Prediction of Delhi's Air Quality Index Using Deep Learning Approach

Aniket Shakya[1], Jay Singh Rajput[2],
[1]Student of M. Tech, [2]Research Scholar
(Environmental Engineering),
Madhav Institute of Technology & Science,
Gwalior, Madhya Pradesh, India

Prof. A. K. Saxena[3]
[3]Associate Professor,
Madhav Institute of
Technology & Science,
Gwalior, Madhya Pradesh, India

Abstract:- Air pollution is becoming a concern issue as it deteriorates the ambient atmosphere. On account of rapid industrialization and urbanization & wreaking havoc on the nature, this situation is getting worsen diurnally. Consequently, it causes health issues to human & animals, potential damage to historical monuments, and imparts negative impact on plant growth. Numerous initiatives have been made by state and federal governments to reduce air pollution. Globally, air quality is measured by the "Air Quality Index (AQI)" that further used to educate the public about the current status of air pollution in their surroundings. This suggested research explores an effective method for predicting Delhi's air quality index. In this study, the 'Gated Recurrent Unit (GRU)' model is used to forecast daily air quality index. In this model, air quality index isused as a 0dependent variable & daily pollutant concentration as an independent variable. Furthermore, hyper-parameter settings have been also assigned such as No. of neurons - 50, epochs - 100, learning rate - 0.001, loss function –Mean square error (MSE), training & testing dataset – 80 – 20 (in %), andoptimization function–'ADAM'. The accuracy of model represents through Mean square error (MSE)- 21.89, Root mean square error (RMSE)- 4.69, Mean absolute error (MAE)- 3.43, Correlation coefficient (R^2 value)- 0.99. Therefore, this study is proposing "GRU" model to produced precise &accurate daily estimates of air quality index.

Keywords-Air Quality Index(AQI),Deep Learning Model, GRU, Prediction Model.

I. INTRODUCTION

In recent decades, energy consumption has expanded quickly due to accelerated rate of urbanisation and industrialization, that resulting in severe air quality problems. Air pollution incidents have been more common in China recently, posing a major threat to inhabitants health[1]. Air pollution introduces hazardous substances in the form of particulates pollutants and gaseous pollutants in the ambient atmosphere. These air pollutants are harmful to the environment, such as particulates, tropospheric ozone, carbon dioxide, or sulphur dioxide etc. Consequently, air quality has drastically disrupted the ecosystem's balance, resulting in the greenhouse gases, acid rain, and depletion of ozone layer, among other things.Also, multiple studies have governed that long-term air pollution spread variety of negative health effects[2], including respiratory and cardiovascular problems.

World Health Organization (WHO) report shows air pollution causes 4.2 million deaths every year, and more than 90% of the world's largest population that lives in areas where air quality is hazardous. Seven of the 10 most polluted cities in the world are in India, according to a study of'IQ-Air', 'Air Visual', and 'Greenpeace'[3][4].Central Pollution Control Board (CPCB), India is monitoring the air quality throughout the country. These networks of air monitoring stations have been set up by the CPCB to keep tabs on local air quality. There are approximately 46 monitoring stations in and around Delhi (34 in Delhi and 12 in NCR). The air pollutants concentration level depend on its location in relation to nearby stations[5][6].The elders and children, who may be more susceptible to the air pollution-related illnesses, they will be benefitted from accurate prediction of air quality for the city[5]. An accurate assessment of air quality and a forecast based on that information can help decision-makers to enhance air quality at a lower cost and with greater efficiency come up with new and innovative solutions for the future[7]. Therefore, a prior air quality alert system is required for properly and truly predicting ambient air & its information[8].

The Air Quality Index (AQI) concept has been developed in many industrialised nations from last three decades[9][10][11][12]. It is a comprehensive method that transformed weighted values of different air pollutant (SO_2, CO, NO_x, etc.) into a single number, which represent the status of air quality.The objective of AQI has (i) Filtering the scientific and technical complexity into simple understandable information, (ii) Communicated with citizens for their present and future reference. Additionally, forecasting of AQI can be emerging as efficient tool for individuals and local authorities to combat with air pollution problem at local level.

