Forecasting Hazard Level of Air Pollutants Using LSTM's

Saba Gul and Gul Muhammad Khan(✉)

National Center of AI, University of Engineering and Technology, Peshawar, Pakistan
{sabagul,gk502}@uetpeshawar.edu.pk

Abstract. The South Asian countries have the most polluted cities in the world which has caused quite a concern in the recent years due to the detrimental effect it had on economy and on health of humans and crops. PM 2.5 in particular has been linked to cardiovascular diseases, pulmonary diseases, increased risk of lung cancer and acute respiratory infections. Higher concentration of surface ozone has been observed to have negatively impacted agricultural yield of crops. Due to its deleterious impact on human health and agriculture, air pollution cannot be brushed off as a trivial matter and measures must be taken to address the problem. Deterministic models have been actively used; but they fall short due to their complexity and inability to accurately model the problem. Deep learning models have however shown potential when it comes to modeling time series data. This article explores the use of recurrent neural networks as a framework for predicting the hazard levels in Lahore, Pakistan with 95.0% accuracy and Beijing, China with 98.95% using the time series data of air pollutants and meteorological parameters. Forecasting air quality index (AQI) and Hazard levels would help the government take appropriate steps to enact policies to reduce the pollutants and keep the citizens informed about the statistics.

Keywords: Air pollution · AQI · Forecasting · LSTM's

1 Introduction

Air pollution has been brushed off for quite some time as a trivial subject but the current research suggest the relentless damage it can cause humans and crop yield. In particular the cities of South Asian countries such as China and India have made to the list of most polluted cities and the cities of Pakistan are joining the list due to increase in levels of particulate matter and toxic fumes from the industries [17].

*Supported by NCAI, UET-P.

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 584, pp. 143–153, 2020.
https://doi.org/10.1007/978-3-030-49186-4_13

Exposure to higher concentration of surface ozone can trigger allergic reactions such as asthma and cause inflammation of air ways due to oxidative stress [11,12]. PM2.5 has been associated with 4 to 8% increase in cardiopulmonary diseases and lung cancer [10]. Air pollution has been linked to cardiovascular diseases in urban communities. Most of the hospitalized patients suffering from diseases like angina, myocardial infarction and heart failure have been put in such a situation, due to the long-term exposure to combustion-derived nanoparticles that incorporate reactive organic and transition metal components [13]. Moreover, studies suggest that high concentration of surface ozone has a detrimental effect on crop yield [14,15]. In recent years, due to increase in awareness of the bleak consequences of air pollutants, forecasting of air pollutants and their impact on human and crops has become an active area of research. Several deterministic and non-deterministic models were explored to model the behavior of pollutants [7,16]. Deep leaning models have had quite some success when it comes to modeling the problem and forecasting air pollutants. The meteorological parameters due to their conducive behaviour in pollutant dissemination and pollutant concentrations were used to forecast Hazard levels. Since the parameters used for modeling are time series, so the recurrent neural networks, Long Short Term Memory (LSTM) networks are employed due to their ability to accurately capture temporal trends [5–7].

The two major contributions of this article are the following:

1. Provide a dataset comprising of Lahore, Pakistan meteorological and pollutants statistics.
2. Employ deep learning model to develop a forecasting and classification system for assessing air quality.

2 Literature Survey

An LSTM model is trained in [1] on sensor data of Aerosol Optical Depth (AOD), meteorology and particulate matter which can provide quite accurate prediction of the concentrations of harmful gases (80% $PM_{2.5}$ variability). The system has been successfully deployed in Beijing, China and has helped in bringing down the pollution in Beijing by 23%.

A supervised regression model is developed based on historical data of air pollution in Sydney [2] which surpassed its contemporary ANN's in terms of accuracy in prediction and has high spatial resolution.

Forecasting air pollution is done through Multi-channel Ensemble framework through supervised extraction and learning which out performs its contemporary state of the art systems [3]. $PM_{2.5}$, PM_{10}, SO_2, CO, NO_x and ozone levels are predicted quite accurately.

In [4] attempts are made to model the complex relation between different parameters and its individual impact on pollutant levels using deep distribution fusion network while the spatial correlation is modelled using deep neural network. The system, deep air out performs ten state of the art baseline models

and achieves an average accuracy of 81.1%, 63%, 46% in 1–6 h, 7–48 h, sudden changes when deployed in 300+ cities of China.

