Bima Journal of Science and Technology, Vol. 4(2) Dec, 2020 ISSN: 2536-6041

DEEP RECURRENT NEURAL NETWORK FOR POLLUTION FORECASTING IN SMART CITIES

[1*]RUFAI DUWAP BELLO, [2]HAKEEM ADEWALE SULAIMON, [3]MUSTAPHA LAWAL ABDULRAHMAN AND [1]MUHAMMAD AMINU AHMAD

[1]Department of Computer Science, Kaduna State University Kaduna State, Nigeria
[2]Department of Computer Science, Federal College of Education, Zaria Kaduna State, Nigeria
[3]National Center for Remote Sensing Jos, Rizek Village Plateau State, Nigeria

Corresponding Author: elroofey@gmail.com

ABSTRACT

Air pollution has been marked as one of the major problems of metropolitan areas around the world. According to The Health Effects Institute (HEI), over 95% of the world's population is breathing polluted air, which contributed to the death of 6.1 million people across the world in 2016. These health complications can be avoided or diminished through raising the awareness of air quality conditions in urban areas, which could allow citizens to limit their daily activities in the cases of elevated pollution episodes, by using models to forecast or estimate air quality in regions lacking monitoring data, New automated paradigms base on artificial intelligent have to be thought to improve the prediction performance. Many influencing factors make the prediction complex. Traditional approaches depend on numerical methods to estimate the air pollutant concentration and require lots of computing power. Moreover, these methods cannot draw insights from the abundant data available, thus, Neural networks are recently applied in this context due to their wide application. Deep recurrent neural network models provide a practical approach to air pollution and air-pollution prediction. To address this issue, this paper puts forward a deep learning approach using multivariate LSTM for the prediction of air pollution concentration in smart cities. The City Pulse EU FP7 Project smart city data set was used in this study. The proposed model was evaluated against state-of-the-art prediction techniques used in pollution forecasting using RMSE, MAE, and R on Matlab 2018a. The proposed system model for multiple pollutants forecast concurrently, which significantly increase the accuracy by 20% and 36% in terms of RMSE and 26% and 47% in terms of MAE respectively.

Keywords: Artificial Neural Network, Deep Learning, Deep Recurrent Neural Network, Recurrent Neural Network, Multi layer Perceptron

INTRODUCTION

Monitoring and preserving air quality have become one of the most essential activities in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, and fuel uses. The deposition of harmful gases is creating a serious threat to the quality of life in smart cities (Kang *et al*., 2018). Thus, an early warning system based on accurate forecasting tools must be implemented to avoid the adverse effects of exposure to major air pollutants (Lazrak *et al*., 2018). These health complications can be avoided or diminished through raising the awareness of air quality conditions in urban areas, which could allow citizens to limit their daily activities in the cases of elevated

pollution episodes, by using models to forecast or estimate air quality in regions lacking monitoring data (Rybarczyk and Zalakeviciute, 2018).

According to (Mocanu, *et al*.,2016), forecasting can be grouped into either one of these three groups, they include (i) short term forecast usually ranging from a day to a week (ii) medium-term forecast ranging from a week to a year and (iii) long term forecast usually ranging from a year and above. There are a few main approaches to air pollution modeling—atmospheric chemistry, dispersion (chemically inert species), and machine learning. However, recent studies show that the traditional deterministic models struggle to capture the non-linear relationship between the concentration of contaminants and their sources of emission and dispersion (Rybarczyk and Zalakeviciute, 2018). To tackle the limitations of traditional models, the most promising approach is to use statistical models based on machine learning (ML) algorithms (Rybarczyk and Zalakeviciute, 2018). Statistical techniques do not consider physical and chemical processes and use historical data to predict air quality.