To predict small changes inair pollution concentrations are very crucial and difficult, especially in case of limited data inputs and high parameter variability[13]. In such situation, high computational power is required, which can be achieved through machine learning and deep learning approach. Machine learning is a subset of Artificial Intelligence (AI) that enables a machine to learn from data and information. On the other hand, deep learning (DL) has the capacity to learn from enormous volumes of data. Apart from others, the DL has wide application in air pollution field too [14]. Since the AQI follows a periodic pattern, deep learning models can be used to effectively predict the AQI values. There are various deep learning algorithms like

CNN, LSTM, GRU, RNN and DNN used to predict AQI[15].

Now a days, researchers are more interested in solving many air-pollution related problems with the computational techniques. In this context [16] have found a significant impact of NO, CO, and particulate matter on respiratory mortality by using multilayer perceptron (MLP) or non-linear autoregressive (NAR) models.[17]–[20] have evaluatedvarious machine learning and deep learning algorithms such as multiple linear regression, Support Vector Machine, Random Forest, Extreme Gradient Boosting, computer vision algorithm and Light GBM, etc. for predicting the local AQI level. Even, statistical models are frequently employed in the analysis and forecast on air quality such as the AQI was predicted using a statistical approach based on linear multiple linear regressions[21], an AQI monitoring station in Northern Californiawas used to develop a multiple regression model that successfully predicted day-ahead AQI levels based on three years of daily data[22].

In this study, an accurate and precise model is aimed to be developed for air quality index prediction for Delhi. In this regard, data has to be imported from the CPCB official website and then make it available for computational purpose in the deep learning algorithm with data pre-processing techniques. A deep learning model "Gated Recurrent Unit(GRU)" has been applied on the dataset and optimize their hyper parameters up to desired outcome that will lead to achieve better results.

II. STUDY AREA

Delhi is the capital of India spread over 1483 km^2& has coordinates 28.7041° N, 77.1025° E.Delhi has one of the highest road densities of 1749 km of road length per 100 km^2 in India. Its high population growth rate, & high economic growth rate, has resulted in increasing demand for transportation creating excessive pressure on the city's existent transport infrastructure. The city faces acute transport management problems leading to air pollution, congestion and resultant loss of productivity. The air quality index (AQI) in Delhi was generally moderate (101–200) from January to September of the prior five years. Several factors cause the air quality index to fall to dangerously low levels from poor (201–300) to very poor (301–400) or severe (400$^+$) throughout the months of October to December. On a scale of contaminants and control strategies, Delhi's air pollution status has altered significantly. Many districts in Delhi's National Capital Region (NCR) are among the most polluted areas in the city.

III. RESEARCH METHODOLOGY

To accomplish this task, stepwise process is adopted as mentioned in the flowchart shown in Fig. 1. Initially, datasets were collected and prepared for further analysis. Here, a statistical analysis represents characteristic of datasets. Subsequently, datasets are further utilized to develop a proposed deep learning model "GRU". Further, detailed information about each stepis mentioned in the subsequent sections.

