International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

Forecasting of water quality index using long short-term memory (LSTM) networks

[1]Odubela Christiana. A, [2]Balogun Wasiu A, [1]Akinpelu, T. A, [3]Abioye Mayowa, [4]Oluwe Musbau Olajide

[1]*Civil Engineering Dept. Lagos State Polytechnic, Ikorodu, Lagos State.*
[2]*Mechatronics Engineering Dept, Lagos State Polytechnic, Ikorodu, Lagos State.*
[3]*Mechanical Engineering Dept. Kogi State Polytechnic, Lokoja, Kogi State.*
[4]*Electrical Electronic Engineering Dept., AkanuIbiam Federal Polytechnic, Unwana, Ebonyi State*

Submitted: 15-10-2021 Revised: 26-10-2021 Accepted: 28-10-2021

ABSTRACT
Investigating the quality of water is crucial towards the prevention of outbreak of water borne disease as well as its applicability of water in the area of road construction, agriculture, ad fishery. The main focus of this paper is to build up a water quality forecasting model with the assistance of water quality parameters utilizing Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN). This study uses the water quality index data of six years from the Kaggle online database. For this paper, data incorporates the estimation of four parameters namely pH, water temperature, water conductivity, and ORP which impact and affect water quality index. To assess the performance of the developed LSTM model and benchmarked with RNN, the metrics used are Regression coefficient and Root Mean Squared Error. The prediction performance shows that the LSTM out performed that of RNN for the prediction of water quality index (WQI).

Keywords: Water Quality, LSTM, RNN, Regression, Model, Forcasting

I. INTRODUCTION

Water is crucial for all types human and physical activities such as agriculture, construction transportation etc. The nature of water helps in controlling the biotic diversity, vitality, and rate of succession. The impacts of unclean water are broad, affecting each part of life affects not just aquatic lives but a greater percentage of human life and sustainability [1]. The integrity of water quality genuinely influences human wellbeing, fishery economy, and agricultural activities. In this way, the administration of water assets is pivotal so as to upgrade the nature of water so as to be sure that the water consumed meets the standard of approved agency for water quality[2].

Water assumes a crucial role in our day-by-day life and the nature of water in a region intensely influences the practical improvement of nearby ordinary industrial, agricultural and other anthropogenic activities. Common water resources like groundwater and surface water have dependably been the least expensive and most broadly accessible sources of freshwater. In any case, these assets are destined to progress toward becoming defiled because of different variables including human, industrial and commercial activities just as common procedures. Notwithstanding that, poor sanitation foundation protocols and absence of mindfulness additionally contribute enormously to drinking water defilement. A considerable lot level of the water pollutants have long haul long term negative effects on water quality, causing hazard to human wellbeing [3].

Poor water quality influences the earth activities and human well being. Additionally, contaminated water can prompt some waterborne ailments and furthermore impact child mortality. As indicated by the United Nations, waterborne infections cause the death of 1.5 million children for every year. The World Health Organization says that consistently more than 3.4 million individuals die because of water-related ailments. In this way, it is extremely essential to devise novel methodologies and techniques for checking the level of water deterioration deteriorating water quality and to figure out future water quality patterns. So as to complete valuable and productive water quality analysis and foreseeing the water quality examples, it is important to incorporate a

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

temporal dimension to the analysis, with the goal that the seasonal variation of water quality is tended to. Distinctive approaches have been proposed and applied for analysis and checking water quality and time series analysis.

1.1 Water Quality Parameters

1.1.1 pH

pH is the measure of the degree of the acidity or alkalinity of a water solution. The pH scale is a logarithmic value with the range 0 to 17, whose neutral point is 7.A water solution above 7 is alkaline or basic solution, while the water solution below the value of 7 is acidic. The effect of temperature on Ph is of inverse relationship. That is, as the temperature of water increases, so do it pH value decreases and vice versa.

1.1.2 Temperature of Water

Temperature is another important water quality parameters which is need to be considered alongside other water properties. Temperature has a effect on the conductivity and pH of water.

