for the valuation of the air quality index analysis are nitrogen dioxide (NO2), sulfur dioxide (SO2), oxygen (O3), ammonia (NH3), and lead (Pb). Because of these toxic gases, if the air quality index crosses a value of more than 100, it will start affecting human and animal health severely
Each of these contaminants has a sub-index determined by taking into account the measured ambient concentrations, the relevant standards, and the anticipated health impact. All AQI sub-indices are reflective of the worst one. The medical experts who make up the organization have also contributed significantly to the process of predicting the health effects of various AQI categories and contaminants. Here are the air quality index (AQI) values, ambient concentrations (health breakpoints), and anticipated health implications for the eight contaminants that were identified:
Good (from 0 to 50) Low-Level Effect ,Very good (51–100)} In certain individuals, it may induce a slight annoyance during breathing. Pollution levels are moderate (101-200). People with heart conditions, children, and the elderly may experience difficulty breathing, and those with lung diseases, such as asthma, may also experience difficulty breathing. Scathing (201-300) Some individuals may experience difficulty breathing after extended exposure, and those with cardiac conditions may also feel pain. Decent (301-400) by all accounts Prolonged exposure could lead to respiratory illnesses in humans. Individuals predisposed to heart and lung conditions may see a more severe impact.Very serious (401–500) Potentially harmful to the respiratory system of otherwise healthy individuals and extremely dangerous to those who already have heart or lung conditions. Even brief periods of physical activity might have negative effects on health.

With the assistance of intelligent grid systems and the powerful predictive capabilities of artificial intelligence, it is possible to control the supply and demand of renewable energy in an effective manner. One example of anything that has the potential to optimize efficiency while simultaneously reducing expenditures and unnecessary carbon pollution is the improvement of weather forecasting.There are a number of applications for artificial intelligence (AI) in the subject of environmental, social, and governance (ESG) studies. These applications include the tracking and monitoring of water usage, air pollution, and other environmental challenges. The use of artificial intelligence has the potential to be an effective weapon in the struggle for social justice and equality.There is the potential for the application of machine learning algorithms to the prediction of air pollution levels and quality indices. Additionally, artificial intelligence-enabled equipment such as motion-sensing cameras and drones can be deployed to collect and analyze large amounts of data pertaining to biodiversity.

An aerial-ground air quality sensing framework using UAV swarms for fine-grained monitoring and forecasting of air quality with privacy-preservation federated learning is proposed by Yi Liu et al. in this research [1]. To achieve energy-efficient AQI scale inference, the author first uses the lightweight Dense-MobileNet model to learn the haze feature of UAV-taking haze photos. The second point is that the author suggests a graph-topology based GC-LSTM model to achieve both current and future AQI predictions, which would further enhance the aerial sensing system's inference accuracy. While comparing the suggested framework against 3D CNN, 2D CNN, and SVM methods—all of which could jeopardize privacy while forecasting—the author assesses its performance on a real-world dataset. The outcomes demonstrate that the suggested approach outperforms the current ones. This is one of the first studies on air quality forecasts using federated deep learning, as far as the author is aware.

Prior data-driven deep-learning methods mainly disregarded domain-specific knowledge-integration and uncertainty metrics, as described by [2] Yang Han et.al., when making air pollution predictions. This research delves into the air pollution prediction issue using a domain-specific Bayesian deep-learning strategy. Case studies of Beijing, China, and London, UK demonstrate that, on average, domain-specific knowledge and Bayesian approaches can reduce prediction errors by 12.4% for London and 3.7% for

Beijing, respectively. In addition, the best hybrid Bayesian models can enhance the traditional hybrid baselines for London by 9.5% and Beijing by 1.8%, respectively, and they can attain the lowest prediction errors. The results presented by the author emphasize the value of incorporating domain-specific knowledge and imply that the use of Bayesian approaches enhances the performance of conventional deep-learning models. Additionally, it enables the integration of several forecast tactics, leading to even more precise predictions.

In [3] their explanation of the problems with air pollution forecasting, Jovan Kalajdjieski et al. pointed out that the system is dealing with multivariate data that frequently lacks geographical and temporal relationships. There is a significant risk to the reliability of the conclusions formed from data due to transmission interruptions and failures of IoT sensors. An equally important goal, which often goes hand-in-hand with predictive analytics in many fields, is to address these longstanding problems about missing data that decrease the effectiveness of the prediction models. This study presents the author's overall framework for an Internet of Things (IoT) air pollution monitoring and forecasting system, which includes generative and predictive models. The LSTM encoder-decoder architecture forms the basis of the suggested models for air pollution forecasting. All sorts of LSTM encoder and decoder variations were investigated and tested, including bidirectional, stacked, and attention-based ones. For example, in the author's case study on PM2.5 prediction for the city of Skopje, meteorological conditions, timeframe, and location information were included as auxiliary variables that are substantially connected with pollution. The author's observations align with the relevant research that highlights the importance of training data sizes as a critical component influencing the prediction capabilities of deep learning systems. When it comes to problems with the amount and quality of training datasets, data augmentation approaches are thought of as a generic solution. The author suggests four methods for data augmentation of time series data using state-of-the-art adversarial networks to overcome the issues of missing sensor data.

Environmental monitoring, climate forecast, and control methods are among the many uses of deep learning models. More accurate and versatile drought, cyclone, typhoon, rainfall, and air quality index predictions have been made using deep learning models. In order to assess and manage the most catastrophic climate crisis and save life on Earth, this is a crucial component. When predicting the level of the air pollution index, most deep learning models—including ANN, LSTM, RNN, and many more—are used as a single model network. When it comes to air pollution index prediction, a single neural network is clearly insufficient, according to the data. Predicting the air quality index, thus, is an essential task for hybrid models. In order to forecast the air quality index, this study primarily makes use of weather data in conjunction with air pollution data. The air quality indices can be impacted by the meteorological parameters that the artificial neural network (ANN) model produces from weather data. The Air Quality Index (AQI) is effectively predicted using a Long Short-Term Memory (LSTM) Neural Network that is trained on data from these weather parameters.