Models are trained on existing measurements and are used to estimate or forecast concentrations of air pollutants according to predictive features (e.g., meteorology, land use, time, planetary boundary layer, elevation, human activity, pollutant covariates, etc.) example includes Regression model (Chen, Chen, Wu, Hu, and Pan, 2017), time series model (Nhung *et al*., 2017), Autoregressive Integrated Moving Average (ARIMA) (Zafra, Ángel, and Torres, 2017). These analyses describe the relationship between variables based on possibility and statistical average. Well-specified regressions can provide reasonable results. However, the reactions between air pollutants and influential factors are highly non-linear, leading to a very complex system of air pollutant formation mechanisms. Therefore, more advanced statistical learning (or machine learning) algorithms are usually necessary to account for proper non-linear modeling of air contamination.

The smart city presents a suitable platform that collects open and big data precisely, with integration and direct analysis done on its main system. Forecasting and analysis of the air quality parameters represent the most important topics of environmental and atmospheric research in the smart city. The data of the smart city are created in the form of a time series. Time series analysis is the main task in prediction modules (Ghoneim and Manjunatha, 2017). Time series forecasting utilizes information with historical values to predict future activities by associating patterns. Specifically, the studies in Rao, *et al*. (2019) point out those conventional methods need prior knowledge about the model structure and are based on the theoretical hypothesis. Also, they work with various data constraints. Statistical models apply simple parameter-based methods surpassing the complicated structures. Hence, these models may not reveal valuable insights into the data. (Rao *et al*., 2019). Deep learning approaches have emerged as powerful solutions to mitigate these limitations over conventional methods (Qi *et al*., 2018). Thus, this paper puts forward a deep learning approach to enhance the prediction performance using multivariate LSTM for the prediction of air pollution concentration in smart cities.

MATERIALS AND METHODS

Data Collection

The City Pulse EU FP7 Project smart city data set was used in this study, to build an hourly prediction model for air pollution concentration level using the prescribed data from City Pulse Project smart city data sets the data has to be preprocessed. Since the observations of Pollutants recorded every five minutes then these data have to be aggregated to be suitable for the hourly prediction model. The pollution data sets, measurements were recorded every five minutes. Pollutants measured include air pollution, particulate matter, carbon monoxide, Sulphur dioxide, and nitrogen dioxide. These measurements were collected by 449 sensors located in different places inside the Aarhus city providing 449 files composing the pollution data set.

Model Development

Framework for the Proposed Model

To implement the model, the proposed framework is described below so that the architecture must be able to process the data with a high throughput with good performance by introducing an LSTM sequence input layer to the conventional LSTM network and we further deepened the LSTM layer with multiple hidden flayers.

