IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Spatial-temporal, terrain forecasting of air quality model by multiple Deep Neural Networks

S. Jeya[1*], L. Sankari [2]

[1, 2] Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India

Corresponding Author: jeya.s.2000@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.472477 | Available online at: www.ijcseonline.org

Accepted: 20/May/2019, Published: 31/May/2019

***Abstract*—**In order to maintain the air quality, continuous monitoring and analysis of the air pollution data is necessary; especially in areas where industrial and vehicular emissions contribute more to poor air quality. Inhalation of high concentration of fine particulate matter (PM2.5) causes lung, heart and various other diseases, which increase hospital visits and mortalities each day. The focus of this paper is to analyse the historical pollution data and corresponding meteorological data from selected areas, and to forecast PM2.5 over the next 48 hours by using multiple neural networks. In this proposed model, experiment is conducted by including pollutant and meteorology data recorded for every hour from 14 places, which includes northern, southern, western and eastern parts of India. Spatial temporal relations and terrain impact are then extracted. The proposed system applies multiple neural networks including convolutional neural network, artificial neural network, long short-term memory and adaptive neuro-fuzzy inference system to predict the air quality. The proposed model - ANFIS prediction performance - is better than the existing ANN model.

***Keywords*—**Convolutional neural network; LSTM; adaptive neuro-fuzzy inference system; air quality forecast; dynamic time warping; Euclidean distance.

I. INTRODUCTION

PM2.5 is a tiny particle of about diameter 2.5 micro meters, and originates from various sources like dust storms, forest and agricultural fires, vehicular emissions, industrial emissions, biomass cooking etc.,[1]. Due to their small size, they penetrate deep into the lungs during breathing. Long-time exposure of PM2.5 causes ailments to humankind like lung cancer, COPD (chronic obstructive pulmonary disease), heart diseases, asthma and emphysema. Ambient particulate matter annual exposure, as a population weighted mean in India in 2017, was 89·9 μg/m³, one of the highest in the world [2]. Forecasting air quality is essential to find patterns, extract information which would help guiding public from exposure to PM2.5, by reducing the individual's outdoor or indoor activities. This study forecasts 48 hours of PM2.5 density from historical data. The existing model spatial-temporal deep neural network (ST-DNN), to forecast 48-hour air quality by applying artificial neural network along with Long Short Term Memory (LSTM) to extract spatial temporal relationship and terrain impact on air quality prediction by CNN was proposed by Ping-Wei Soh et al. The existing model applied K-Nearest Neighbour by Euclidean Distance (KNN-ED) to extract spatial relationship among related locations, by using latitude longitude co-ordinates on Taiwan and Beijing datasets. The model used KNN Dynamic Time Warping (KNN-DTW) to extract temporal dependency and terrain impact on air quality, by using terrain related information for the area around the locations selected for this study. Finally, artificial neural network predicts the 48-hour forecasting. Due to high cyclicality of PM2.5, it is necessary to explore the correlation with space on one hand and temporal dependency at a given location on the other hand in order to precisely model the PM diffusion [3]. In this proposed model hourly recorded meteorological data such as wind speed and wind direction, relative humidity, temperature and pollutants such as PM2.5 and PM10 from north, east, south, west part of India are taken for analysis. Data related to elevation space is also included. This study extracts temporal information, spatial relationship and terrain information for the area around the locations and applied multiple neural network architectures LSTM, ANN, CNN and ANFIS for predicting the air quality. The results of the proposed ANFIS model have shown better prediction performance when compared to the existing ANN model. The organisation of this paper is as follows. Section II deals with related study, section III identifies the problem, narrates overall framework of the proposed system, materials and methods used, section IV Illustrates the results, section V concludes the paper.