The second section of this paper analyses the studies that were taken into consideration in the past. Section 3 lays forth the plan of action by narrating the steps in detail. Section 4 covers the experimental evaluation, and Section 5 discusses possible modifications to end the paper by concluding the current proposed idea.

II LITERATURE SURVEY

In their study, Jingyang Wang et al. [4] suggest a novel CNN-AGU-based AQI prediction model. The AGU's input data in this model are the data features extracted by the CNN. By incorporating the attention mechanism within the gated unit, the AGU improves the gated unit's memory capability. When the AGU incorporates the DAM, the gated unit becomes even more responsive to learning from past data. When compared to the other nine models, CNN-AGU performs better in the comprehensive evaluation, according to the experiments. The following two areas will constitute the bulk of future research on AQI prediction: (1) The experiment primarily processes time series data via the temporal attention mechanism. Time series data, such pollution levels, weather conditions, and traffic flow, and spatial data, like Points of Interest (POI) and road networks, are both real-world aspects that impact AQI.

A hybrid model is suggested by Yuxuan cao et al. [5] to forecast air quality indicators from several monitoring sites in the coming hours. If you want more accurate predictions of air quality, try the suggested model that combines the extended ARIMA model with the EMD technique and purged SVD. To be more specific, EMD is employed to extract smooth subseries from the initial non-stationary air quality indicator series. Also, all air quality indicators from different monitoring stations are predicted at the same time using an extension of the classic ARIMA model. The experimental findings show that the author's model achieves better accuracy and lower time cost than the state-of-the-art models for predicting air quality. But this study isn't without its flaws. The author's reliance on past air quality measures for forecasting is the study's primary shortcoming. The author's analysis does not take into account a number of elements that impact air quality, including transportation and weather. To improve accuracy in the future, the author will fully utilize that information. In addition to studying additional models (such as deep models) to enhance the author's proposed model, the author will also use a correlation graph to examine the causal relationships between air quality and different factors influencing it, and anomaly detection to further correct the predicted values. When writing about real-world applications with dispersed monitoring stations, the author can think about how to use technologies like the internet of things (IoT) and edge computing to cut down on data transmission delays so that diverse stations can provide real-time predictions.

Economic development brought forth by fast urbanization is causing serious environmental pollution and endangering people's lives, according to Chunhao Liu et al. [6]. Consequently, attaining the goal of sustainable development is contingent upon precise analyses and predictions of air quality. The optimization of the model's parameters is the subject of this paper's investigation into the prediction model design challenge. It is conceived, implemented, and tested to be a GA-KELM model. It can successfully investigate and understand the interplay of multivariate air quality correlation time series such temperature, humidity, wind speed, SO2, and PM10, and it has been demonstrated experimentally to be more efficient than the traditional shallow learning. Consequently, the study's GA-KELM model can be utilized to alert vulnerable populations to impending air quality disasters and offer helpful assistance to those groups. But there's always room for improvement and additional research.

In order to make better predictions about air quality, the authors of the study by Hongqian Chen et al. [7] suggested a model that uses an integrated dual LSTM approach. Here is how the integrated approach came to fruition and what it accomplished. The first step in making predictions from air quality data is to set up a single-factor prediction model. Next, a multi-factor prediction model is set up to forecast the present station's data using a combination of the present station's and nearby stations' previous data, together with weather information. The next step is to use XGBoost regression to construct the ideal boost tree, which combines the single-factor and multi-factor models to produce the most accurate predictions. To get the most accurate predictions, the approach presented in this research integrated two models that had been set up in two spatial and temporal dimensions. The time dimension was initially used to construct the one-factor models for all factors. A one-factor model takes a single variable, such as PM2.5, as input. Using the properties in the temporal dimension, author can get the forecast outcomes. After that, the spatial dimension is used to establish the multi-factor model. Inputs to the multi-factor model are chosen from a variety of sources, including data from the present station and nearby stations, meteorological data, and more.

As stated by Tongyno Yang et al. in [8]. This study presents and validates, in various simulation settings, a method for predicting the flight path of air targets using a divide-and-conquer strategy using two variables' weights. The findings of the simulation demonstrate that, under the assumption of reliable path prediction, adaptive decomposition of airspace based on obstacle variables can decrease the number of sub-regions and enhance the search efficiency of those sub-regions. Furthermore, the course convergence speed is enhanced by an average of 34.78% when using the upgraded ant colony algorithm to forecast the flight paths of air targets, in comparison to ACO. Important to the deployment choice, the simulation results demonstrate that the suggested approach can account for the precision and real-time of course prediction. The worldwide course prediction approach that takes macro modal limitations into account is examined in this research. Using the air targets' motion performance restrictions and the global course sequence, future research will focus on studying the local 3D trajectory of air targets.

The work of Elham Ghaffari et al. is cited as [9]. Finding the best shortest path routing for smart cities is the focus of this research. Using image-processing techniques, this method transforms a Google Map into a weighted graph. When it comes to traffic, we rely on GPS devices, and when it comes to air pollution, we rely on monitoring stations. Factors such as air pollution and traffic congestion are used to determine the graph's weights. locations with poor air quality and heavy traffic receive lower weights, whereas locations with high population density or contamination receive higher weights. Afterwards, a linear programming approach is employed to resolve the routing problem, which is recast as an optimization procedure. Unless there is no other option, people will not choose overcrowded and polluted paths. The efficiency of the suggested technique is tested by determining the shortest path using the Tehran city map.