Real time air-pollution predication is carried out in Daegu city, Korea [5] by processing the big data received from the air quality sensing modules installed on taxis. The spatial distribution of the pollutant levels is fed to a CNN model. For accurate processing of the temporal data; LSTM is used with a NN in parallel to cater for the meteorological factors effecting pollutant concentration. The testing results in an accuracy of 74% in real time over the data collected over a span of four months.

Spatial-temporal information is used in [6] to predict air quality using a combination of neural networks called ST-DNN which attempts to model the correlations between several meteorological conditions, elevation space and PM levels. LSTM is used to model long term temporal relations i.e. historical time series relation; CNN extracts the relationship between terrain information and pollutant levels while ANN is used with the current data and thereby models high frequency information. When evaluated on Taiwan and Beijing dataset, the network outperformed the baseline and comparative networks under consideration.

In order to enact policies to alleviate the pollution levels, accurate prediction is needed to carry out informed decisions. The temporal data of pollutants along with meteorological data is processed by a recurrent model [7], LSTM to forecast air pollutants since LSTMs have the ability to capture sequential relations. The frame work can predict air pollution 5–10 h in the future quite well but as the future time steps are increased beyond 10 h, we see degradation in performance. Since short term data of 6–10 h is needed to predict future time steps, power consumption can be reduced by turning the sensors on at specific intervals to collect data.

Artificial neural network (ANN) is used to predict PM_{10} concentration at 6 subways in Seoul, Korea [8]. Due to impracticality of monitoring PM_{10} directly at the crowded stations, PM_{10} concentrations are obtained from public data service near subway stations (PM_{10} out). In addition, it is observed that the shape and depth of the platform at the subway stations play an important role in influencing the model performance. The framework was able to predict PM_{10} concentrations at the platforms with an accuracy 67–80% depending upon parameters; inflow of PM_{10} (PM_{10} in), outflow of PM_{10} (PM_{10} out), ventilation operation, shape and depth of platform.

In [9], air pollution is forecasted using spatio-temporal data of city of Tehran, Iran obtained over a span of 10 years. Several machine learning methods such as; regression support vector machine, geographically weighted regression, artificial neural network and auto-regressive nonlinear neural network are evaluated on two datasets, one of which is cleaned via Savitzky-Golay filter while the other dataset was noisy due to missing entries. On both datasets, nonlinear autoregressive exogenous (NARX) neural network displays superior performance with exceptional performance over the former dataset.

3 Prediction Model Framework

In this article, we use a recurrent neural network that is; long short-term memory (LSTM), to capture the temporal trends of pollutant data. LSTM's perform better on sequential data as it takes the historical events into account by taking the output at instant t-1 as input in addition to inputs at t. This characteristic introduces the concept of Memory in neural networks which is of import when it comes to analyzing data of pollutants as it varies temporally.

Equation 1 and 2 describe the working of an RNN; where H is the tanh activation function, W defines the weight matrices between hidden and input layer (W_{xh}), hidden and hidden layer (W_{hh}), hidden and output layer (W_{hy}), x_t the input sequence, h_t the hidden vector of a module at instant t and b the bias to compute output y_t by iterating across these equations from t = 1 to T.

$$h_t = H(W_{xh}h_{xt} + W_{hh}h_{t-1} + b_h) \tag{1}$$

$$y_t = W_{hy}h_t + b_y \tag{2}$$

Though, RNN perform better when the sequences are short but suffer inherently from exploding gradient problem when working with data having long term dependencies. This problem is tackled by LSTM's which due do its gated memory architecture resolves the issue of vanishing and exploding gradients and is able to retain information for an extended period of time. Equation 3, 4, 5, 6 describes the input, forget and output gate and cell activation vectors of LSTM architecture respectively. Where σ is the sigma activation function.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{4}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{5}$$

$$c_t = f_t * c_{t-1} + i_t * tanh(W_{xg}x_t + b_g) \tag{6}$$

$$h_t = o_t * tanh(c_t) \tag{7}$$

3.1 Employed Datasets

The architecture was evaluated using two datasets, the modified UCI dataset published by [7] and on a dataset we introduced with parameters recorded in Lahore, Pakistan. The modified UCI dataset has meteorological data of wind speed, direction, air pressure, temperature, dew point, wind speed, cumulative rain hours and cumulative snow hours. Pollutant data of only $PM_{2.5}$ is recorded 25 times throughout the day. The parameters are collected over a span of 7 years with 43,825 samples from 2010 to 2017 across 35 different stations in Beijing, China. We have taken average of the data per day and on the basis of $PM_{2.5}$ concentration, we calculate the AQI value which is determined by the standard formula developed by environment protection agency (EPA), US. Based on the AQI value, a column of hazard level is added to the dataset. The information

of date, hour, day, month, and year in the dataset is removed and pre-processed using normalization.