1.1.3 Water Conductivity

Water conductivity is the ionic strength of a water solution to conduct electricity with the typical unit for measurement being micro-Siemens per centimetre. The conductivity of water increases as the dissolved ions increases. An increase in conductivity of value of water, may signify that the water is polluted, such as sewage leaks, chemical waste flooding into water. Water conductivity is directly related with the salinity, denoting the fact that conductivity of water increases as the salinity increases.

1.1.4 Oxidation and Reduction Potential (ORP)

Oxidation-Reduction potential is the measure of a water solution oxidizing power. That is the potential of a water solution ability to sanitize itself. The more the oxidizers the more the ORP values. Similarly, the lower the ORP, the more the reducers

The paper is focus on the use of LSTM to forecast the water quality parameter using the dataset from Kaggle database. These parameters incorporate physical, biological and chemical factors which impact water quality. The outcome shows that the Machine Learning procedures so as to anticipate the future water quality patterns of a specific region with the assistance of historical water quality data. LSTM model is utilized to build up a methodology for viable water quality forecast and analysis. The model performance based on LSTM and RNN for water quality forecast is compared. The results verify the efficiency of the model that is proposed.

II. LITERATURE REVIEW

The study of accuracy of different machine learning model for the classification of water quality is proposed by [4], the outcome of the classification performance shows that MLP and IBk classifiers outperformed others with regression coefficient of 91.57% for the Kinta rivers dataset used to train the models. Similar research work was implemented by [5] where series of analysis was carried out to investigate the different parameters of water quality collected from an online database as well as the use of principal component analysis and artificial neural network (ANN) to predict the quality of water index parameters. The significant result shows that ANN was able to predict the quality of water for different usage. Another crucial aspect of water quality investigation is in the area of monitoring using either internet of things (IoT) and other remote sensing technology. In order to prevent outbreak of water borne disease through improve water quality measurement, the use of IoT was proposed by [2], to improve the quality of water in Fiji Island. Similar Sigfox based IoT monitoring approach with water quality prediction was investigated by [6]. The monitoring platform leverage of a low power wide area network technologies to transmit sensed was quality parameters used for training a deep learning algorithm for the prediction of water quality.

III. METHODOLOGY

Machine learning is an aspect of Artificial Intelligence (AI) that gives a machine the capacity to consequently learn and improve from experience without being unequivocally programmed. It centres on the development of computer programs that can get the data and use it for learning processes themselves. The process of learning starts with observations or information, such as examples, direct expertise, or instruction, so as to look for patterns in data and settle on better choices later on based on the examples we give. The main aim is to enable the computers to adapt consequently without human interference or help and alter the activities as required. The methodology utilized in this research involves Machine Learning with training and testing data from Kaggle online data repository. The theoretical background of the methodology is as follows:

3.1 Long Short-Term Memory (LSTM)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

Recurrent neural networks (RNN) are networks with loops in them, enabling the information to persevere as shown in figure 1. When the gap between the related information and the place it is required is small, RNNs can learn to be programmed to utilize the past information [7]. Unfortunately, as the gap increases, RNNs become unfit to learn to associate the information.

LSTMs are an improved type of RNNs, equipped for adapting long term conditions. Recollecting information for long periods purposes their default behaviour. LSTMs also have a chain like structure, yet the repeating module has an alternate structure, not at all like RNNs [8]. Rather than having a single neural network, there are four layers, cooperating in a unique manner. The way to LSTM is the cell state. The cell state is somewhat similar to a conveyor belt. It runs straight down the whole chain, with some minor linear connections. It is extremely simple for information to the cell state, carefully controlled by structures called gates as represented by figure 2.