Citation: [10] Al-Eidi Shorouq et al. In order to forecast smart city air quality, this study offers a thorough comparison of various regression models. Among the several regression models tested, the Decision Tree model stood out for its exceptional performance. Improving model accuracy through optimizing feature selection and correcting data imbalances was greatly aided by including Exploratory Data Analysis and the SMOTER approach. Utilizing cloud computing for regression modeling offers many benefits, as highlighted in the study. The model execution time was lowered by using cloud resources, which improved efficiency and made it more scalable. This sped up the process of testing, training, and deploying the models, which improved their usefulness in the actual world.

[11] Yuting Yang et al. The author of this research uses Shapley Additive explanations, an explainable deep learning method, to show how weather affects air quality forecast. The main objective is to examine the impact of weather circumstances on air quality forecast by interpreting the well-established LSTM and GRU models using the SHAP interpretation approach.

[12] Ning Jin et al. An new multi-variate MTMC learning framework for predicting concentrations of air pollutants is the main contribution of this paper. With the help of federated learning, the multivariate AQI data is trained and forecasted simultaneously. With MAE, RMSE, and MAPE values all below 3, the suggested technique can closely track the actual AQI, as shown in the testing findings. By giving the government up-to-date environmental information, an accurate forecast of air pollution is critical for human health and supports decisions on environmental management. The experimental phase makes use of a real-world dataset gathered from weather stations in the Beijing, China area.

Referenced in [13] Wu Zhiyuan et al. The authors of this work suggest a meta-adversarial learning strategy for adaptive and probabilistic air pollution prediction problems. In order to train an implicit conditional generator that may deliver useful task-specific predictive distribution, the author's model engages in adversarial three-player game training with a give backbone predictor. In addition, a theoretical model is presented by the author, which views the suggested model as an approximated minimizer for the Wasserstein distance, which measures the separation between a latent generative model and the actual distribution of data. Experimental results on both simulated and real-world datasets demonstrate the enhancement in providing more meaningful adaptive and probabilistic predictions while maintaining satisfactory metrics for point prediction errors. It may be concluded that the author's model is more capable of handling the complicated uncertainty estimation and adaptation abilities needed for air pollution prediction in the actual world.

[14] Huynh A. D., Nguyen et al. This research introduces a novel deep learning model called LSTM-BNN, which aims to enhance the accuracy and reliability of air pollution forecasts in the Australian state of novel South Wales. Specifically, it targets the two main air pollutants, PM2.5 and ozone. By combining observations with data from the existing CCAM-CTM, the suggested network makes use of both one-step recursive forecast and multistep forward direct forecast methods. Instead of providing point-wise estimates like deterministic models do, the resultant model gives the prediction distributions as posteriors at each time step. To measure uncertainty in both the actual data and the model that was created, the Monte-Carlo dropouts are used to approximate Bayesian judgments. In order to reduce computational latency and

achieve improved forecast accuracy compared to Gaussian-based inference, the author developed KDEST, an algorithm that uses subdivision tuning to estimate kernel density with a significantly smaller number of distribution samples.

[15] Mokhtari Ichrak et al. The authors of this work have taken up the challenging task of air pollution forecasting in this article. To start, the author evaluated and discussed numerous high-quality studies that covered the current state of the art in air pollution prediction systems. Secondly, in order to forecast highly dynamic air pollution, such as in the event of unexpected pollution episodes, the author introduced a ConvLSTM-based spatio-temporal deep learning model. The model that was constructed can predict the output of several nodes simultaneously with just one framework. Without knowing the connections between nodes beforehand, it learns which ones are more significant for predicting a single node. Third, the author benchmarked seven state-of-the-art methods, including both traditional deep learning and more recent approaches to machine learning, to see how well author's model performed. Two real-world, highly dynamic data sets were used to conduct the testing. Regardless of the training amount taken into account, the experimental findings showed that the author's forecasting model outperformed the baseline methodologies and was useful for short-term air pollution forecasting. Finally, the author's forecasting model used two methods—MC dropout and quantile regression—to calculate uncertainty estimates.

