# An ILSTM Machine Learning Model for Accurate AQI Prediction in IoT Application



**Sivakumar S[1], Ramya V[2]**

[1]Department of Computer Science and Engineering, FEAT, Annamalai University, Annamalainagar, Tamilnadu,India.

[2]Department of Computer Science and Engineering, FEAT, Annamalai University, Annamalainagar, Tamilnadu,India

## Abstract

Prediction of time series is an intriguing topic to many researchers as there exist inherent complexities in time dependant data. The conventional methods of analysing the time series data need pre-processing and interpretation of specific time dependence. Recurrent Neural Network serves as the solution for representing the time series reliance in Neural Network framework. But the RNN's had complexity as the gap separating the context information and the application. This problem is solved by proposing the LSTM variant termed as Improved Long Short-Term Memory (ILSTM). These cells are capable of learning the long-lasting dependencies which RNN cells found difficulty in learning. The prediction resulting in an LSTM model is having notable change in analytical ability over a plain LSTM with dropout towards prediction. Eventually, present model is used in data store for Air Quality Index (AQI) prediction. To create an alertness among people about the AQI degradation and its effect on human health, the predicted AQI is hosted in social networking sites.

**Keywords**: *RNN, LSTM, ILSTM model, AQI prediction*

**DOI Number: 10.14704/nq.2022.20.9.NQ44256** **Neuro Quantology 2022; 20(9):2184-2191**

## 1.Introduction

Industrialization has infused both sophistications and hard challenges in human life. Air pollution is such a threatening challenge faced by the world lately. The air quality is degraded due to the fatal gases emitted from industries, global urbanization, ever growing automobiles and consumption of fossil fuels. Recent research reports have probed the impact of deadly particles from harmful gases over human 's health. It states that higher atmosphere contamination levels results in respiratory tract infections, cardiovascular diseases and lung cancer. Consequently, research in the area of air pollution becomes mandatory in environmental protection. The need for evaluating and forecasting the air quality is expanding in order to take necessary precautions by both civilians and any administration in emergency management. It helps the government in framing appropriate contingency measures like restricting the production and usage of motor vehicles and reducing the emissions from highly polluting industrial plants. However, prediction of air quality, forecast of accuracy and reduction of training time are complex tasks need to be addressed immediately in prevention of air pollution. Figure 1.1 depict the architecture diagram of proposed air quality monitoring and prediction system which is sequence modelling from data collection to data visualization.

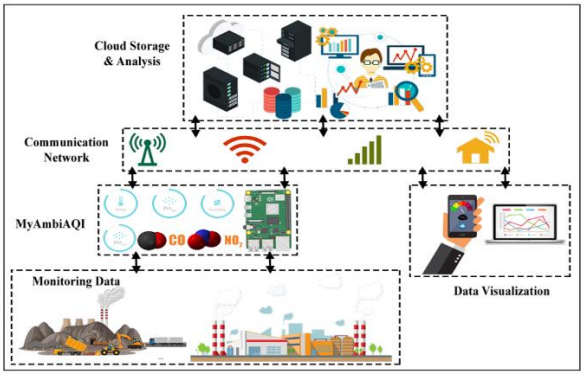


Figure 1.1. IoT based AQI monitoring and prediction system using ILSTM

## 2. Related work

The major motivation for this work is [1], in which an analysis on DJIA is performed and found reassuring values using de-noised data over economic time-series, having the pre processed data resulting in an LSTM model having around 50% enhancement in prediction via a simple LSTM with few eliminations. Some other works suggest ways of enhancing LSTM using methods such as [2] and [3]. There are numerous techniques to pre-analyse and re-analyse the data using conventional statistical methodologies. These techniques are using the exclusively statistical analysis with some precise data set.

### 2.1 LSTM based model

The LSTM models cover error analysis in primary statistical models and this method utilizes the LSTM analysis for predicting the errors found by the first model and used to accurate the predictions associated with original model.

The time series models executes hybrid models using multiple approaches that show consistent results. Methods in [6-9] have shown deep learning and LSTM models excel in recognizing the data patterns whereas in complex cases non stationary,they fail in performing the noising methods. Therefore, usage of added approaches for extracting additional relevant information and analysis with a deep approach provide the effective performance [10-12].

