International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

A Deep Learning based Air Quality Prediction

B. Lakshmi Sravya [1], A.S. MahaLakshmi [2], D.Balaji Bhavya Swarupini[3], B.V. Sai Jaswanth [4]

[1,2,3,4] *Lendi Institute of Engineering & Technology, Jonnada, Vizianagaram, Andhra Pradesh*

Abstract - *Industries are the major means of air pollutants. Air pollution in the form of carbon dioxide and methane raises the earth's temperature, the less gasoline we burn, the better we do to reduce air pollution and harmful effects of climate change. Especially at metropolitan cities, the change in the temperature combined with harmful chemicals may lead to dangerous signs of air pollution. Quality of air prediction techniques has a major importance in the current learning world. Many machine learning algorithms done a lot of research in identifying the air quality index. Applying deep learning models on these data can show great difference in predicting the quality of air. We proposed an LSTM based deep learning technique in evaluating hourly based encompassing air quality. The proposed results outperformed the existing model results through predicting RMSE value.*

Key Words: Pollution, Prediction, LSTM, Deep Learning.

1. INTRODUCTION

Industrialization and urbanization have intensified environmental health risks and pollution, especially in developing countries like India. Study shows that air pollution poses a major health risks such as stroke, heart disease, lung cancer, and chronic and acute respiratory diseases. According to the World Health Organization (WHO) report [1], 14 out of the top 15 most polluted cities in the world are in India (in which Delhi is among the top list), an estimated 12.6 million people die from environmental health risks annually. According to the WHO, 92% of the world's population lives in areas where the air quality is below the WHO standards [2].

About 88% of premature deaths occur in the low and middle-income countries, where air pollution is escalating at an alarming rate. India is the third largest producer of greenhouse gases after China and the United States [3]. The severity of air pollution is so much that as per 2016 study conducted by the Indian Institute of Tropical Meteorology (IITM) and Atmospheric Chemistry. Observations and Modelling Laboratory, National Centre for Atmospheric Research, Boulder, Colorado, USA [4], life expectancy among Indians reduces by 3.4 years on an average while among the residents of Delhi it reduces by almost 6.3 years.

There are 6 prominent air pollutants present in the air, Particulate Matter (PM2.5 and PM10), Carbon Monoxide (CO), Ozone (O3), Nitrogen dioxides (NO2), Sulphur dioxide (SO2). Table 1 shows the sources of air pollutants and their major effects on human health and environment [5]. To track the rising pollution trend in India, the government of India has installed pollutant's measuring sensors at various stations covering major pollution prone areas. Multiple steps have been taken by the government to control pollution such as metro facility, increase in public transport, and laws such as even-odd system for personal vehicles. Considering the current trend of pollution growth, these solutions are bound to fail in future. Therefore, air pollution forecasting and generating solutions to control it are today's need.

1.1 CRITERIA POLLUTANTS

Table -1: Emission Sources and Major Effects

Criteria pollutants	Emission sources		Major effects	
	Natural sources	Anthropogenic sources	Health effects	Environmental effects
Sulphur Dioxide (SO2)	Volcanic emissions	Burning of fossil fuels, metal melting etc.	Respiratory problems, heart and lung disorders, visual impairment	Acid rain
Nitrogen dioxide (NO2)	Lightning, forest fires etc.	Burning of fossil fuels, biomass & high temperature	Pulmonary disorders, increased susceptibility to respiratory infections	Precursor of ozone formation in troposphere, aerosol formation.

		combustion process		
Particulate matter (PM)	Windblown dust, pollen spores, photochemically produced particles	Vehicular emissions, industrial combustion processes, construction industries	Respiratory problems, liver fibrosis, lung/liver cancer, heart stroke, bone problems	Visibility reduction
Carbon monoxide (CO)	Animal metabolism, forest fires, volcanic activity	Burning of carbonaceous fuels, emission from IC engines	Anoxemia leading to various cardiovascular problems. infants, pregnant women and elderly people are at higher risk.	Effects the amount of greenhouse gases which are linked to climate change and global warming.
Ozone (O3)	Present in stratosphere at 10-50 km height	Hydrocarbons and NOx upon reacting with sunlight results in (O3) formation.	Respiratory problems, asthma, bronchitis etc.	O3 in upper troposphere causes green house effects, harmful effects on plants, death of plant tissues.