III PROPOSED MODEL METHODOLOGY

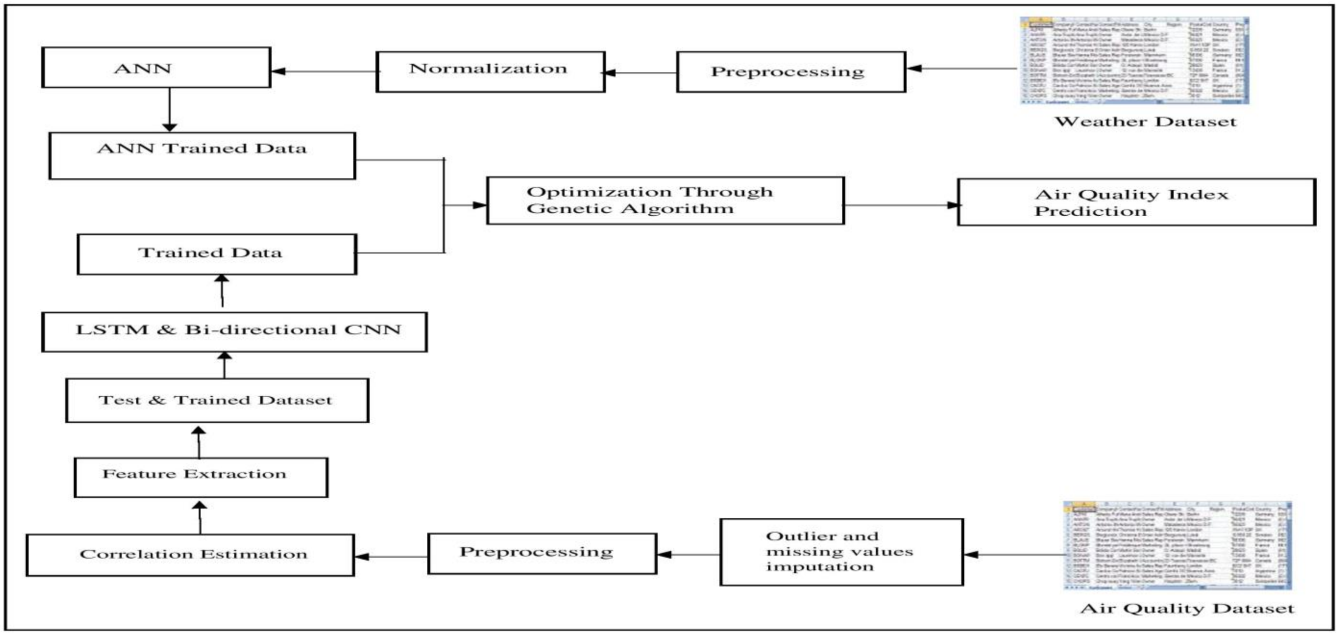


Figure 1: Proposed model for Air Quality Index prediction through hybrid neuron network

Figure 1 above illustrates the proposed model for predicting the air quality index using a hybrid neural net. The below-mentioned steps describe each phase in detail.

Step 1: Dataset collection and Preprocessing LSTM and Bi Directional CNN – This is the first step in the proposed model, which begins with the dataset collection from the URL [16]. The dataset, which originates from India, includes the following attributes: city, date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3,O3,Benzene, Toluene, Xylene, AQI, and AQI_Bucket. The system downloads and stores this dataset attributes in a CSV file. The Pandas library of the Python programming language reads this dataset once the system receives them. The system describes the object of the dataset and displays a histogram for parameters such as mean, standard deviation, 25%, 50%, and 75% of the maximum count, as well as the maximum value of the attribute.

Following this step, the entropy of the dataset's data types—such as strings and floats—is asserted by obtaining information about the dataset's attributes. The dataset labels are analyzed and collected in a list. Once the dataset has been labeled, the frequency of the labeled classes is approximated to ensure that the dataset is balanced. To ensure that these identified classes are distributed evenly, they are oversampled. In the end, this procedure improves the preprocessing step, which leads to accurate predictions of the Air Quality Index.

Step 2: Data Imputation - With the oversampled data, we estimate a heat map for each attribute. This is accomplished by calculating the total amount of missing data and the percentage of that data that is missing by comparing the transition data throughout the sorting process. The fillna() function is used to imputation these missing data points for essential attributes.

Imputation makes use of a label encoder, which is formalized as an argument to the fit transform function, for each object in the attributes. The goal of mouse imputation is produced by iteratively applying multiple imputation by chained equations after the characteristics have been transformed using the fit transformer. To fill in missing values in a dataset, the multiple imputation method called MICE is employed, under certain data missingness mechanism assumptions. If the model is given a dataset with missing values in at least one variable, it can generate multiple copies. After mice imputation is used, the specific combinations' missing values are found, and those values are filled in using known values from other characteristics in the data. After that, we use the acquired imputated values to estimate the interquartile range (IQR), which is between 0.75 and 0.25. Using this method, we can create a dataset that includes missing value imputation.

Step 3: Pearson Correlation - Mice imputation is followed by the collection of a complete dataset list, which is then subjected to Pearson Correlation analysis. We make use of the values of the Pearson relationship in order to locate the characteristics that have the least amount of relationship. We can use Pearson's correlation to discover whether or not two features are related to one another. As a result, a correlation matrix has been produced, which has the potential to be helpful in selecting the appropriate combination of characteristics. Following the revelation of this new knowledge, we are now able to compute correlation values and disregard the characteristics that have a lower correlation. It is possible to determine a Pearson correlation by utilizing Equation 1 as shown below.

$$= \frac{\sum(x_i-\bar{y})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sum(y_i-\bar{y})^2} \text{-------} (1)$$

Where,

x_i=values of x (independent) variable

y_i= values of y (Dependent) variable

$\bar{x}$= mean of x variable values

$\bar{y}$= mean of y variable values

Following this process, unnecessary attributes are eliminated from the obtained dataset list to extract the feature list in the next phase of the proposed system.

Step 4: Feature extraction selection and Data Segregation - For the imputated and preprocessed data, the minMaxscaler function is applied in this feature selection procedure. The minmaxscaler Transform the features by scaling them to a range we specify. This estimator applies separate scaling and translation to each feature in order to get it inside the specified range on the training set, for instance, between 0 and 1. The effect of outliers is not mitigated by MinMaxScaler, even if it linearly scales them into a predetermined range, with the biggest data point representing the highest value and the smallest the minimum. This process yields two lists, as one contains all other feature as X and the other one contains the labels as Y. The minmaxscaler transformation is shown in the below equation 2 and 3.

$$\mathrm{x_{std}} = \frac{(\mathrm{x-x.min(axis=0)})}{(\mathrm{x.max(axis=0)-x.min(axis=0)})} \text{-----}(2)$$

$x_{scaled} = x_{std} * (\max - \min) + \min$—(3)

where min, max = feature_range.

The Sklearn library offers a function named train_test_split(), which divides the parameters, X and Y feature list data, into four lists: y_train, y_test, X_train, and X_test. We use the Y_train and X-train data to train the model, while we autocorrect the model by testing it during the training process with the Y_test and x_test data. In order to construct an accurate neural network model, we utilize 80% of the data for training and the remaining 20% for testing. Next, we use the Sklearn library's minmaxscaler () function to scale the data within the range of 0 to 1. We use bi-directional CNN -LSTM model to train the lists X_train and Y_train efficiently.

Step 5: LSTM and Bi-directional CNN- We build a neural network model from the obtained split dataset lists for training and testing. We implement a convolution layer with 32 kernels of size Uno in the first and second layers of the model, followed by a Max pooling layer of one dimension with a dropout rate of 30%. In the third layer, there is a convolution layer with 64 kernels of size 1, followed by a max pooling layer of one dimension with a dropout ratio of 35%. Equation 4 mentions that the Relu activation function powers all the first, second, and third layers. The tanh activation function, which can be seen in equation 5, powers the fourth convolution layer. It has a 128-kernel size 1 and comes after a one-dimensional max pooling layer with a 40% dropout rate.