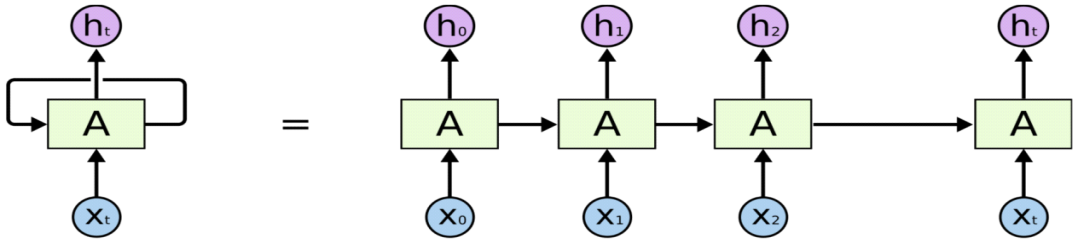

Figure 1: Recurrent Neural Network (RNN)

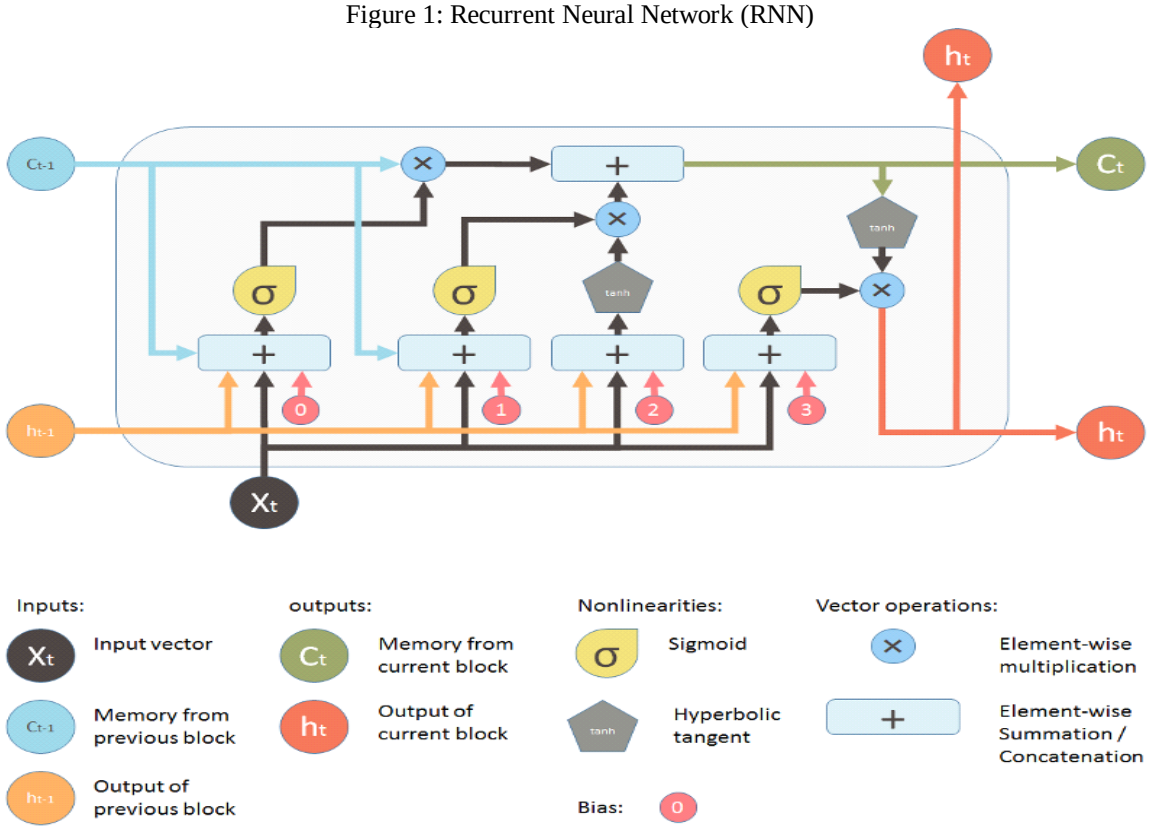

Figure 2: Basic LSTM Memory Block

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

LSTM networks are appropriate for classifying, processing and making predictions based on time series data, since there can be lags of obscure duration between important events in a time series [9]. They were created to manage the exploding gradient and vanishing gradient problems that can be experienced when training traditional RNNs. The activation function of the LSTM gates is frequently the logistic function. The weight of these connections, which need to be learned during training, decide how the gates operate.

