TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

Deeper into Air Pollution Forecasting: A Comparative Study of Deep Learning Models on Air Pollution Forecasting

Naresh Kumar, Jatin Bindra, Rahul Mattoo, Rajat Sharma, Varun Taneja
Department of CSE, Maharaja Surajmal Institute of Technology, Delhi
Deepali Gupta, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.

Article Info
Volume 83
Page Number: 1378 - 1383
Publication Issue:
May - June 2020

Article History
Article Received*: 11August 2019*
Revised*: 18November 2019*
Accepted*: 23January 2020*
Publication*: 10 May2020*

Abstract:

Air pollution is a matter high concerns for urban areas. Outdoor air pollution is at a level which can seriously threaten and harm the human health and life in major cities, especially to elderly and children. Therefore, many countries in the world have constructed stations for monitoring air pollution around major cities to observe air pollutants such as PM 2.5, PM 10, CO, NO2, SO2 and to alert their citizens if there is a pollution index which excesses the quality threshold. Also, air pollution is impacted by the meteorological factors of local place such as temperature, humidity, rain, wind, etc. This research aims in comparative study to forecast the reading of air pollution by using various deep learning models. Finally, it does the investigation on accuracies of various models in predicting air-pollution and discussion on draw backs of models used. Through this project, there will be a significant motivation for not only continuing research on urban air quality but also help the government leverage these insights to enact beneficial policies.

Keywords: *Deep learning models, Air pollution, long short term model, convolutional neural networks, Air quality monitoring, Air pollution prediction.*

1 Introduction

Air pollution and prevention of it are the areas of constant scientific challenges during last decades. However, there still remain huge global problems. Air pollution affects a person's respiratory and cardiovascular system, it is causing an increase in the mortality rate and an ever increasing risk of diseases and problems for the population. There have been many efforts by both local and state government to understand and predict air quality index with aims of improving the public health.

As per the paper [1], major air pollution causing materials also called air pollutants are PM2.5, PM10, CO, NO2, and SO2. Following is a brief description these air pollutants. PM2.5 is fine atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers. PM10 is coarse particulate that is 10 micrometers or less in diameter. CO refers to Carbon Monoxide, which is produced by burning and combustion of fuels such as natural gas, coal or wood. Exhaust by the Vehicles contributes to the majority of carbon monoxide let into the atmosphere. NO2 refers to Nitrogen Oxides, expelled from high-temperature combustion. SO2 is Sulphur Oxides, which are expelled by volcanoes and in various industrial processes. Sulphur compounds are found in Coal and petroleum, and their combustion generates Sulphur dioxide [1].

Since deep learning models form the basis of this project. Deep Learning is defined as " Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks."

2 Related Work

People around the world breath air which is not fit according to World Health Organisation Air Quality Guidelines. Various research has been conducted to

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

evaluate the reasons and further to predict the air quality in future.

.

Deep In [1], study has been conducted on real-time air quality data gathered via sensors installed on taxis running across Daegu city, Korea .The data used was vast and collected within an interval of 1 minute, in both spatial and temporal format. The study proposed a hybrid architecture which was a combination of convolutional neural networks (CNN) and LSTM model to forecast air quality. This architecture used a novel approach but still produced good prediction results. Authors have implemented a CNN unit to generate grid images for air pollution distribution at multiple locations within the city. These grid images internalized the air quality relationships among various locations. The proposed model was combined with a LSTM unit that forecasts air pollution with the help of time series in data. Apache Spark was used as a big data computing engine due to massive scale of the data and the model was built upon TensorFlow deep learning framework. The paper claimed the accuracy for prediction to be approximately 74%.

The literature in [2] proposed a similar model as in [1] but distinguished its work by focussing more on air quality relationships among multiple locations within the region. The study focussed on determining the most related locations to a particular location using the k-nearest neighbours approach. The generated training dataset contained the top k related locations to any given location. The results produced by this method on the Taiwan and Beijing datasets inferred that the LSTM unit enhanced the first hour predictions and the inclusion of CNN unit boosted the prediction performance over longer time frame. This was possible as the model was able to extract temporal delay factor from surrounding location features by learning spatial information. This method enabled better prediction for time periods up to next 48 hours.

Authors of [3] have suggested that MAE (Mean Absolute Error) served better as an error metric than MSE (Mean Squared Error). They conducted a study using the deep learning model defined in [2] on multiple datasets. The model was first trained on Air Korea dataset, the model with the learned weights was further re-trained on Daegu dataset and then on Seoul Clean Air dataset. This model was also re-trained on a European dataset which removed all China-related features.

