http://ijea.jctjournals.com/ ISSN(ONLINE):2320 – 0804

Volume-11, Issue-7, July, 2022

International Journal ofEngineering Associates (IJEA)

Page Number: 01-06

Prediction of Air Quality Efficient Model By Deep Learning-Based LSTM and GRU

Manish Kumar[1], Prof. Amit Namdev[2]

[1]M.Tech Scholar, [2]Assistant Professor

[1,2]Mittal Institute of Technology, Bhopal, M.P., INDIA

[1] manishsingh12101991@gmail.com, [2]amit.namdev1811@gmail.com

***Abstract*— In today's world, air pollution is one of the most pressing environmental problems. As a result, both the environment and human health are at risk. Air quality in urban areas is deteriorating at an alarming rate. Air pollutants affect both land and water, and the latter is particularly vulnerable. In recent years, air pollution has become a major environmental protection problem due to the growth of urbanization and industrialization. Air quality forecasting has become an urgent and critical issue because it has a direct impact on people's daily lives. Predicting air quality is a difficult task due to the interdependence of many complex factors. Air quality, a vital natural resource, has been negatively affected by economic activity. The lack of long-term data makes it impossible to account for seasonal and other variables in most forecasts of poor air quality, although much research has been done on the topic. The State Pollution Control Board's AQI data was used to develop a number of different prediction models (CPCBs). It is necessary to pre-process the data obtained from various ground stations. Consequently, the raw data had to be filtered before using deep learning (DL) methods for prediction. For forecasting Air Quality Index (AQI) levels, deep learning approaches such as LSTM and GRU show promising results. In order to obtain the best parameter prediction results in terms of R2, RMSE and MAE, this study uses the Delhi dataset. The RMSE and MAE of the proposed model are 1.41 and 1.84 for the LSTM model with GRU. Using these algorithms to predict the air quality index can be quite accurate, according to the findings.**

Keywords - deep learning (DL), R2, Air Quality Index (AQI), RMSE and MAE

1. INTRODUCTION

In today's world, air pollution is one of the most pressing environmental problems. As a result, both the environment and human health are at risk. Air quality in urban areas is deteriorating at an alarming rate. Air pollutants affect both land and water, and the latter is particularly vulnerable. In recent years, air pollution has become a major environmental protection problem due to the growth of urbanization and industrialization. Air quality forecasting has become an urgent and critical issue because it has a direct impact on people's daily lives. Predicting air quality is a difficult task due to the interdependence of many complex factors. Air quality, a vital natural resource, has been negatively affected by economic activity. The lack of long-term data makes it impossible to account for seasonal and other variables in most forecasts of poor air quality, although much research has been done on the topic. The State Pollution Control Board's AQI data was used to develop a number of different prediction models (CPCBs). It is necessary to pre-process the data obtained from various ground stations. Consequently, the raw data had to be filtered before using deep learning (DL) methods for prediction. For forecasting Air Quality Index (AQI) levels, deep learning approaches such as LSTM and GRU show promising results. In order to obtain the best parameter prediction results in terms of R2, RMSE and MAE, this study uses the Delhi dataset. The RMSE and MAE of the proposed model are 1.41 and 1.84 for the LSTM model with GRU. Using these

algorithms to predict the air quality index can be quite accurate, according to the findings.

II. LITERATURE REVIEW

In the last several years[1], in metropolitan areas, the level of pollution in the air has been constantly rising. Some of the most polluted cities in the world, including Gurugram, Faisalabad, Delhi, and Beijing, have seen a worrying increase in the amount of air pollution in their environments. Forecasting is essential because to the human, ecological, and economic toll that pollution exacts, and it is an investment that is beneficial on both the individual and communal levels. If we have accurate forecasts, they will be able to prepare ahead, which will reduce the negative impacts on health and the expenditures connected with them. The levels of air pollution in a given area are highly influenced by the local meteorological conditions. In the field of environmental science, the generation of deterministic models to investigate the behaviour of air pollutants is often not particularly accurate since these models are complicated and need simulation at the level of the interactions between molecules.

In recent years[2], the prevention & control of environmental pollution attracted much attention, and the haze weather directly affects people's travel health. In order to effectively prevent and control air pollution, optimize the air quality evaluation system. In this study, PM2.5, PM10, SO2, NO2, CO and O3_8h are used as characteristic factors, and air quality indexis used as a decision factor. A variety of regression algorithms are selected to establish a prediction model, and the accuracy and generalization ability of various algorithms are compared. The results show that the Random Forest Regression algorithm (RFR) and the Gradient Boosting Regression algorithm (GBR) can effectively predict the AQI and the air quality level. This study provides a reference for the establishment of the air quality model.