### 2.2 Neural network-based models

Similarly, neural networks are used for generating huge amount of data through Internet of Things. This work considered the following research contribution to identify the research gap in the fields of AQI prediction.

In time series prediction, conventional regression models like RFR,SVR and MLP [13] utilized SVR for predicting the Air Quality Index of Delhi . Due to the unsteadiness in AQI data characteristics, achieving a high fitting degree becomes difficult for SVR. Researchers of [14]and [15] proposed the RFR for air quality prediction based on Spark clustering. The author in the article [16] utilized MLP for forecasting the PM10 concentration and O3 in urban areas. For most of the non-stationary time series data, conventional MLP prediction model possess low prediction fitting degree.

### 2.3 Recurrent based Neural Network

RNN shows improvement in time series prediction with respect to fitting degree. Using this method, RNN's calculation result possess the features of remembering the previous results. For instance, [17,18] utilized RNN to predict air quality. The problems of "gradient explosion" and "gradient disappearance" arrive while training the model due to the RNN's nature of historical dependence on data.



The problems of "gradient explosion" and "gradient disappearance" caused by the RNN's is reduced by the gated technology. For instance, [19] utilized LSTM to predict the modifications in air pollutants. The extraction of eigenvalues in single prediction model is insufficient thereby it becomes hard to obtain high precision prediction. Dsa et al. applied LSTM for predicting the air quality and Alhirmizy et al. used the same for forecasting. A researcher [2,20] utilized assorted data with increased data volume for improving the prediction accuracy which in turn created the problem of increased model training time. The article [20] combines CNN and LSTM based AQI prediction model to improve the extraction of features and integration of air quality data thereby improving the accuracy in prediction. The article [21] used CNN to overcome the insufficiency in feature extraction of LSTM to a considerable level.

## 3. Long Short Term Memory (LSTM)

The LSTM network is a subset of RNN specialised in knowledge and perceiving long term dependencies. In a RNN model, a recapping chain of modules are present to process the sequence passed as a data point while in an LSTM model, the modules are intricate than RNN, with four interacting layers.

### 3.1 LSTM gates

An LSTM module has lot of main parts of which the major part is cell state that flows the whole part of the chain with limited communications. It optionally let the information through by employing gates to add/remove information. LSTM module has four stages:

**Forget Gate Layer**: It is a sigmoid layer that decides the amount of earlier information to be retained by looking at the earlier prediction and the actual value.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (3.1)$$

**Input Layer Gate:** It is a sigmoid layer creates a vector of new candidates by deciding which values will be updated and combined with a tanh layer. The typical LSTM architecture is presented at Figure 3.1.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (3.2)$$

$$\tilde{C}t = tanh(W_c . [h_{t-1}, x_t] + b_c) \quad (3.3)$$

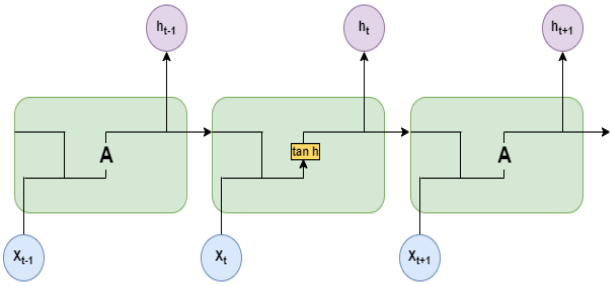


Figure 3.1. LSTM architecture

**Cell State Update:** Point wise multiplication is used for updating the cell values and keeping it as a percentage of new information to include.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}t \quad (3.4)$$

**Output Gate:** It is a sigmoid gate followed by tanh layers for deciding how and which parts of the cell state it would output and normalizing the cell state. The outcome from multiplying these two elements result in a selective cell state in the accurate scale for prediction.