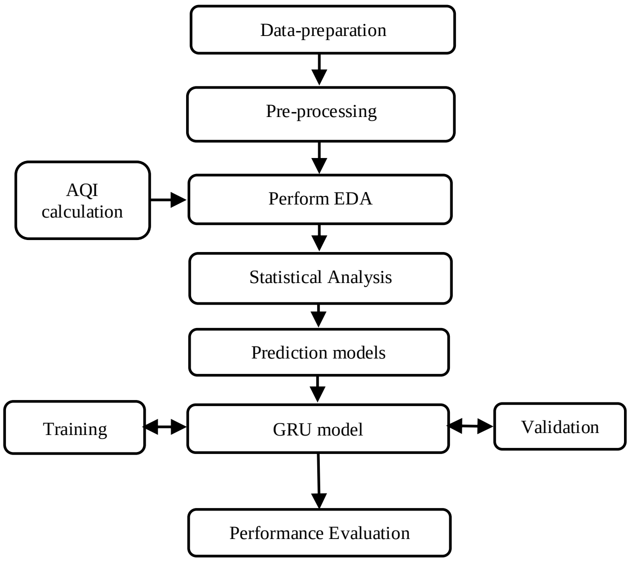


Fig. 1: Flowchart of the proposed GRU model

A. *Data Acquisition, preparation, and Normalization*

The dataset was obtained from the official website of the Central Pollution Control Board (CPCB) https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/dataover a period of nine years (2012 to 2020) from all 34 monitoring stations in Delhi (India). Within the study, seven pollutants such as $PM_{2.5}$, PM_{10}, NO_X, NH_3, SO_2, CO, and O_3 are used as independent variables, whereas the AQI as a dependent variable. These pollutants were chosen as they were use in the AQI calculator provided by the CPCB, which is a part of the Ministry of Environment Forestand Climate Change in India. Also, the measuring units of 24-hourly averaged concentrations of all parameters are same (in µg/m³), except CO (inmg/m³). In this manner, a total of 110032 data points were collected under each parameter from all 34 monitoring stations of Delhi.Furthermore, collected raw data were prepared as several types of ambiguity were present in the raw datasets such as null values, duplicates values, and outlier values etc. In this way, a total of 23449 data points were remained after data pre-processing. In general, each air quality parameters have their own dimensions to represent air quality. This condition may rise problem in the prediction of dataset in prediction model. Therefore, data normalization is necessary to perform by using Equation (1) to normalize the values of the measured parameters that will be used in the computations in order to increase prediction accuracy. In this way, the entire variable comes under a same dimensionless category.

$$Normalized\ Data\ (x') = \frac{x - min(x)}{max(x) - min(x)} (1)$$

B. *AQI calculation*

The AQI technique combines air pollutant concentrations into a single value in such a way that the general public can understand about air quality status. In order to create an AQI, there are two stages: (i) Sub-indices creation (for every pollutant) (ii) Aggregation of these sub-indices to get an overall AQI[23]. Further, the sub-indices of each pollutant were calculated with the linear segmentation principle-based equation (see eqn. 2). The relevant data used in this equation were taken from Table 1.Thereafter, the maximum sub-indices value of any pollutants (I_{PM10}, $I_{PM2.5}$, I_{NOX}, I_{NH3}, I_{SO2}, I_{CO}, I_{Ozone}) represent the value of AQI.

$$\text{Sub Index,}\mathbf{I_P} = \frac{\mathbf{I_{HI} - I_{LO}}}{\mathbf{BP_{HI} - BP_{LO}}}(\mathbf{C_P} - \mathbf{BP_{LO}}) + \mathbf{I_{LO}} \qquad (2)$$

Where,

BP_{HI} = Breakpoint concentration given in the table that is $\geq C_P$.

BP_{LO} = Breakpoint concentration given in the table that is $\leq C_P$.

I_{HI} = Sub-Index value corresponding to B_{PHI}.

I_{LO} = Sub-Index value corresponding to B_{PLO}.

C_P = Actual ambient concentration of pollutant 'P'.

P = Anypollutant either of PM2.5, PM10, NO_X, NH_3, SO_2, CO, Ozone.

The AQI can be estimated using a variety of methodologies provided by the CPCB. However, this estimate is based on the United States Environmental Protection Agency's (USEPA) method and CPCB recommended this approach to compute AQI since, it can be used to a wide range of contaminants and is more reliable than other AQI formulations provided by various agencies because it is(i) free of eclipsing or ambiguity (Ott 1978) and (ii) the health effects of combining contaminants 'synergistic effects' are unknown. Further, a health-based score cannot be merged or weighted; therefore, the aggregation of sub-indices is based on the maximal operator approach. In this regard, Sharma et al. (2001, 2002, and 2003) used the maximum operator approach to build an AQI index for IIT-Kanpur and the entire country.

