International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Air Quality Index Prediction using LSTM

Nithyashree K R[1], S Bhumika[2], Sahana R[3], Ranjitha V[4]

[1,2,3,4]*Department of Information Science and Engineering, Sri Venkateshwara College of Engineering, Bangalore-562157, Karnataka, India*

Abstract - *Air Pollution is one of the major problems across all countries in the world. Over several years, countries are finding various means of battling against air pollution and reduce the number of pollutants in air. Air pollution can cause several types of respiratory problems, heart diseases and lung cancer. As the saying goes "Prevention is better than cure", it is always better to know the level of pollution at the earliest and take preventive measures. Air Quality Index is the measure of air pollution. This paper provides machine learning model to predict the air quality index for the next one hour based on the major pollutants like: SO2, NO2, O3, CO and environmental factors such as temperature, pressure, rainfall, wind speed per minute and wind direction. This paper provides the model that predicts the amount of any pollutant (PM 2.5, PM10, SO2, NO2, CO, O3)in air for the next 1 hour.*

***Key Words*: AQI (Air Quality Index), LSTM (Long short term memory), Recurrent neural networks, RMSE.**

1. INTRODUCTION

Air Quality Index is the measure of how much the air is polluted in a particular region. This can be determined by the amount of fine particles known as particulate matter 2.5 (PM 2.5) present in air. Major pollutants of air are Sulphur dioxide (SO2), nitrogen dioxide (N02), carbon monoxide (CO) and ozone(O3). Other environmental factors such as temperature, pressure, rain, wind direction, wind speed and wind direction as well play a vital role in polluting air.

Machine learning and deep learning concepts have proven to be powerful and most useful in making complex decisions and predictive analysis. Artificial Neural networks have proven to be the best among various machine learning algorithms for predictive analysis. However, normal feedforward neural networks suffer from vanishing gradient or exploding gradient problems. Various gradient algorithms have been introduced to address these problems. But, none of them completely eliminate the problem. On the other hand, each neuron in a feedforward network has independent calculations of weights and bias. This cannot be used for sequential prediction where the output depends on the previous input data. Also, the neurons must be capable of storing the previous state input data to predict the next state output. Recurrent neural networks are thus used for sequential data and solve the problem caused due to independence of neurons in the feedforward neural networks

1.1 Recurrent neural networks

Recurrent neural networks are a form of neural networks in which the output of one layer of neurons is fed as an input to the next layer. Each neuron in RNN is associated with a memory. Recurrent neural networks are widely used for problems including sequential data or time series data. However, though the RNNs are used for sequential analysis, they still suffer from vanishing gradient and exploding gradient problems. LSTMs are a class of recurrent neural networks used for sequential prediction and to address the vanishing gradient problem in RNN.

A) LSTM: Long short term memory is one of the powerful RNN used for prediction including sequential data. Problems such as stock market prediction, speech recognition, character recognition are being addressed by LSTMs. Since LSTMs are widely used for sequential analysis, they can be trained to predict air quality index levels for the next hour or even next month by the historical data obtained through sensors at various weather stations. The proposed system was trained with the dataset at one of the weather stations at the Beijing City, China. The basic operation of each unit in LSTM involves three gates:

a) Input gate: This gate serves as the entry point to the LSTM. Inputs from the dataset are fed to the network via this gate.

b) Forget gate: This gate decides what data should be stored or is important and what data is to be discarded or forgotten by the network. Several activation functions are used to decide upon the data to be stored namely sigmoid function, relu or tanh.

c) Output gate: Not all the data that is stored in LSTM can be the output. This gate selects the appropriate output of the unit and produces it to the next unit.