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (3.5)$$

$$h_t = o_t * \tanh(C_t) \quad (3.6)$$

LSTM layer is designed by the above equations. It remembers the information using an arbitrary number of modules in a selective way**.**

### 3.2 Improved Long Short-Term Memory (ILSTM)

However, The LSTM gives the average accuracy with real time data specially prediction of Air Quality Index (AQI). Prediction of air quality is an effort demanding task involving a lot of research over the last decade in devising several methodologies for ensuring accuracy in prediction. This work uses an Improved ILSTM model to retain the historical data since with initial state to current state. The proposed ILSTM model is developed as per the algorithm given under.

### 3.3 ILSTM Circuit

LSTM make small modifications to the information by multiplications and additions. LSTMs can selectively remember or forget things using cell state through which the information flows. The information at a particular cell state has three different dependencies.

These dependencies can be generalized to any problem as:

1. The previous cell state having the information available in the memory after the previous time step
2. The previous hidden state has the same as the output of the previous cell
3. The input at the current time step - the new information fed in at that moment

As per the present work scenario LSTM has been enhanced to improve the accuracy of prediction. The pseudo code the proposed ILSTM is designed and the architecture of ILSTM is given under with figure 3.2 similarly, equations 3.7,3.8 and 3.9 define the functionalities.

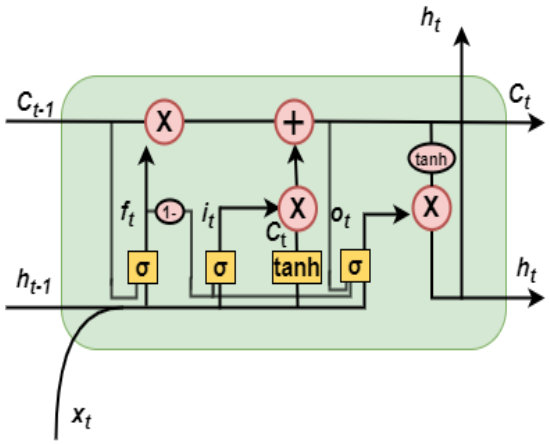


Figure 3.2 ILSTM architecture

$$f_t = \sigma(W_f \,.\, h_{t-1}, x_t + b_f) \quad (3.7)$$

$$i_t = \sigma(W_i \,.\, h_{t-1}, 1 - f_t, x_t + b_i) \quad (3.8)$$

$$o_t = \sigma(W_o \,.\, h_{t-1}, 1 - f_t, x_t + b_o) \quad (3.9)$$

### 3.4 ILSTM Algorithm

1. *ILSTM is built using a series of gates.*
2. *Design the gates to control the sequence of information*
3. *Gates are named as forget, input and output*
4. *Apply the filters in every gates*
5. *Process cell state $c_t$ is designed by previous cell state $c_{t-1}$*
6. *Hidden gate ($h_t$) is designed as per the inputs of input data ($x_i$) and previous $h_{t-1}$*
7. *The forget gate ($f_t$) is designed by pointwise multiplication with sigmoid function and $h_{t-1}$,$c_{t-1}$ and $x_t$ as inputs*
8. *Define a cell state for memory network*
9. *The cell state is defined by both pointwise multiplication and addition with three inputs namely tanh, sigmoid and rememberable forget output (1-$f_t$)*
10. *Output gate is defined by pointwise multiplication with sigmoid of $x_t$,$h_{t-1}$ and 1-$f_t$*

## 4. Results and Discussion

The proposed model has been compared with SARIMA, ARIMA and basic LSTM model to identify the efficiency in terms of accuracy with real time data. Furthermore, the present work is used to predict the AQI from the past set of data. To predict the AQI some of the major air pollutants and gas pollutants were considered. The same has been predicted using ILSTM model to improve the prediction accuracy. Eventually, the multisensory handheld kit was deployed to sense the pollutants from the environment.

### 4.1Prediction of Fine PM.

In this research, the PM concentration data and gaseous pollutants is used as input values. The proposed model is implemented in cloud store for prediction of individual pollutant then the AQI is calculated as per the conventional method and the result is hosted in the web server for visualization. The data is divided in two proportions of 70% for training and 30% for testing. The raw data obtained from the MyAmbiAQI is combined and pre-processing is carried out to check missing values and categorical values within the dataset. Then ILSTM model is applied to the training and testing datasets for predicting $PM_{10}$,$PM_{2.5}$,$SO_2$,$NO_x$and CO concentrations for a period of 10 days succeeding the study period.