We obtained the time series pollutants data from environmental protection agency (EPA) Punjab, Pakistan for a span of 2 years from 2017 to 2019. The data of air pollutants is received from 6 stations across the city which includes particulate matter (PM_{10}, $PM_{2.5}$), Nitrogen dioxide, Sulphur dioxide and surface ozone. The meteorological parameters play an instrumental role towards pollution dissemination and concentration in a particular region, thus the meteorological department of Pakistan was contacted to obtain the statistics of wind direction, temperature, barometric pressure, humidity, visibility and type of weather. The data of air pollutants and meteorological statistics are combined and pre-processed to form a dataset of 1500 samples for monitoring and predicting the hazard levels in the form of AQI. We have categorized the hazard into six levels according to the pollutants concentration defined by air quality index (AQI) values set by EPA, US as described in Fig. 1.

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Fig. 1. Air Quality Index set by environment protection agency, US

3.2 Network Architecture

The frame work comprises of three layers; a single LSTM layer followed by two dense layers with activations of Tanh and softmax respectively.

The network is evaluated on Lahore dataset by using metrics of sparse categorical cross entropy and accuracy. Batch size of 16 is used with adam as an optimizer and the network is trained for 300 epochs with a data split of 70/15/15 for training, validation and testing.

For modified UCI dataset, the network is trained for 300 epochs with a data split of 70/15/15, batch size of 8 and adamax as an optimizer. Python packages of Keras, tensorflow, Scikit-Learn and Pandas are used to model the network. Early stopping techniques are used by observing the loss on the validation data to reduce over-fitting by curtailing the training period.

3.3 Tuning Network Hyper-Parameters

The hyper-parameters of LSTM model is then tuned based on data to configure optimal parameters. Batch size, Numbers of training epochs, optimizer, learning rate and type of activation function are some of the hyper-parameters tuned by employing grid search algorithm (GSA) to improve performance of the model. We started with tuning the number of training iterations and batch size simultaneously. The model was modified based on these optimal hyper-parameters and the grid search algorithm was run again to find an appropriate optimizer. The model was then tuned based on these parameters to find an activation function that boosts the performance of the LSTM model using grid search algorithm.

According to the results of grid search algorithm, for Lahore dataset, the optimal hyper-parameters for the LSTM model are listed in Table 1, 2, 3 and 4. In Table 2, we select tanh as an activation as it gives better performance with all the other parameters tuned.

Table 1. Selection of training iterations and Batch size using GSA on Lahore dataset

Batch size	Epoch number	Accuracy
16	*300*	*0.92832*
8	350	0.92115
16	350	0.91398
8	300	0.91398
32	350	0.91039
32	300	0.91039
64	350	0.89964

Table 2. Optimal activation function selection using GSA on Lahore dataset

Activation function	Accuracy
softsign	0.92832
tanh	*0.92473*
hard sigmoid	0.92115
linear	0.92115
relu	0.91039
softplus	0.91039
sigmoid	0.90323
softmax	0.82079

The results of grid search algorithm for modified UCI dataset are tabulated in Table 5, 6, 7 and 8.

The optimized hyper-parameters highlighted in italics are reconfigured by incorporating early stopping criterion using the validation set which improves the performance of the model employed.

Table 3. Results of optimizer selection using GSA on Lahore dataset

Optimizer	Accuracy
Adam	*0.931899*
Adadelta	0.92473
Nadam	0.91756
Adamax	0.91398
RMSprop	0.91039
Adagrad	0.88889
SGD	0.55197

Table 4. Optimal learning rate selection using GSA on Lahore dataset

Learning rate	Accuracy
0.002	*0.935454*
0.001	0.921169
0.01	0.914026
0.2	0.896169
0.1	0.889026
0.3	0.462922

Table 5. Selection of training iterations and Batch size using GSA on modified UCI dataset

Batch size	Epoch number	Accuracy
8	*300*	*0.98609*
32	300	0.98609
64	500	0.98510
16	300	0.98411
64	300	0.98361
32	500	0.98312
16	500	0.98262
8	500	0.98213
8	100	0.97120

Table 6. Optimal activation function selection using GSA on modified UCI dataset

Activation function	Accuracy
tanh	*0.98709*
hard sigmoid	0.98560
sigmoid	0.98560
linear	0.98312
relu	0.98262
softsign	0.98262
softplus	0.98064
softmax	0.87388