1.2 OBJECTIVE

Figuring out these problems in the environment and applying techniques to improve the efficiency and predicting the air index using different learning problems.

2. RELATED WORK

Reddy et al. [6] investigate the use of LSTM framework for forecasting pollution in future based on time series pollutant and meteorological data in Beijing area. The main aim of this paper is the application of LSTM sequence to scalar model to forecast pollution. Zheng et al. [7] address the issue of air quality inference based on air quality reported by existing sensor stations. Meteorological data, traffic flow, human mobility, point of interests (POIs) are other features used to infer AQI at non-sensor locations.

He et al. [7], Roy et al. [8] provide a method to predict PM10 concentration. Pérez et al. [9] proposed a method to predict PM2.5 concentration for the next 24 hours. In our work, we are providing a method to predict each pollutant concentration and AQI up to next 12 hours.

Roy et al. [8] used Mill tailings at Kolar Gold Fields data for their experiments. Monitoring was carried out at the National Institute of Rock Mechanics (NIRM), Kolar Gold Fields (KGF). Pérez et al. [9] performed their experiments for Malaysia. The data was provided by Malaysian Meteorological Department (MMD) and Department of Environment (DOE).

He et al. [10] develop a hybrid methodology to forecast PM10. The paper combines both Autoregressive Integrated Moving Average (ARIMA) and ANN models to improve forecast accuracy. The paper used ARIMA to model the linear component and then ANN model is used to take care of the residuals from ARIMA model. They report that hybrid model can be a effective way to improve PM10 forecasting accuracy compared to single ARIMA model. Roy et al. [8] present an ANN based approach as predictive and data analysis tool for the evaluation of air pollutant. The paper proposes a multilayer feed forward network to predict PM10 concentration using meteorological data. Pérez et al. [9] proposed a three layer neural network to predict PM2.5

concentration. They used previous 24 hours PM2.5 data for prediction. In this work, we have developed a time series based stacked LSTM model to forecast air pollutant concentration.

2.1 RESULTS AND DISCUSSIONS

The main aim of the study is to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city. Deep learning is employed to predict future values of PM2.5 based on the previous PM2.5 readings. This can be done by using Long Short-Term Memory (LSTM) which is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. The output is the result of calculated Root Mean Square Error (RMSE). Low RMSE value indicates that the model has accurate results.

