# Air Quality Prediction (PM2.5 and PM10) at the Upper Hunter Town - Muswellbrook using the Long-Short-Term Memory Method

Alexi Delgado[1], Ramiro Ricardo Maque Acuña[2]
Department of Engineering, Mining Engineering Section
Pontificia Universidad Católica del Perú
Lima, Peru

Chiara Carbajal[3]
Administration Program
Universidad de Ciencias y Humanidades
Lima, Peru

***Abstract*—Air quality is crucial for the environment and the life quality of citizens. Therefore, in the present study a software application is developed to predict air quality on the basis of 2.5 particulate matter ($PM_{2.5}$) and 10particulate matter ($PM_{10}$), in the city of Upper Hunter, Australia, as it is considered to be one of the cities with the lowest air quality levels worldwide. For this purpose, it has been decided to use the methodology of long-short term memory (LSTM) from data collected by NSW department of planning industry and environment during the period of 30 September 2012 to 30 September 2019, to predict the behavior of the mentioned particulate matter during the month of October 2019. A comparison between the average and maximum values suggested by the software and the actual values has been made and it is shown that the predicted results of the study are quite close to reality. Finally, the results obtained in this study may serve as a basis for local authorities to proceed with the necessary protocols and measures in case an alarming prediction occurs.**

***Keywords*—Air quality; long-short term memory (LSTM); 2.5 particulate matter ($PM_{2.5}$); 10 Particulate matter ($PM_{10}$)**

## I. INTRODUCTION

The effects of air pollution on health have been the focus of study in the past decades [1]. In the late eighties approximately, epidemiological studies have proved a relationship among air pollution levels and cardiovascular mortality, as well as hospital admissions and emergency room visits [2] in both developed and developing countries [3]; which leads to the recognition of air pollution as an influential and changeable determinant of cardiovascular disease in urban communities [4]. Furthermore, it has been estimated, according to the World Health Organization, that environmental air pollution is responsible for about 4.2 million premature deaths worldwide annually by 2018 [5]. This is why more citizens are recognizing the importance of air quality to their health nowadays [6]; as this not only impacts on the quality of life of the population, but also on their productivity, or school absenteeism in the case of young people, and therefore on their nation's GDP. As Xiang et al [7], quoted by [6], pointed out, high-resolution air quality data in the urban context are essential for the management of cities.

Therefore, the present study aims to propose a software to predict the air quality, based on data previously collected by NSW department of planning industry and environment [8]. For this purpose, Long Short Term Memory (LSTM), a particular kind of Recurrent Neural Networks (RNN) [9], will be used since its effectiveness in air quality prediction has been demonstrated in various studies such as [9], [10] as well as its capability of learning long-term dependencies unlike RNN methodology itself [11], [12]; by adding memory cell into hidden layer, so as to control the memory information of the time series data [13]. LSTM has the form of a repeating block chain for learning the time series information having three basic "gate", named input gate, output gate, and forget gate [14], [15]; its steps will be further explained in the next section.

The present study will be accomplished through the collection of data from the city of Upper Hunter, Australia, since it is recognized as one of the cities with the greatest negative impact in terms of air quality; therefore, a prediction of the behavior of air particles is one of its most urgent needs to address this problem. Given that air pollution, in Australia, has been estimated to be responsible for more deaths than road accidents [16], likewise Australia has been considered one of the countries with the highest levels of asthma in the world due to this air quality [17], [18]. As a result, there have been public complaints recently where residents of Upper Hunter, Muswellbrook claim that air pollution in their city has become part of their daily lives, getting worse every day and affecting the public health of citizens [19]–[21].

Ultimately, the purpose of the study is to suggest a software capable of predicting air quality through the behavior of $PM_{2.5}$ and $PM_{10}$ during a period of 30 days. After that period, a contrast with the actual air quality level will be made to evaluate the accuracy level of the software.

The structure of this investigation will be divided as follows. In Section II, the methodology in conjunction with its steps to be developed will be presented. In addition, the case study, in which the research was applied, and the corresponding explanation of the object of study are found in Section III. Subsequently, the data already processed will be shown in comparative graphs between the predicted data and the actual measured value in the Results and Discussion section. Finally, the corresponding conclusions are given in Section V, indicating the advantages of the method and the proposals for improvements on the future.