A RNN utilizing LSTM can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, gradient descent, joined with back propagation through time to calculate the gradients needed during the optimization process, in order to change weights.
Gates are an approach to alternatively let data through. They are made out of a sigmoid neural net layer and a point wise multiplication operation as illustrated by figure 3.

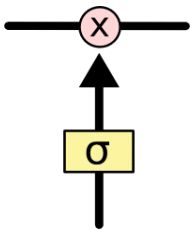

Figure 3: Block diagram of a gate

The sigmoid layer yields numbers somewhere in the range of 0 and 1, depicting the amount of every component ought to be let through. A value of 0 signifies "let nothing through", while a value of 1 signifies "let everything through". An LSTM has three of these gates, to secure and control the cell state as shown in figure 4.

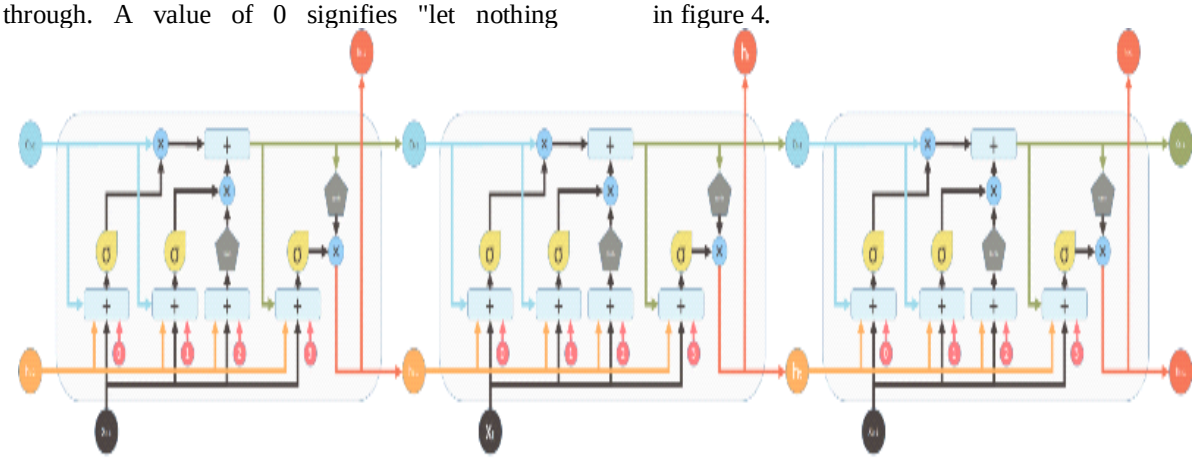

Figure 4: LSTM network with memory blocks

The initial phase in our LSTM is to choose what data we are going to discard from the cell state. This choice is made by the sigmoid layer, called the "forget gate" layer. It looks at h_{t-1} and x_t and yields a number somewhere in the range of 0 and 1 for every cell state C_{t-1}. 1 signifies "totally keep this" and 0 implies “totally dispose of this".

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

f_t is the value of the forget gate at t^{th} time. W_f is the weight between the forget gate and the input layer. h_{t-1} is the output of the previous memory block. x_t is the input vector. $b_f$is the bias vector.

Following stage is to choose what new data we are going to store in the cell state. This has two sections. Initial, a sigmoid layer, called the input gate layer, chooses which values should be updated. Next, a tanh layer makes a vector of new candidate values, C’t, which could be added to the state.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (2)$$

$$C'_t = \tanh(W_c.[h_{t-1}, x_t], b_c) \quad (3)$$

i_t is the value of the input gate. $W_i$is the weight between the input gate and the input layer. $b_i$is the bias vector. W_c is the weight between the input gate and the hidden layer.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

After that, we update the old cell state, C_{t-1}, into the new cell state, C_t. We multiply the old state by f_t, overlooking the things we chose to overlook before. Then we add $i_t * C'_t$. This is the new candidate values, scaled by the amount we chose to update each state value.

$$C_t = f_t * C_{t-1} + i_t * C'_t \quad (4)$$

C_t is the memory from the current block. C_{t-1} is the memory of the previous block.