II. RELATED STUDY

Ping-Wei Soh applied artificial neural networks to forecast the air quality up to 48 hours. General predictive model called spatial-temporal deep neural network (ST-DNN) was applied to meteorology and PM2.5 data and the model extracted the spatial relationship by k-Nearest Neighbour - Euclidean Distance (KNN-ED) method to calculate location. Temporal distance by Dynamic Time

Warping Distance (KNN-DTWD) along with KNN was used to identify the most similar temporal behaviour locations and terrain impact on air quality by using Taiwan and Beijing data sets. Air quality correlations for similar locations are explored, and temporal dependencies at a given location are identified by multiple neural networks. Long short term memory is used to model historical time series behaviour to help capture target location time series trends, while artificial neural network uses only current data, which is sensitive to rapid changes. The following processes used in this existing model are 1. Temporal information of the target location. 2. Related locations spatial relations. 3. The area around the locations terrain information. Final forecast is done by artificial neural network [3].

Chiou-Jye Huang et al. experimented convolutional neural network (CNN) and Long Short Term Memory (LSTM) to forecast PM2.5. Historical data of cumulated hours of rain, cumulated wind speed, PM2.5 concentration, are all used for developing the deep neural network model which combines CNN and LSTM architecture to forecast the next hour concentration of PM2.5. MAE and RMSE were lowest when compared with the other models such as SVM, RD, DT and MLP [4].

Athira V et al. applied recurrent networks such as Recurrent Neural Network (RNN), along with LSTM and Gated Recurrent Unit (GRU) on AirNet data to forecast PM10 concentration based on pollution and meteorological time series AirNet data. For their analysis, they have gathered air quality data from China National Environmental Monitoring Centre, and meteorological data gathered from Global Forecasting System. GRU performance is slightly better than the other two [5].

Guyu Zhao et al. established air quality spatiotemporal model by combing the temporal and spatial correlation. Diffusion of pollutants is influenced by spatial distance and wind. The air quality spatiotemporal network model is used to explore the local similarity and regional interaction by community detecting algorithms. Their contribution is mainly to identify the source and spreading of pollutants [6].

A. Nazif et al. conducted the analysis with the daily average PM10, temperature (T), humidity (H), wind speed and wind direction data for 5 years (2006–2010), from two industrial air quality monitoring stations, for seasons during south-west monsoon and north-east monsoon from two different states in Malaysia, and used principal component analysis (PCA), lognormal regression (LR), multiple linear regression (MLR) and principal component regression (PCR) to forecast next-day average PM10 concentration level. PCR has better predictability and lower error rate than LR and MLR [7].

III. MATERIALS AND METHODS

The proposed system with geographical coordinates calculates Euclidean distance between locations and identifies most strongly related locations with influential spatial-temporal relationships to the target location. Air quality at one location can be spatially correlated with another location. Not only local emissions, but also emissions from the surrounding areas impact the air quality of a particular location. Dynamic time warping method is used to extract the locations with highly related temporal relationship to target location and is also used to calculate the time series feature distances between locations. After sorting distances, top k most similar locations are used to predict target location sequence. For low and high frequency information, LSTM and ANN are used in order to match trend and rapid changes in the input data. Terrain related data are included and CNN is used to extract the obscured terrain relationships. Finally, the ANFIS model is used for predicting the PM2.5.

For the proposed predictive model, related spatial and temporal parameters are identified by using KNN-ED (K-Nearest Neighbour by Euclidean Distance) and KNN-DTWD (K-Nearest Neighbour by Dynamic Time Warping Distance). Terrain information for the area surrounding the locations are also extracted, because PM2,5 distribution has a close relationship with elevation. A prediction model is framed by multiple neural networks CNN, LSTM, ANN and ANFIS.

A. K-Nearest Neighbour by Euclidean Distance (KNN-ED)

KNN is a non-parametric conventional classifier, which generally calculates distance between two data points by using Euclidean distance function. Geographical co-ordinates such as latitude and longitude of the locations are used for Euclidean distance calculation [8].

$$\text{dist(A,B)}=\sqrt{\frac{\sum_{i=1}^{m}(x_i-y_i)^2}{m}} \qquad (1)$$

B. K-Nearest Neighbour by Dynamic Time Warping Distance (KNN-DTWD)

DTWD is used to identify similarity and Global optimal alignment between two time series. DTW is a time series alignment algorithm used to find an optimal alignment between two given (time-dependent) sequences. By mapping data points to corresponding intervals, DTW is used to exploit temporal distortions between two time series and by the use of distance matrix, best alignment is achieved [9]. DTW eliminates time shift and scaling effects. DTW identifies the most strongly related temporal relationships to the target location based on historical data, and time series feature distances between locations are then calculated. After sorting the distances, top k most similar locations are chosen for the target location sequence [3].

$$\text{DTW(X,Y)}=\min\{\sqrt{\sum_{k=1}^{k} w_k/K} \qquad (2)$$

Adjacent locations with similar temporal patterns have high correlation with target location. These become the inputs to LSTM and ANN. In order to extract the terrain related features, 42 square sections of terrain data are used. CNN is used to extract the obscured terrain relationship. Figure 1, shows the input output relationships between various neural network models in the architecture.