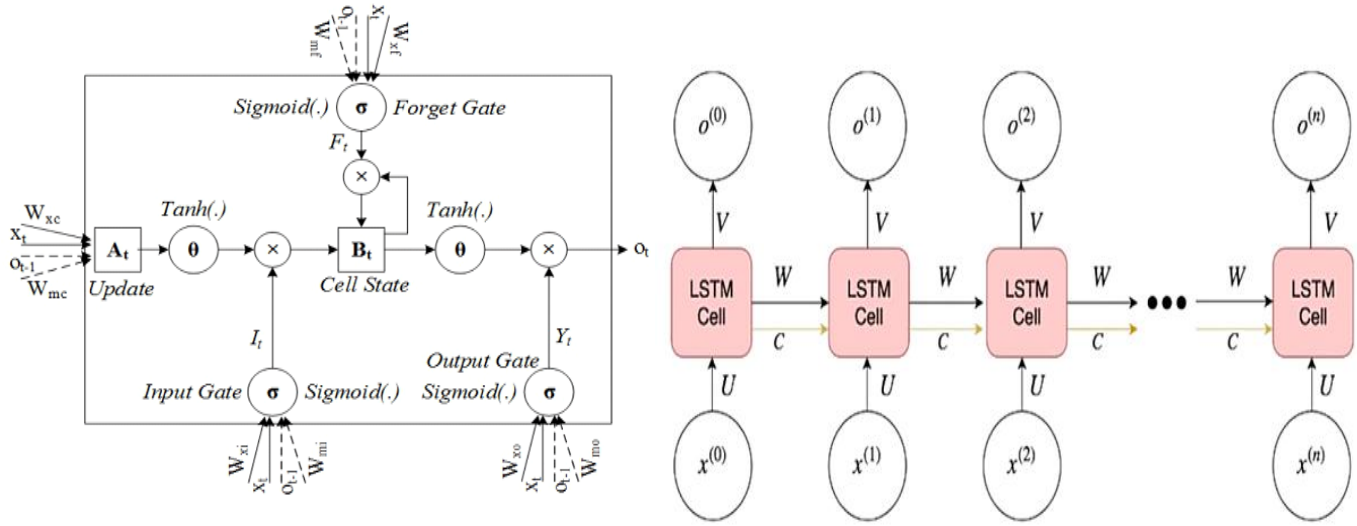


Fig. 2.1 LSTM and Nerve Cell

2.2 METHODOLOGY

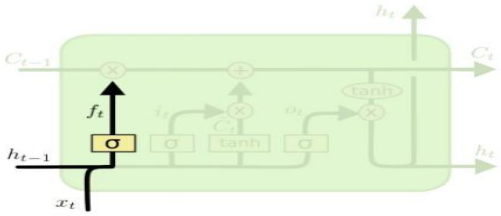
Step 1: The data is taken in the form of csv file. (data.csv)
Step 2: After the input dataset is given, the data will be preprocessed by

- Removing Null values from a data frame and replace NaN values with default values.
- Sometimes our data will be qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.
- Then it fit the model to the data, then transform the data according to the fitted model.

Step 3: After the preprocessing, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers. Then using s_to_super function the first column of row(t) is shifted to last column of row(t-1) and concatenated. This act transforms a normal preprocessed dataset to recurrent dataset.
Step 4: Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e. our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import **test_train_split** from **model_selection** library of scikit.
Step 5 : Now to build our training and test sets, we will create 4 sets— X_train (training part of the matrix of features), X_test (test part of the matrix of features), Y_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices) , Y_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices). We will assign to them the test_train_split, which takes the parameters — arrays (X and Y), test_size.
Step 6: Now, we need to build a model to train the data. Here the model used is Long Short-Term Memory.

Step 7: An LSTM has a similar control flow as a recurrent neural network. It processes data passing on information as it propagates forward. The differences are the operations within the LSTM's cells. These operations are used to allow the LSTM to keep or forget information.

Step 8: The first step in LSTM is to decide what information you are going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer." It gives a value between 0 and 1, where a 1 represents "keep this as it is" while a 0 represents "get rid of this."

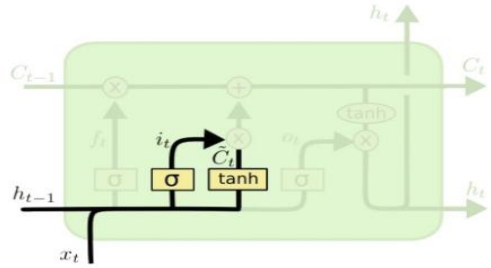


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Fig 2.2 Forget Gate

This step has two parts:

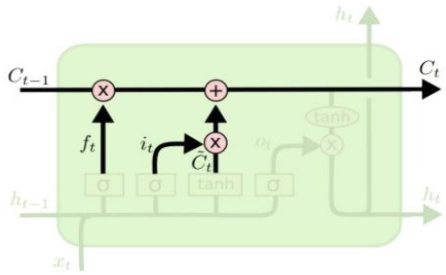
- First, a sigmoid layer called the "input gate layer" decides which values we'll update.
- Next, a tanh layer creates a vector of new candidate values that could be added to the state. In the next step, by combining these two layers, a new update is being created.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Fig 2.3 Input Gate

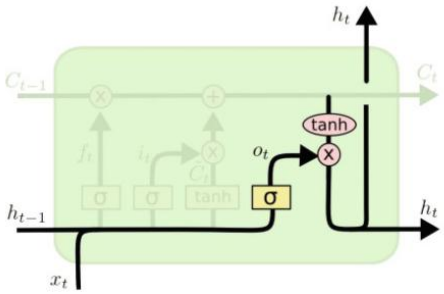
Step 9: It is now time to update the old cell state, Ct−1, into the new cell state Ct. The last step has already created an update. We only need to update it.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Fig 2.4 Current state

Step 10: Finally, we need to decide what we're going to output based on the context that we have selected.