AQI Category Range	**PM_{10} 24-Hr**	**$PM_{2.5}$ 24-Hr**	**NOx 24-Hr**	**O_3 8-Hr**	**CO 8-Hr**	**SO_2 24-Hr**	**NH_3 24-Hr**
Good (0-50)	0-50	0-30	0-40	0-50	0-1	0-40	0-200
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400
Moderate (101-200)	101-250	61-90	81-180	101-168	2-10	81-380	401-800
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200
Very poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1201-1800

Table 1: Sub-index and Break-point concentrations of each pollutant given by CPCB

C. Prediction Model: Gated Recurrent Unit (GRU)-

The gated recurrent unit is a type of LSTM-based RNN model that is more specialised. The internal unit of the GRU is similar to that of the LSTM, with the exception that the GRU combines the forgetting and incoming ports into a single update port. The multi-GRU prediction system was built using GRU models for power generation scheduling and monitoring. It is considered to be more straightforward in computation and execution, while being influenced by the LSTM unit. It keeps LSTM from yielding to the problem of disappearing gradients. Its internal structure is simplified, making it easier to train because less mathematics is required to improve its internal states. The updated port specifies how much of the previous moment's status data should be kept in the current state, whereas the reset line specifies whether the present state should be combined with the previous data or not. The inner layer of a GRU unit cell is depicted in the diagram. These are the mathematical procedures required to control the locking mechanism of the GRU cell[24][25]. GRU's architecture is depicted in Fig. 2.

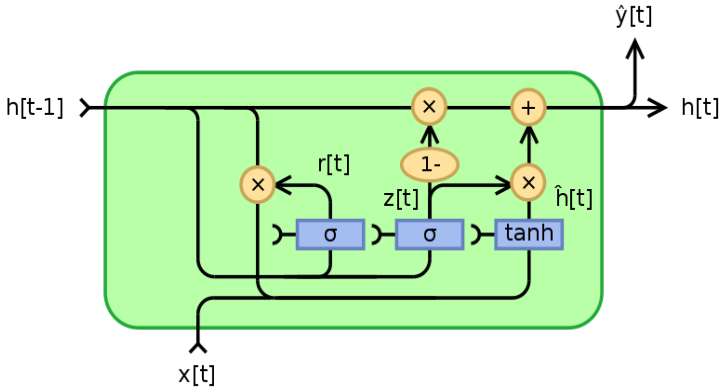


Fig. 2: A general architecture of GRU.

Equations are used in the GRU to calculate the next output or state value (1) (2) (3) and (4)[26].

$$z_t = \sigma(W_z * [x(t), h(t-1)]) \quad (1)$$
$$r_t = \sigma(W_r * [x(t), h(t-1)]) \quad (2)$$
$$h(t) = \sigma(W_h * [x(t), (r_t * h)(t-1))]) \quad (3)$$
$$h(t) = (1 - z_t) * h(t-1) + z_t * h(t) \quad (4)$$

Where, σ = Activation function, x(t) = Input, h(t-1) = Previous output, W_z, W_r, W_h= Weights. GRU is preferred over LSTM because it has lesser number of parameters and hence requires fewer training devices. GRU networks outperformed LSTM networks on lesser datasets when it came to music and speech signals.

IV. RESULTS AND DISCUSSION

A. *Statistics analysis-*

a) Descriptive statistics

The 24- hour averaged concentration of $PM_{2.5}$, PM_{10}, NO_x, NH_3, SO_2, O_3, and CO are 109.79, 229.01, 60.95, 37.72, 14.58, 37.15, and 1.45, all parameter have same measurement unit i. e. µg/m³, except CO concentration is measured in mg/m³. Further, the positive value of kurtosis of each parameter shows peak at a certain concentration level in the data distribution see Table 2. Specially, CO concentration has very specific peak value and does not vary so much. However, the data distribution of AQI are sufficient varied and represent nearly about normal distribution. Furthermore, the positive value of skewness shows left skewness in the data distribution, which shows that mode of the data is on higher side and represent that repetitive value of parameter concentrations are more. However, the data distribution of the AQI shows normal distribution with skewness 1.05 and having a symmetrical data distribution.

b) Exploratory Data analysis

Exploratory Data Analysis (EDA) is a technique for identifying a dataset's most important characteristics. It's used to understand data, put it into perspective, analyse components and their relationships, and offer recommendations that could help with forecasting model development. The correlation between variablesis depicted in (Figure 3).