In this study[3], using a variety of machine learning ensemble approaches, a job that involves the near future forecasting of fine-grained levels of air quality is investigated. Approaches such as majority voting, averaging, weighted averaging, and sixteen various stacking strategies are included in the ensemble methods. Comprehensive experimental comparisons are carried out in order to examine the performance of these different ensemble approaches. Models such as the traditional Autoregressive Integrated Moving Average (ARIMA), the widely used deep learning model Long Short-Term Memory (LSTM) neural network, and nine of the base- level models such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and several boosting models are included in the comparison models.

According to hedonic price theory[4], demonstrate the internal mapping relationship between feature variables such as housing prices as well as air quality by using classification algorithms from machine learning. These algorithms include SVM, NB, and KNN. Feature variables that are included in this relationship are including air quality as well as housing prices. Using these feature factors, one may forecast the air quality in an urban residential area. In conclusion, studies have been carried out using the dataset consisting of residential districts. It was shown that the maximum accuracy was achieved in Tianhe, which is located in Guangzhou city. When it comes to accurately predicting the quality of the air in urban residential neighbourhoods, this strategy is both practical and reliable. The new technique offers significant value to purchasers as a reference method in a practical sense.

This study [5], presents a model for predicting the grade of air quality relying on the K neighbour technique. Initially, the appropriate meteorological website's historic air quality monitoring info is crawled & stored to a local CSV file; the information is then read as well as the statistical method is being used to graphically present the six variables that impact the air quality level assessment; last, the K closest neighbour method is picked and the discrepancy is corrected. By parameters training and subsequent testing, a 95.10 percent accuracy rate was achieved. It is now possible to forecast the air quality level based on a random collection of data, which is in accordance with the predicted findings.

III. PROPOSED METHOD

The methods of data collection and preprocessing, along with model construction, constitute the methodology. Scripts written in Python will be used to create all of the deep learning models that will be tested in this investigation. There will be a detailed breakdown of each process in this section.

A. Data Collection

Data from Delhi, India, are used in this study to monitor and forecast air quality in the city. For nine years, data was collected from Delhi's 40 pollution monitoring stations using the official website of the Central Pollution Control Board (2012 to 2020). The AQI is the dependent variable in this study, with seven factors serving as independent variables. AQI calculator provided by the CPCB, which is a part of Ministry of Environment, Forests, and Climate Change, Government of India, was used to select seven pollutants (PM2.5, PM10, NOX, NH3, SO2, CO and Ozone) from the daily air quality data. PM2.5- g/m3, PM10- g/m3, NOX- g/m3, NH3- g/m3, SO2- g/m3, CO- mg/m3, O3- g/m3 are the 24-hourly averaged surface pollutant concentrations and units for the relevant metrics.

B. Data Pre-processing

A forecasting model's accuracy and generalizability are both reliant on the quality of the data it receives during collection. The most common methods for analysing data include the following: (a) Filling in missing information as well as deleting or changing exceptional case data points; (b) Maintaining uniform data distribution by normalising or standardising data; (c) extracting features from a dataset to reduce and condense it.

Fig.1 : Correlation between input dataset

A correlation matrix is a table which displays the correlations for several variables. The correlation between the other number of values shown in the table is illustrated in the matrix. Figure 10 on the right shows how it can be used to summarise large datasets and identify patterns in data.

A. Proposed Models (LSTM and GRU)

Model strategies deep learning models are discussed in the following section of the article (LSTM and GRU).LSTM Model :-The term "RNN" refers to a type of recurrent neural network, which is distinct from ANNs and CNNs. There is no previous output taken into account when ANN and CNN make a prediction about the final product's output. Despite this, with RNNs, the previous iteration's results are also taken account. In situations in which there is a steady stream of data, an RNN can be a valuable tool. It is a subclass of the RNN algorithm, the LSTM model. While learning, RNNs take into account their previous output. In contrast, it does not have a mechanism for removing irrelevant characteristics from the previous iteration's output. With the help of the Forget Gate, however, LSTM is able to accomplish this task. LSTM is constructed from LSTM Cells. The requirements may dictate how many LSTM cells are used. Figure 2 depicts the LSTM Cell's structure.