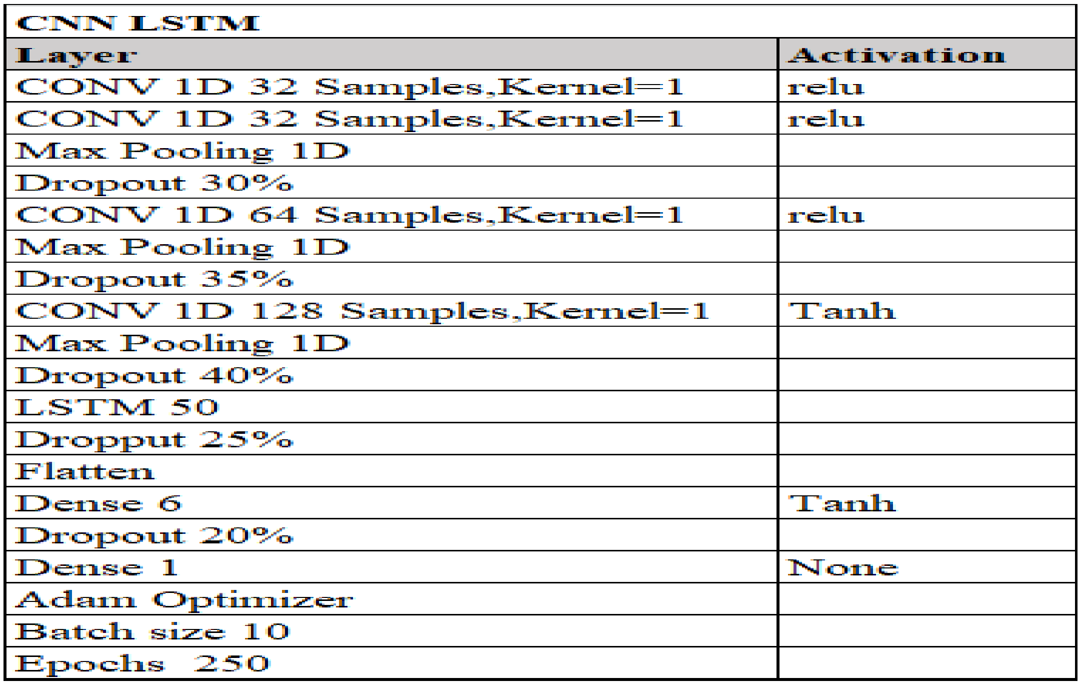
We deploy an LSTM model with 50 layers and a dropout rate of 25%, followed by a dense layer of 6, indicating the classification labels powered by the tanh activation function. Finally, we add this LSTM model with a single dense layer, a dropout ratio of 20%, and an Adam optimizer with 250 iterations and a batch size of 10. The trained data is then produce the required prediction data that will be further used by the next process.

The used tanh and Relu activation function are depicted in equation 1 and 2.

$tanh = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$ ______(4)

Relu=max(0,x) _____(5)

The architecture of Bi-directional CNN and LSTM can be shown in the figure 2.



CNN LSTM	
Layer	**Activation**
CONV 1D 32 Samples,Kernel=1	relu
CONV 1D 32 Samples,Kernel=1	relu
Max Pooling 1D	
Dropout 30%	
CONV 1D 64 Samples,Kernel=1	relu
Max Pooling 1D	
Dropout 35%	
CONV 1D 128 Samples,Kernel=1	Tanh
Max Pooling 1D	
Dropout 40%	
LSTM 50	
Dropput 25%	
Flatten	
Dense 6	Tanh
Dropout 20%	
Dense 1	None
Adam Optimizer	
Batch size 10	
Epochs 250	

Figure 2: Architecture for CNN LSTM

Step 6: Training with ANN- A weather dataset is downloaded from the URL as mentioned in [17]. This dataset contains some attributes like Precip Type,Temperature (C),Apparent Temperature (C),Humidity, Wind Speed (km/h),Wind Bearing (degrees),Visibility (km),Cloud Cover, Pressure (millibars),Daily Summary and label .

After acquiring the dataset, it typically undergoes the pre-processing stage. During the pre-processing step, the attributes are gathered and stored in a two-dimensional list after reading the dataset from the designated path. The dataset has been imported successfully from the specified path using the Pandas library in Python. Using a two-dimensional list, the initial parameters of the attributes are estimated to describe the characteristics of the dataset, such as the mean and standard deviation. With this procedure, the system is currently analyzing the dataset to gather attribute information for various data types, such as strings and floats. Our objective is to determine the entropy of the dataset's data types in order to generate a well-prepared list.

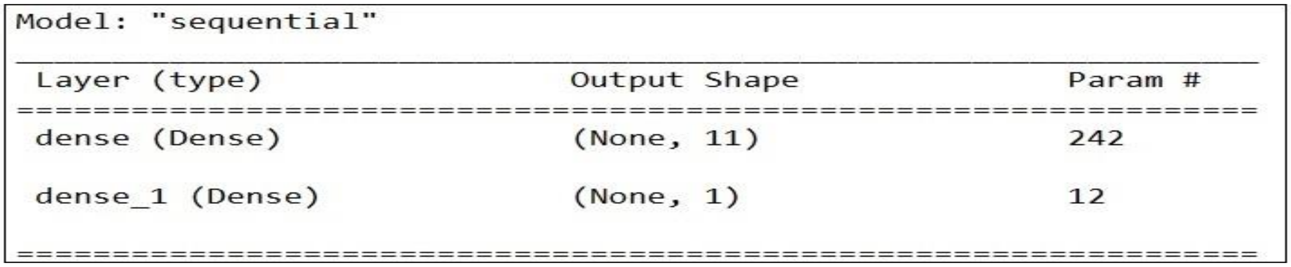
To modify the attributes and scale them to a specified range, the minMaxscaler function is used. To ensure that all features fall inside the specified range on the training set—say, from 0 to 1—this estimator adjusts and tweaks them one by one. Gradually, MinMaxScaler brings outliers within a specified range by assigning the highest value to the biggest data point and the lowest value to the smallest. The minmaxscaler transformation can be expressed using equations 2 and 3.