This research focussed alternately on improving the reliability of the studied models in the real world scenario.

Alternative models studied in [4], [5] and [6] were also considered appropriate for air pollution forecasting. The various models which were implemented by the authors include - spatio-temporal deep learning (STDL) based air quality predictors, generative adversarial networks (GANs), deep air learning (DAL) and convolutional neural network (CNN) model, spatio- temporal GRU-based prediction frameworks, change point detection Model with RNM (CPDM), sequential network construction model (SNCM), and self-organizing feature maps (SOFM). Authors of [7] also tested multiple error metrics for understanding the reliability of these models. But these models proved to be very complex from implementation perspective. Y. A.Ayturan et al in [4] inferred that these models required too much computational power and superior understanding of meteorological sciences for full scale implementation.

Additional literature such as [5] that used model also proposed in [1] and [2] were able to successfully impute missing data an

d outliers using a novel technique such as decision trees. This literature claimed 8 hours averaged surface ozone concentration prediction using deep learning methods.

Literatures [8], [9] and [10] proposed ideas that assumed air quality in a region is unaffected by sudden global changes in weather and air pollution. They proposed only LSTM models that predicted air quality by memorizing historical data. Authors in [7] compared this to the baseline support vector regressor used for predicting single time step. Hyper-parameter tuning was used effectively in these papers to deliver results on Beijing and CityPulse EU FP7 dataset respectively.

Study in [11] showed that deep air learning (DAL) model was able to solve the problems in the topics of interpolation, prediction, and feature analysis of fine- grained air quality and was better than peer models. But as [4] concluded that implementation of this model provided numerous challenges. Thus, the scope of [12] was intended for future implementations only.

Rizwan et al.[13] provided insights into air pollution in Delhi city of India which has air pollution which has exceeded the maximum limit by about 10 times as declared by world health organization. The

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

research also provided evidence for their study. They showed that the government of Delhi has taken several steps including introduction of metro, CNG for public transport buses but still there is no major improvement. Kelly and Fussel[14] have discussed the effect that air pollution can have on health of the individuals. They have also discussed the possible ways to reduce the air pollution in the coming future. Qu et al.[15] analysed the air pollution in Hong Kong using various visual analysis tools. They have integrated several techniques like s parallel coordinates and polar systems into their system.

There have been several attempts to monitor air pollution and predict it using technologies like Deep learning, Internet of Things etc. For instance, Dhingra et al. [16] proposed a air pollution monitoring system that operates in three phases. They also developed an Android application to display the air quality data from the sensors. Hu et al. [17] designed a low-cost participatory sensing system (HazeWatch). The system used combination of various tools and technologies which includes cloud computing, mobile sensors, smart phones etc. The research was carried in three phases. The phases include: Collection of data by using various methods, web based tools and mobile apps for visualization and validation of the methods adopted. Gu et al.[18] proposed a heuristic recurrent air quality predictor. One hour prediction model is applied recurrently in which for prediction of air quality one hour later then estimation was carried on for several hours. Boubrima et al.[19] proposed two optimization models which ensured pollution coverage and network connectivity with the minimum possible cost. The model was tested on dataset of Greater London. Hu et al. [20] introduced HazeEst. HazeEst is a model based on machine learning that combines mobile sensor data with fixed station data to estimate air pollution in Sydney at given time of the day. Phala et al [21]. Proposed an air quality monitoring system.The system measured pollutants: CO2, CO, NO2, and SO2 in real-time. A graphical user interface was also developed for ease of interaction for end users. Wand et al. [22] used technologies which include internet of things(IOT) and Neural networks. The research proposed a two-layer model which is based on Long Short Term Memory Neural Network and Gated Recurrent Unit. Jaimini et al. [23] conducted the experiment for analyzing the indoor air quality which is essential for Asthma management in children. The approach used was data driven. It covers impact of cooking, smoking etc in indoor pollution. Zhang et al. [24] processed high-dimensional large-scale data by using the LightGBM model. This model was used for forecasting data for predicting the air quality. The research was further extended by use of spatial data. Data was collected from the 35 air quality stations in Beijing

The general steps [Fig 1] followed in prediction of air pollution include:

- **Data Collection:** This is the most crucial and time consuming step in which data is gathered which will be used to train the model for prediction. The data is gathered using various sensors and devices.
- **Data Pre-processing**: In this step the collected raw data is converted in the form that can be used for prediction. Any missing data, corrupted data, repeated data is handled in this step.
- **Analysing data**: After pre-processing the data is analysed to bbring out any special patterns that can help in prediction. Also, after analyzing and understanding the data applying of deep learning models becomes relatively easier.
- **Training model:** In this step, the cleaned data after pre-processing is fed into the training data to learn features of data.
- **Testing the model:** Finally, the trained model is tested on the testing data which model has not seen previously.

