ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse115912020.pdf
https://doi.org/10.30534/ijatcse/2020/115912020

Forecasting of the Air Pollution Based on Meteorological Data and Air Pollutants using Deep Learning: A Novel Review

Keerthana R[1] Dr. Varghese S Chooralil[2]
[1]Rajagiri School of Engineering and Technology, India, kr.keerthana99@gmail.com
[2]Rajagiri School of Engineering and Technology, India, varghesee.kutty@gmail.com

ABSTRACT

Air contamination has become a major issue in India. Cities like Delhi, Lucknow, and Agra are suffering from severe air pollution. So to protect people from air pollution is by predicting the concentration of air pollutant in air in future years. In this work, a functional model is built to predict the concentration of air pollutants like SO_2, NO_2, PM2.5, CO and O_3 in the air in Indian cities like Delhi, Agra, Lucknow. We take the meteorological data like temperature, relative humidity, wind speed and direction for the accurate prediction of pollutants.

Key words: Functional Model, Fully Connected Network, Long Short Term memory, Recurrent Neural Network.

1. INTRODUCTION

Due to industrialization and urbanization air pollution has become a major problem in India. The WHO report [8] says that, 14 out of 15 most polluted cities in the world are in India. Air pollution causes severe health problems like cancer, respiratory problems, heart attack etc. Most of the Indian cities are polluted, Delhi and Kanpur are top among them. In 2016 Kanpur is the most polluted city in India in which PM2.5 on an average of 173 micrograms per year. The major pollutants in India are SO_2, NO_2, PM2.5, CO and O_3.

- SO_2: India is the largest emitter of SO2 in the world [11]. The source of SO2 in the atmosphere is the burning of fossil fuels in power plants and other industrial factors.

- NO_2: Nitrogen oxide is not directly emitted from any source, but is formed from reaction between gases in the air under the influence of sunlight and high temperature [12]. NO2 is a dangerous pollutant because of, it is the reason for the formation of PM2.5 and ozone. Thus, breathing the air with a high convergence of NO2 can cause diseases in the human respiratory system.

- CO: Carbon monoxide (CO) [13] is a colourless, odourless, and non-irritating gas which is difficult to detect which leads to unexpected death. CO is formed due to incomplete combustion of organic material, for example, gasoline, coal, wood, propane, and natural gas. CO [13] concentrations reach a maximum in the early morning hours due to heavy morning traffic and then fall to elevated levels during the day. A second peak of CO concentration is usually observed corresponding to the late afternoon traffic period, and decreases to low levels during the night. Almost three years the CO will remain in the air. The sources of carbon monoxide have been continuously increasing during the last few years in most of the urban areas.

- O_3: Surface level ozone [14] is one of the important air pollutants. It is formed by the reaction of atmospheric pollutants in the presence of sunlight. The surface ozone shows variation in the country. The levels are high during summer and low in monsoon seasons.

- PM2.5: The particulate matter [13], [15] refers to the solid or liquid particles in its form of dust, mist or smoke etc. originated by the dispersion of particles from breakdown of solid material. The particle size is one of the most important characteristics of airborne particulate matter because it controls the residence time of particles in the ambient air. If the particle that has a diameter less than 2.5 micrometers then these particles can enter the lungs through our nose and some may even enter the circulatory system this causes many respiratory problems, lung diseases etc. Because of its small size, light and fine particle they can stay longer in the air.

These are the major pollutants that can cause severe health problems. The PM2.5 is the dangerous air pollutant because of its small size it can easily entered and penetrate deeper into the lungs than

larger particles [13]. So most of the previous studies are based on the concentration of PM2.5. This work is focus on the prediction of different air pollutants like SO_2, NO_2, PM2.5, CO and O_3 using RNN and the prediction of meteorological factors like temperature, relative humidity, wind speed and direction using LSTM then concatenate the output of these models and build a functional model to predict the accurate values of the air pollutants. For our research study and experiments, we choose Delhi, Lucknow, Agra etc. India is one of the most populated country in the world, therefore forecasting air pollution in Indian cities would have a high impact on people's lives. The data is collected from the Central Pollution Control Board website.

