

## Air quality predictions of Ulaanbaatar using machine learning approach

---

Otgonsuud Badrakh *a,\**, Lodoiravsal Choimaa *b*

*a* Institute of Mathematics and Digital Technology, Mongolian Academy of Sciences,  
54b, Ulaanbaatar, Mongolia

*b* National University of Mongolia  
54b, Ulaanbaatar, Mongolia

E-mail: [otgonsuud@mas.ac.mn](mailto:otgonsuud@mas.ac.mn)

Predicting and forecasting air quality is the one of the most essential activity in the Smart City. Recently, there are many study to use the machine learning approaches for evaluating and predicting air quality using big data. The aim of this study is to obtain machine learning model for air quality forecasting using previous air quality station data and the weather data. The air quality depends on multi-dimensional factors including location, time, weather parameters, such as temperature, humidity, wind direction and force, air pressure, etc. There are many machine learning approaches, but artificial neural Network model tries to simulate the structures and networks within human brain. It is convenient for working to find relation between multi parameters. If the neural network could determine the relation of the air quality using the weather and air quality data of last year, it is possible to predict approximately air quality of Ulaanbaatar city. We used features including parameters of temperature, humidity, wind direction, air pressure, PM2.5 and PM10, NO2, CO, SO2 and measuring time to build recurrent neural network model that is the class of artificial neural network. In this work we did machine learning test of neural network algorithm for the air quality prediction using LSTM /long short term memory/ model and discussed machine learning test results.

---

\*Speaker

*International Symposium on Grids & Clouds 2021, ISGC2021  
22-26 March 2021  
Academia Sinica, Taipei, Taiwan (online)*

POS ( ISGC2021 ) 012

## 1. Introduction

Air pollution is most prevalent in densely populated cities, especially in developing countries with poor environmental regulations. However, even in densely populated areas of developed countries, the level of pollution has reached dangerous levels. Air pollution is usually defined by the following six main indicators. Lead (Pb), Carbon Dioxide (CO), Sulfur Oxide (SO<sub>2</sub>), Nitric Oxide (NO) and Nitrogen Dioxide (N<sub>2</sub>O<sub>5</sub>), Ozone (O<sub>3</sub>), Particles (PM<sub>2.5</sub> and PM<sub>10</sub>). There are 16 air quality stations in Ulaanbaatar that measure air quality and weather information. The Institute of Meteorology provides weather forecasts, but does not provide air quality forecasts. However, machine learning, which has been the main focus of recent research, is able to predict the amount of air pollution, thus avoiding exposure to air pollution by preventing the effects of air pollution [1].

Machine learning is method of artificial intelligence, and it is possible to learn the best algorithms from the given data without having to write them into a program to perform any task using machine learning methods. This means that the program learns from previous experience as a human being and is able to work better. As learning data increases, algorithm performance improves. Any machine learning algorithm has the ability to determine patterns and relationships within data and predict unknown states [6].

The main purpose of this study is to train machines to predict the level of air quality indicators based on air quality station data using an LSTM model of machine learning algorithms [2].

### 1. Machine learning approach

Machine learning algorithms "learn" information directly from data, without predefined formulas or algorithms, based solely on computational methods. There are few main methods of machine learning:

1. Supervised learning is a method that uses both input and output data and is based on the classification and regression techniques. Classification is used to predict discrete values. Regression is used to predict continuous values. This is a training method when the training output data is given. The main idea of the supervised learning method is that we teach how to perform the task given to the machine. In other words, the machine learns how to complete the task by looking at the answers in the training data we enter.

2. Unsupervised learning is a method that uses only input data. This method is useful when you do not know exactly what to look for in the data. Often used to understand raw data. Most unsupervised learning is based on a technique called clustering analysis. Clustering analysis is the process of measuring the properties of data and grouping them accordingly. This is a way of extracting a structure from a data when the response to the data is unknown. In other words, the structure is derived by creating groups based on the relationships between the variables in the data.

3. Semi-supervised learning. This is a training method when some training data is not given. And some of data is given.

4. Reinforcement learning. Reinforcement learning is when a program improves itself by learning how the environment responds (rewards) its actions.

## 1. Methods

We have used these following methods and technologies.

*Programming language:* Python 3.7

*Applications:* Google CoLAB, Jupiter Notebook

*Libraries:* math, numpy, matplotlib, pandas, sklearn.preprocessing, keras

*Training data:* 2018 Ulaanbaatar air quality station data or air pressure, carbon dioxide, ozone, particles, relative humidity, sulfur dioxide, temperature, wind direction, wind speed with date, time and total of 8,760 rows and 10 columns of data were prepared and tested as training sets for one hour.