The dataset list is then separated into features(X) and labels(Y), this X and Y list is used to form 4 lists like train_X, test_X, train_Y, and test_Y through train_test_split(). A sequential layer is formed for the ANN neural network model which is followed by a Dense layer with outshape for mentioned parameter. The next step is to create an input layer dense layer with 11 kernel units by selecting the 'uniform' option. The initializers are responsible for setting the Keras layers' initial random weights. A "relu" activation function with 20 input dimensions is also included in this layer. After that, the "uniform" parameter is used to initialize a dense layer with one unit of kernel, and the output layer is enlarged to include it. To make the output layer stronger, the sigmoid activation function is utilized. The deployed model is using adam optimizer for 200 epochs to get the desired predicted values. The sigmoid equation can be seen in the (6)

$$S(X) = \frac{1}{(1+e^{-x})} \_\_\_\_\_\_(6)$$

Where,
X is the input to a neuron
$S(x) = \text{Sigmoid Activation Function}$
e= Euler's Number

The summary of the constructed ANN model can be seen in the below figure 3.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 11)                242

 dense_1 (Dense)             (None, 1)                 12

=================================================================
```

Figure 3: ANN Model Summary

Step 7: AQI index prediction through Genetic algorithm - Once we obtain the prediction data for both air pollution and weather datasets, we consider these prediction values as the initial population of the genetic algorithm. Next, the selection process uses if-then rules to select a fitness value that meets the accuracy requirement. This genetic algorithm process produces the best air quality index prediction for the given dataset.

IV RESULTS AND DISCUSSIONS

We developed the proposed model for air quality index prediction through a hybrid neural network in the Python programming language using the Jupyter IDE. The model uses a Core i5-equipped machine with 8 GB of primary memory. We extensively use two parameters, such as root mean square error (RMSE) and accuracy, to assess the error rate and accuracy of the proposed model.

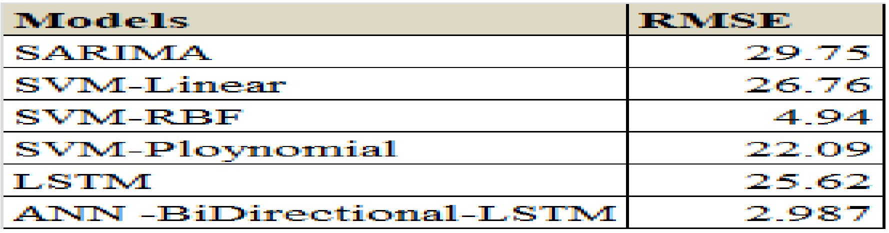
A popular metric for assessing the error rates of prediction is the root-mean-squared error or root-mean-square deviation. With the use of the distance measure, it reveals the extent to which predicted values deviate from the actual values. The root-mean-squared error (RMSE) can be calculated by first calculating the residual (the difference between the prediction and the truth) for each data point, then calculating the norm of the residual for each data point, and finally taking the square root of the mean of the residuals. Since RMSE relies on and requires actual measurements at every anticipated data point, it finds widespread usage in supervised learning contexts. In this research, the root mean square error (RMSE) is used to calculate the margin of error between the expected and actual AQI detection made by the ANN and Bidirectional CNN -LSTM models. To see the RMSE method in action, investigate equation 1 that is given below. AQI estimation identification degrees of correctness and inaccuracy are examined by the models. After these data have been analyzed, the disparity is calculated using Equation 7.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^{2}}{n}} \qquad \_\_\_\_\_\_(7)$$

Where,
∑ - Summary
(x1 – x2)2 – Disparities Squared for the total of the differences

N - The Count of Trails

To calculate the approach's error rate using root-mean-squared (RMSE), we must first find the mean squared error (MSE). The mean square error (MSE) is the discrepancy between the predicted and observed AQI values. Table 1 below shows the results for the ANN and Bidirectional CNN-LSTM models for the respective dataset compared to other models presented in [18], and figure 3 shows the corresponding graph.



Models	RMSE
SARIMA	29.75
SVM-Linear	26.76
SVM-RBF	4.94