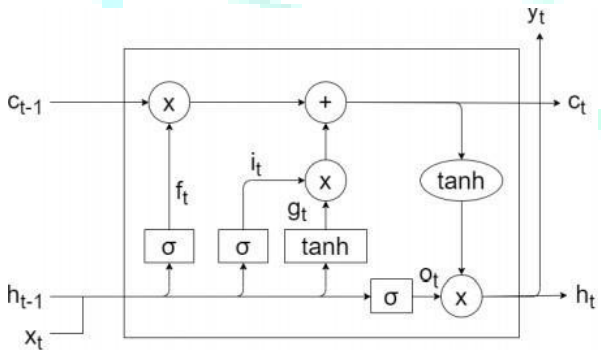


Fig. 2 : Long Short-Term Memory

In comparison to a standard ANN, the LSTM's architecture is more complex. In total, there are four entrances. Each gate utilises a neural network to carry out a specific function. An information gate can be used to learn with input and hidden layer weights. As a final step, the forget gate erases any information that is no longer relevant. The following LSTM layer makes use of the cell state as a cell state. Last but not least, the Output Gate serves as an input to the next layer.

This algorithm's first step is known as the forgetting gate. While simultaneously erasing unnecessary data, the Forget Gate keeps and erases useful information from previously hiddenstates and current input. A forget gate is constructed using the sigmoid function. For critical data, forget gate's value is 1 and 0 for non-critical data.

Using the input gate, the next cell's state can be updated. It is possible to manipulate the hidden state and the current input with the help of the tanh function. The previous concealed state and the new cell state are cleaned up using a sigmoid function. The tanh output is then multiplied by the sigmoid and tanh functions to keep the important data there.

The cell's status is used to transmit data. Input gate outputs are used to determine the current state of the cell. Cell state modifications are then sent as input to the next layer.

In order to generate a new hidden state, the output gate is used. The tanh function is applied first to the newly updated cell state. The current input and the previous hidden state are subjected to a sigmoid function, and the output of that function is then multiplied by the output of that function. Using a hidden state, input data and forecasts can be gathered.

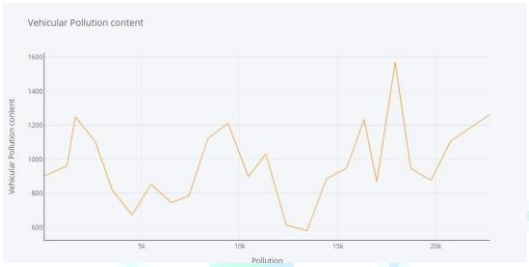
```
Model: "sequential_31"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_25 (LSTM)               (None, 50)                10400
_________________________________________________________________
dense_51 (Dense)             (None, 1)                 51
=================================================================
Total params: 10,451
Trainable params: 10,451
Non-trainable params: 0
```

Fig. 3: LSTM Model Summary

IV. RESULTS & DISCUSSION

Results and a description of the model can both be found in this section. Python is used throughout the experiment. Pandas, a Python library that supports multidimensional arrays, can be used for data analysis. For the Jupyter notebook, NumPy, Pandas, Matplotlib, Keras, TensorFlow, and seaborn were just a few of the Python libraries used. Many different performance matrices were employed (described below). Using the dataset that will be discussed in more detail below allowed us to reach the conclusions we did.

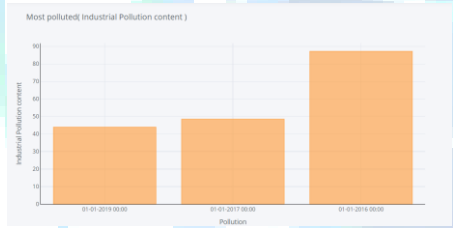
After parameterization and imputation, the section delves into the specifics of building AQI prediction models. Section four evaluates how well the AQI forecasting models performed.



