A recurrent neural network forecasting technique for daily PM$_{2.5}$ concentration level in Southern Kerala

Sherin Babu$^1$, Binu Thomas$^2$

$^1$Assistant Professor, Department of Computer Science, Assumption College Autonomous, Changanacherry, Kottayam, Kerala, India
$^2$Associate Professor, Department of Computer Applications, Marian College Autonomous, Kuttikkanam, Idukki, Kerala, India

Abstract. The natural environment and public health are seriously impacted by air pollution. One of the most dominant contributors to air pollution around the world is particulate matter PM$_{2.5}$. Predicting air pollution in advance has considerable importance for the regulation of people's health and to implement pollution control strategies for air quality management. The study was conducted for Thiruvananthapuram district, the southernmost region of Kerala. The data for the period from 1st July 2017 to 31st December 2019 were collected from the Central Pollution Control Board (CPCB) website. To predict the daily PM$_{2.5}$ concentration, Recurrent Neural Network (RNN) based Long Short-Term Memory (LSTM) was used here. LSTM was built on the inputs of four meteorological parameters, namely average temperature, wind speed, wind direction and relative humidity and air pollutant parameter PM$_{2.5}$ values. Performance evaluation of the prediction model has conducted, and the results showed that the model attained considerable prediction accuracy.

1. Introduction

Air pollution means contamination of the air in the earth's atmosphere by harmful gaseous substances, particulates and biological molecules [1]–[3]. Air pollution is considered a severe threat to public health across the globe. As a result of civilization advancement, globalization and rapid industrial development, the impact of human activities on the natural environment is rising and, unfortunately, in most cases, unfavourable for the existence of the environment. Both temporal and spatial properties are dependent on air pollution. The key one is factors related to meteorology, which vary in spatiotemporal form. The temperature, humidity, amount of rainfall, wind speed, wind direction, etc. make air pollution differ from place to place. Traffic density and traffic congestion are other important causes of air pollution. The zone with a higher level of traffic has lower ambient air quality. Particulate matter is the sum of all solid and liquid particles suspended in the air such as dust, pollen, soot, smoke and liquid droplets[4]. PM$_{2.5}$ refers to the atmospheric particulate matter that has a diameter of fewer than 2.5 micrometres, which is about 3% of the diameter of human hair. PM$_{2.5}$ particles in the air reduce visibility and cause the air to appear hazy when its levels are elevated. PM$_{2.5}$ particles can penetrate deeply into the lungs, and hence exposure to high concentrations of PM$_{2.5}$ will cause respiratory and cardiovascular diseases. Several studies indicated that long-term exposure to PM$_{1.5}$ has resulted in elevated mortality rates. To control parameters within reasonable limits, prediction mechanisms for
pollutant concentration PM$_{2.5}$ in advance have become necessary [5]. We need to carry out the PM$_{2.5}$ prediction to provide an early air quality warning.

2. Related Work

Several PM$_{2.5}$ prediction methods are developed by researchers, based on statistic models and machine learning techniques. A deep learning model to forecast PM$_{2.5}$ concentrations based on an auto-encoder and bidirectional long-term short-term memory (Bi-LSTM) was explored to demonstrate the association between PM$_{2.5}$ and multiple climate variables by Zhang et al.[6]. Wen et al. proposed a spatiotemporal convolutional long-term short-term memory neural network extended (C-LSTME) model to predict PM$_{2.5}$ concentration in China [7]. Park et al., developed a convolutional neural network (CNN) model for estimating daily 24-h averaged ground-level PM$_{2.5}$ values in United States of America, by incorporating aerosol optical depth (AOD) data, meteorological fields, and land-use data [8]. Kaur and Mandal established that non-linear autoregressive network with exogenous input (NARX) is superior in predicting daily PM$_{2.5}$ concentrations of Delhi area, compared to other neural network models [9].