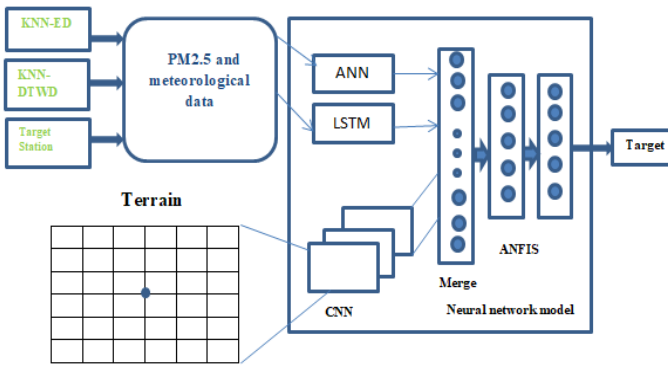

Fig. 1- Proposed predictive model architecture

.

C. Data Description

As per the WHO report, the impact of pollution is heavy in low- and middle-income countries, where the quality of air exceeds the World Health Organisation's limits. Fourteen locations are selected from India representing northern, southern, western and eastern parts of the country. The dataset contains the meteorology data, along with PM2.5 concentration, on hourly basis from Jul 2018 to Dec 2018 for this multidimensional time series forecasting. The features of the dataset are PM2.5, PM 10, Barometric Pressure, Temperature, Relative humidity, wind speed, wind direction, solar radiation and atmospheric temperature.

D. Data pre-processing

Sensor recorded data generally have missing values due to various reasons. Pre-processing the data is an important and preliminary step for any data analysis in order to avoid ambiguity. Data cleansing for missing or null values, by proper imputing technique is important as the missing data can contribute to skewed results for the model. For imputing the missing values, Mean Imputation method is applied.

Table I - National air quality index-CPCB

AQI	Remark	Colour Code	Possible Health Impacts
0-50	Good		Minimal impact
51-100	Satisfactory		Minor breathing discomfort to sensitive people
101-200	Moderate		Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor		Breathing discomfort to most people on prolonged exposure
301-400	Very poor		Respiratory illness on prolonged exposure
401-500	Severe		Affects healthy people and seriously impacts those with existing diseases

E. Convolutional neural network

For time series data, forecasting through Long Short Term Memory is considered as the state-of-the-art model due to its recurrent connections. A convolutional neural network with multiple layers of dilated convolutions, in which filters are applied, excels in training and prediction performance, due to the small number of trainable weights.