### 4.2 Descriptive Data Analysis

In this section, this work presents the data analysis collected by the MyAmbiAQI sensor kits in neyveli, in various days but approximately the same period of time. It also presents the data acquired by the kit over a one-year period. Figure 4.1, shows the prediction of the concentrations of $PM_{10}$ based on the obtained values with MyAmbiAQI kit. The values of concentrations are generally between the limits of (103 µg/m3 and 205 µg/m3 minimum and maximum respectively). Although high concentrations were observed if the kit is close to a specified pollution source like near to mine; peaks is observed in ILSTM machine learning model.

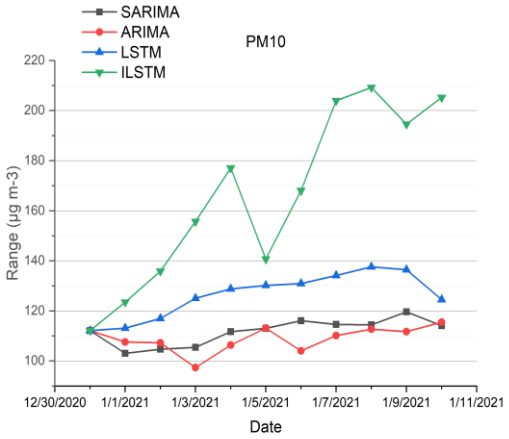


Figure 4.1 . Particulate matter 10 prediction.

Figure 4.2, depicts the prediction of the concentrations of $PM_{2.5}$ based on the obtained values with MyAmbiAQI kit. It is shown that these concentrations generally between the limits of (14 µg/m3 and 51 µg/m3 minimum and maximum respectively). Although high

concentrations were observed if the kit is close to a specified pollution source like near to mine; peaks is observed in SARIMA machine learning model.

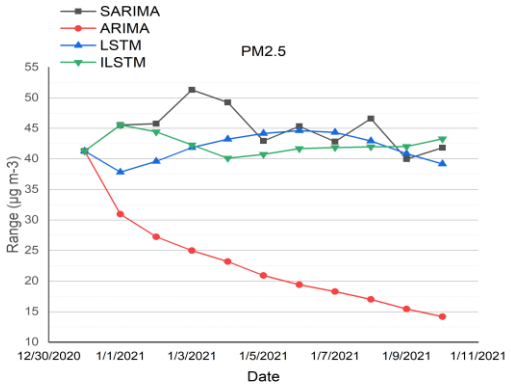


Figure 4.2. Particulate matter 2.5 prediction.

Figure 4.3. denotes the gaseous pollutants ($SO_2$) considered as a prediction target in this research work. These concentrations are generally between the limits of (8 µg/m3 and 12 µg/m3 minimum and maximum respectively). Although high concentrations were observed if the kit is nearer to a specified pollution source like near to mine; peaks is observed in ARIMA machine learning model

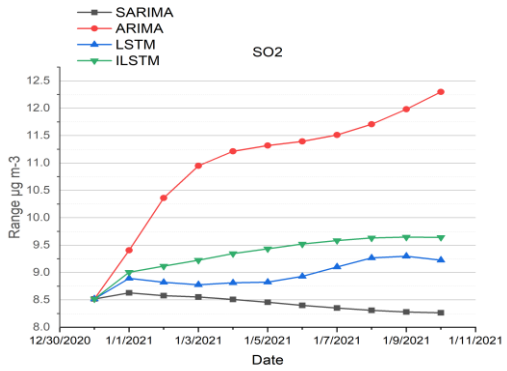


Figure 4.3. $SO_2$ prediction using ILSTM and other time series models.

Figure 4.4. shows the gaseous pollutants (NOx) noted as a prediction target in this research work. It is shown that these concentrations generally between the limits of (9 µg/m3 and 12 µg/m3 minimum and maximum respectively). Although high concentrations were observed if the kit is in the proximity of a specified pollution source like mine; peaks is observed in ARIMA machine learning model.