A hybrid deep learning method based on multiple attention-LSTM neural networks was presented by Yuan et al., to predict the PM$_{2.5}$ concentration in Hefei city [10]. Xing et al. created a deep learning model termed as temperature-based deep belief networks (TDBN) to predict the daily concentrations of PM$_{2.5}$, in Chaoyang Park in Beijing of China [11]. Geetha and Prasika developed a deep learning technique Long-Short Term Memory (LSTM) used for predicting various air pollutants, including PM$_{2.5}$ [12].

Lu et al. developed novel deep belief network (DBN) integrated with multilayer restricted Boltzmann machines and a single-layer back-propagation network model for the haze level prediction. This model used both air quality and the meteorology data, and then the haze prediction was carried out by using a competitive adaptive-reweighed method [13]. Another way aimed at PM$_{2.5}$ forecasting is developed by Zhou et al. Here Multivariate Bayesian Uncertainty Processor (MBUP) is combined with an artificial neural network (ANN) to perform the PM$_{2.5}$ prediction [14]. A combination of Convolution Neural Network and Long Short-Term Memory, called the ConvLSTM model, was proposed by Le et al., which automatically manipulates the spatial and temporal characteristics of the PM$_{2.5}$ air pollution data in Beijing city [15]. In [16], the authors claimed that the LSTM model outperformed all other neural network models in forecasting hourly PM$_{2.5}$ values in two big cities of South Korea. Ma et al. proposed a Geo-LSTM that have taken into account of the spatial-temporal correlation of air pollutant concentrations from different monitoring stations and then run the forecast model to predict the PM$_{2.5}$ values [17].

3. Methodology

This work aimed to develop an RNN model for predicting the daily PM$_{2.5}$ value for the southern part of Kerala. RNN is a type of deep learning based artificial neural network designed to process sequential data and recognize patterns in it. The Recurrent Neural Network is a type of feedforward neural network with an internal memory capable of processing input sequences. RNNs remember the inputs and the context as they have internal memory, enabling users to have more flexibility in the types of data that models or networks can process. RNNs are a powerful tool when data is sequential, and the next data point depends on the previous data point [18], [19]. Since they understand the context in a sequence, RNNs can produce better results/predictions. Several studies show that the current PM$_{2.5}$ concentration has a clear association with past results.

A type of recurrent neural networks is Long Short-Term Memory (LSTM) networks, which makes it easier to recall past data in memory. Hochreiter & Schmidhuber introduced LSTM to prevent the issue of capturing long-term dependencies by an RNN [20]. In each unit of the hidden layer, the LSTM network comprises a memory block containing three types of gate functions, namely the input gate, forget gate, and output gate. A block operates on an input sequence and each gate within a block uses
the sigmoid activation function to control whether or not they are activated. By using back-propagation, it trains the model. As current pollutant concentrations are influenced by the past concentration of pollutants and past meteorological factors, LSTM was used to derive the time-series features. Also, LSTMs are popular and efficient for analyzing sequential data and time series, and it is used in this research study. Before feeding the data to the RNN model, data collection, analysis for feature correlation, and data preprocessing were carried out to get better prediction from the RNN model. All these data were collected from the Central Pollution Control Board (CPCB) website. The data collected contained daily PM$_{2.5}$ values and various meteorological parameters. The best meteorological parameters for the study were selected using correlation method. Then data preprocessing was carried out to handle missing data in the dataset. After that, by setting the best hyperparameters, the LSTM-RNN model was constructed and optimized. Then, the model was trained on the training dataset and then predictions were made on the test dataset. Finally, the model was evaluated using statistical measurement parameters to verify it’s performance.

3.1. Study Area
This study is conducted for the southernmost district of Kerala named Thiruvananthapuram. For the analysis, the monitoring station located at Plammoodu (Latitude: 8.51N, Longitude: 76.94), Thiruvananthapuram, Kerala's capital, was selected. The Kerala Pollution Control Board owns and maintains this station. The rapidly rising urban areas and the development of housing, business and industrial units have caused the air to become increasingly polluted, thus raising the area's pollution rate.