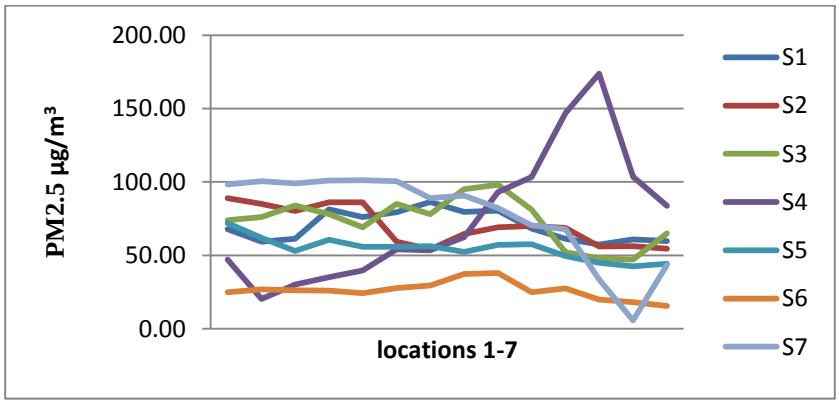

Fig. 2-Time series hourly PM2.5 data for seven locations

Application of Convolutional Neural Network for time series prediction will perform well on datasets that have a spatial relationship. CNN layers comprise of convolutional layer, pooling layer, RELU layer and fully connected layers. CNN uses convolution, instead of general matrix multiplication in at least one of its layers. Usually a CNN model has three stages which are learnable filters each performing a convolution in parallel. The second stage is an element-wise non-linearity like fully-connected layer. The third stage is called pooling. Pooling is nothing but down sampling the output vector of the second stage. Altogether, a convolutional layer tries to find the local patterns in the input [10].

F. Long Short Term Memory (LSTM)

Persistence of information differentiates traditional neural networks with the recurrent neural networks. RNN internal memory's capability of remembering the inputs helps in precise prediction of the future. With the help of loop, information is passed from one step to the next in recurrent neural networks.

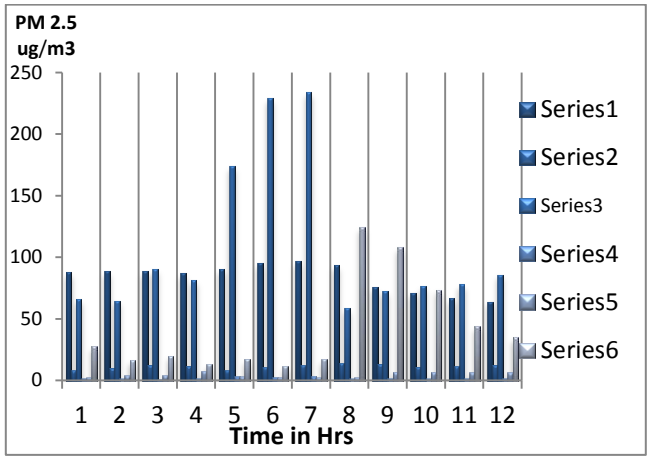

Fig. 3 - Similar locations pollutant prediction

At the same time to overcome the short term memory problem faced by Recurrent Neural Networks, LSTM with gates which can regulate the flow of information in order to carry information from earlier time steps to later ones are very much useful. LSTM are special kind of recurrent neural networks capable of learning long-term dependencies. Remembering information for longer periods make them tremendously good at solving variety of problems. RNN's remember their inputs over a long period of time with the help of LSTM. The three gates of LSTM are input, forget and output gate. Vanishing gradients is solved through LSTM, it keeps the gradients steep and therefore the training is relatively short and the accuracy is high [11].

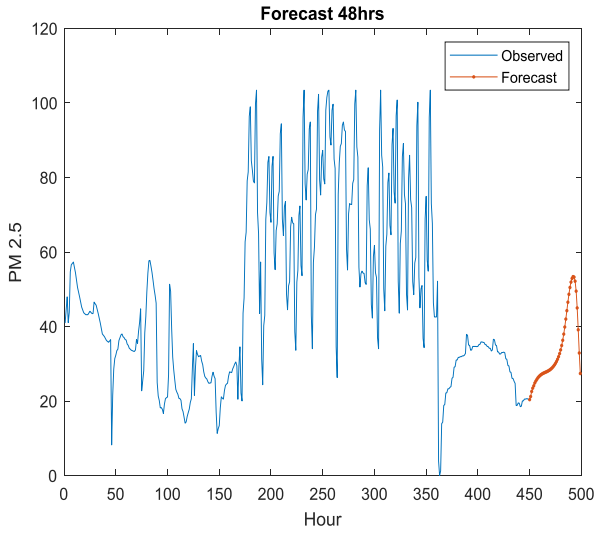

Fig. 4 – Forecasting PM2.5 48 hours

LSTM is perfectly suitable for training time-series/sequential data. Elementwise sigmoid function, σ, and hyperbolic tangent function is used to scale each element of the gate vector, with a value between 0 and 1. Forget gate (ft) is responsible to forget information which is not important from memory cells, based on the error propagation. Memory cell state is maintained over long sequences by forgetting old information and combining new information.