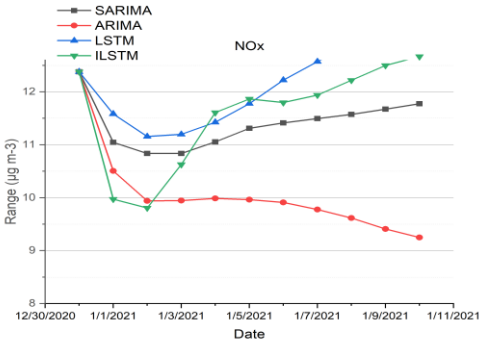


Figure 4.4 . NOx prediction using ILSTM and other time series models.

Figure 4.5. shows the gaseous pollutants (CO) which were considered as a prediction target in this research work It is shown that these concentrations generally between the limits of (0.6 µg/m3 and 0.8 µg/m3 minimum and maximum respectively). Although high concentrations were observed if the kit is in proximal to a specified pollution source like near to mine; peaks is observed in ILSTM machine learning model.

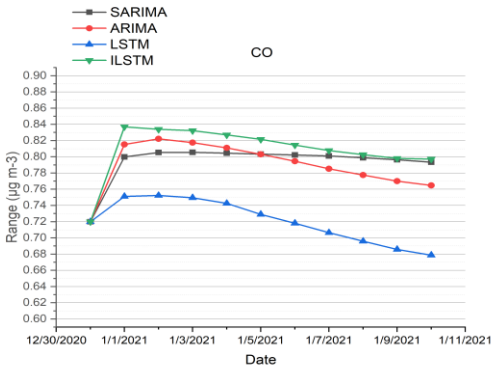


Figure 4.5. CO prediction using ILSTM and other time series models.

The mean value of both gaseous and particulate matter pollutants is shown in Table 4.1. The proposed and conventional machine learning model's outcomes (i.e prediction) has been considered from the specified periods.

Table 4.1. Mean of prediction of pollutants using numerous machine learning models

| Mean | SARIMA | ARIMA | LSTM | ILSTM |
|---|---|---|---|---|
| **$PM_{10}$** | 111.719 | 108.613 | 127.808 | 171.435 |
| **$PM_{2.5}$** | 45.144 | 21.169 | 41.874 | 42.405 |
| **$SO_2$** | 8.434 | 11.215 | 8.996 | 9.416 |
| **$NO_x$** | 11.301 | 9.831 | 12.095 | 11.501 |
| **CO** | 0.801 | 0.796 | 0.721 | 0.817 |

Typical ILSTM AQI is obtained from the table 4.2. shown.

| AQI | | | | |
|---|---|---|---|---|
| Time | SARIMA | ARIMA | LSTM | ILSTM |
| 12-31-2020 | 108 | 108 | 108 | 108 |
| 01-01-2021 | 102 | 105 | 109 | 116 |
| 01-02-2021 | 103 | 105 | 111 | 124 |
| 01-03-2021 | 104 | 97 | 117 | 137 |
| 01-04-2021 | 108 | 104 | 119 | 151 |
| 01-05-2021 | 109 | 109 | 120 | 127 |
| 01-06-2021 | 111 | 103 | 121 | 145 |
| 01-07-2021 | 110 | 107 | 123 | 169 |
| 01-08-2021 | 110 | 108 | 125 | 173 |
| 01-09-2021 | 113 | 108 | 124 | 163 |
| 01-10-2021 | 109 | 110 | 116 | 170 |

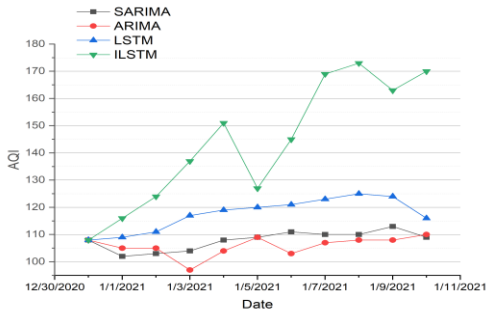


Figure 4.6. AQI prediction index by different machine learning models.

The ILSTM AQI calculation is shown in figure 4.6 as per the table 4.2. The proposed ILSTM Model gives the accurate air quality information compares with other conventional machine learning model's outcomes (i.e prediction).