At last, we have to choose what we are going to yield. This output will be based on our cell state however will be a filtered form. To begin with, we run a sigmoid layer which chooses what parts of the cell state, we are going to yield. At that point, we put the cell state through tanh (to push the values in the range of - 1 and 1) and multiply it by the result of the sigmoid gate, with the goal that we just output the parts we chose to.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

$o_t$is the value of the output gate. W_o is the weight between the hidden layer and the output gate. b_o is the bias vector. h_t is the output of the current block.

Thus, this single unit settles on choice by thinking about the present information, past output and past memory. What's more, it produces new output and adjusts its memory [10].

3.2 Data Collection

This step is essential in light of the fact that the quality and amount of information that we accumulate will legitimately decide how great your predictive model can be. The information for this examination has been gathered from Kaggle, an online information repository supporting the procurement, handling and long haul storage of water quality over the world. Four parameters have been decided for this study, i.e. Temperature, pH, dissolved oxygen (DO) and turbidity. The time interval of 15 minutes has been gotten to complete a successful predict process using this time series that incorporates date/time, parameters and their measurements along with their measurement units.

3.2 Model Training

The training dataset contains 1,339 rows, which is 80% of the total dataset. The second piece of the dataset will be utilized for assessing the model. The testing dataset contains 263 columns, which is 20% of the original dataset. The LSTM model training using the dataset was ccarried outto steadily improve the model's capacity to predict the water quality index. The contrast between the model's expectations and the output delivered with the output that it should produce and modify the number of neurons in the output layer, to such an extent that we will have progressively accurate prediction. This process is repeated to train the model. Each cycle of modifying weights and biases is called as a "training step". The LSTM model is trained for 250 epochs, adam solver and a batch size of 1 is used as seen in Table 1.

3.4 Model Evaluation

This enables the testing of the proposed LSTM model against the information that has never been utilized for preparing. This helps to acertain how the model may perform against information that it has not yet observed. This is intended to be illustrative of how the model may perform in reality. We will utilize Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) to assess the execution of our models.

3.4.1 Mean Squared Error (MSE)

MSE calculates the averages of the squares of the error, that is, the average squared difference between the evaluated qualities and what is assessed. It is a risk function, comparing to the estimation of the squared error loss. It evaluates the nature of a predictor (i.e. a capacity mapping discretionary contributions to an example of estimations of some arbitrary variable). In the event that a vector of n forecasts produced from an example of n data pints focuses on all variables and Y is the vector of observed estimations of the factors being anticipated, at that point the within-example MSE of the predictor is figured as: -

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y'_i)^2 \quad (7)$$

3.4.2 Root Mean Squared Error (RMSE)

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is an as often as possible utilized measure of the differences between values (example or populace values) anticipated by a model or an estimator and the values observed. The RMSD speaks to the square root of the second sample moment of the differences between forecasted values and observed values or the quadratic mean of these differences. RMSD is a proportion of accuracy, to contrast between the error of various models for a specific dataset and not between datasets, as it is scale-dependent. RMSD is the square root of the average of squared errors.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y'_i)^2} \quad (8)$$

3.4.3 Regression Coefficient

The coefficient of determination, indicated by R^2 or r^2 and pronounced "R squared", is the extent of the variance in the dependent variable that is expected from the independent variable(s).

A data set has n values indicated as y_1 ..., y_n (collectively known as y_i or as a vector $y = [y_1,...,y_n]^T$), each associated with a predicted (or modelled) value $f_1,...,f_n$ (known as f_i, or sometimes $\hat{y}_i$, as a vector f).

Define the residuals as $e_i = y_i - f_i$ (forming a vector e).