Input gate:

$$it = \sigma i(WxiXt+Whiht-1+bi) \quad (3)$$

where xt is the input tensor, ht is the output tensor, and w,b are the weight and bias functions respectively.

Forget gate:

$$ft = \sigma f(WxfXt+Whfht-1+bf) \quad (4)$$

Wxf: weight matrix from input x, ft is a forget gate, Xt input x at time t

Whf: weight matrix from previous hidden vector ht−1

bf: forget gate bias.

Output gate along with hidden vector:

$$ot = \sigma o(WxoXt+Whoht-1+bo) \quad (5)$$

$$ht=ot*\tanh(st) \quad (6)$$

LSTM with previous information, updates the current

state and also computes the gradient by truncating Back Propagation through time [12].

G. *Terrain extraction*

Terrain features greatly influence different locations. Exploring the relationship between barriers and altitude differences on one hand and different locations on the other hand, is done by using terrain related data to improve location correlations and to extract hidden terrain relationships between locations. Convolutional neural network is used to extract obscured terrain relationship and spatial correlations between locations.

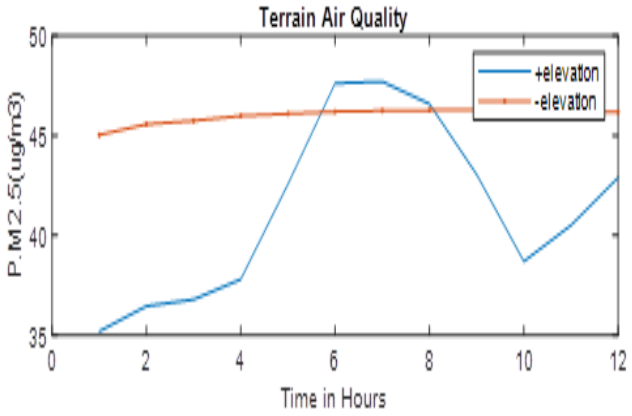

Fig.5 - Terrain elevation impact on PM2.5.

IV. RESULTS

Model performance is measured by Root Mean Square Error (RSME) - a standard statistical metric error in air quality, meteorology and climate studies. Another metric is Mean Absolute Error (MAE) which is also used for model evaluation.

$$MAE = \frac{1}{n} = \sum_{j=1}^{n} |y_i - \hat{y_i}| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2} \quad (8)$$

Predictor accuracy can be measured by measuring the difference from the predicted value and actual known value, i.e. how far off from each other. The test error is the average loss over the test set. The comparison is done for prediction accuracy by mean absolute error and root mean squared error, and the proposed method has low MAE and RMSE and thereby shows higher prediction accuracy.

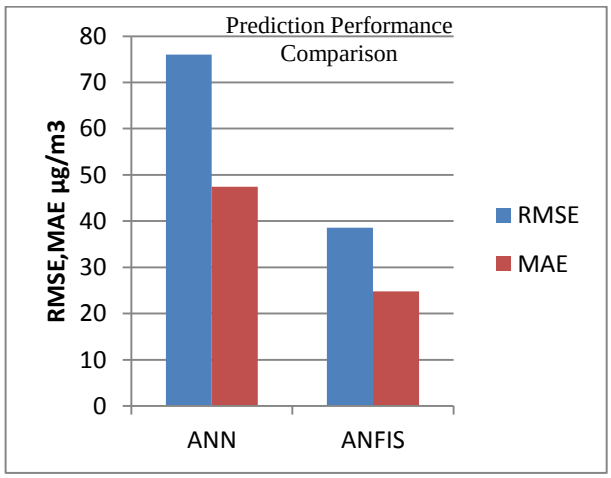

Fig. 6 –Performance Comparison

The proposed model's (Adaptive Neuro-Fuzzy Inference System) evaluation performance is compared by root mean squared error and mean absolute error functions with artificial neural network performance in fig. 5 and the proposed model shows superior performance than ANN. Terrain air quality of PM2.5 with +elevation and –elevation is shown in figure 5. Forecasting PM2.5 48 hours will not only create awareness to public but limit their outdoor activities at days when the pollution level is severe.