2. METHODOLOGY

The proposed system is divided into the following modules to predict air quality index.

a) Data cleansing: For any machine learning or deep learning model, it is very important to train the model with appropriate dataset that includes almost no null or NaN values. This module involves removing null and NaN values from the raw dataset, preparing it for processing.

b) Data preprocessing: String values as such for wind direction are encoded using suitable encoder.

c) Data visualization: Various pollutants can be visualized using a graph that provides insights about the rise and fall of their concentration in air. A graph for each of the pollutants is plotted with x axis representing number of samples and y axis representing concentration in µg/m^3

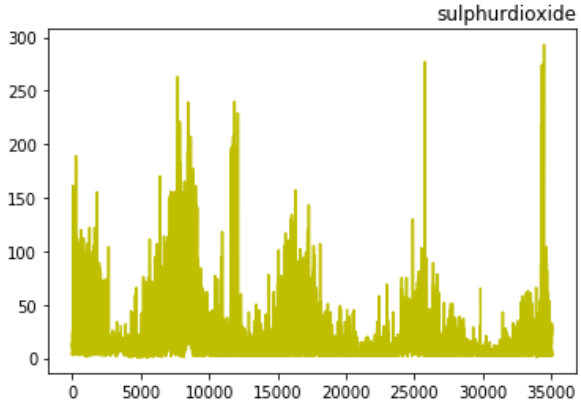

Fig 1. SO2 concentration in air

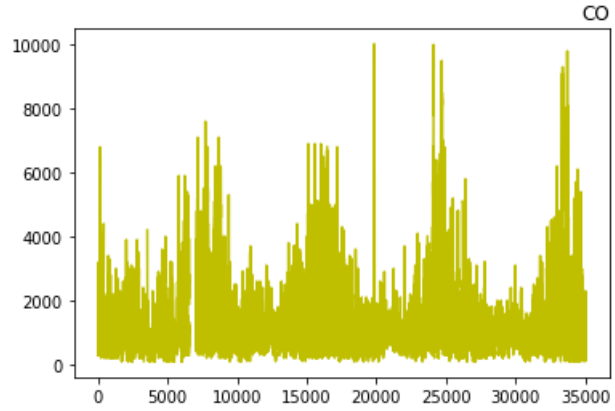

Fig 2. CO concentration

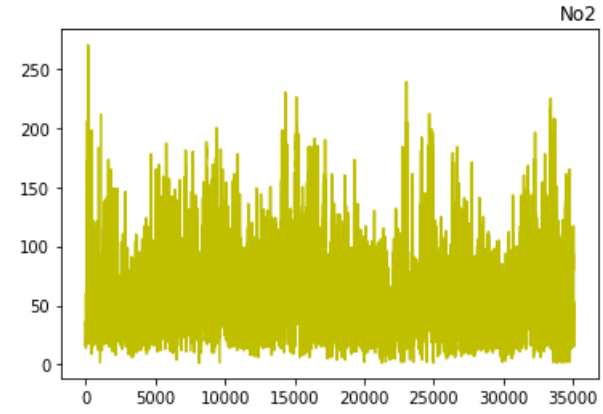

Fig 3. NO2 concentration

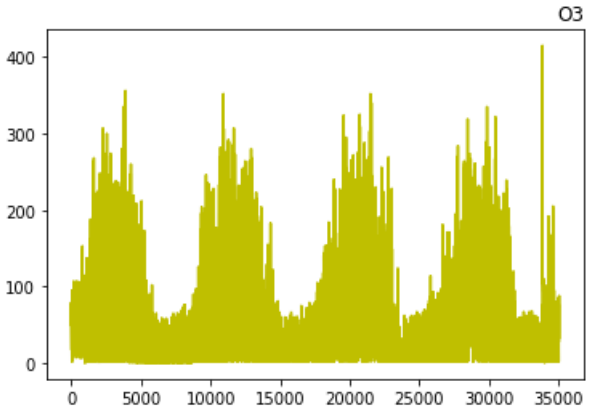

Fig 4. O3 concentration

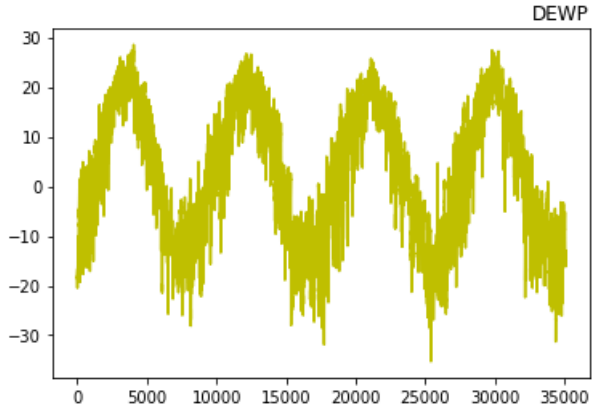

Fig 5. Dew point concentration

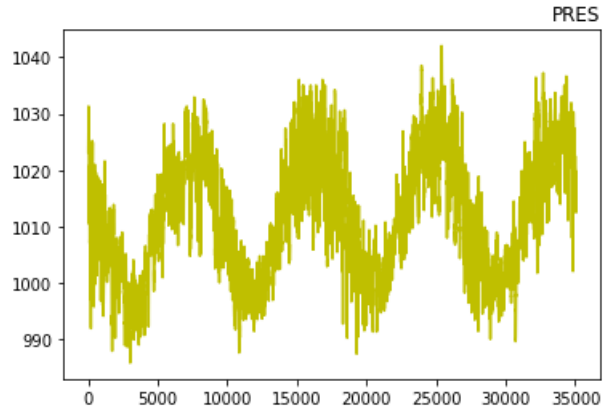

Fig 6. Pressure

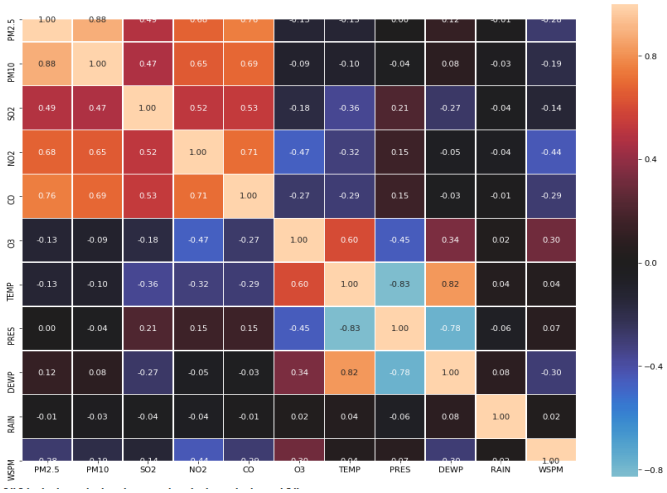

Fig 7. Correlation heatmap representing correlations between various pollutants.