3.2. Data Set
The dataset consists of the daily values of meteorological parameters and daily values of PM$_{2.5}$ pollutant concentrations (μg/m$^3$). The meteorological parameters were selected after correlation based feature selection process. The parameters chosen are average temperature, wind speed, wind direction and relative humidity. The data of 914 days, from 1st July 2017 to 31st December, 2019 have been collected. The details of the data set are listed in Table 1.

| Location                  | Plammoodu, Thiruvananthapuram |
|---------------------------|-------------------------------|
| Unit                      | Daily                         |
| Dataset span              | 914 days                      |
| Meteorological Parameters | Average temperature, wind speed, wind direction and relative humidity |
| Air pollution parameter   | PM$_{2.5}$                    |

The actual distribution of PM$_{2.5}$ values over the above specified period is given in figure 1. The figure showed that the peak values of PM$_{2.5}$ are observed in the winter season, i.e., from November to January. It is also noted that the concentration of PM$_{2.5}$ is small during the rainy season, i.e., in the months of June, July and August.
Clearly, it can be stated that the PM$_{2.5}$ distribution is not stationary, and the data values are time-dependent. Year-wise distribution of PM$_{2.5}$ in 2017, 2018 and 2019 are given in figure 2, figure 3 and figure 4 respectively. From the year wise distribution of PM$_{2.5}$ depicted in the plots, showed that there is slight decline in the daily PM$_{2.5}$ values over the years from 2017 to 2019. This is a positive indication that air pollution in the area is getting reduced.

3.3. Data Pre-processing
The data is pre-processed before it is given for RNN model training. The pre-processing steps used in this study are described here:

Missing value handling - The dataset contained missing values due to instrument malfunction, power failure or connection issues. Those values are imputed using iterative imputation model. Iterative imputation refers to a process where each feature is modelled as a function of the other features. Each function is sequentially imputed, one after the other, enabling previous imputed values to be used in the prediction of subsequent features as part of a model.

Feature scaling - It is a method used to normalize the range of independent variables or features of data.
The MinMaxScaler class is used to transform the data. On the training dataset values, the scaling coefficients called Min and Max values are determined and then applied to scale the test dataset. The default value range returned by MinMaxScaler is 0 to 1. Google Colab environment is used for developing the model. In the model, there are five layers in total, including an input layer, three intermediate layers and an output layer. Activation function used here is relu and the optimizer used is Adam. The output layer contained only one node that produces daily PM$_{2.5}$ values. The whole PM$_{2.5}$ dataset is divided into training dataset and test dataset. Out of 914 total instances, 80% of data (731 cases) is contained in the training dataset, and 20% data (189 cases) is used for testing the model. The LSTM model is made to fit on all of the training data. Here the LSTM is made to run for 1500 epochs. Then at each step, a prediction is made based on the test data. The PM$_{2.5}$ concentration has a range from 1 to 96, and its histogram plot is as shown in figure 5. The histogram is right-skewed.

![PM 2.5 Histogram Plot](image)

**Figure 5.** PM$_{2.5}$ values and frequency – histogram plot

### 4. Results and Discussion

After the training process, the model is used for making predictions. As the first stage, the model is trained on the training data, and then in the next stage, for each test record in the test data set, PM$_{2.5}$ value is predicted. The testing data are used to predict the PM$_{2.5}$ concentration values for a given day, and to test the performance of the proposed model. The root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient ($R^2$), and mean absolute percentage error (MAPE) are employed to evaluate the performance of the model and the prediction accuracy. $R^2$ denotes the coefficient of determination, which is a statistical measure that denotes the amount of the variation for a dependent variable on an independent variable in a regression model. RMSE is a standard way to measure the error of a model in predicting quantitative data. The mean absolute error of a model is the mean of the absolute values of the individual prediction errors over all instances in the test data set. The models with lower RMSE and the higher $R^2$-values are said to be acceptable. Table 2 shows the values of the above said evaluation parameters on the LSTM model.

| Model | $R^2$  | RMSE  | MAE   |
|-------|--------|-------|-------|
| LSTM  | 0.872  | 7.618 | 5.106 |

Table 2. Performance of the LSTM PM$_{2.5}$ prediction model.
This is the initial machine learning-based research study conducted for the estimation of air pollutant PM$_{2.5}$, done in the Southern region of Kerala. Also, the study utilized the correlation of meteorological parameters in predicting the air pollutant concentrations.