Statistic Parameters	**$PM_{2.5}$**	**PM_{10}**	**NOx**	**NH_3**	**SO_2**	**CO**	**O_3**	**AQI**
Mean	109.79	229.01	60.95	37.72	14.58	1.45	37.15	236.72
Standard Deviation	90.51	144.31	63.27	21.45	9.58	1.05	23.24	147.42
Kurtosis	3.89	1.17	8.87	10.80	11.77	87.50	2.03	1.61
Skewness	1.75	1.04	2.58	2.33	2.09	5.27	1.15	1.05
Count	23449	23449	23449	23449	23449	23449	23449	23449
Confidence level (95%)	1.16	1.85	0.81	0.27	0.12	0.01	0.30	1.89

Table 2: Descriptive statistics of pollutants and AQI

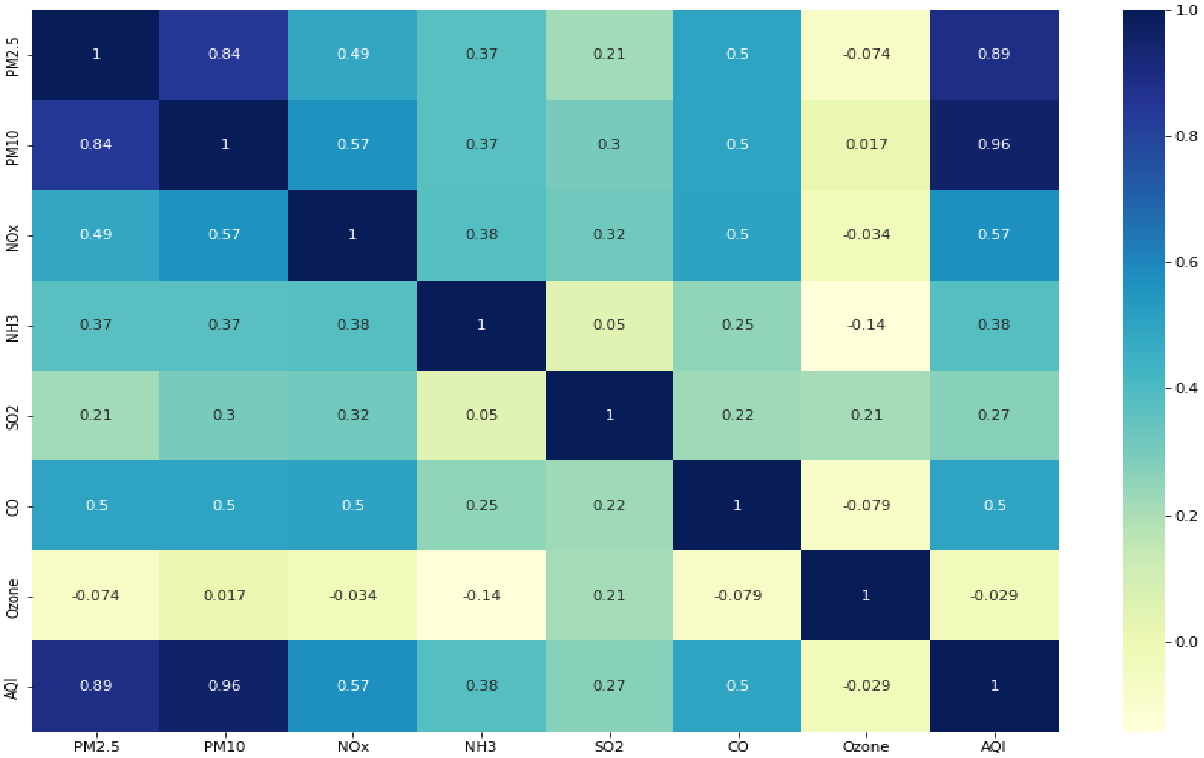
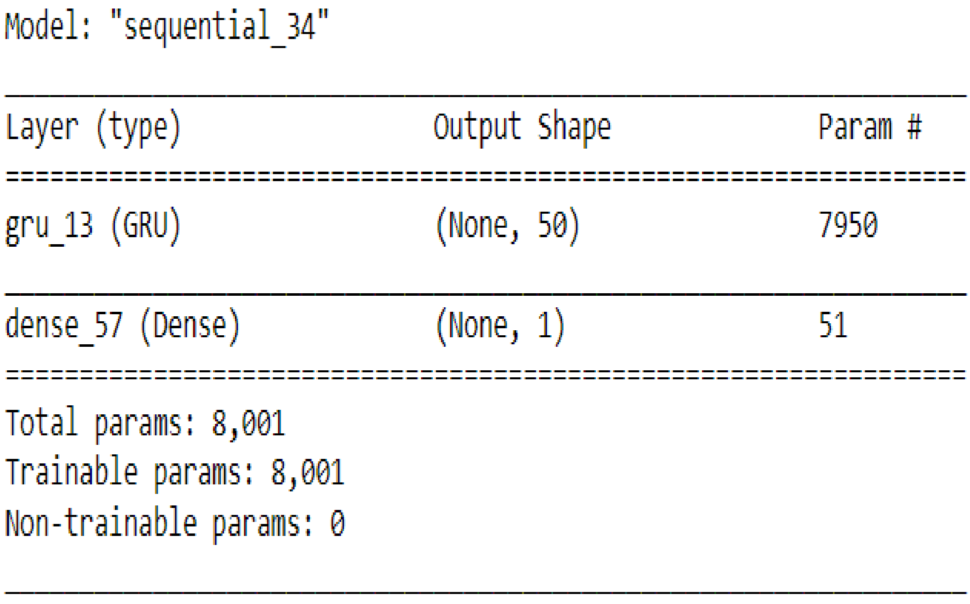