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Fig 2.5 Output layer

Step 11: The prediction class is given to the model with the input data instances. With the help of those input instances the model predicts our required output. here the input instances are given from test_X data(test part of the matrix of features)

Step 12: To Predict a model we took model_predict ().
Step 13: And we calculate Root Mean Square Error (RMSE).It is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results. It is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:

It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. ... Lower values of RMSE indicate better fit.

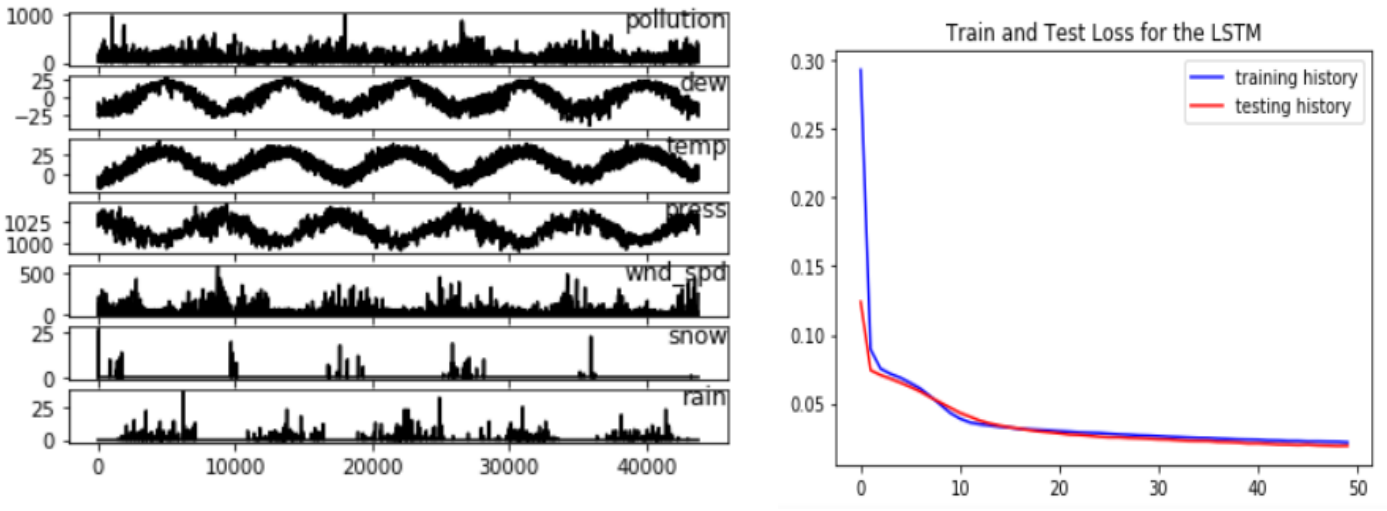


Fig. 2.6 Data Visualization

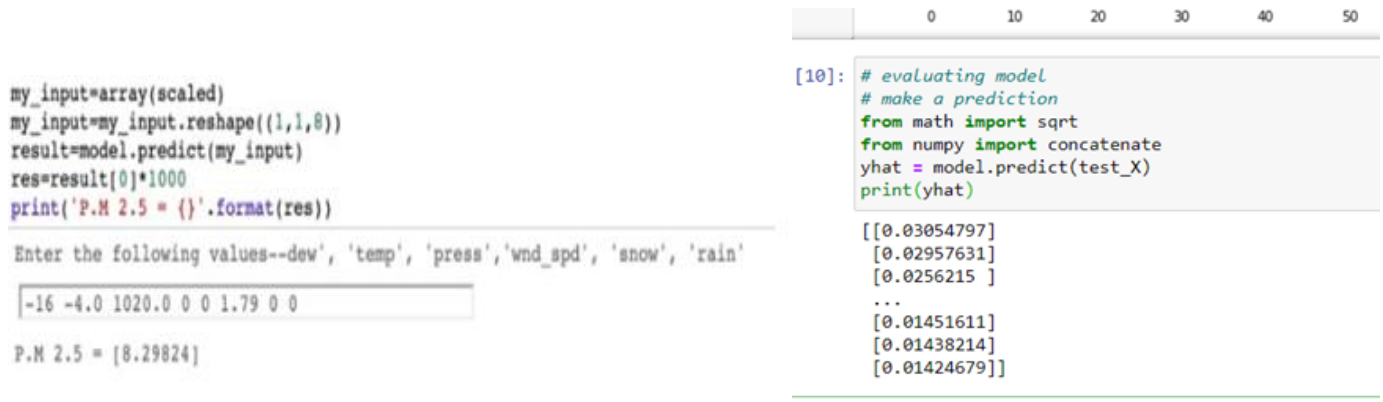


Fig. 2.7 Predicted RMSE and P.M.2.5 Values

3. CONCLUSION

Deep learning method is gradually developing as a promising technique for forecasting non-linear time series information like meteorological and pollution data. In this paper, we used different deep learning models for the prediction of air quality. Here we trained LSTM network on AirNet data to predict the future PM 2.5 and Calculated RMSE (Root Mean Square Error). The analysis can be further extended by utilizing methods like Convolution Neural Network (CNN) to catch the uneven changes happening in the air pollution data. The connection between different features can likewise be assessed hence enabling us to see whether there is any hidden parameter which will correlate the performance of features that appears to be different from the first peek.

REFERENCES

[1] IndiaToday.in. 2018. 14 of world's most polluted 15 cities in India, Kanpur tops WHO list. India Today (2018). https: //www.indiatoday.in/education-today/gk-current-affairs/story/ 14-worlds-most-polluted-15-cities-india-kanpur-tops-who-list-1224730-2018-05\-02