## II. Methodology

Long Short Term Memory, usually known as LSTM, was introduced for the first time in 1997 by Hochreiter and Schmidhuber [22]. Its main function is to remember information for long periods of time [11], having their internal memory for processing sequences of inputs, by recording old and current data [23]. One of its advantages is that to address long time lag issues, LSTM can manage noise, spread patterns and constant variables, as will be used in the present study. And compared to finite-state automata [24] or hidden Markov models [25], LSTM does not demand prior selection of a limited set of states. In principle, it can deal with unlimited state numbers. Furthermore, as opposed to conventional methods, LSTM is able to distinguish rapidly from two or more separate occurrences of a specific item in an entry sequence, without relying on appropriate examples of short-term training [22].

The following steps will be carried out [11]:

Step 1: The decision on which information will be removed from the processing cell will be made at this stage. This decision is made by a sigmoid layer ($\sigma$) called the "forgotten gate layer" as shown in Fig. 1 [11]. This gate gives values between 0 and 1, where 1 represents keeping the value and 0 represents removing the value completely. This model basically tells us that the value to be predicted will be clearly linked to the previous values.

Step 2: The new information is decided to be stored in the processing cell. This is divided into two parts, in the first one a sigmoid layer called "input gate layer" will decide which values will be updated. Subsequently, a tanh layer will generate a vector of new candidate values, which will be added to the prediction cell. Finally, both steps were combined to update the prediction cell. Fig. 2 [11] is presented for further details.

Step 3: At this stage the old cell status, $C_{t-1}$, is updated to the new cell status $C_t$. All the previous steps have already decided how to proceed, the only thing necessary is to execute them.

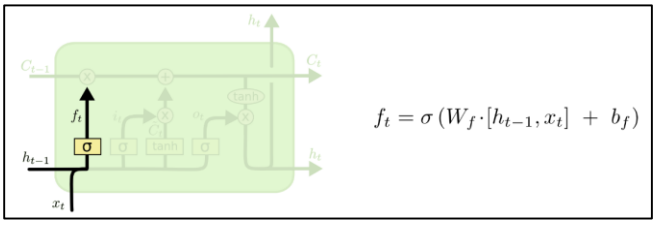


Fig. 1. Forgotten Gate Layer Equation ($f_t$) [11].

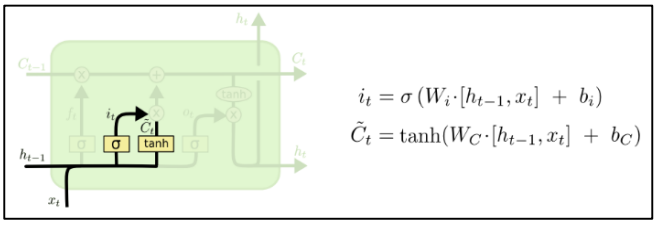


Fig. 2. Equation of the Input Gate Layer ($i_t$) and the Vector of the New Candidate Values ($\tilde{C}_t$) [11].

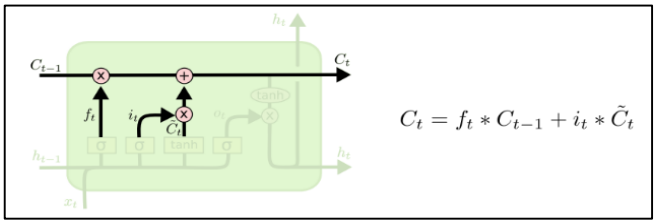


Fig. 3. Update layer equation ($C_t$) [11].

The old state is multiplied by $F_t$, leaving behind the things previously decided to forget. Next step is to add $i_t * \tilde{C}_t$. These are the new candidate values, scaled according to what we decided to update each state value. The representation can be seen in Fig. 3 [11].

Step 4: The final step is to decide about what we will produce. That output should be on the basis of our cellular state, but it will be a filtered version. In order to do this, a sigmoid layer will be executed, which will decide those parts of the cellular state that we are going to produce. Next, in order to push the values to be between -1 and 1, the cell state will be passed through tanh and multiplied by the output of the sigmoid gate, resulting in only the parts we decide to generate. For a better understanding observe Fig. 4 [11].