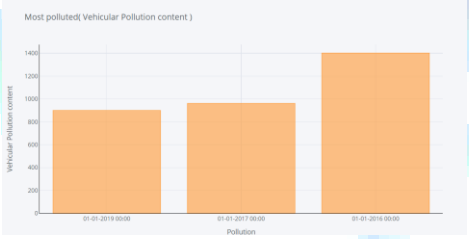
(a) Vehicular Pollution Contact

(b) Industrial Pollution Contact

Figure 4: Vehicular and Industrial Pollution Contact

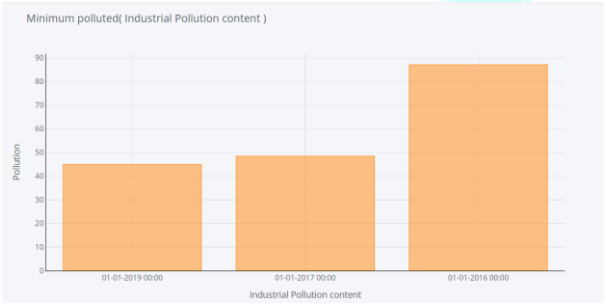


(a) Most polluted Industrial Pollution Contact

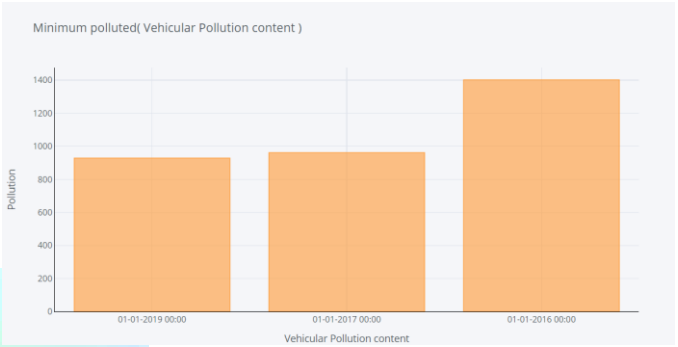


b) Most polluted Vehicular Pollution Contact

Figure 5 : Most polluted vehicular and industrial pollution content

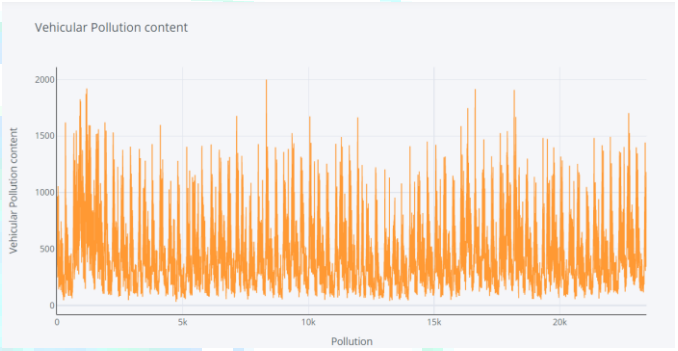


(a) Minimum polluted industrial Pollution Contact

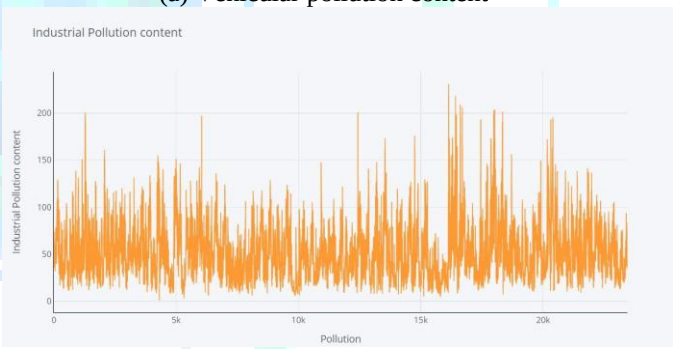


(b) Minimum polluted Vehicular Pollution Contact

Figure 6 : Minimum polluted industrial and Vehicular Pollution Contact

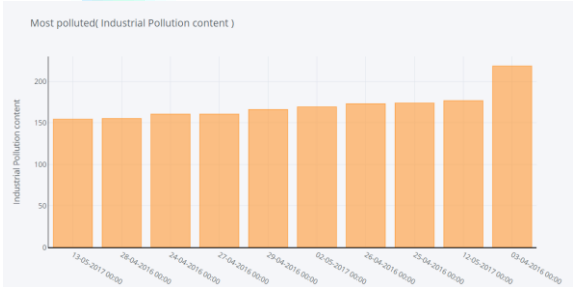


(a) Vehicular pollution content

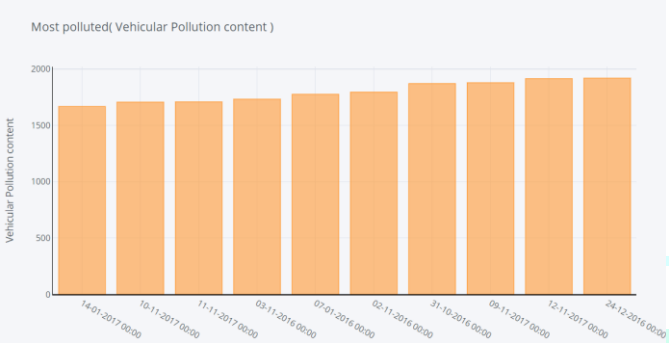


(b) industrial pollution content

Figure 7 : Vehicular & Industrial pollution



(a) Most polluted industrial pollution content



(b) Most polluted Vehicular pollution content

Figure 8 : Most polluted industrial and Vehicular content

Figures 4 and 5 show how air quality forecasting and modelling data is presented and displayed in a sample. It's important to keep in mind that, as the graph shows, the greater the time step of the prediction, the more tolerance is needed to ensure the estimate. Dataset performance is depicted in the following figures 6 and 7 . As depicted in these charts, industrial and vehicle traffic polluted content varies during the various phases. Images 6, 7, 8, and 9 each show a different type of vehicular pollution, as well as a different type of industrial pollution. Figure 9 shows the most polluted industrial and vehicular pollution content, while Figure 9 shows Vehicular & Industrial pollution.

A. Performance Evaluation

It is important to note that the most commonly used metrics are RMSE (root mean squared error), MAE (mean average error), and R 2 (R-squared), which are used to determine the difference between the prediction result and its true value. For each model and approach, these three indicators serve as a baseline to compare alternative parameter values. Performance validation introduces a bias when data is divided, trained, and tested only once. So the results obtained through testing dataset may no longer be valid if testing subset is altered. In order to deal with this problem, each model is reconstructed 20 times using random subsets of the train and test data. In other words, the 80:20 split ratio hasn't changed.