Problem Statement: To predict the concentration of air pollutants like SO_2, NO_2, CO, PM2.5, O_3 in industrial cities like Agra, Delhi, Lucknow in India using deep learning models.

Objective:

- Predict the air quality of different cities using RNN and LSTM.
- Predict the air pollutant concentration using RNN and predict the temperature, wind speed, wind direction and relative humidity using LSTM.
- Build a functional model by concatenating the output of LSTM and RNN and give to a Fully Connected layer for accurate prediction

The main contribution of this paper is to predict the air pollution of cities in India and then say what will be the condition of air in those cities in future. The figure 1 [19] shows different levels of air pollutants level.

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600
Severe (401–500)	430+	250+	400+	748+	34+	1600+

Figure 1: Pollutants levels.

2. LITERATURE SURVEY

In this section, we discuss the related works on predicting the air pollution using various deep learning methods. In previous papers, most of them predict the concentration of PM2.5 using various factors and methods.

In paper [4] it uses one dimensional convolutional networks and bidirectional Gated Recurrent Unit (GRU). The use of convolutional network is to reduce the size of input data by extracting its features. It process the meteorological data and PM2.5 time series data. The data preprocessing is done using 1D convolutional networks and prediction of PM2.5 time series data by using bidirectional GRU.

In paper [1], [6], it predicts the concentration of pollutant PM10 using deep learning models like Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU). The input data is normalized to set in the range of 0 to 1. The six dimensional input such as temperature, Humidity, U-wind component, v-wind component, Precipitation and Total cloud cover is fed into the neural network and gives the predicted value of PM10 in the air.

In paper [2], the bidirectional LSTM model is used to predict the concentration of PM2.5. The input given to Model is PM2.5, and different meteorological factors. Because of taking the bidirectional LSTM it takes the input in forward and backward layer, and output from both layer are concatenated and given to an activation function for prediction.

The paper [3], [21] is to predict the concentration of PM2.5 using the seq2seq model. The sequence to sequence has the encoder and decoder both contains RNN and it runs in a step by step manner i.e. it is not parallel. So sequence to sequence has slow training speed, because of the RNN in the encoder. Another problem in error accumulation. So the RNN is replaced with a fully connected layer in order to increase its speed. In the decoder part N-step recurrent prediction is applied to improve the accuracy.

The time series based LSTM Model [5], this is to predict the concentration of air pollutants in India using meteorological factors, traffic data, festivals and national holidays. So here prediction of concentrationof air pollutants is done and also it calculates the air quality index in future. It takes the current input data and predicts the next hour pollutant. These four dimensional data are fed into the LSTM model.

3. METHODOLOGY

There are six prominent pollutants in the air, Particulate Matter (PM2.5 and PM10), Carbon Monoxide (CO), Ozone (O_3), Nitrogen dioxides (NO_2), Sulphur dioxide (SO_2). India's Central Pollution Control Board now routinely monitors these pollutants. The observing of meteorological parameters like wind speed and direction, relative humidity, and temperature has been integrated along with the monitoring of air quality. The aim of this work is, to predict the concentration of air

pollutants like SO_2, NO_2, PM2.5, CO, O_3 in the air, in industrial cities in India. The proposed method consists of three phase:

- Prediction of Air Pollutants Using RNN.
- Prediction of Meteorological Data using LSTM.
- A Deep Learning Prediction Model by combining multiple input.