*Test data:* In 2018, the above data of Ulaanbaatar city air quality station was prepared as a test set of 8760 rows and 10 columns of data with one hour accuracy.

*Training model:* Long short term memory /LSTM /

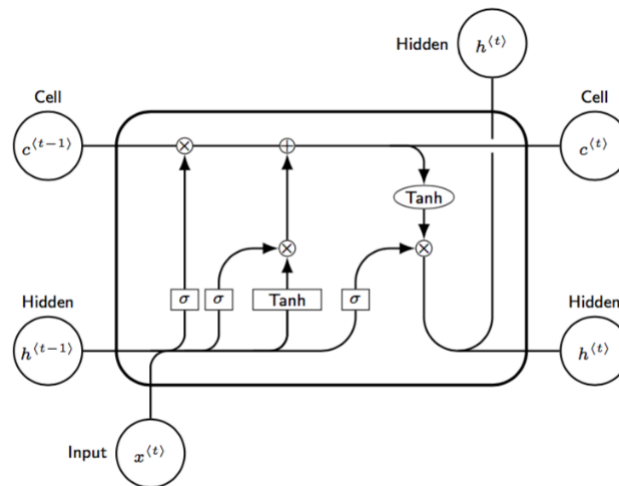


Figure 1. Principle of LSTM model

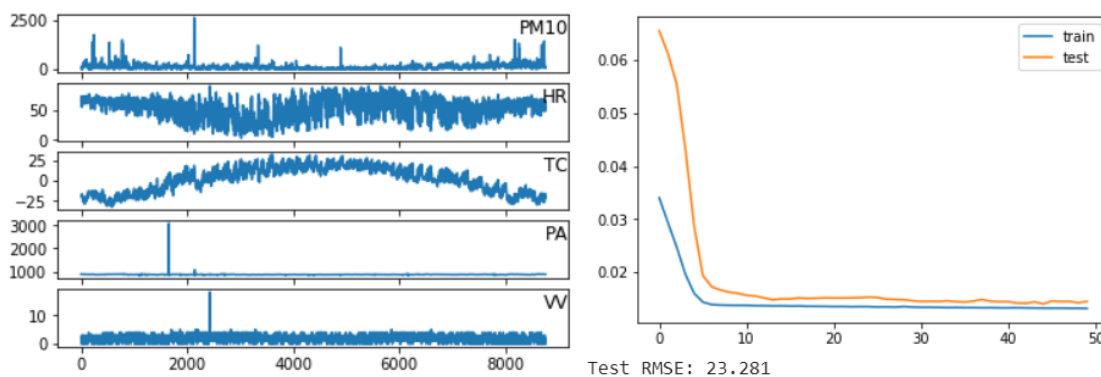
Long short term memory model is advanced recurrent neural network, which gives opportunity to save information. The LSTM consists of three parts, the first part chooses whether the data previous came is to be remembered or forgotten. In the second part, the cell tries to learn new data from the input. The last cell passes the updated information from the current time step to the next time step. These three parts of an LSTM cell are called as gates. The first one is called forget gate, the second one is called input gate and the third one is the output gate. LSTM also has a hidden state where  $h(t-1)$  represents the hidden state of the previous time step and  $ht$  is the hidden state of the current timestamp. In addition to that LSTM also have a cell state represented by  $C(t-1)$  and  $C(t)$  for previous and current time step [2],[7].

## 1. Results and discussion

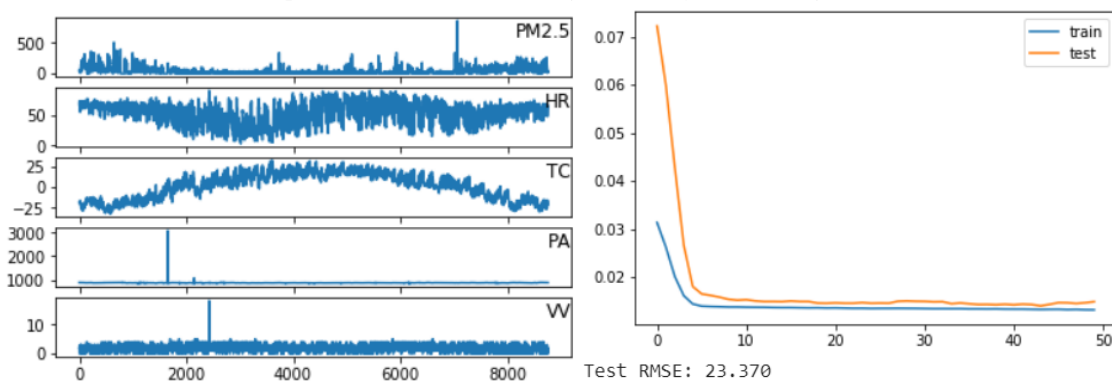
The main purpose of this study was to use an LSTM model to predict air pollution parameters, and to provide machine learning with a neural network structure with 7 input nodes and 1 output node.