[2] weforum.org. 2016. 92 % of us are breathing unsafe air. This map shows just how bad the problem is. (2016). https://www.weforum.org/agenda/2016/09/ \92-of-the-world-s-population-lives-in-areas-with-unsafe-air-pollution-levels\ -this-interactive-map-shows-just-how-bad-the-problem-is/

[3] reuters.com. [n. d.]. India says is now third highest carbon emitter. ([n. d.]). https://www.reuters.com/article/us-india-climate/india-says-is-now-third-highest-carbon-emitter-idUSTRE6932PE20101004

[4] hindustantimes.com. 2016. Air pollution shortens your life by 3.4 years, Delhiites worst hit. Hindustan Times (2016). https://www.hindustantimes.com/mumbai/air-pollution-shortens-your-life-by-3-4-years/story-L9VOawHyX4PCMfCuAjv4ML.html

[5] Central pollution Control Board. [n. d.]. ENVIS Centre on Control of Pollution Water, Air and Noise. ([n. d.]). http://cpcbenvis.nic.in/envis_newsletter/Air%20pollution%20in%20Delhi.pdf

[6] Vikram Simha A Reddy, Pavan S. Yedavalli, Shrestha Mohanty, and Udit Nakhat. 2017. Deep Air: Forecasting Air Pollution in Beijing, China.

[7] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air: when urban air quality inference meets big data. In KDD.

[8] Surendra Roy. 2012. Prediction of Particulate Matter Concentrations Using Artificial Neural Network. 2 (03 2012), 30–36.

[9] Patricio Perez and Jorge Reyes. 2001. Prediction of Particlulate Air Pollution using Neural Techniques. Neural Computing & Applications 10, 2 (01 May 2001), 165–171. https://doi.org/10.1007/s005210170008

[10] G. He and Qihong Deng. 2012. A Hybrid ARIMA and Neural Network Model to Forecast Particulate Matter Concentration in Changsha, China.