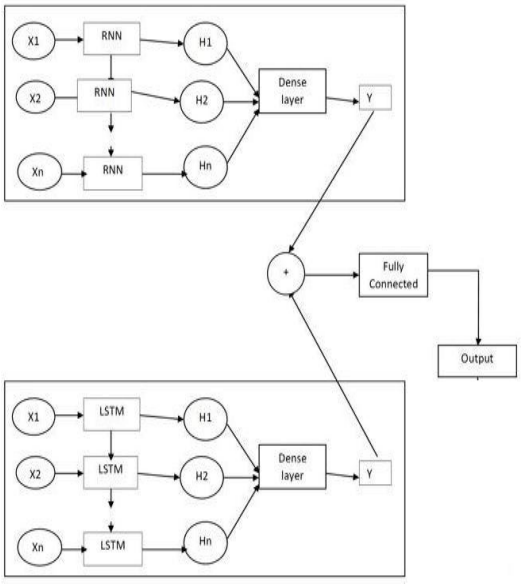

Figure 2: Proposed Architecture.

The proposed model shown in figure 2, is used for the prediction of air pollutant concentration is designed using deep learning methods like Recurrent Neural network (RNN), Long Short Term Memory (LSTM), and Fully Connected Network (FC). A functional model is built by concatenating the output of RNN and LSTM, and then giving it to a fully connected layer for more accurate prediction. The real time data of pollutants is downloaded from the Central Pollution Control Board (CPCB) website [7] every day from 2016 to 2018. The data is normalized using Min-Max Scalar function in python. The input to the Recurrent Neural Network is a time series data, the values of air pollutants like SO_2, NO_2, CO, PM2.5, and O_3. The day to day input is given and predicts the next day's value. Similarly the input to LSTM is time series data of temperature and relative humidity and predicts the next day temperature and relative humidity. The output from RNN and LSTM is concatenated and given to a fully connected layer for more accurate prediction.

3.1 Prediction of Air Pollutants Using RNN

In this section, the RNN is used to predict the concentration of air pollutants like SO_2, NO_2, CO, PM2.5, and O_3. For data preprocessing, first step is, remove the columns that is not need, check missing values if there is any missing value (0 or None) then replace it with a quartile value. Quartiles [16] in statistics are values that divide your data into four equal parts. The first quartile (also called the lower quartile) is the number below which lays the 25 percent of the data. The second quartile (the median) divides the data in the middle and has 50 percent of the data. The third quartile (also called the upper quartile) has 75 percent of the data. And find the mean of each quartile values then replace zeros with mean of quartiles that zeros comes in position between the quartiles. Min-Max Scalar [5], [20] of sklearn library is used to convert whole dataset into float data type. When scaling, the value varies between 0 and 1.

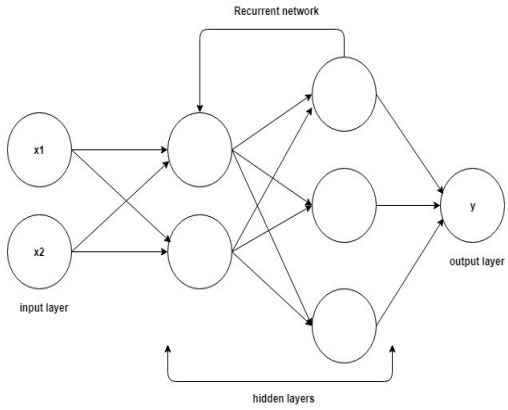

Figure 3: Recurrent Neural Network.

RNN [10] is a deep neural network, whose basic structure is unit. The unit of vanilla RNN is shown in Figure 4 [3]. At each time step, the input consists of the data of the current time step and the hidden state of the previous time step [3], [10]. This current time step input and previous time step input is multiplied with a weight and given this to a tanh function and get the output of the current time step. The hidden state at time steps can be calculated in equation (1) [3]:

$$h_s = \tanh\left(W * [h_{s-1}, x_s] + b\right) \quad (1)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2)$$

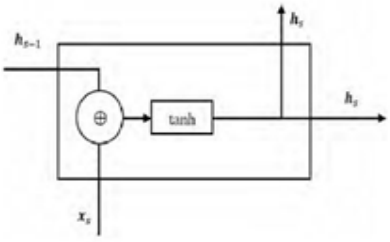

Figure 4: The structure of vanilla RNN unit.