B. Experimented Results

GRU and Long Short-Term Memory are two well-known deep learning classification approaches that we used in this study (LSTM). Moreover, when compared to other methods currently employed. Graphs, metrics, and tables illustrate the experiment's findings. In the time since the experiment, we have conducted a thorough evaluation of the results. A deep learning model for classification and feature extraction was developed as part of this project. An improved predicate air quality can be achieved by adjusting certain model parameters, as shown in Table I.

Table I: Parameter Setting

Model	Sequential
RNN	GRU and LSTM
Neuron	50
Loss	MAE
Activation function	ADAM
Epoch	100
Metrics	MAE

Table II: Comparison between the base and proposed model using performance parameters

Model	RMSE	MAE	R^2
Base Stacking Ensemble	17.18	11.08	0.76
Propose LSTM	1.41	2.0	0.99
Propose GRU	1.84	3.38	0.99

A comparison of proposed and base models' performance is shown in Table II, using the three performance metrics described above. For example, the base model's RMSE is 17.18; its MAE is 11.8; and its R square is 0.76 percent. The GRU's Rsquare is 99; the LSTM model's RMSE is 1.41; and the second proposed GRU model's RMSE is 1.84; and its MAE is 3.38.

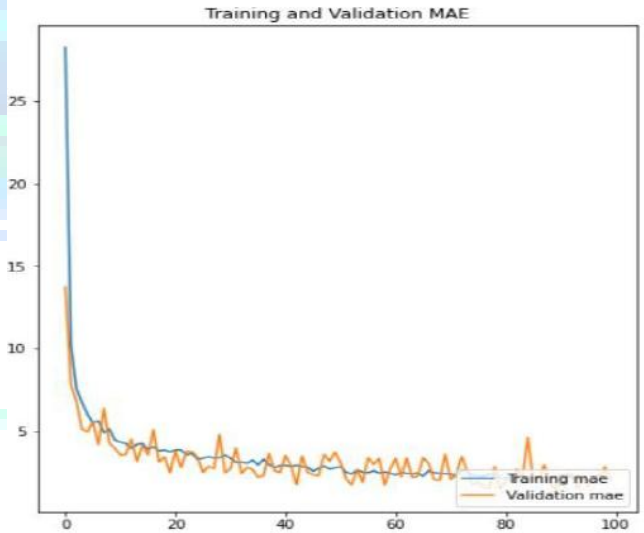


Figure : 9 MAE graph of LSTM and GRU

The proposed LSTM and GRU model's MAE performance is shown in the figure 20 using training and validation data. 2.0 and 3.38 respectively, for the LSTM and GRU models. There are exactly 100 epochs in this graph, and the y-axis displays the MAE value. To begin, the model received sequence data rather than random data from the input layer and the following LSTM and GRU layers. To avoid overfitting the model, a dropout layer is used. Finally, a single hot encoded output is generated using a dense layer. In

addition to creating the model, begin training by stopping at a checkpoint and then restarting. An early termination of training and a saving of model weight occurs when observed losses exceed tolerance levels.