If is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (9)$$

Then the change of the dataset can be measured using three sum of squares formulas:

- The total sum of squares (proportional to the variance of the data)
  $SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (10)$
- The regression sum of squares, also called the explained sum of squares
  $SS_{reg} = \sum_i (f_i - \bar{y})^2 \quad (11)$
- The sum of squares of residuals, also called the residual sum of squares

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \quad (12)$$

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (13)$$

Table 1: LSTM architecture and set options for the training model

Model Training Parameters	**Values**
Solver	Adam
Maximum number of epochs	250
Gradient threshold	1
Initial learn rate	0.005
Learn rate schedule	piecewise
Learn rate drop period	120
Learn rate drop factor	0.25
Verbose	0
Number of features	1
Number of responses	1
Number of hidden units	150

IV. RESULTS

The basic line plot for all the four parameters required for predicting water quality is showing in figure 5.The plot contains temperature shows a variation within the range of 0 to 30 degree Celsius. In the line plot of pH, we can see that the values are in the range of 6.7 to 7.3. This shows that the pH of water from a dataset from Kaggle is in the ideal range. (6.5 to 7.5). Similarly, for trend and variation was observed for dissolved oxygen and turbidity.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

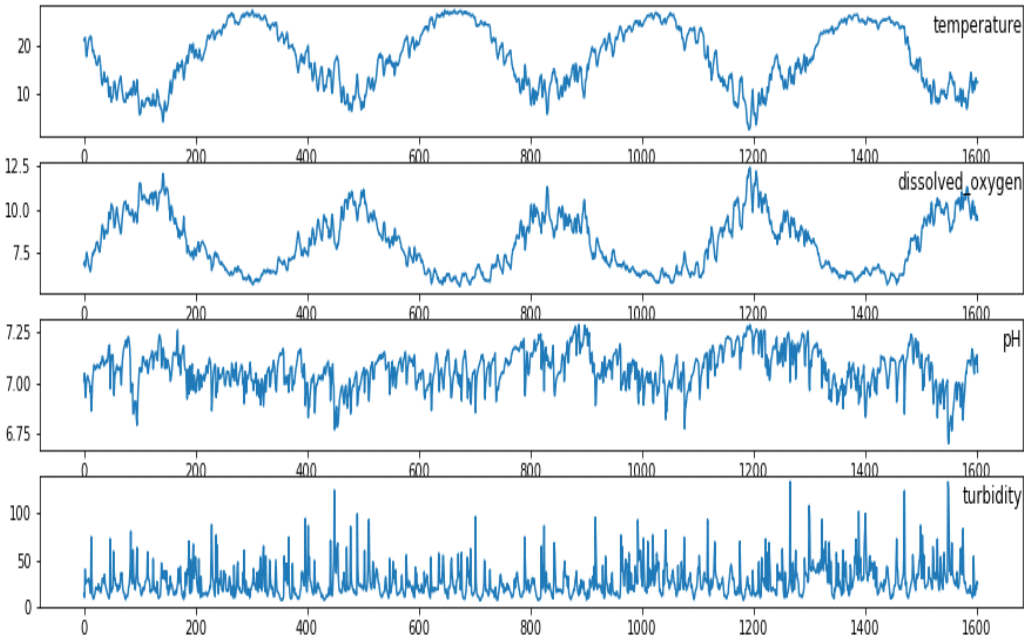

Figure 5: Line plot of four parameters

The model was trained and used for prediction of water quality index. Here, it can be seen that the proposed LSTM model that was trained is better than the other models that was used to benchmark it in terms of performance. The performances of the two models have been calculated using the three metrics, R^2, MSE and RMSE as seen in Table 2. Here, we can see that LSTM has performed better than the other conventional RNNmodel.

The performance evaluation of a typical regression algorithm such as employed in the models presented using LSTM is measured by its root mean square error (RMSE). This is an indication of how close the forecast values are to the observed values.