Where h_{s-1} the hidden state of previous time step is, h_s ishidden state of current time step, W and b are the weights and biases, respectively. For a

regression problem such as air quality prediction, the final hidden state can be used to obtain the final prediction by [3], [10]:

$$p = W_p * h_s + b_p \quad (3)$$

Where p is the final prediction, h_s is the final hidden state of RNN. W_p and b_p are weights and biases.

3.2 Prediction of Meteorological Data Using LSTM

Here temperature, relative humidity, winds speed and direction data are used. So the prediction of meteorological factors like temperature, relative humidity, wind speed and direction are takes place here using LSTM. For data pre-processing, first step is, removing columns that is not need, replace the missing value with a quartile value by finding the mean of each quartile values. When a missing value comes in position between the quartiles replace with its mean. Here also scaling is performed because, LSTM model works best on values in 0-1 range.The LSTM architecture is shown in figure 5. The structure of LSTM unit is shown in Figure 6 [3], [9]. LSTM [9] is comprised of cells, which memorize the information. It consists of:

- Forget gate: Current input and previous output are multiplied with weight followed by the addition of bias, which is given to a sigmoid function and get a value between 0 and 1. If the output is 0, the piece of information is forgotten else keep.

- Input gate: It decide what information are going to store in the cell state. First we pass current input and previous output to a sigmoid function it gives a value between 0 and 1. After that, the current input and previous output is given to a tanh function and gets a value between -1 and 1. Then both are multiplied and decide which information is to be keep in the cell.

- Output gate: First a vector is generated by applying tanh function on the cell. At last, the values of vector and the regulated values are multiplied to be sent as an output and input to the next cell.

After computing output of LSTM [3] are in the range of (0, 1) if it is close to 0 almost nothing enters the cell. If it is 1

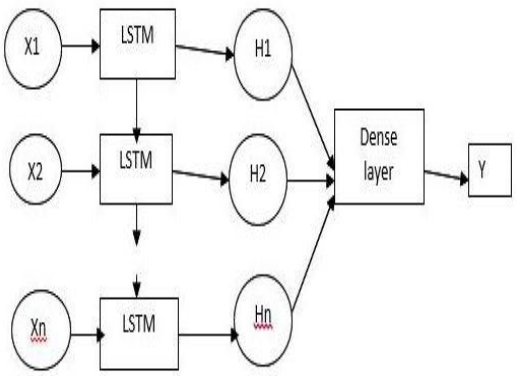

Figure 5: Long Short Term Memory.

The hidden state of s^{th} time step can be computed as [3]:

$$i_s = \sigma(W_i*[h_{s-1}, x_s]+b_i) \quad (4)$$

$$f_s = \sigma(W_f*[h_{s-1}, x_s]+b_f) \quad (5)$$

$$o_s = \sigma(W_o*[h_{s-1},x_s]+ b_o) \quad (6)$$

$$\tilde{C}_s = tanh(W_c*[h_{s-1}, x_s]+b_c) \quad (7)$$

$$C_s = f_s*C_{s-1} + i_s * \tilde{C}_s) \quad (8)$$

$$h_s = tanh(C_s)* o_s \quad (9)$$

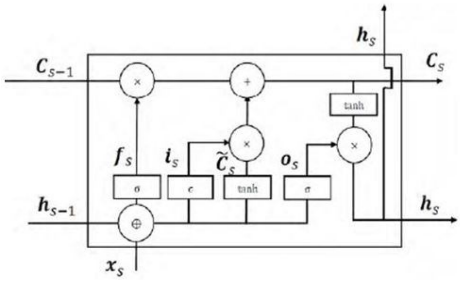

Figure 6: The structure of LSTM unit.