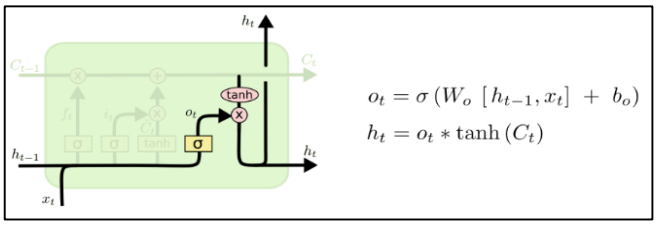


Fig. 4. Output layer equation ($h_t$) [11].

## III. Case Study

The present study is focused on analyzing the data recorded by the NSW department of planning industry and environment [8] regarding $PM_{2.5}$ and $PM_{10}$ in the context of Upper Hunter during the period of 30 September 2012 to 30 September 2019; through which it will be sought to create software capable of predicting the behavior of these particulates matter through to October 2019. In other words, the matter particles for this study were classified into two groups: particulate matter with a diameter of up to 2.5 µm solid dust particles, soot, among others; and metal particles whose diameter varies between 2.5 and 10 µm [26]. These two groups are very fine particles in the air that are measured by micrometers [27]; as a reference, it is need to be taken into account that human hair is about 100 micrometers [19].

Nevertheless, there is a distinction between $PM_{2.5}$ and $PM_{10}$, as the first group is a stronger threat to public welfare than the second group [28]. As confirmed by studies that show that these particles are more likely to penetrate the respiratory system and deposit in the alveoli of the lungs with the possibility of reaching the bloodstream because of their small size [29]. Converting it as one of the main health hazards in large cities around the world [30]. Therefore, according to [31], several studies on the relationship between $PM_{2.5}$ and the mortality rate have been promoted, such as the one carried out

in the United States by [32]. However, this does not imply that $PM_{10}$ is not considered harmful to health, since it also greatly affects the eyes, nose and throat of citizens who are exposed to high levels of air and/or dust, where these particles are transported [33].

Regarding the data obtained, a total of approximately 5000 inputs were collected, of which the days that did not report values were removed from the calculation. Likewise, the high levels of non-standard values were used, as well as the low values for the origin of the measures, thus no data processing was carried out.

The raw data were used to create a neural network, in order to do this the annual trend of air quality parameters was analyzed, according to the steps described in the methodology each value that entered through the forgotten gate layer will be iterated 60 times for each new value, namely, for each value the previous 60 data will be used as reference. This process created a new value that will depend on the previous data and will be used as input data for the calculation of the next value, as well as successively until all the existing data are used.

## IV. Results and Discussion

The results are presented in Fig. 5, 6, 7 and 8.

The comparison between the actual and predicted average and maximum values of $PM_{10}$ in the period October 2019 are shown in Fig. 5 and Fig. 6, respectively.

Similarly, the actual and predicted $PM_{2.5}$ average and maximum values for the month of October 2019 can be observed in Fig. 7 and Fig. 8.

These results were obtained from the model implemented using advanced neural network tools such as Keras [34] and Tensorflow [35]. In this model, 130 iterations were used for the data relation for every 60 values analyzed. Nevertheless, due to the complexity of the neural equations of LSTM, without the assistance of a powerful GPU each training cycle can require 2 to 3 hours for a cycle of 130 iterations. For the present case, it lasted about 45 minutes for each prediction.

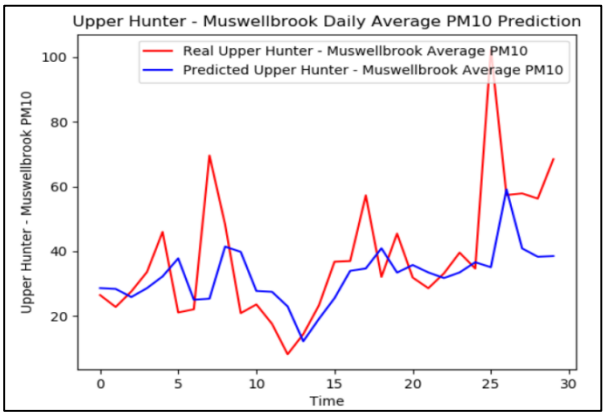


Fig. 5. Comparison between Actual and Predicted Average Values of 〖PM〗_10 During October 2019.