In the above graphs, the results of the LSTM-controlled machine learning on each of the above air pollution parameters were shown in the following graphs using a Pyplot on the Jupiter Notebook [3], [4], [5]. The graph shows the 2018 weather data from the WFP, such as PA-air pressure, TC-air temperature, VV-wind direction, HR-relative humidity, CO-carbon oxide, NO<sub>2</sub>-nitrogen dioxide, NO-nitrogen oxide, O<sub>3</sub>-ozone, PM<sub>10</sub>-particulate matter, SO<sub>2</sub>-sulfur dioxide levels with each of the six air quality indicators, and the training and test set using these data to

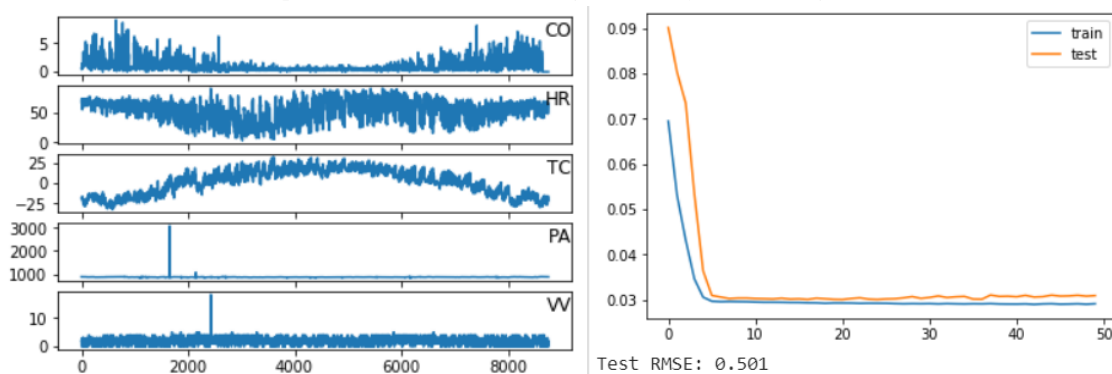
determine the RMSE / Root mean square error / shows the calculation. The training was conducted by fifty Epochs.