Fig. 3: Correlation between the variables of dataset

B. Prediction of AQI: Gated Recurrent Unit (GRU) model

The prediction of AQI was obtained through GRU model. Actually, the prediction ability of the model depends upon the set of hyper parameters related to GRU model. Therefore, a hit and trial approach were executed to obtained optimized value of the hyper parameter. These optimized values of hyper parameter are shown in Table 3. In this model, 80:20 ratios were chosen for the training and testing of dataset respectively. For the GRU model, the 'Mean square error (MSE)'is used as loss function, and the GRU &dense layers were trained.The loading parameters of several pollutants and AQI are shown in (Figure 4).



```
Model: "sequential_34"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
gru_13 (GRU)                 (None, 50)                7950
_________________________________________________________________
dense_57 (Dense)             (None, 1)                 51
=================================================================
Total params: 8,001
Trainable params: 8,001
Non-trainable params: 0
_________________________________________________________________
```

Fig. 4: Shows the loading parameters of GRU model

Also, the "ADAM" optimization function was used in conjunction with a 50-encodingdimension and a learning rate 0.001. Further, a minimum value of loss function was obtained (see fig. 5) with good correlation (R^2 = 0.99) shown in Table 4. Also, the data distributions of the actual and predicted AQI are shown in Fig. 6.

Hyper parameters	Sequential
RNN Layers	GRU
Neurons	50,1
Optimizer	ADAM
Learning rate	0.001
Metrics	MSE
Loss	MSE
Epochs	100

