# Air Quality Estimation Using LSTM and An Approach for Data Processing Techniques

Minh-Anh Ton-Thien[1,2,3,*], Chuong Thi Nguyen[1,4,*], Quang M. Le[1,2,3], Dat Q.Duong[1,2,3]
[1]AISIA Research Lab, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4] iLotusLand, Vietnam
[*]Two first author have equal contribution
minhanhtt2000@gmail.com
chuong.nguyen@vietan-enviro.com

## ABSTRACT

This paper describes our approach for the MediaEval2021 "Cross-Data Analytics for (transboundary) Haze Prediction" subtask1. The objective of this subtask is to predict PM10 values at different locations in multiple countries using data only from each country itself. In addition, we have applied XGBoost to deal with missing PM10 values on the training dataset and Long Short-term Memory (LSTM) [2] models to predict air pollution.

## 1 INTRODUCTION

Nowadays, air pollution leads to increasing cases of cardiovascular and respiratory diseases. It also affects social and economic activities. By using data from the last several days to predict air pollution for upcoming days, we can plan appropriate activities to protect our health.

As given in the task description [3] of the MediaEval2021, subtask 1 provides time-series datasets collected from different air quality and weather stations in Brunei, Singapore, and Thailand. Therefore, we decided to use LSTM models to predict air pollution of the next day from weather features and air quality of 10 previous days. For Brunei, we built and compared different variants of the LSTM model, i.e., the LSTM, Bidirectional LSTM, and Stacked LSTM. On the other hand, for Singapore and Thailand datasets, because of the lack of time and many PM10 values that need to be predicted hourly, we only used Bidirectional LSTM.

## 2 OUR APPROACH

### 2.1 Missing Values Imputation

Each data point in the training dataset represents the information of a location of one day in the country. We observed many missing values in the datasets; thus, we decided to employ two different methods to impute the missing values according to the value type. For data extracted from weather stations (i.e., temperature, rainfall, humidity, and wind speed), we filled the missing values with the mean values of its stations. For data related to air quality from monitoring stations, i.e., PM10, we employed XGBoost [1] to impute the missing values from the weather features.

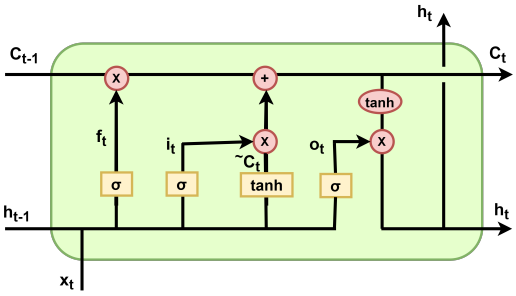


**Figure 1: Architecture of a LSTM cell**

First, the missing values of weather features on the training dataset were filled using the first method. Next, we created a new dataset from the original training dataset by dropping the rows where PM10 values are missing. Then, the new dataset was used to build the XGBoost model to predict missing PM10 values on the original training data from weather features.

It is worth noticing that all-weather features of Thailand collected in 2015 are missing. Therefore, to avoid interference when filling missing values, we dropped all data points in that year.

### 2.2 Models

Research studies have shown that LSTM is suitable for time series data [8, 9]; it is good at solving long-term memory problems, especially predicting n-th samples using many time steps before. Thus, we applied LSTM models in our study to predict air pollution.

*2.2.1 LSTM.* One of the disadvantages of Recurrent Neural Network (RNN) is it can not process long sequences; LSTM architecture is proposed to solve that problem. An LSTM cell, as shown in Figure 1, has a cell state C that allows the information to flow through for long-term memory. It also includes three gates: forget gate-decide what information should be kept or discarded by looking at the previous state and current input. Input gate decides what information is essential at the current step and how to add to the cell state; output gate decides what the output should be.


*MediaEval'21, December 13-15 2021, Online*

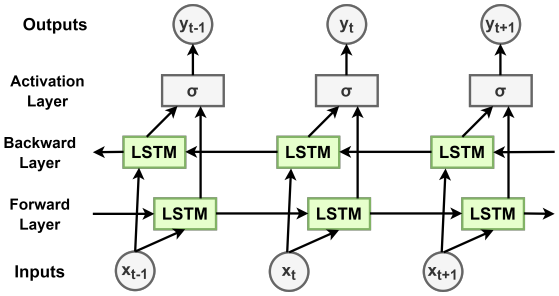


**Figure 2: Architecture of an unfolded Bi-LSTM**

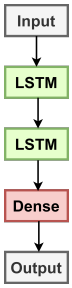


**Figure 3: Architecture of a Stacked LSTM**

*2.2.2 Bi-LSTM.* Bi-LSTM model, which was developed from Bidirectional Recurrent Network [6], consists of two LSTM layers: one taking the input in a forward direction, and the other in a backward direction. The architecture of an unfolded Bi-LSTM, as depicted in Figure 2, helps the network go through the input at the same time so that it can recognize the pattern of our data better.