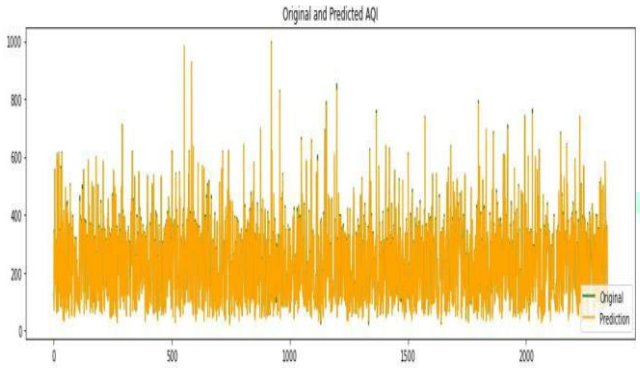


Figure 10: Prediction graph of proposed models

V. CONCLUSION AND FUTURE WORK

Because of the data's volatile and dynamic nature, as well as its unpredictable nature in space and time, predicting pollutant levels is inherently difficult. Nevertheless, the importance of predicting pollution concentrations has increased as a result of pollution's negative effects on people and the environment. It is possible to conclude, based on a review of specific publications that use deep learning methods to predict air pollution in smart cities, that deep learning is an overall technology which can be applied in a variety of research disciplines. In the early stages of air pollution forecasting, deep networks that best describe the characteristics of contaminant progression across a wide range of scales are needed to better describe how pollutants spread. Deep learning has the potential to improve air quality prediction in a number of ways, including forecasting, estimating sources, and filling in missing data. We used GRU and LSTM to predict levels of pollutants like NO2, SO2, Pm10 and Pm2.5, and the Air Quality Index, using publicly available data for Delhi (AQI).

The achievement of machine Learning Techniques, such as Artificial Neural Networks (ANN) and genetic algorithms, should be investigated and evaluated in the future as a next step. For larger datasets, we'd like to look into different hyper parameter optimization techniques and variable selection.

REFERENCES

[1.] J. Cai, X. Dai, L. Hong, Z. Gao, and Z. Qiu, "An Air Quality Prediction Model Based on a Noise Reduction Self-Coding Deep Network," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/3507197.

[2.] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep Learning for Air Quality Forecasts: a Review," *Curr. Pollut. Reports*, vol. 6, no. 4, pp. 399–409, 2020, doi: 10.1007/s40726-020-00159-z.

[3.] G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and Forecasting of Air Quality Index in Chennai using Regression and ARIMA time series models," *J. Eng. Res.*, 2021, doi: 10.36909/jer.10253.

[4.] P. Patil, P. Modi, and N. Kamble, "DETERMINATION OF AIR QUALITY MONITORING & PREDICTION," pp. 1146–1148, 2020.

[5.] S. Bhattacharya and S. Shahnawaz, "Using Machine Learning to Predict Air Quality Index in New Delhi."

[6.] R. M. Patil, D. H. T. Dinde, and S. K. Powar, "A Literature Review on Prediction of Air Quality Index and Forecasting Ambient Air Pollutants using Machine Learning Algorithms," *Int. J. Innov. Sci. Res. Technol.*, 2020, doi: 10.38124/ijisrt20aug683.

[7.] Xie Yu, "Deep Learning Architectures for PM2.5 and Visibility Predictions," 2018.

[8.] S. V. Kumar *et al.*, "Data analysis for predicting air pollutant concentration in Smart city Uppsala," *Proc. 2nd ACM SIGSPATIAL Int. Work. GeoStreaming*, vol. 4, no. 6, pp. 664– 672, 2015.

[9.] R. G L, "Air Quality Index Detection Using Machine Learning and IOT," *Interantional J. Sci. Res. Eng. Manag.*, vol. 06, no. 02, pp. 3068–3071, 2022, doi: 10.55041/ijsrem11605.

[10.] B. Chakravarthi, "Prediction an air quality index data using machine learning and deep learning MSc Research Project Data Analytics RuchitaPatil."