The architecture of the proposed framework is shown in Figure 1

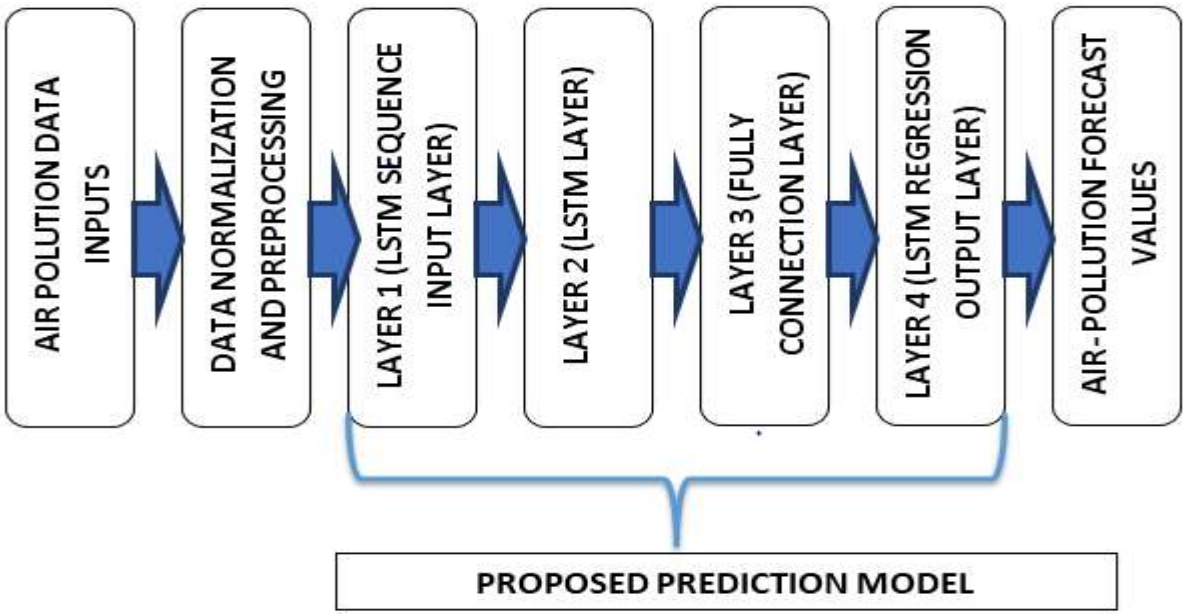

Figure 1: Proposed Framework.

The framework consists of two main stages as discussed below.

Stage 1: Data set Collection and Preprocessing Stage
This stage combines the data set identification, collection, and preprocessing of the data set to remove inconsistencies.

Stage 2: LSTM Stage
The dataset is fed to the sequence input layer of the LSTM network. The core components of an LSTM network are a sequence input layer and an LSTM layer. A sequence input layer inputs sequence or time-series data into the network. An LSTM layer learns long-term dependencies between time steps

of sequence data. The LSTM network starts with a sequence input layer followed by an LSTM layer. The network ends with a fully connected layer and a regression output layer as shown in Figure 2.

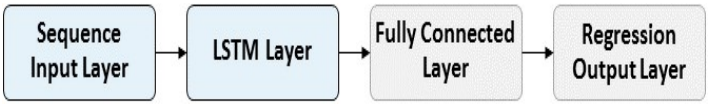

Figure 2: Structure of the LSTM Network with the proposed Sequence input layer

Choice of Metrics

Testing the Neural Network is a vital step in the design process. In this experiment, A Standard Statistical forecasting metric will be used to evaluate the network performance in other to select the model with the best performance. The developed network will be tested and evaluated by using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) in the MATLAB Neural Network (NN-tools) package.

Evaluation Setup

In this experiment, the dataset was imported and Assigned variables such that the input dataset was equated to the target data and then we transpose the input data set from columns to rows and we convert concurrent vectors to sequential vectors. The Data was split into training, testing, and validation. 70% of data was set for training, 15% for testing, and 15% for validation.

RESULTS AND DISCUSSION

In this research, the accuracy of the estimated forecasts of the proposed model will be compared with the other models to ascertain which model gives a more accurate forecast through the use of RMSE and MAE. We will now present the result based on forecasting using all the algorithms and present the results by the performance standard of each of the models and discuss the findings.

From the results of our simulation, we obtain the following plots which will demonstrate how perfectly fit is our model. To validate the developed network, the Error histogram was used. In this research, four multiple pollutants were model which includes: Particulate matter, Carbon Monoxide, Sulphur Dioxide, and Nitrogen Dioxide. Figure 3 and 4 shows graphical representations of the trend in the multiple pollutants that was model in this research work.