Where i_s is the input gate, f_s is the forget gate, and $o_s$is the output gate. $\tilde{C}_s$ is a candidate cell value that represents new information, $C_s$represents the cell value, $h_s$is the current hidden state, $h_{s-1}$is the previous hidden state. $x_s$is the data of s^{th} time step. $C_{s-1}$is the cell value of the previous time step, and it represents old information. W and b are weights and biases respectively.

3.3 A Deep Learning Prediction Model Combining Multiple Inputs

The output (Predicted value) from the LSTM and RNN is concatenate and give to a fully connected layer to improve the accuracy prediction. Fully Connected network is shown in figure 7.

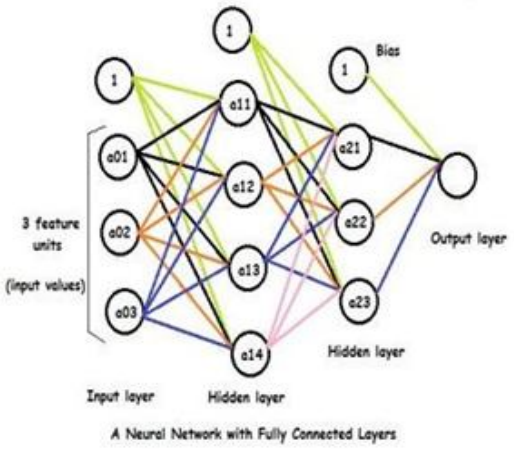

Figure 7: Fully Connected Network.

Fully Connected [17] layers in a neural network are those layers where all the inputs from one layer are connected to every activation unit of the next layer.

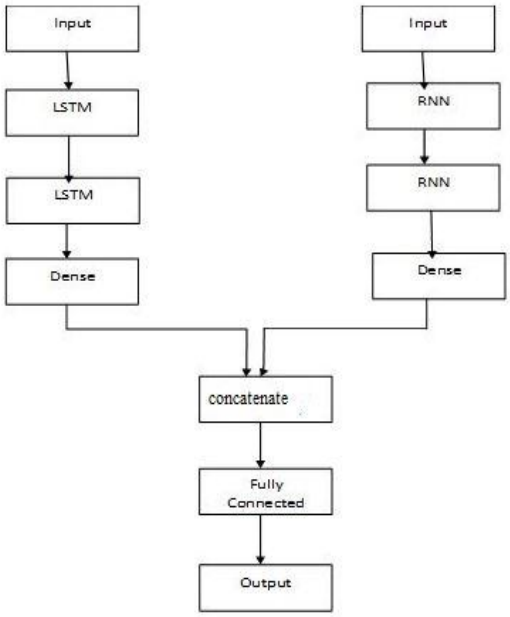

Figure 8: Functional model for prediction by combining multiple inputs.

In most of the machine learning models [17], [22] the last few layers are full connected layers, which compile the data extracted by previous layers to form the final output. Let x be the input to a fully connected layer. Let y_i be the i^{th} outputfrom the fully connected layer. Then y_i is computed as:

$$y_i = \sigma(W_1 * x_1 + \cdots + W_m * x_m) \quad (10)$$

The functional model is shown is figure 8. It is the combination of multiple deep learning model that is, the output of two or more models is combined and given as an input to another model. Here we take LSTM and RNN as two deep learning models.

4. EXPERIMENTS AND RESULTS

The experiment was setup in Ubuntu version 18.04 and python 3.5with tensorflow 1.8 and with the help of Keras the neural networkcan be build.

The model used for the prediction of air pollutant concentration is designed using deep learning methods like Recurrent Neural network (RNN), Long Short Term Memory (LSTM), and Fully Connected Network (FC). A functional model is built by concatenating the output of RNN and LSTM, and then gives it to a fully connected layer for more accurate prediction. The real time data of pollutants is downloaded from the Central Pollution Control Board (CPCB) website [7] [23] of every day from 2016 to 2018. The data is normalized using Min-Max Scalar function in python. The input to the Recurrent Neural Network is a time series data, the values of air pollutants like SO_2, NO_2, CO, PM2.5, and O_3. The day to day input is given and predicts the next day value. Similarly the input to LSTM is time series data of temperature and relative humidity and predicts the next day temperature and relative humidity. The output from RNN and LSTM is concatenated and given to a fully connected layer for more accurate prediction.