Table 7. Results of optimizer selection using GSA on modified UCI dataset

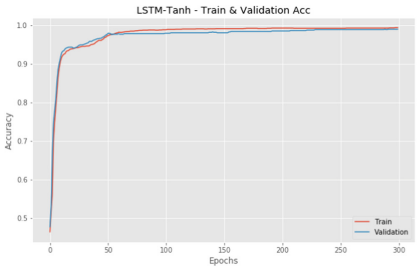
Optimizer	Accuracy
Adamax	*0.98759*
Adadelta	0.98461
RMSprop	0.97617
Adam	0.97567
Nadam	0.97368
Adagrad	0.95680
SGD	0.94836

Table 8. Optimal learning rate selection using GSA on modified UCI dataset

Learning rate	Accuracy
0.002	*0.972691*
0.001	0.951837
0.01	0.943893
0.1	0.936941
0.2	0.935452
0.3	0.767130

4 Result and Analysis

The hyper-parameters are tuned using grid search algorithm on the training set and on the validation data with respect to the categorical cross entropy error. It is observed that the model performs best with batch size of 8 with training of 300 epochs on the Beijing dataset and batch size of 16 with training of 300 epochs on Lahore dataset. Moreover, adam and adamax are employed as an optimizers for Lahore and Beijing datsets respectively which helps in convergence at a faster pace. Figure 2 shows that model when trained for 300 epochs on the modified UCI dataset attains a maximum validation accuracy of 98.9583% at epoch 288 and an accuracy of 98.95% on the test set. Thus the temporal characteristic of the data is modeled quite accurately using the recurrent network architecture. Figure 3 depicts the prediction model performance on the test set.



(a) Training and Validation accuracy

val_loss	val_acc	loss	acc	epoch
0.035727	0.989583	0.022682	0.993049	295
0.035713	0.989583	0.022635	0.993049	296
0.035673	0.989583	0.022582	0.993545	297
0.035627	0.989583	0.022534	0.993545	298
0.035617	0.989583	0.022476	0.993545	299

(b) Loss and accuracy results of training/dev set during the final epochs of modified UCI dataset

Fig. 2. Network training results on modified UCI dataset

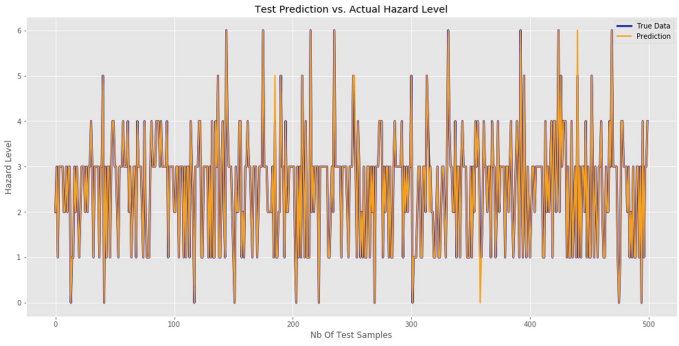
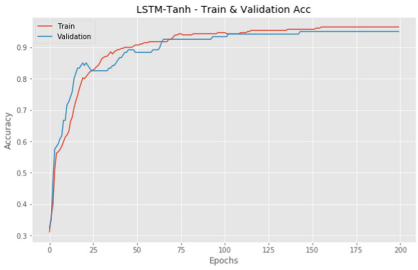


Fig. 3. Actual Vs. Predicted values of employed architecture on modified UCI dataset

The second dataset comprises of parameters recorded over a span of 2 years, thus after tuning the hyper-parameters, we train the network with a batch size of

16 for 300 epochs with early stopping criterion to avoid over-fitting. The results of training are described in Fig. 4 with maximum accuracy achieved at epoch 143 on the validation set. On the second dataset, an accuracy of 95.0% is achieved on the test set as depicted in Fig. 5. The deterioration in performance of the LSTM model for the Lahore dataset is due to the limited time series data required to infer the trends.