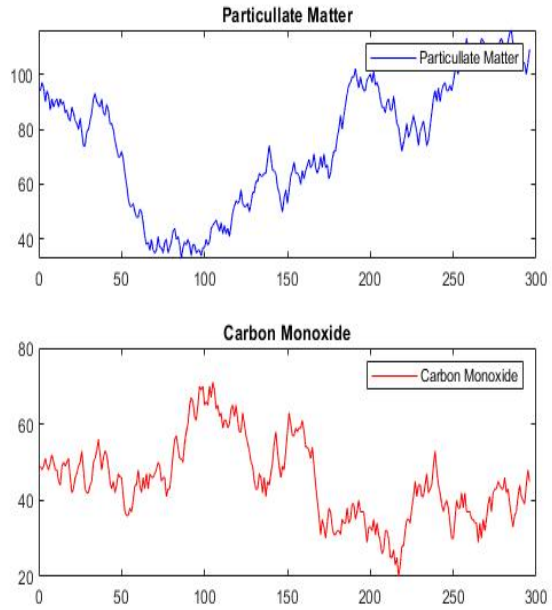

Figure 3: Trend of particulate matter and carbon monoxide

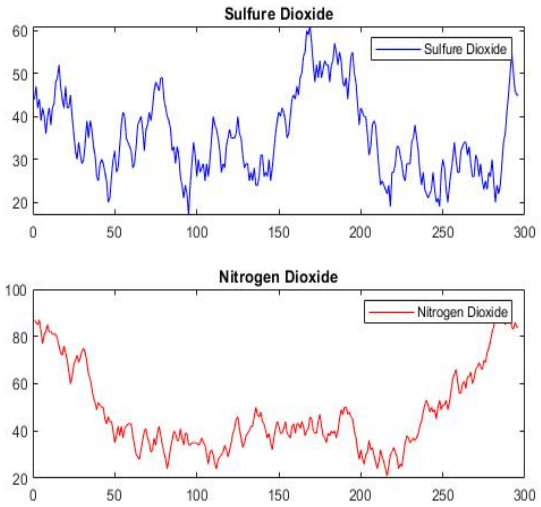

Figure 4: Trend of Sulphur Diaoxide and Nitrogen Diaoxide

Similarly, figure 5 shown the error histogram for the fit set of the proposed model. It shows how the error sizes are slightly well distributed. Typically, when most errors are near zero, it has been observed a better-trained model. In this case, however, it is confirmed that the network also has errors near zero.

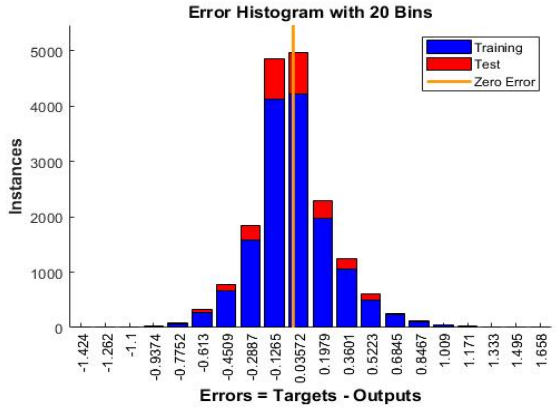

Figure 5: Error histogram for the fit set of the proposed model

The training algorithm used in this research work to train our model is Bayesian regularization back propagation algorithm which converged after 203 epochs, and it showed stability (no increase after converging) and no overshoot (no increase before converging), as shown in Figure 6.

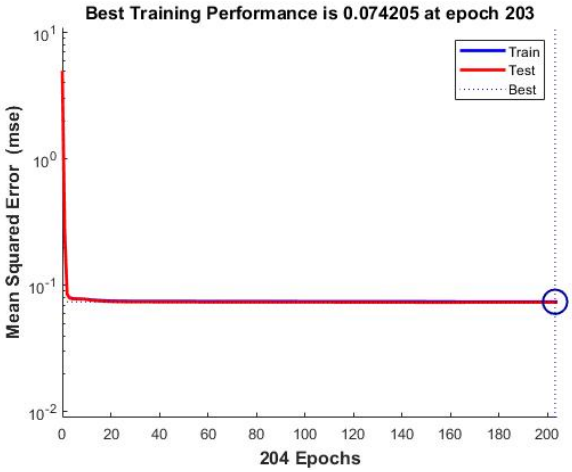

Figure 6: Training Performance for the proposed Model

In table 1, the performances of the proposed deep RNN models compare with those of the existing model for the forecast of air pollutants is tabulated. For the Validation RMSE and MAE the smaller the value the better the accuracy, in this case, the proposed Multivariate-LSTM model has RMSE of 0.003840 and MAE of 24.28 respectively. This is by far the lowest value when compared against the existing model i.e. LSTM- RNN which has RMSE of 17.9700 and MAE of 32.81 and MLP which has RMSE of 26.5471 and MAE of 46.0001 respectively. Thus, our proposed Multivariate-LSTM has clearly enhanced the forecasting accuracy as against all the existing models that were used for the evaluation by achieving the lowest values in terms of validation RMSE and MAE. Thus, our proposed model demonstrates superiority in forecasting air-pollution with high accuracy as against the state of the art.