Graph 1. PM10- Prediction from one year data of 2018

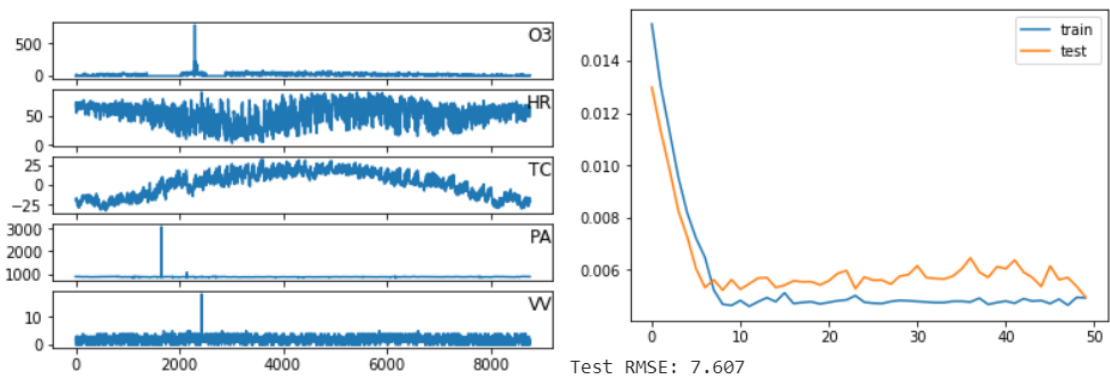
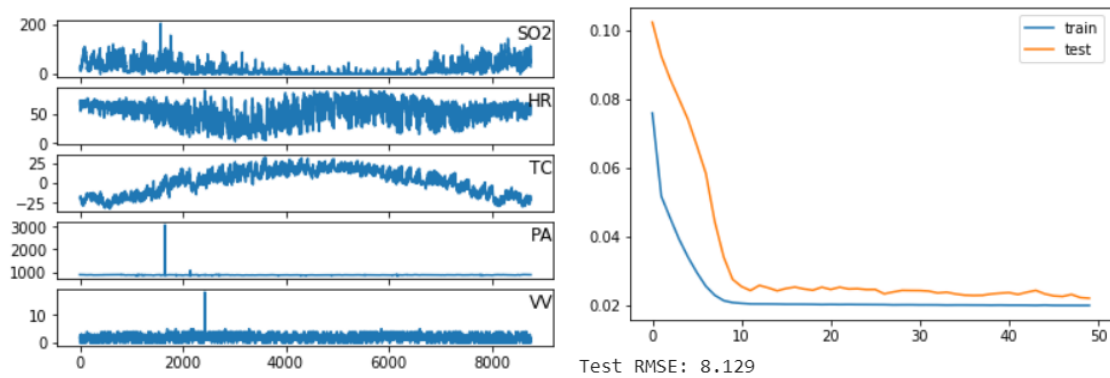
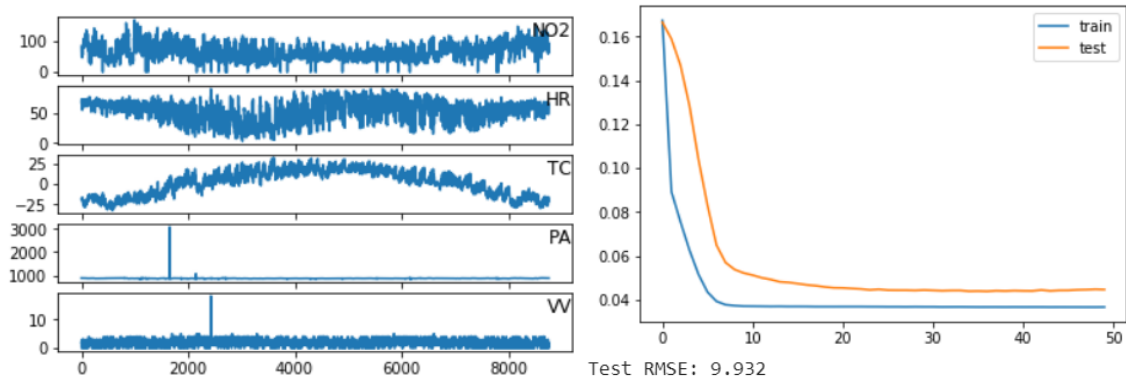


Graph 2. PM2.5- Prediction from one year data of 2018



Graph 3. CO- Prediction from one year data of 2018

POS (ISGC2021) 012



In the results, evaluating the RMSE of PM<sub>10</sub> and PM<sub>2.5</sub> were high value as others, SO was the lowest, and the RMSE of other three parameter showed medium value. However, this values are acceptable and if the training error is less than the test error, the training is considered successful and it means overfitting. In this article shows the results by 50 epochs, as the training is generally stable after 50.

## 1. Conclusion

In this study, we studied the machine learning and conducted machine learning using the LSTM model to predict air quality. Initially, we experimented with one-year data from one station. The experiments found the correlation between the weather data and each of the six air quality indicators, including PM<sub>10</sub>, PM<sub>2.5</sub>, CO, NO<sub>2</sub>, SO<sub>2</sub>, and O<sub>3</sub>, and estimated the possible values. This is based on the method of regression analysis of controlled learning, and the appropriate

equation is derived from statistically relevant phenomena by determining the degree to which any form of correlation is consistent with the given experimental and measurement values. The results show that as the training period increases, the standard error decreases and the difference between the test and measurement values decreases, so we conclude that the air quality prediction machine training was successful. In the future, we will be able to analyze large amounts of data by adding experimental data, and we can experiment with different predictions using different types of data.

## References

- [1] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, “*Air Quality Prediction: Big Data and Machine Learning Approaches*”, International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018.
- [2] V. M. Niharika and P. S. Rao, “*A survey on air quality forecasting techniques*”, International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014.
- [3] E. Kalapanidas and N. Avouris, “*Applying machine learning techniques in air quality prediction*”, Sept. 1999.
- [4] S. Deleawe, J. Kuszniir, B. Lamb, and D. J. Cook, “*Predicting air quality in smart environment*”, J Ambient Intell Smart Environ., pp. 145-152, 2010.
- [5] L. Xiang, L. Peng, Y. Hu, J. Shao, and T. Chi, “*Deep learning architecture for air quality predictions*”, Environmental Science and Pollution Research, vol. 23, no. 22, pp. 22408-22417, 2016.
- [6] JürgenSchmidhuber, “*Deep learning in neural networks: An overview*”, Neural Networks Volume 61, pp.85-117, January 2015.
- [7] Pranjal Srivastava, “*Essentials of Deep Learning: Introduction to Long Short Term Memory*”, Analytics Vidhya, 2017.