Table 3: Hyper parameters for GRU model

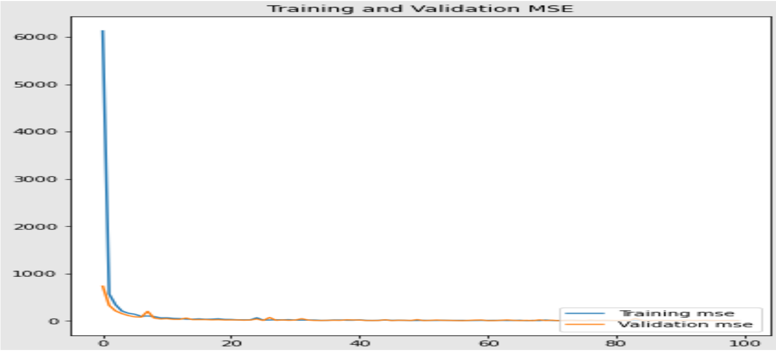


Fig. 5: Shows MSE of Training and Validation

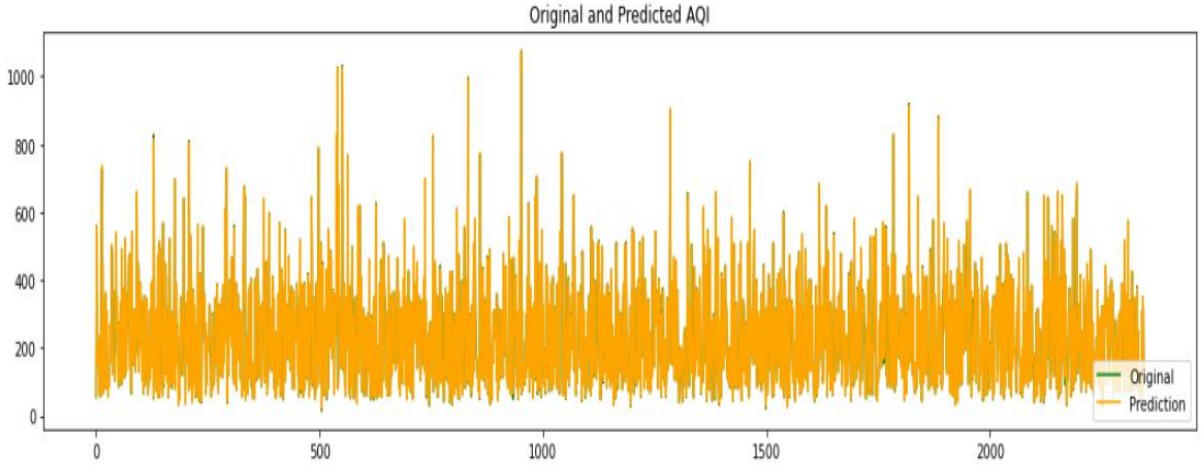


Fig. 6: Actual and predicted AQI data distribution

The above model is used for daily ambient air quality predictions. Once model have indeed been trained then the daily AQI values predicted by the model are produced by comparing them to the daily actual air pollution concentrations. A quantitative evaluation of an output's accuracy was undertaken using statistical criteria. The outcomes of the GRU model's trainingare presented in Table 4.

Model	MSE	RMSE	MAE	R^2
Gated Recurrent Unit (GRU)	21.89	4.68	3.43	0.99

Table 4: Illustrates the evaluated results of proposed model in this study

V. CONCLUSION

Predicting air pollution is a critical analytical issue that has the ability to aid decision-making and alleviate air quality-related difficulties. In this paper, we looked at how deep learning architectures could be used to effectively handle this problem. This research focuses on AQI prediction for different air pollutant concentrations& gave the accuracy of model represent through Mean square error (MSE)- 21.89, Root mean square error (RMSE)- 4.68, Mean absolute error (MAE)- 3.43, Correlation coefficient (R^2 value)- 0.99. The proposed strategy is meant to be applicable to a wide variety of applications; the methodology's phases can be used to perform similar tasks, such as anticipating additional air pollution concentrations. Further, use of hybrid model along with meteorological parameters so this model can aid in achieving better results and accuracy.

REFERENCES

[1.] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, “Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models,” in *Procedia Computer Science*, 2020, vol. 171, no. 2019, pp. 2057–2066, doi: 10.1016/j.procs.2020.04.221.

[2.] S. Bali, “Indian Air Quality Prediction and,” vol. 14, no. 11, pp. 181–186, 2019.

[3.] R. Murugan and N. Palanichamy, “Smart city air quality prediction using machine learning,” *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1048–1054, 2021, doi: 10.1109/ICICCS51141.2021.9432074.

[4.] H. Mahboubi *et al.*, “Maximum Lifetime Strategy for Target Monitoring With Controlled Node Mobility in Sensor Networks With Obstacles,” *IEEE Trans. Automat. Contr.*, vol. 61, no. 11, pp. 3493–3508, 2016, doi: 10.1109/TAC.2016.2536800.

[5.] S. Abirami and P. Chitra, “Regional air quality forecasting using spatiotemporal deep learning,” *J. Clean. Prod.*, vol. 283, p. 125341, 2021, doi: 10.1016/j.jclepro.2020.125341.