3 Challenges

The objective of Based on the various literatures from year 2007 to 2019 discussed in section 3, some of the challenges
are explained below:

- The data used in [1] is not adequate with just 4 months of data samples. At least 1 to 2 years of data is required to cover all patterns in air pollution values.
- The grid size used in [1] is just 32x32 which could be quite small for a CNN model to find valuable patterns.
- Monthly average data used in the models in [2] and [9], do not take into account the

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

weather and pollution patterns during various seasons.

- In P.W. Soh et al [2], along with the weather and climatic factors, the authors could have also considered another feature like spatial data that affect PM2.5 components for better real time results.
- In literature [3], to enhance the accuracy of the model, it needs to be trained more in the future by adding features like satellite imaging that incorporate spatial data.
- Measurements of performance of the model in [5] were persistent in form, but the RMSE was always biased higher than the MAE measurement.
- The performance of the PM2.5 prediction model for air quality emergency management in [6] can be improved with the inclusion of more comprehensive information such as real time sensor data and weather API.
- The authors of [8] designed and implemented a very simple structure for the LSTM model. This model could not establish relationships between data samples captured at multiple locations within the city. Collection of data from various heterogeneous sources posed a challenging task during the pre-processing and analysis phase [8].
- Models implemented in [9] could not explain spatial patterns explaining co-relations between different locations within the dataset.
- The literature [10] focusses only on the empirical comparison of different LSTM variants and not implementing best predictive models.

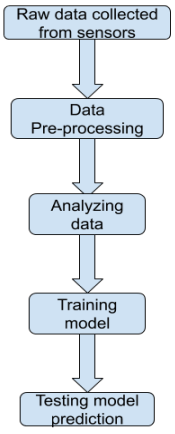

Fig 1. General steps followed in air pollution prediction

4 Results

The various air pollution models discussed in Section 2 were studied in terms of the accuracy of models using matrices available. The performance of the models are listed in the Table 1.

Table 1: Previous studies conducted for air pollution estimates.

Research Paper Reference Number	Model Used	Pollutants	Prediction Performance
[1][Fig 1]	LSTM and CNN	PM 2.5	74 %
[2] [Fig 2]	kNN-ED, kNN-DTWD, LSTM and CNN	PM 2.5	NA
[3]	RNN and LSTM	PM 2.5	RMSE 26.27 (8 hours) RMSE 27.28 (12 hours) RMSE 31.29 (24 hours)
[5]	GRU	PM 2.5	MAE 4.6147 MSE 15.7878 MAPE 6.29%
[7]	LSTM	O_3 Concentration	MAE 0.235 (24 hours)
[8]	LSTM	PM 2.5	RMSE 44.15 (5 hours) R2 0.689 (5 hours) RMSE 108.14 (120 hours) R2 -0.328 (120 hours)
[9]	LSTM	O_3 and NO_2	95%
[11]	DAL	PM 2.5	RMSE 0.0667 (1- 12 hours) RMSE 0.0877 (37-48 hours)

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

Table 1 shows the comparison between various models used in theory to predict the air pollutants along with the results. The models used include Knn, LSTM, CNN and other state of the art Machine Learning algorithms. The root mean square error(RMSE) depicts performance of these models.

5 Conclusion

In this study, various state of the art models were studied and compared to provide a clear view on the progress of air quality prediction in literature. Out of various machine learning and deep learning models, LSTM which is a deep learning model seems to give more promising results.
The study can be carried on by fine tuning the deep learning models and predicting the air pollutants. More data and faster computing power can further help in increasing the accuracy.

6 Future Scope

Based on literature review the future needs are:
1.Data quality and validation issue- There are lots of data quality issues that affect the accuracy of air quality evaluation and assessments due to device faults, multiple entries, and inconsistent formatting problems. There is a strong need for data quality assurance research. The data collected should be accurate as it forms the foundation of Deep learning models.
2.Real-time air quality monitor and supervision- As the advance of smart sensing and IoT, more and more environmental sensors have been installed and being used for many air monitoring resources. However, there is a lack of dynamic air quality evaluation which takes place in real-time and instant report result is produced and shared. Air in a city could be considered as a multi-level air system which is impacted by multiple factors like direction, temperature, location, time, wind speed etc. The air quality on all levels usually affects each other.