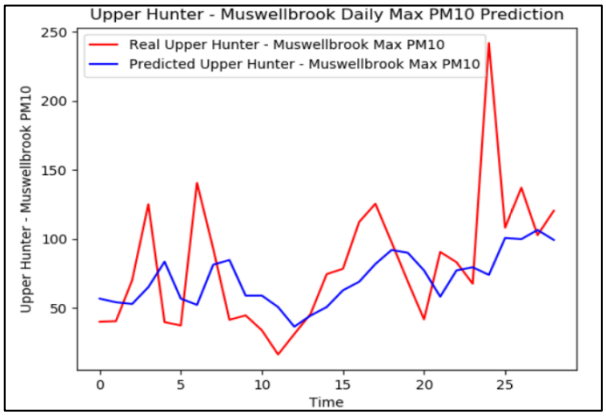


Fig. 6. Comparison between the Real and Predicted Maximum Values of 〖PM〗_10 During October 2019.

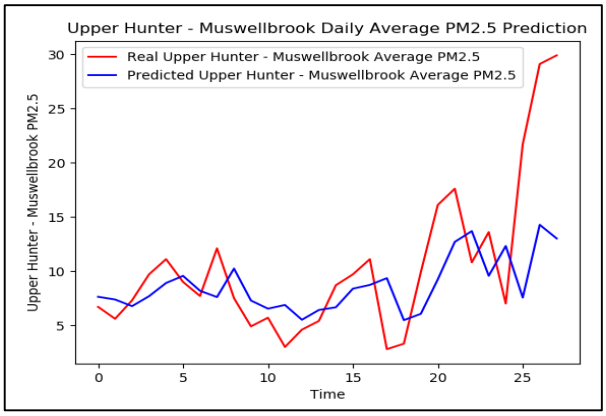


Fig. 7. Comparison between Actual and Predicted Average Values of 〖PM〗_2.5 During October 2019.

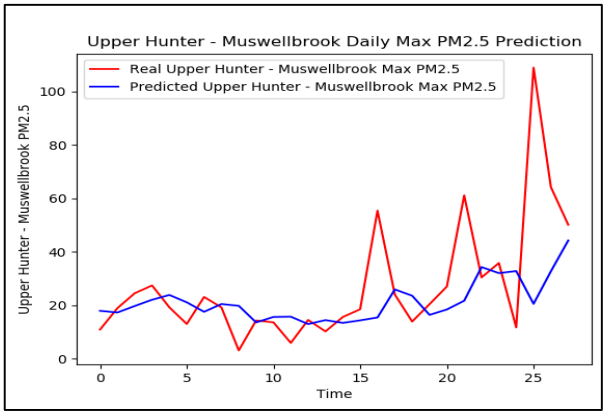


Fig. 8. Comparison between the Real and Predicted Maximum Values of 〖PM〗_2.5 During October 2019.

However, it should be noted that these surveillance data were worked on as they were found (raw data), i.e. no treatment was done if there were illogical values. The prediction in Fig. 5, 6 and 7 was almost exact at the moment of finding the trend, since we can observe very noticeable changes in the air quality during the 30 days; whereas, for Fig. 8, the variation during the first days was not much for what the software considered to be an almost constant trend.

The predicted values were found to be quite close to the actual behaviour of the particulate matter, therefore the trend predicting ability of the variation in air quality of the actual sampling is confirmed for both the average and maximum values of $PM_{2.5}$ and $PM_{10}$ as the more abrupt the variations in air quality, the more accurate the prediction is in assessing these changes.

## V. Conclusion

Air quality prediction is of great importance for environmental protection. Considering the multivariate data in terms of $PM_{2.5}$ and $PM_{10}$ values of the information collected by the Upper Hunter station during the years 2012 to 2019 it was possible to verify that the LSTM method is valid for predicting behavior of the mentioned parameters in the future, allowing the development of protocols or procedures in case of an alarming prediction.

In the present work it is demonstrated that the development of a computer code whose purpose is to predict air quality can be developed from a raw data base. The prediction model based on LSTM makes good use of the time sequence of air quality information and, at the same time, allows its prediction accuracy to be improved. However, its limitation is that a large amount of historical monitoring data is required to train the prediction models. In addition, the training time is long depending on the quality of the prediction.

Similarly, the application potential of the LSTM method can be used for different needs, as presented in previous research. It can also be used as a management tool for evaluation projects or prevention measures.