The result from our experiment shows that in general, the proposed models perform better than the LSTM base RNN in

forecasting multiple pollutants simultaneously. The result shows that both models have performed comparatively in predicting air-pollutions.

Table 1: The performances of the proposed deep RNN models compare with those of the existing model

	Proposed Model	RNN	MLP
RMSE	0.003840	17.9700	26.5471
MAE	24.28000	32.8100	46.0001

Similarly, Table 2 illustrates how the performances of the proposed model when compared with those of the Multilayer perceptron and LSTM base RNN model in terms of training performance and running time (cost of computation).The proposed model achieved the best training performance with MSE of 0.074205 at epoch 203 which was lower than the existing system which achieved the best value of 0.29754 since the smaller the value the better the performance results.

Table 2: Performance for the training sets of the proposed model against the existing models.

Model	Training	Computation Time
Proposed Model	0.74205	2.25
RNN	0.29754	1.15
MLP	0.12054	1.05

But one of the major drawbacks of the proposed system as against the existing system is the implementation and computational complexity, this was further proven in Table 2 above, the proposed system has the highest computational time of 2.25 as against the RNN, which has 1.15 and MLP has 1.05. Thus, the proposed model has a high cost of computation when compared to the existing models. This is because the existing system models a single pollutant predictor while the proposed model tries to predicts multiple pollutants simultaneously thus, increasing the computational time of the model.

Clearly, from the aforementioned result analysis, the proposed model in general was able to enhance the prediction accuracy significantly compared to state-of-the-art prediction techniques. Although, these results were achieved at the expense of the significant cost of computation. Hence, the proposed prediction model demonstrates the suitability of application in the context of air-pollution forecasting.

CONCLUSION

Conclusively, the proposed study puts forward a deep learning approach using multivariate LSTM for the prediction of air pollution concentration in smart city. The proposed neural network model presented performs well in forecasting air pollution for the medium to long-term time horizon. Further, the proposed deep recurrent neural network model performs better in terms of validation RMSE and MAE than the existing model in forecasting air-pollution in the smart city for a medium-term time horizon. It was able to account for long-term dependencies in the context of air-pollution forecasting. Similarly, the proposed prediction model for multiple pollutants forecast concurrently, which significantly increase the accuracy in terms of RMSE and MAE which makes the prediction more reliable and robust.

REFERENCES

Chen, J., Chen, H., Wu, Z., Hu, D., and Pan, J. Z. (2017). Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems, 64*, 281-291.

Ghoneim, O. A., and Manjunatha, B. (2017).

Forecasting of ozone concentration in smart city using deep learning. Paper presented at the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., and Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development, 9*(1), 8-16.

Lazrak, N., Zahir, J., and Mousannif, H. (2018). *Air Quality Monitoring Using Deterministic and Statistical Methods.* Paper presented at the International Conference on Big Data and Smart Digital Environment.

Mocanu, E., Nguyen, P. H., Gibescu, M., and Kling, W. L. (2016). Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks, 6*, 91-99.

Nhung, N. T. T., Amini, H., Schindler, C., Joss, M. K., Dien, T. M., Probst-Hensch, N., . . . Künzli, N. (2017). Short-term association between ambient air pollution and pneumonia in children: A systematic review and meta-analysis of time-series and case-crossover studies. *Environmental pollution, 230*, 1000-1008.

Qi, Z., Wang, T., Song, G., Hu, W., Li, X., and Zhang, Z. (2018). Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Transactions on Knowledge and Data Engineering, 30*(12), 2285-2297.

Rahman, A., Srikumar, V., and Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied energy, 212*, 372-385.

Rao, K. S., Devi, G. L., and Ramesh, N. (2019). Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks. *International Journal of Intelligent Systems and Applications, 11*(2), 18.

Rybarczyk, Y., and Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied Sciences, 8*(12), 2570.