### 4.3 Model Performance Evaluation

RMSE (equation number 4.1 and 4.2) is applied to evaluate the proposed model relating to other existing forecasting models. The prediction capability of ILSTM is evaluated using the MAE and R-square ($R^2$ equation 4.3). The calculation formula for MAE and RMSE is as follows:

$$MSE = \frac{\sum_{t=1}^{T}(\hat{y}t - y^t)2}{T} \quad (4.1)$$

$$RMSE = \sqrt{\frac{1}{n}} \sum_{i=1}^{n}(\hat{y}_I - y_I)^2 \quad (4.2)$$

$y_i$ = predictive value , $\hat{y}_i$ =the actual value. The MAE value increases with the decreasing values of prediction in both MSE and RMSE.

The formula for calculating $R^2$ is as follows:

$$R^2 = 1 - \frac{(\sum_{i=1}^{n}(y_i - \hat{y})^2)/n}{(\sum_{i=1}^{n}(\hat{y}_i - t\hat{y}_i)^2)/n} \quad (4.3)$$

$Y_i$ = predictive value, $\hat{y}_i$ = true value, and $t\hat{y}_i$ = average value. The $R^2$ range is (0,1). If the values of MSE and RMSE are closer 0 then accuracy of the prediction process is increased. Then only we can get the smaller error values for the predicted and real values. The compatibility of the ILSTM model becomes better if $R^2$ value is closer to 1 and therefore it gains the betterment in time series models. The closer $R^2$ is to 1, the better machine learning time series model is ILSTM. Table 4.3. Shows the calculation of RMSE, MSE and $R^2$ where the proposed ILSTM algorithm neared to 1 in $R^2$, hence it evident that the ILSTM is most effective model for AQI prediction.

Table 4.3. MSE, RMSE and $R^2$ comparison with machine learning models.

| ML. Alg | MSE | RMSE | $R^2$ |
|---|---|---|---|
| SARIMA | 6.556 | 42.989 | 0.913 |
| ARIMA | 6.554 | 42.957 | 0.952 |
| LSTM | 6.403 | 41.003 | 0.962 |
| ILSTM | 6.299 | 39.688 | 0.973 |

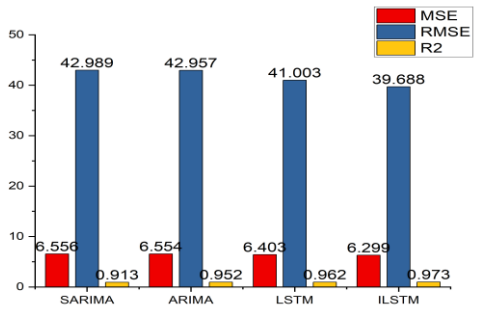


Figure 4.7. MSE,RMSE and $R^2$ evaluation index.

From figure 4.7. After analysing the various machine learning models, it is inferred that ILSTM model surpassed the other algorithmic models in time series. Also, this model is showing best results with $R^2$ of 0.973 which is closer to 1.

## 5.Conclusion

The proposed model is compared with SARIMA, ARIMA and Long Short-Term Memory algorithm (LSTM). SARIMA and ARIMA proved to be efficient in extracting data from inaccurate and non-linear information. Trained dataset has the best values of input. Network will fail in deducing the values precisely outer training set. The investigational outputs prescribed that the proposed model outperformed the traditional prediction algorithms. The model obtained the lowest MSE and RMSE as 6.299, 39.688 wise versa. Similarly, this model brings very closest $R^2$ value than other models used in real time prediction.



## References

1: Qi Tang et al. "Prediction of Financial Time Series Based on LSTM Using Wavelet Transform and Singular Spectrum Analysis". In: Math. Prob. Eng. 2021 (May 2021), p. 9942410. ISSN: 1024-123X. DOI: 10.1155/2021/9942410.

2: Wuwei Liu and Jingdong Yan. "Financial Time Series Image Algorithm Based on Wavelet Analysis and Data Fusion". In: J. Sens. 2021 (Mar. 2021), p. 5577852. ISSN: 1687-725X. DOI: 10.1155/2021/5577852.