Air Quality Standards [7] are the standards for ambient air quality which is by the Central Pollution Control Board (CPCB) in India. The major air pollutants Indian Standards are described in the figure 1.

The whole dataset is divided into train data and test data. The deep learning model will learn with the help of train data and error is evaluated using Root Mean Square Error (RMSE). RMSE is the difference between air pollutant values predicted by a model or estimated value and the value actually observed. The model is trained on past data.

The dataset contains the values of air pollutants like SO_2, NO_2, CO, PM2.5, and O_3, wind speed, wind direction, temp, and relative humidity of cities Agra, Delhi, Lucknow etc. of the year 2016 to 2018.

The data is preprocessed by removing columns that is not need, check missing values if there is any missing value (0 or None) then replace it with a quartile value. Quartiles [16] in statistics are values that divide your data into four equal parts. The first quartile is the number below which lies the 25 percent data. The second quartile is the middle value. The third quartile has 75 percent of the data. And find the mean of each quartile values then replace zeros with mean of quartiles, which zeros comes in position between the quartiles. MinMax Scaler [5], [20] of sklearn library is used to convert whole dataset into float data type. In scaling, the value varies between 0 and 1.

	t	t+1	t+2	t+3	t+4	t+5
0	0.143645	0.143645	0.135800	0.159541	0.140493	0.112950
1	0.143645	0.135800	0.159541	0.140493	0.112950	0.133402
2	0.135800	0.159541	0.140493	0.112950	0.133402	0.132100
3	0.159541	0.140493	0.112950	0.133402	0.132100	0.106269
4	0.140493	0.112950	0.133402	0.132100	0.106269	0.123535
5	0.112950	0.133402	0.132100	0.106269	0.123535	0.143337
6	0.133402	0.132100	0.106269	0.123535	0.143337	0.103494
7	0.132100	0.106269	0.123535	0.143337	0.103494	0.110312
8	0.106269	0.123535	0.143337	0.103494	0.110312	0.086228
9	0.123535	0.143337	0.103494	0.110312	0.086228	0.084550
10	0.143337	0.103494	0.110312	0.086228	0.084550	0.104796
11	0.103494	0.110312	0.086228	0.084550	0.104796	0.117712
12	0.110312	0.086228	0.084550	0.104796	0.117712	0.123741
13	0.086228	0.084550	0.104796	0.117712	0.123741	0.079993

Figure 9: After applying shift function.

After that convert the data into data frame by, applying shift operation in python which is shown in figure 9. A Data frame [18], [23] is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas [18] Data Frame consists of three principal components, the data, rows, and columns. When dealing with time-series data, the shift method is used to shift values in a column up or down. After the shifting operation the output is shown in figure 9.

5. CONCLUSION

The model for air quality prediction, proposes a functional model for more accurate prediction. This model is the concatenation of RNN and LSTM outputs. With the help of RNN and LSTM we can predict the pollutants concentration correctly.