d) Series to supervised conversion: Since LSTM is a supervised learning algorithm, labelled input and output data is to be fed to the model. Thus, the series data is converted to supervised data before fitting to the LSTM model.

e) Model fitting: The dataset is split into training and testing data. The preprocessed dataset is then fed to the model prior to which network parameters are to be set. The important parameter to be set for a neural network is an optimizer. An optimizer is a method or algorithms which is used to set various attributes of neural networks such as weights, bias, learning rates etc. There are various optimizers for neural networks and the choice depends on the problems addressed by each of them. Since various pollutants are used to predict the PM 2.5 concentration in air, it is important to provide different learning rates for each of these features. This purpose is served by an optimizer called Adam.

f) Error calculation: Mean square error (RMSE) is calculated for the model to evaluate the accuracy of prediction. The proposed model produced an error of approximately 0.1.

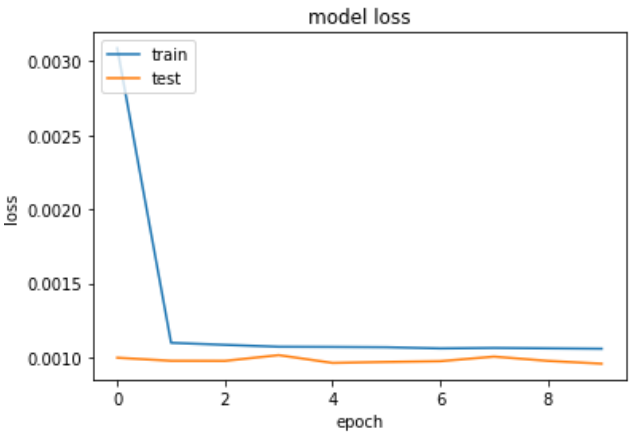

Fig 8. MSE error plot

```
PM10 concentration for next hour would be: 27.149012
Mean square error 0.0009564670581444191
Train: 0.00004, Test: 0.00017
```