Table 2: Performances comparison of the LSTM and RNN model

MODEL	**R^2**	**RMSE**
LSTM	0.725	1.1
RNN	0.719	0.5

The trained LSTM network with the specified training options as described in Table 1. The figure 6 below the training trend of the LSTM model in terms of the RMSE and Loss Function of the training model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

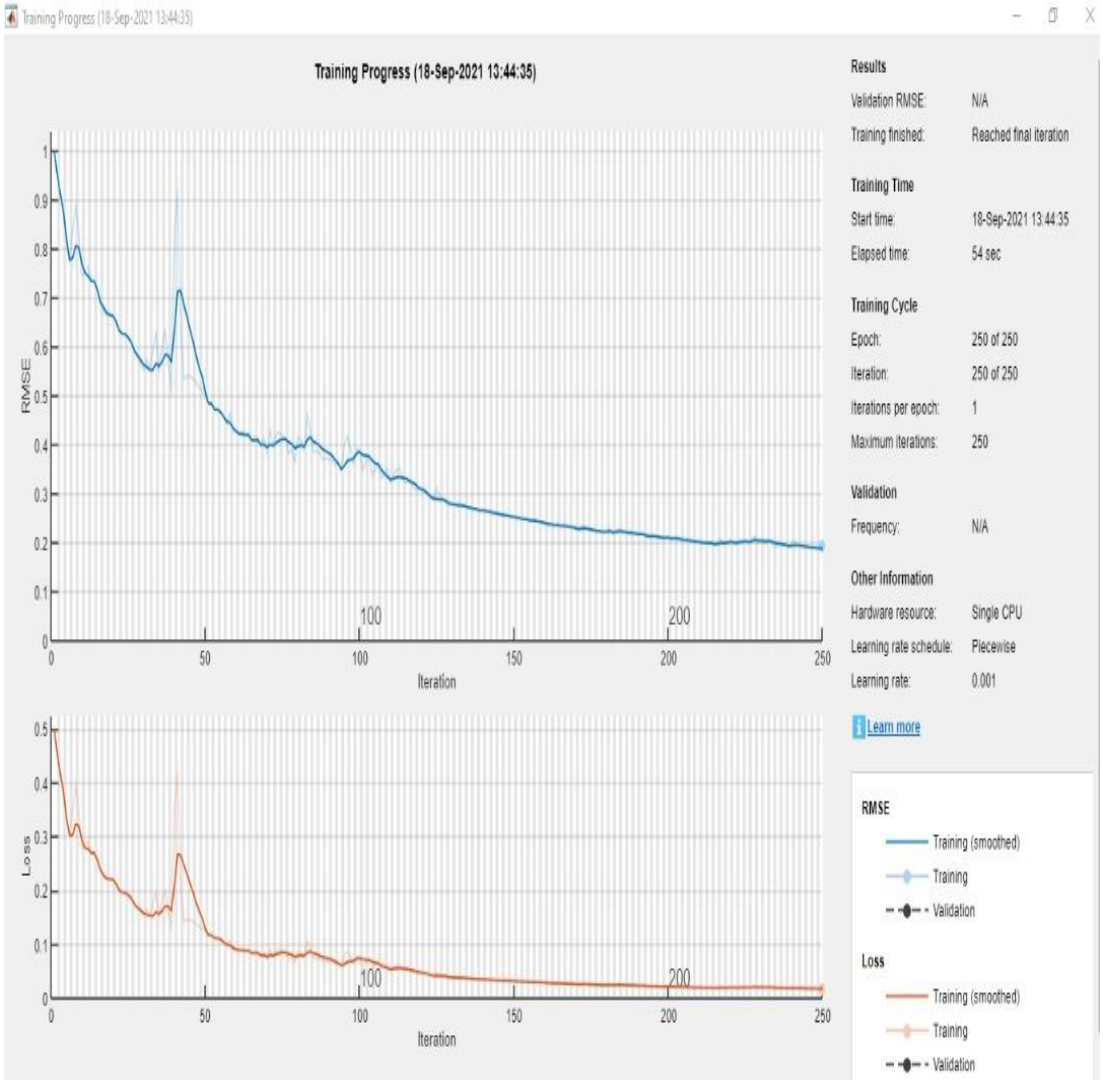

Figure 6: The RMSE and Loss Function of the training model

To forecast the water quality, the values of multiple time steps in the future is shown in figure 7, where the prediction and update-state of the state function is used to predict time steps one at a time and update the network state at each prediction. For each prediction, use the previous prediction as input to the function. The test data was standardize using the same parameters as the training data. The figures shows that the LSTM model could accurately predict the water quality which can guide scientist and engineers of the usage of water for their various application.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