3: Sidra Mehtab and Jaydip Sen. "Analysis and Forecasting of Financial Time Series Using CNN and LSTM-Based Deep Learning Models". In: ResearchGate (Dec. 2020). URL: https://www.researchgate.net/publication/346647852_Analysis_and_ Forecasting _ of _ Financial _ Time _ Series _ Using _ CNN _ and _ LSTM - Based _ Deep_Learning_Models

4: Angelo Garangau Menezes and Saulo Martiello Mastelini. "MegazordNet: combining statistical and machine learning standpoints for time series forecasting". In: arXiv (May 2021). eprint: 2107.01017. URL: https://arxiv.org/abs/2107.01017v1.

5: Colah. Understanding LSTM Networks – colah's blog. [Online; accessed 29. Oct. 2021]. Oct. 2021. URL: https://colah.github.io/posts/2015-08-UnderstandingLSTMs

6: Zielak. Zielak's Bitcoin Historical Data w/o NaN. [Online; accessed 29. Oct. 2021]. Oct. 2021. URL: https : / / www . kaggle . com /

kognitron / zielaks - bitcoin - historical-data-wo-nan.
7: kieferk. pymssa. [Online; accessed 18. Nov. 2021]. Nov. 2021. URL: https://github.com/kieferk/pymssa.
8: Colah. Understanding LSTM Networks – colah's blog. [Online; accessed 29. Oct. 2021]. Oct. 2021. URL: https://colah.github.io/posts/2015-08-UnderstandingLSTMs.
9: A guide for using the Wavelet Transform in Machine Learning. [Online; accessed 5. Nov. 2021]. Aug. 2021. URL: https://ataspinar.com/2018/12/21/a-guide- forusing-the-wavelet-transform-in-machine-learning.
10: Sidra Mehtab and Jaydip Sen. "Analysis and Forecasting of Financial Time Series Using CNN and LSTM-Based Deep Learning Models". In: ResearchGate (Dec. 2020). URL: https://www.researchgate.net/publication/346647852_Analysis_and_ Forecasting _ of _ Financial _ Time _ Series _ Using _ CNN _ and _ LSTM - Based _ Deep_Learning_Models.
11: Z. Wang and Y. Lou. "Hydrological time series forecast model based on wavelet denoising and ARIMA-LSTM". In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Mar. 2019), pp. 1697–1701. DOI: 10.1109/ITNEC.2019.8729441.
12: Jason Brownlee. Making Predictions with Sequence. 2019. URL: https : / / machinelearningmastery.com/sequence-prediction/. (accessed: 10.02.2021).
13: Ganesh, S. S., Modali, S. H., Palreddy, S. R. & Arulmozhivarman, P. Forecasting air quality index using regression models: A case study on Delhi and Houston. ICOEI. 248–254. (2017).
14: Tan, Y. & Mao, X. Assessment of the policy effectiveness of Central Inspections of Environmental Protection on improving air quality in China. J. Clean. Prod. 288, 125100 (2020).
15: Zhang, Y., Yu, D. & Chen, G. Advanced recurrent neural networks for automatic speech recognition. New Era Robust Speech Recognit. (2017).
16: Sun, L. et al. A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. Health Inf. Sci. Syst. 8, 1–7 (2020).
17: Deng, H., Zhang, L. & Shu, X. Feature memory-based deep recurrent neural network for language modeling. Appl. Soft Comput. 68, 432–446 (2018).
18: Yu, Z., Moirangthem, D. S. & Minho, L. Continuous TiRMSEcale long-short term memory neural network for human intent understanding. Front. Neurorobot. 11, 42 (2017).
19: Ysc, A. et al. An LSTM-based aggregated model for air pollution forecasting. Atmos. Pollut. Res. 11, 1451–1463 (2020).
20: Cheng, J., Liu, Y. & Ma, Y. Protein secondary structure prediction based on integration of CNN and LSTM model. J. Vis. Commun. Image Represent. 71, 102844 (2020).
21: Yin, Q. et al. CNN and RNN mixed model for image classification. MATEC Web Conf. 277, 02001 (2019).