SVM-Ploynomial	22.09
LSTM	25.62
ANN -BiDirectional-LSTM	2.987

Table 1: RMSE Comparison

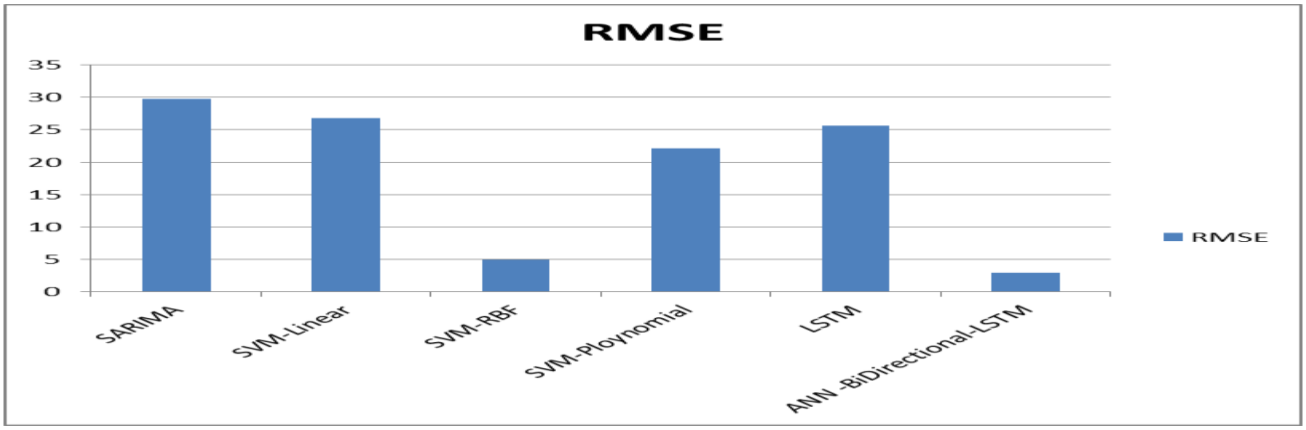


Figure 3: Comparative Analysis of RMSE score of Different models

The main objective of [18] is to forecast the air quality index for the city of Ahmedabad in the Indian state of Gujarat, this study aims to analyze different machine learning approaches like SARIMA, SVM, and LSTM. In order to prepare the data for the machine learning models, this study employs a variety of preprocessing techniques. This research uses the RBF kernel model-based support vector machine technique using data supplied by the Central Pollution Control Board of India. The SVM-RBF of [18] yields the best RMSE of 4.94, On the other hand our model ANN-Bidirectional-LSTM yields the RMSE of 2.987 which is far better than the other models due to hybridization of the neural networks.

To test the deployed system, we utilize the following equations to express the confusion matrix's accuracy score parameter. The four variables Accuracy, Precision, Recall and Macro F1 are represented with the equations 8, 9, 10, and 11, respectively.

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad -(8)$$

$$\mathrm{Precision(P)} = \frac{TP}{TP+FN} \quad -(9)$$

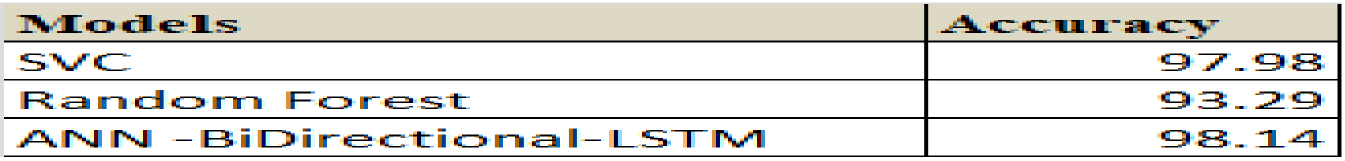
$$\mathrm{Recall(R)} = \frac{TP}{TP+FP} \quad -(10)$$

$$\mathrm{Macro-F1} = \frac{2*P*R}{P+R} \quad -(11)$$

Here, TP is True positive cases, TN is True Negative cases, FP is False positive cases and FN is False Negative cases.

The obtained accuracy scores are comapred with that of [19], authors have conducted an extensive investigation of the air pollution levels in two areas of Kolkata, specifically Rabindra and Victoria, which encompass six main air pollutants, such as PM2.5, NO2, and PM10. The analysis and research of the data included standardizing the data, filling in missing numbers, deleting duplicates, and removing outliers. After running five primary classification algorithms with the right hyperparameter tuning, we found that the SVC model achieved the best accuracy on the Rabindra dataset at 97.98%, while the random forest model achieved the best accuracy on the Victoria dataset at 93.29%.

Table 2 tabulates the accuracy of the SVC model [19] and the accuracy of the bi-directional CNN-LSTM, and Figure 4 below shows the corresponding graph.



Models	Accuracy
SVC	97.98
Random Forest	93.29
ANN -BiDirectional-LSTM	98.14

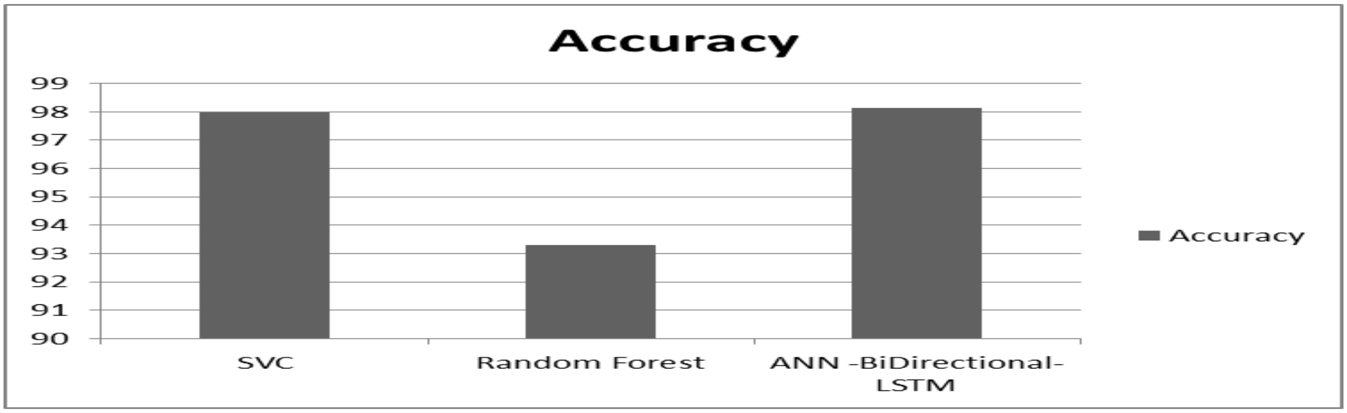


Figure 3: Comparative Analysis of Accuracy score of Different models

Evidently, the results show that the CNN-LSTM model outperforms the SVC model from [19] in the long run. This is because, compared to the Solo model, neural network hybridization on a dispersed dataset yields superior results.

V CONCLUSION AND FUTURE SCOPE

Air quality index prediction is a much-needed solution to curb air pollution so that we can make the earth liveable. Artificial intelligence is crucial in this process. Without the deployment of hybrid neural networks, single-model neural networks would not be as effective in achieving greater accuracy. Hence, this research work incorporated a hybrid neural network with the combination of an ANN for the weather dataset and a bidirectional CNN-LSTM for the air quality index (AQI) dataset. The bidirectional CNN-LSTM model initially computes the dataset for the missing values to select the required attributes. The Pearson correlation model checks the attributes for correlation, leading to the feature selection process. We then split the selected features for the bidirectional CNN-LSTM model's training and testing process. We preprocess and divide the weather dataset into separate training and testing lists to implement the ANN model. The obtained predicted results from Bidirectional CNN-LSTM and ANN model are fed to the Genetic algorithm step to catalyze the prediction accuracy. The obtained results are tested for RMSE and accuracy. Our model, ANN-Bidirectional-LSTM, yields an RMSE of 2.987, which is far better than the other models due to the hybridization of the neural networks. ANN-Bidirectional-LSTM yields an accuracy of 98.14, which is also better than that of the SVC model and other models. Future work the proposed model can deploy the transformers to enhance the training process by adding to the earlier epochs. This process can be massive and can predict the global AQI based on huge parameter across the globe.