(a) Training and Validation accuracy

val_loss	val_acc	loss	acc	epoch
0.123860	0.95	0.117020	0.964158	195
0.123876	0.95	0.116676	0.964158	196
0.123878	0.95	0.116334	0.964158	197
0.123885	0.95	0.115992	0.964158	198
0.123915	0.95	0.115654	0.964158	199

(b) Loss and accuracy results of training/dev set during the final epochs

Fig. 4. Network training results on Lahore, Pakistan dataset

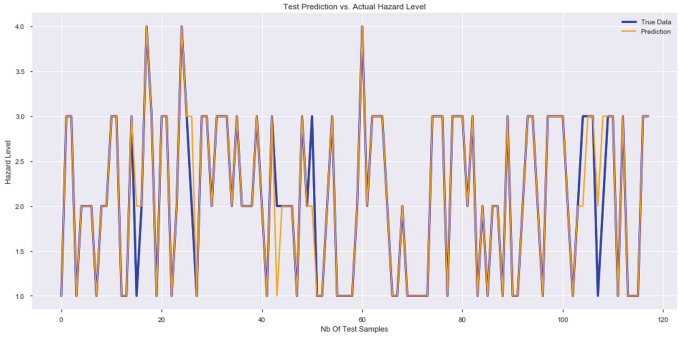


Fig. 5. Actual Vs. Predicted values of employed architecture on Lahore, Pakistan dataset

5 Conclusion

A model for forecasting hazard level has been devised and its performance is evaluated on the meteorological and pollutant data of two of the most polluted cities in the world; Beijing, China and Lahore, Pakistan. It is observed that

despite different topography and meteorological information, the proposed network models the complexity of the diverse temporal information quite well. The proposed architecture after employing GSA optimization is able to forecast the hazard levels of the next 24 h with an accuracy of 95.0% on the data recorded in Lahore, Pakistan and 98.95% on Beijing, China dataset due to ability of LSTM's to model temporal data and is thus able to learn the trends of air pollutants. This is an effective measure for the people going out to take necessary precautions and assist the environment protection agencies to enact policies and take steps towards reducing the health and economic risk caused due to high level of pollutants.

Acknowledgment. We would like to thank NCAI for funding this study. The modified UCI data-set employed in our study have been acquired from [7]. The second dataset was created by the data of pollutants taken from EPA lahore, Pakistan and meteorological parameters from Pakistan meteorological department.

References

1. Han, Y., Lam, J.C.K., Li, V.O.K.: A Bayesian LSTM model to evaluate the effects of air pollution control regulations in China. In: 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, pp. 4465–4468 (2018)
2. Hu, K., Sivaraman, V., Bhrugubanda, H., Kang, S., Rahman, A.: SVR based dense air pollution estimation model using static and wireless sensor network. In: 2016 IEEE SENSORS, Orlando, FL, pp. 1–3 (2016)
3. Zhang, C., et al.: Early air pollution forecasting as a service: an ensemble learning approach. In: 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, pp. 636–643 (2017)
4. Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y.: Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 965–973. ACM (2018)
5. Le, D.: Real-time air pollution prediction model based on spatiotemporal big data. arXiv preprint arXiv:1805.00432 (2018)
6. Soh, P., Chang, J., Huang, J.: Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access **6**, 38186–38199 (2018)
7. Reddy, V., Yedavalli, P., Mohanty, S., Nakhat, U.: Deep air: forecasting air pollution in Beijing, China (2018)
8. Park, S., et al.: Predicting PM10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). J. Hazard. Mater. **341**, 75–82 (2018). https://doi.org/10.1016/j.jhazmat.2017.07.05010.1016/j.jhazmat.2017.07.050. ISSN 0304–3894
9. Delavar, M., et al.: A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS Int. J. Geo Inf. **8**, 99 (2019). https://doi.org/10.3390/ijgi8020099
10. Pope III, C., et al.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA **287**(9), 1132–1141 (2002)
11. Kim, K.-H., Jahan, S.A., Kabir, E.: A review on human health perspective of air pollution with respect to allergies and asthma. Environ. Int. **59**, 41–52 (2013)

12. Kelly, F.J.: Oxidative stress: its role in air pollution and adverse health effects. Occup. Environ. Med. **60**(8), 612–616 (2003)
13. Mills, N.L., et al.: Adverse cardiovascular effects of air pollution. Nat. Rev. Cardiol. **6**(1), 36 (2009)
14. Chuwah, C., van Noije, T., van Vuuren, D.P., Stehfest, E., Hazeleger, W.: Global impacts of surface ozone changes on crop yields and land use. Atmos. Environ. **106**, 11–23 (2015)
15. Lin, Y., et al.: Impacts of O3 on premature mortality and crop yield loss across China. Atmos. Environ. **194**, 41–47 (2018)
16. Bai, L., Wang, J., Ma, X., Haiyan, L.: Air pollution forecasts: an overview. Int. J. Environ. Res. Public Health **15**(4), 780 (2018)
17. World air quality report. https://www.iqair.com/world-most-polluted-cities