*2.2.3 Stacked LSTM.* Stacked LSTM is a model that includes multiple LSTM layers. By making the model deeper, it has proved its effection in sequence data [5, 7]. In our study, we used 2-layers Stacked LSTM architecture, as described in Figure 3.

## 2.3 Models description

For each country, we did not develop multiple models for each station but combined data from all the stations into one complete dataset and build models to recognize the pattern form the dataset. The first 80% of the data is used for training sets, and the last 20% of the data is used for validation sets. Information of ten previous days, which includes weather features and air quality, is used to predict the PM10 values of the upcoming day. The Adam optimizer [4] is employed for model training, and the number of epochs used during training is 100 with a batch size of 512.

For Brunei, we experimented with LSTM, Bidirectional LSTM (Bi-LSTM), and Stacked LSTM. The organizers provided the PM10 values hourly for Singapore and Thailand, and we only tested all three LSTM models for three first hours PM10 values (i.e., *PM*10_1, *PM*10_2, and *PM*10_3). The results show that Bi-LSTM is slightly better than LSTM and Stacked LSTM, so we chose to employ the Bi-LSTM model for each hourly PM10 value for the final predictions.

# 3 RESULTS AND ANALYSIS

We evaluate the proposed models using RMSE metric calculated as follow:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} ||y(i) - \hat{y}(i)||^2}{N}}$$

where $N$ is the number of data points, $y(i)$ is the i-th measurement and $\hat{y}(i)$ is its ground truth.

The experimental results on validation sets of Brunei, Singapore, and Thailand datasets are shown in Table 1. In Brunei, Bi-LSTM achieves a slightly better result than LSTM, with a score of 3.625. For Singapore and Thailand, the average scores are 5.821 and 10.624.

**Table 1: Test results from best run submission on validation sets for Subtask 1**

| | RMSE | | |
|---|---|---|---|
| Model | Brunei | Singapore | Thailand |
| Bi-LSTM | **3.625** | 5.821 | 10.624 |
| LSTM | 3.629 | - | - |
| Stacked LSTM | 3.921 | - | - |

Table 2 described the evaluation results of our submitted run on the test datasets. Compare with the results on validation sets, it shows that our models did not work well for Brunei and Singapore on the test sets and overfitting has occurred.

**Table 2: Results that task organizers provided on the held-out test datasets**

| | RMSE | | |
|---|---|---|---|
| Model | Brunei | Singapore | Thailand |
| Bi-LSTM | 10.967 | 10.248 | 9.762 |

In this work, Bi-LSTM initially has shown some promising results for Brunei and Singapore on the validation sets. However, it did not perform as expected in the test sets. It might be because our missing values imputation technique is not good enough and the models can't fully recognize the pattern of the datasets.

# ACKNOWLEDGMENTS

Finally, we would like to send our thanks to AISIA Research Lab for supporting our team; the Organization Board of MediaEval 2021 and the Task Organizer for providing us with an opportunity to participate in the competition.

# REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[2] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[3] Asem Kasem, Minh-Son Dao, Effa Nabilla Aziz, Duc-Tien Dang-Nguyen, Cathal Gurrin, Minh-Triet Tran, Thanh-Binh Nguyen, and Wida Suhaili. Overview of Insight for Wellbeing Task at MediaEval 2021: Cross-Data Analytics for Transboundary Haze Prediction. Proc. of the MediaEval 2021 Workshop, Online, 13-15 December 2021.

[4] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[5] H. Sak, Andrew Senior, and F. Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (01 2014), 338–342.

[6] M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681. https://doi.org/10.1109/78.650093

[7] Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems* 4 (09 2014).

[8] Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. 2018. Air Pollution Forecasting Using RNN with LSTM. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. 1074–1079. https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178

[9] Thanongsak Xayasouk, HwaMin Lee, and Giyeol Lee. 2020. Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. *Sustainability* 12 (03 2020), 2570. https://doi.org/10.3390/su12062570