REFERENCES

1. V. Athira, P. Geetha, R. Vinayakumar, and K. P. Soman. **DeepAirNet: Applying recurrent networks for air quality prediction**, International Conference on Computational Intelligence and Data Science, Dec. 2018.
2. Srinivasa Rao Kurapati, Lavanya Devi G and RameshNeelapu. **Bidirectional Long Short Term Memory based Recurrent Neural Networks for Air Quality Prediction: Case of Visakhapatnam,** International Journal of Scientific Research and Reviews Jan 2019.
3. Bo Liu, Shuo Yan, Jianqiang Li, GuangzhiQu, Yong Li1,Jianlei Lang, And RentaoGu. **A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction**, IEEE Access, March 2019. https://doi.org/10.1109/ACCESS.2019.2908081
4. Qing Tao, Fang Liu 1, Yong Li , And Denis Sidorov. **Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU**, IEEE Transaction, April, 2019.
5. V. Chaudhary, A. Deshbhratar, V. Kumar, and D. Paul. **Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India**, Samsung Research Institute India, Aug 2019.
6. J. Zhao, F. Deng, Y. Cai, and J. Chen. **Long short-term memory-fully connected (LSTM-FC) neural network for PM2.5 concentration prediction,** Chemosphere, Apr. 2019. https://doi.org/10.1162/neco.1997.9.8.1735
7. Indian government, Central Pollution Control Board (CPCB), https://data.gov.in/catalog/ historical-daily-ambient-air-quality-data.
8. IndiaToday.in. 2018. 14 of world's most polluted 15 citiesin India, Kanpur tops WHO list. India Today (2018), https://www.indiatoday.in/ education-today/gk-current-affairs/story/-14-wo rlds-most-polluted-15-cities-india-kanpurtops- wholist-1224730-2018-0502.
9. Sepp Hochreiter and J¨urgen Schmidhuber. Long Short Term Memory, https://doi.org/10.1162/neco.1997NeuralComputation, Nov, 1997.
10. Z. C. Lipton, J. Berkowitz, and C. Elkan. **A critical review of recurrent neural networks for sequence learning**, https://arxiv-.org/-abs/1506.00019, 2015.
11. India largest SO2 emitter in the World says Green peace's new analysis, https://www.green-peace.org/india/en/press/4015/india-largest-so2-emitter-in-the-world-says-green-peaces-new-analysis/.
12. India has three of the world's 50 nitrogen Emission hotspots https://www.down-to-earth.org.in/news/air/india-has-three-of-the-world-s-50-nitrogen-emission-hotspots-6198.
13. **7** Air Pollutants Commonly Found in Urban Atmosphere of India http: //www.you-rarticlelibrary.com/air-pollution/7-airpollutant-commonly-found-in-urban-atmosphere-ofindia-/19768.
14. Characteristics of the Ozone pollutionand its health effects in India https://www.teriin.org-/researchpaper/characteristics-ozone pollution-and-its-health-effects-india.
15. S. Bull, O. Gaemperli, and P. Kaufmann. **Impact of a new motion-correction algorithm on image quality of low-dose coronary CT angiography in patients with insufficient heart rate control**, Acad. Radiol., vol. 21, pp. 312–317, 2014.
16. What are Quartiles? https://www.statistics-how-to.data-science-central.com/what-are-quartiles/.

17. Fully Connected Layer: The brute forcelayer of a Machine Learning model, https://iq.open-genus.org/fully-connected-layer/.

18. Python: Pandas Data Frame, https://www.geeksforgeeks.org/python-pandas-data-frame/.

19. Air quality index, https://en.wiki-pedia.org/wiki/AirqualityindexIndia.

20. Fabian Pedregosa, Ga¨el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. **Scikit-learn: Machine Learning in Python**. J. Mach. Learn. Res. 12 Nov. 2011, 2825–2830. http://dl.acm.org/citation.cfm?id1953048.2078195.

21. Thomas M, Chooralil DVS. **Security and privacy via optimized block chain** International Journal of advanced trends in computer science and engineering. 2019; 8(3):415-418. https://doi.org/10.30534/ijatcse/2019/14832019

22. Jisna Jaison, Chooralil DVS, John J. **Efficient Route Recommendation System Based On Keyword Using Candidate Route Generation And Travel Route Exploration**, International Journal of Advanced Trends in Computer Science and Engineering. 2019; 8(3):419-425. https://doi.org/10.30534/ijatcse/2019/15832019

23. Vijayan, Vinodh, Biju Paul. **Traffic scheduling for Green city through energy efficient Wireless sensor Networks,** International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.4, July – August 2019, ISSN 2278-3091