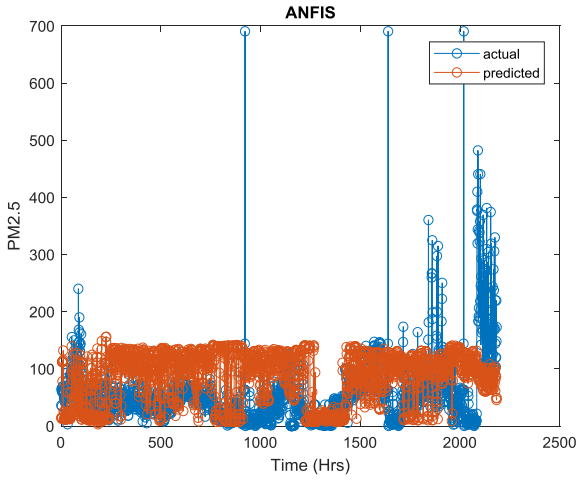

Fig. 7 - Proposed ANFIS Model Prediction

V. CONCLUSION

The adverse health effects caused by PM2.5 can be controlled by forecasting the concentration values of pollutants by intelligent machine learning models. The present study is used to find how air pollution is produced and disseminated. The real time air quality data, meteorological data and terrain information are all collected

from 14 stations from different parts of India to analyse spatial-temporal relationship and the terrain impact on PM2.5 and prediction model is built by ANN, LSTM, CNN and final prediction by ANFIS. The proposed model outperformed the existing model by showing best RMSE and MAE results. By providing a forecast of 48 hours of PM2.5, preventive measures can be taken, to limit the outdoor activities of humans especially children and old age people who are susceptible to diseases very easily.

Acknowledgement: The authors wish to thank the CPCB for the air quality parameters used for this analysis.

REFERENCES

[1] Hao Guo, Sri Harsha Kota, Shovan Kumar Sahu, Jianlin Hu, Qi Ying, Aifang Gao, Hongliang Zhang, "Source apportionment of PM2.5 in North India using source-oriented air quality models", Environmental Pollution 231 (2017) pp.426-436.

[2] The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017, India State-Level Disease Burden Initiative Air Pollution Collaborators, Lancet Planet Health 2019; 3: e26–39.

[3] Ping-Wei Soh, Jia-Wei Chang, and Jen-Wei Huang, "Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations",. Ieee Access,Vol-6, 2018, pp. 38186-38199

[4] Chiou-Jye Huang and Ping-Huan Kuo, "A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities", Sensors 2018, 18, 2220;

[5] Athira V, Geetha P, Vinayakumar R, Soman K P, "Deep AirNet: Applying Recurrent networks for Air Quality Prediction", Procedia Computer Science 132(2018) pp-1394-1403.

[6] Guyu Zhao, Guoyan Huang, Hongdou He, And Qian Wang, "Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality", Ieee Access, vol.7,2019, pp-26241-26254.

[7] Nazif1 · N. I. Mohammed1 · A. Malakahmad1 · M. S. Abualqumboz1, "Multivariate analysis of monsoon seasonal variation and prediction of particulate matter episode using regression and hybrid models", International Journal of Environmental Science and Technology, June 2019, vol.16(6), pp-2587-2600.

[8] Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai, " The distance function effect on k-nearest neighbor classification for medical datasets", Springerplus. 2016; 5(1): 1304.

[9] Duarte Folgado, Marília Barandas, Ricardo Matias, Rodrigo Martins, Miguel Carvalho, Hugo Gamboa, "Time Alignment Measurement for Time Series", Pattern Recognition, Volume 81, September 2018, pp- 268-279.

[10] Sakshi Indolia, Anil Kumar Goswami, S. P. Mishra, Pooja Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science 132 (2018) pp-679–688

[11] Jianfeng Zhang, Yan Zhu, Xiaoping Zhang, Ming Ye, Jinzhong Yang, "Developing a Long Short-Term Memory (LSTM) based Model for Predicting Water Table Depth in Agricultural Areas", Journal of Hydrology, Vol. 561, June 2018, pp- 918-929.

[12] Ryan G.Hefron, Brett J.Borghetti, James C.Christensen, Christine M. Schubert Kabban, "Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation", Pattern Recognition Letters 94 (2017) pp-96–104.