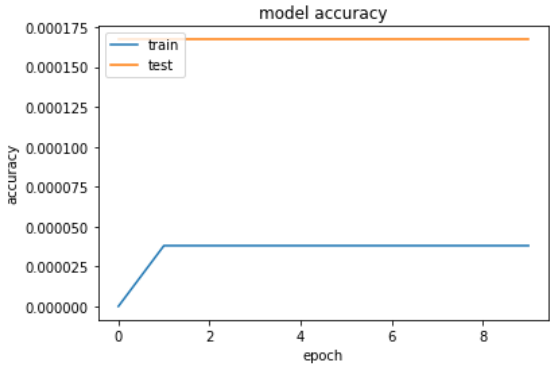

Fig 9. Model accuracy plot

g) Classification: Based on the output from the model it is possible to determine whether the air is going to be rarely, mediumly or highly polluted for the next hour. This classification purely depends on the ranges in the particular region.

2.1 Flowchart

The flowchart representing the model is shown in Fig 10.

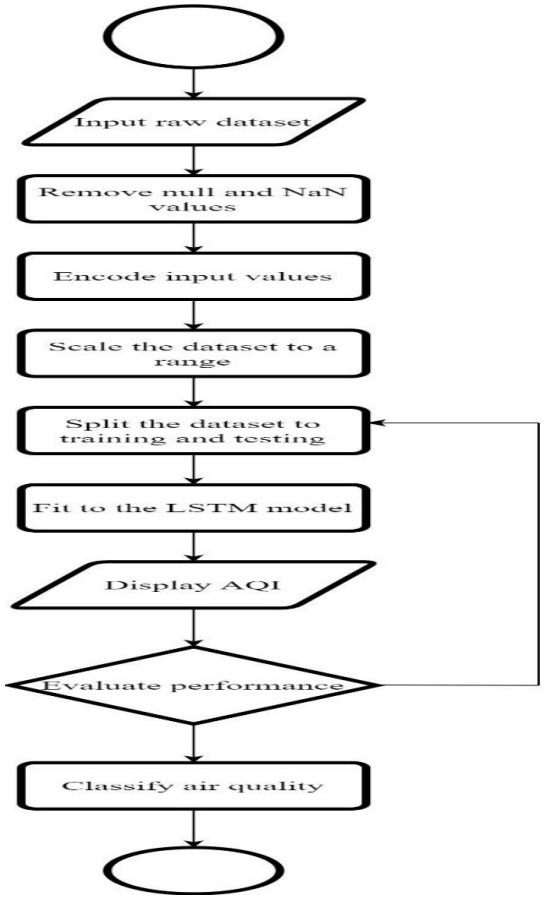

Fig 10. Flowchart representing operations performed.

3. CONCLUSION

This paper provided the efficient method of determining concentration values of various pollutants causing pollution. The proposed model can be used by various weather station and user applications that can predict the pollutant level instantly.

REFERENCES

[1] Temesegan Walelign Ayele, Rutvik Mehta, "Air pollution monitoring and prediction using IoT", IEEE Conference 2018.

[2] Nadjet_Djebbri, Mounira_Rouainia,"Artificial Neural Networks Based Air Pollution Monitoring in Industrial Sites", IEEE Conference 2017.

[3] Xia Xi, Zhao Wei, Rui Xiaoguang, Wang Yijie, Bai Xinxin, Yin Wenjun, Don Jin, "A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method", IEEE Conference 2015.

[4] Peijiang Zhao, Koji Zettsu," Convolution Recurrent Neural Networks Based Dynamic Transboundary Air Pollution Prediction",IEEE Conference 2019