## References

[1] B. Brunekreef and S. T. Holgate, "Air pollution and health," Lancet, vol. 360, no. 9341. Elsevier Limited, pp. 1233–1242, 19-Oct-2002, doi: 10.1016/S0140-6736(02)11274-8.

[2] J. K. Mann et al., "Air pollution and hospital admissions for ischemic heart disease in persons with congestive heart failure or arrhythmia," Environ. Health Perspect., vol. 110, no. 12, pp. 1247–1252, Dec. 2002, doi: 10.1289/ehp.021101247.

[3] J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger, "Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994," N. Engl. J. Med., vol. 343, no. 24, pp. 1742–1749, Dec. 2000, doi: 10.1056/NEJM200012143432401.

[4] N. L. Mills et al., "Adverse cardiovascular effects of air pollution," Nature Clinical Practice Cardiovascular Medicine, vol. 6, no. 1. Nature Publishing Group, pp. 36–44, 25-Nov-2009, doi: 10.1038/ncpcardio 1399.

[5] Organización Mundial de la Salud, "Calidad del aire y salud," Organ. Mund. la Salud, p. 11, May 2018, doi: 10.1016/S2214-109X(16)30143-7.

[6] J. Huang et al., "A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments," IEEE Internet Things J., vol. 6, no. 2, pp. 3240–3247, Apr. 2019, doi: 10.1109/JIOT.2018 .2881240.

[7] Y. Xiang, R. Piedrahita, R. P. Dick, M. Hannigan, Q. Lv, and L. Shang, "A hybrid sensor system for indoor air quality monitoring," in Proceedings - IEEE International Conference on Distributed Computing in Sensor Systems, DCoSS 2013, 2013, pp. 96–104, doi: 10.1109/JIOT.2018.2881240.

[8] NSW department of planning industry and environment, "Upper Hunter - Live air quality data." [Online]. Available: https://www.dpie.nsw.gov.au/air-quality/live-air-quality-data-upper-hunter. [Accessed: 02-Mar-2020].

[9] Y. Jiao, Z. Wang, and Y. Zhang, "Prediction of air quality index based on LSTM," in Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019, 2019, pp. 17–20, doi: 10.1109/ITAIC.2019.8785602.

[10] Y. T. Tsai, Y. R. Zeng, and Y. S. Chang, "Air pollution forecasting using RNN with LSTM," in Proceedings - IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, IEEE 16th International Conference on Pervasive Intelligence and Computing, IEEE 4th International Conference on Big Data Intelligence and Computing and IEEE 3, 2018, pp. 1068–1073, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178.

[11] A. Delgado, A. Aguirre, E. Palomino, G.Salazar, "Applying triangular whitenization weight functions to assess water quality of main affluents of Rimac river," in Proceedings of the 2017 Electronic Congress, E-CON UNI 2017, 2018-January, pp. 1-4.

[12] M. Hajiaghayi and E. Vahedi, "Code Failure Prediction and Pattern Extraction using LSTM Networks," Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol., pp. 55–62, Dec. 2018.

[13] T. Li, M. Hua, and X. Wu, "A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5)," IEEE Access, vol. 8, pp. 26933–26940, Feb. 2020, doi: 10.1109/access.2020.2971348.

[14] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, and Z. Miao, "Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network," IFAC-PapersOnLine, vol. 51, no. 17, pp. 831–836, Jan. 2018, doi: 10.1016/j.ifacol.2018.08.091.

[15] X. Xu and M. Yoneda, "Multitask Air-Quality Prediction Based on LSTM-Autoencoder Model," IEEE Trans. Cybern., pp. 1–10, Oct. 2019, doi: 10.1109/tcyb.2019.2945999.

[16] M. Sacasqui, J. Luyo, A. Delgado, "A Unified Index for Power Quality Assessment in Distributed Generation Systems Using Grey Clustering and Entropy Weight," in 2018 IEEE ANDESCON, ANDESCON 2018 - Conference Proceedings, pp. 8564631.

[17] A. Corderoy, "Australia has one of highest rates of asthma in the world," The Sydney Morning Herald, 2014.

[18] T. To et al., "Global asthma prevalence in adults: Findings from the cross-sectional world health survey," BMC Public Health, vol. 12, no. 1. BioMed Central, p. 204, 2012, doi: 10.1186/1471-2458-12-204.