REFERENCES

[1] Y. Liu, J. Nie, X. Li, S. H. Ahmed, W. Y. B. Lim and C. Miao, "Federated Learning in the Sky: Aerial-Ground Air Quality Sensing Framework With UAV Swarms," in *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9827-9837, 15 June15, 2021.

[2] Y. Han, J. C. K. Lam, V. O. K. Li and Q. Zhang, "A Domain-Specific Bayesian Deep-Learning Approach for Air Pollution Forecast," in *IEEE Transactions on Big Data*, vol. 8, no. 4, pp. 1034-1046, 1 Aug. 2022, doi: 10.1109/TBDATA.2020.3005368.

[3] J. Kalajdjieski, K. Trivodaliev, G. Mirceva, S. Kalajdziski and S. Gievska, "A Complete Air Pollution Monitoring and Prediction Framework," in *IEEE Access*, vol. 11, pp. 88730-88744, 2023, doi: 10.1109/ACCESS.2023.3251346.

[4] J. Wang, L. Jin, X. Li, S. He, M. Huang and H. Wang, "A Hybrid Air Quality Index Prediction Model Based on CNN and Attention Gate Unit," in *IEEE Access*, vol. 10, pp. 113343-113354, 2022, doi: 10.1109/ACCESS.2022.3217242.

[5] Y. Cao, D. Zhang, S. Ding, W. Zhong and C. Yan, "A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition," in *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 99-111, February 2024, doi: 10.26599/TST.2022.9010060.

[6] C. Liu, G. Pan, D. Song and H. Wei, "Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine," in *IEEE Access*, vol. 11, pp. 67086-67097, 2023, doi: 10.1109/ACCESS.2023.3291146.

[7] H. Chen, M. Guan and H. Li, "Air Quality Prediction Based on Integrated Dual LSTM Model," in *IEEE Access*, vol. 9, pp. 93285-93297, 2021, doi: 10.1109/ACCESS.2021.3093430.

[8] T. Yang, F. Yang and D. Li, "An Air Target Course Prediction Method Based on Sub-Regions Divide and Conquer With Double Variable Weight," in *IEEE Access*, vol. 10, pp. 117871-117885, 2022, doi: 10.1109/ACCESS.2022.3220676.

[9] E. Ghaffari, A. M. Rahmani, M. Saberikamarposhti and A. Sahafi, "An Optimal Path-Finding Algorithm in Smart Cities by Considering Traffic Congestion and Air Pollution," in *IEEE Access*, vol. 10, pp. 55126-55135, 2022, doi: 10.1109/ACCESS.2022.3174598.

[10] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," in *IEEE Access*, vol. 11, pp. 115140-115149, 2023, doi: 10.1109/ACCESS.2023.3323447.

[11] Y. Yang, G. Mei and S. Izzo, "Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning," in *IEEE Access*, vol. 10, pp. 50755-50773, 2022, doi: 10.1109/ACCESS.2022.3173734.

[12] N. Jin, Y. Zeng, K. Yan and Z. Ji, "Multivariate Air Quality Forecasting With Nested Long Short Term Memory Neural Network," in *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8514-8522, Dec. 2021, doi: 10.1109/TII.2021.3065425.

[13] Z. Wu, N. Liu, G. Li, X. Liu, Y. Wang and L. Zhang, "Learning Adaptive Probabilistic Models for Uncertainty-Aware Air Pollution Prediction," in *IEEE Access*, vol. 11, pp. 24971-24985, 2023, doi: 10.1109/ACCESS.2023.3247956.

[14] H. A. D. Nguyen *et al.*, "Long Short-Term Memory Bayesian Neural Network for Air Pollution Forecast," in *IEEE Access*, vol. 11, pp. 35710-35725, 2023, doi: 10.1109/ACCESS.2023.3265725.

[15] I. Mokhtari, W. Bechkit, H. Rivano and M. R. Yaici, "Uncertainty-Aware Deep Learning Architectures for Highly Dynamic Air Quality Prediction," in *IEEE Access*, vol. 9, pp. 14765-14778, 2021, doi: 10.1109/ACCESS.2021.3052429.

[16] https://www.kaggle/input/air-quality-data-in-india/city_day.csv

[17] https://www.kaggle.com/datasets/muthuj7/weather-dataset?resource=download

[18] Nilesh N. Maltare, Safvan Vahora ,"Air Quality Index prediction using machine learning for Ahmedabad city", Elsevier,2023.

[19] Mohsin Imam , Sufiyan Adam , Soumyabrata Dev , Nashreen Nesa ,"Air quality monitoring using statistical learning models for sustainable environment ",Elsevier,2023.