5. Conclusion
The findings demonstrated that the LSTM model could be used with proper consistency to predict the daily PM$_{2.5}$ concentration in the Thiruvananthapuram area. This forecasting model data can help the government officials to take the appropriate steps and provide the public with early notices to take safety precautions. As this is a preliminary research work carried out for the estimation of PM$_{2.5}$ using basic LSTM model, there exist opportunities for utilizing other advanced RNN models for PM$_{2.5}$ estimation. The model can be enhanced by the consideration of the multi-station data, different air pollutant concentrations and meteorological data.

6. References
[1] Brunekreef Bert and Stephen T Holgate 2002 *The lancet* 360 9341 1233-42
[2] Yin P, Guo J, Wang L, Fan W, Lu F, Guo M, Moreno S B, Wang Y, Wang H, Zhou M and Dong Z 2020 *Environmental Science & Technology Letters* 7 2 95-101
[3] Madrigano J, Kloog I, Goldberg R, Coull B A, Mittleman M A and Schwartz J 2013 *Environmental health perspectives* 121 2 192-6
[4] Wu X, Wang Y, He S and Wu Z 2020 *Geoscientific Model Development* 13 3 1499-151
[5] Masood A and Ahmad K 2020 *Procedia Computer Science* 167 2101-10
[6] Zhang B, Zhang H, Zhao G and Lian J 2020 *Environmental Modelling & Software* 124 104600
[7] Wen C, Liu S, Yao X, Peng L, Li X, Hu Y and Chi T 2019 *Science of the total environment* 654 1091-9
[8] Park Y, Kwon B, Heo J, Hu X, Liu Y and Moon T 2020 *Environmental Pollution* 256 113395
[9] Kaur M and Mandal A 2020 *Proc. Int. Conf. on Contemporary Computing and Applications* (Lucknow/IEEE) pp 40-3
[10] Yuan H, Xu G, Lv T, Ao X and Zhang Y 2020 *Analytical Letters* 1-2
[11] Xing H, Wang G, Liu C and Suo M 2020 *Neural Networks* 133 157-65
[12] Geetha S and Prasika L 2019 *International Journal of Mechanical Engineering and Technology* 10 1 1026-32
[13] Lu H, Song J, Di T, Kurdestany J M and Wang H 2018 *Tehnički vjesnik* 25 2 519-27
[14] Zhou Y, Chang LC and Chang FJ 2020 *Science of The Total Environment* 711 134792
[15] Le VD, Bui TC and Cha SK 2020 *Proc. Int. Conf. on Big Data and Smart Computing* (Pusan/IEEE) pp 55-62
[16] Qadeer K, Rehman WU, Sheri AM, Park I, Kim HK and Jeon M A 2020 *Applied Sciences* 10 11 3984
[17] Ma J, Ding Y, Cheng JC, Jiang F and Wan Z 2019 *Journal of Cleaner Production* 237 117729
[18] Mandic D and Chambers J 2001 *Recurrent neural networks for prediction: learning algorithms, architectures and stability* Wiley
[19] Rather AM, Agarwal A and Sastry VN 2015 *Expert Systems with Applications* 42 6 3234-41
[20] Hochreiter S and Schmidhuber J 1997 *Neural computation* 9 8 1735-80

Acknowledgement
We thank the Central Pollution Control Board (CPCB), Government of India for providing the dataset, needed to carry out this work.