[19] A. Bernasconi and M. Pritchard, "Deteriorating air quality in Upper Hunter down to weather not politics, says local MP," ABC, Upper Hunter, 27-Nov-2019.

[20] E. Goetze, "'Our pool is black': Upper Hunter residents vent air-pollution fears," ABC, Upper Hunter, 24-Oct-2019.

[21] "Friends of the Upper Hunter hosting air quality meeting at Muswellbrook's Upper Hunter Conservatorium of Music," Muswellbrook Chronicle, Upper Hunter, 21-Nov-2019.

[22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[23] M. Chovatiya, A. Dhameliya, J. Deokar, J. Gonsalves, and A. Mathur, "Prediction of dengue using recurrent neural network," in Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, 2019, vol. 2019–April, pp. 926–929, doi: 10.1109/icoei.2019.8862581.

[24] T. Otsuki, A. Ito, S. Makino, and T. Otomo, "The performance prediction method on sentence recognition system using a finite state automaton," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1994, vol. 1, pp. I397–I400, doi: 10.1109/ICASSP.1994.389272.

[25] Z. Li, L. Liu, and D. Kong, "Virtual Machine Failure Prediction Method Based on AdaBoost-Hidden Markov Model," in Proceedings - 2019 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2019, 2019, pp. 700–703, doi: 10.1109/ICITBS.2019.00173.

[26] C. Linares and J. Díaz, "¿Qué son las PM2,5 y cómo afectan a nuestra salud?," Ecologistas en Acción, 01-Sep-2013. [Online]. Available: https://www.ecologistasenaccion.org/17842/que-son-las-pm25-y-como-afectan-a-nuestra-salud/. [Accessed: 21-Feb-2020].

[27] J. A. R. Montanez, M. A. A. Fernandez, S. T. Arriaga, J. M. R. Arreguin, and G. A. S. Calderon, "Evaluation of a recurrent neural network LSTM for the detection of exceedances of particles PM10," in 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control, CCE 2019, 2019, doi: 10.1109/ICEEE.2019.8884516.

[28] A. Delgado, P. Montellanos, J. Llave, , "Air quality level assessment in Lima city using the grey clustering method," in IEEE ICA-ACCA 2018 - IEEE International Conference on Automation/23rd Congress of the Chilean Association of Automatic Control: Towards an Industry 4.0 – Proceedings, pp. 8609699.

[29] J. Kaiser, "Air pollution. Evidence mounts that tiny particles can kill.," Science, vol. 289, no. 5476. pp. 22–23, 07-Jul-2000, doi: 10.1126/science.289.5476.22.

[30] L. J. Chen et al., "An Open Framework for Participatory PM2.5 Monitoring in Smart Cities," IEEE Access, vol. 5, pp. 14441–14454, Jul. 2017, doi: 10.1109/ACCESS.2017.2723919.

[31] C. J. Huang and P. H. Kuo, "A deep cnn-lstm model for particulate matter (Pm2.5) forecasting in smart cities," Sensors (Switzerland), vol. 18, no. 7, Jul. 2018, doi: 10.3390/s18072220.

[32] M. A. Kioumourtzoglou, J. Schwartz, P. James, F. Dominici, and A. Zanobetti, "PM2.5 and mortality in 207 US cities: Modification by temperature and city characteristics," Epidemiology, vol. 27, no. 2, pp. 221–227, Jan. 2016, doi: 10.1097/EDE.0000000000000422.

[33] MURCIAL+SALUD, "MATERIA PARTICULADA (PM10 Y PM2,5)," Rev. Fac. Nac. Salud Pública, 2011. [Online]. Available: http://www.murciasalud.es/pagina.php?id=244308&idsec=1573#. [Accessed: 21-Feb-2020].

[34] K. Jakhar and N. Hooda, "Big data deep learning framework using keras: A case study of pneumonia prediction," in 2018 4th International Conference on Computing Communication and Automation, ICCCA 2018, 2018, doi: 10.1109/CCAA.2018.8777571.

[35] W. W. T. Fok et al., "Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine," in 2018 4th International Conference on Information Management, ICIM 2018, 2018, pp. 103–106, doi: 10.1109/INFOMAN.2018.8392818.