[6.] S. Yarragunta, M. A. Nabi, P. Jeyanthi, and S. Revathy, “Prediction of air pollutants using supervised machine learning,” *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1633–1640, 2021, doi: 10.1109/ICICCS51141.2021.9432078.

[7.] K. Zhang, J. Thé, G. Xie, and H. Yu, “Multi-step ahead forecasting of regional air quality using spatial-temporal deep neural networks: A case study of Huaihai Economic Zone,” *J. Clean. Prod.*, p. 123231, 2020, doi: 10.1016/j.jclepro.2020.123231.

[8.] Z. Du, J. Heng, M. Niu, and S. Sun, “An innovative ensemble learning air pollution early-warning system for China based on incremental extreme learning machine,” *Atmos. Pollut. Res.*, vol. 12, no. 9, p. 101153, 2021, doi: 10.1016/j.apr.2021.101153.

[9.] USEPA, *Air Quality Index a Guide to Air Quality and Your Health*. 2014.

[10.] USEPA, “Federal register,” vol. 41, no. 174, 1976.

[11.] Ontario, “A review of the Ontario air quality index and air quality health index system,” *Air Resour. Branch, Ontario Minist. Environ. Toronto, Ont., Canada.*, 2013.

[12.] L. Shenfeld, “Note on Ontario’s air pollution index and alert system,” *J. Air Pollut. Control Assoc*, vol. 20, no. 9, 1970.

[13.] K. M. O. Nahar, M. Ashraf Ottom, F. Alshibli, and M. M. A. Shquier, “AIR QUALITY INDEX USING MACHINE LEARNING-A JORDAN CASE STUDY,” 2020.

[14.] V. Pream Sudha and R. Kowsalya, “a Survey on Deep Learning Techniques, Applications and Challenges,” *Int. J. Adv. Res. Sci. Eng. IJARSE*, vol. 8354, no. 4, p. 3, 2015.

[15.] H. Li, J. Wang, and H. Yang, “A novel dynamic ensemble air quality index forecasting system,” *Atmos. Pollut. Res.*, vol. 11, no. 8, pp. 1258–1270, 2020, doi: 10.1016/j.apr.2020.04.010.

[16.] C. Song and X. Fu, “Research on different weight combination in air quality forecasting models,” *J. Clean. Prod.*, vol. 261, p. 121169, 2020, doi: 10.1016/j.jclepro.2020.121169.

[17.] D. Q. Duong *et al.*, “Multi-source Machine Learning for AQI Estimation,” *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 4567–4576, 2020, doi: 10.1109/BigData50022.2020.9378322.

[18.] T. Madan, S. Sagar, and D. Virmani, “Air Quality Prediction using Machine Learning Algorithms-A Review,” *Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020*, pp. 140–145, 2020, doi: 10.1109/ICACCCN51052.2020.9362912.

[19.] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, “Air Quality Prediction of Data Log by Machine Learning,” *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 1395–1399, 2020, doi: 10.1109/ICACCS48705.2020.9074431.

[20.] L. Ma, Y. Gao, and C. Zhao, “Research on machine learning prediction of air quality index based on SPSS,” *Proc. - 2020 Int. Conf. Comput. Network, Electron. Autom. ICCNEA 2020*, pp. 1–5, 2020, doi: 10.1109/ICCNEA50255.2020.00011.

[21.] A. Barve, V. Mohan Singh, S. Shrirao, and M. Bedekar, “Air quality index forecasting using parallel dense neural network and LSTM cell,” *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 51–54, 2020, doi: 10.1109/INCET49848.2020.9154069.

[22.] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, no. Ml, 2020, doi: 10.1155/2020/8049504.

[23.] M. Bansal, A. Aggarwal, and T. Verma, “Air Quality Index Prediction of Delhi using LSTM,” *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 8, no. 5, pp. 59–68, 2019, doi: 10.13140/RG.2.2.26885.70884.

[24.] Alind Gupta, “Gated Recurrent Unit Networks,” *Geeks for Geeks*, 2019. .

[25.] C. et Al., “Gated Recurrent Unit,” *papers with code*, 2014. .

[26.] S. Kostadinov, “Understanding GRU Networks,” *towards data science*, 2017.