6 References

[1] V.Duc Le and S. Kyun Cha, “Real-time Air Pollution prediction model based on Spatiotemporal Big data.”, The International Conference on Big data, IoT, and Cloud Computing, URL: https://arxiv.org/abs/1805.00432 19-08-2019 14:00, pp. 6, 2018.

[2] P.W. Soh et al. , “Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations.”, IEEE Access, ISSN: 2169-3536, Volume: 6 , pp. 38186-38199, 2018.

[3] T. C. Bui et al. ,“A Deep Learning Approach for Air Pollution Forecasting in South Korea Using Encoder-Decoder Networks & LSTM.”, ArXiv, URL:

[4] https://arxiv.org/abs/1804.07891 19-08-2019 14:25, Volume: abs/1804.07891, pp. 6, 2018.

[5] B. S. Freeman et al. , “Forecasting air quality time series using deep learning.”, Journal of the Air& Waste Management Association, ISSN: 1047-3289, Volume: 68, Issue: 8, pp. 866-886, 2018.

[6] K. Ibrahim et al. , “A deep learning model for air quality prediction in smart cities.” IEEE International Conference on Big Data (Big Data), DOI: 10.1109/BigData.2017.8258144, pp. 1983-1990, 2017

[7] Y. A. Ayturan et al. , “Air Pollution Modelling with Deep Learning: A Review.”,International Journal of Environmental Pollution and Environmental Modelling, ISSN:2618-6128 Volume: 1, Issue: 3, pp. 58-62, 2018.

[8] X. Sun et al. , “Spatial-temporal Prediction of Air Quality based on Recurrent Neural Networks.”, Hawaii International Conference on System Sciences, DOI: 10.24251/HICSS.2019.155, pp. 10, 2019.

[9] V. Reddy et al. , “Deep Air: Forecasting Air Pollution in Beijing , China.”, URL: https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/deep-air-forecasting_final.pdf, 2017..

[10] S. V. Barai et al. , “Neural Network Models for Air Quality Prediction: A Comparative Study.”, Soft computing in industrial applications, ISSN: 1615-3871, pp. 290-305, 2007.

[11] G. Klaus et al. , “LSTM: A Search Space Odyssey.” IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2016.2582924, Volume: 28, Issue 10, pp. 2222-2232, 2017.

[12] Q. Zhongang et al. , “Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality.”, IEEE Transactions on Knowledge and Data Engineering 30, DOI:

May – June 2020
ISSN: 0193-4120 Page No. 1378 - 1383

10.1109/TKDE.2018.2823740, Volume: 30, Issue: 12, pp. 2285-2297, 2018.

[13] Rizwan, S. A. et al. ""Air pollution in Delhi: Its Magnitude and Effects on Health"." *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine* (2013).

[14] Kelly, Frank J. and Julia C. Fussell. "Air pollution and public health: emerging hazards and improved understanding of risk." *Environmental Geochemistry and Health* (2015).

[15] Qu, Huamin et al. "Visual Analysis of the Air Pollution Problem in Hong Kong." *IEEE Transactions on Visualization and Computer Graphics* 13 (2007): n. pag.

[16] Dhingra, Swati et al. "Internet of Things Mobile–Air Pollution Monitoring System (IoT-Mobair)." *IEEE Internet of Things Journal* 6 (2019): 5577-5584.

[17] Hu, Ke et al. "Design and Evaluation of a Metropolitan Air Pollution Sensing System." *IEEE Sensors Journal* 16 (2016): 1448-1459.

[18] Gu, Ke et al. "Recurrent Air Quality Predictor Based on Meteorology- and Pollution-Related Factors." *IEEE Transactions on Industrial Informatics* 14 (2018): 3946-3955.

[19] Boubrima, Ahmed et al. "Optimal WSN Deployment Models for Air Pollution Monitoring." *IEEE Transactions on Wireless Communications* 16 (2017): 2723-2735.

[20] Hu, Ke et al. "HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors." *IEEE Sensors Journal* 17 (2017): 3517-3525.

[21] Phala, Kgoputjo Simon Elvis et al. "Air Quality Monitoring System Based on ISO/IEC/IEEE 21451 Standards." *IEEE Sensors Journal* 16 (2016): 5037-5045.

[22] Wang, Baowei et al. "Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things." *IEEE Access* 7 (2019): 69524-69534.

[23] Jaimini, Utkarshani et al. "Investigation of an Indoor Air Quality Sensor for Asthma Management in Children." *IEEE Sensors Letters* 1 (2017): 1-4.

[24] Zhang, Ying et al. "A Predictive Data Feature Exploration-Based Air Quality Prediction Approach." *IEEE Access* 7 (2019): 30732-30743.