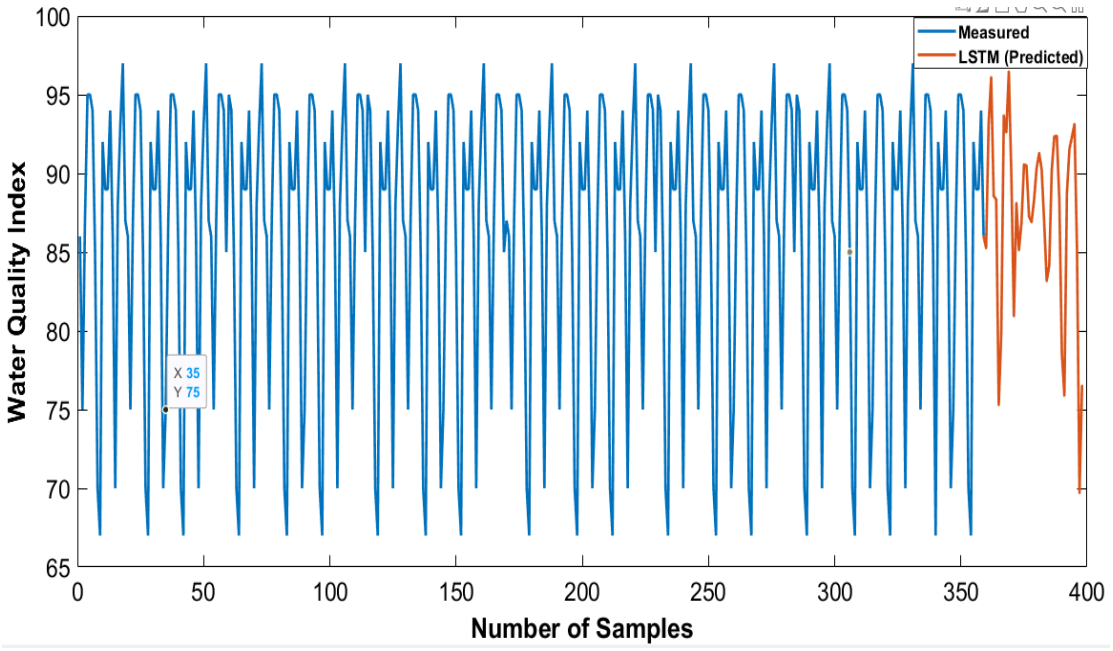

Figure 7: Plot the training time series with the forecasted values

The RMSE values for all the models expressed as error functions are visibly provided in the Error plot (appearing as a subplot) to all the models. Lower values of RMSE compared to the distance from line of best fit indicates how accurate the prediction is. Consequently, RMSE is a measure of the prediction error which by inference indicates a measure of anomaly [11].

To make predictions on a new sequence, reset the network state using reset State. Resetting the network state prevents previous predictions from affecting the predictions on the new data. Reset the network state, and then initialize the network state by predicting on the training data. Figure 8 denotes the comparisons of the forecast with observed values with higher RMSE of 11.01 While, figure 9 describes the comparisons of the forecast with the observed values 5.24.Here, the LSTM prediction model is more accurate, due to reduced value of the RMSE and the fact that the model was updated with the observed values instead of the predicted values.

However, in the plots for their respective updated forecast, the RMSE is visibly improved. This is because during the updated forecast, the forecast values of the previous 10% have been replaced with the observed values. It suffices therefore, to conclude that the performance of the LSTM models is reliably drawn based on the RMSE.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 10 Oct 2021, pp: 1007-1017 www.ijaem.net ISSN: 2395-5252

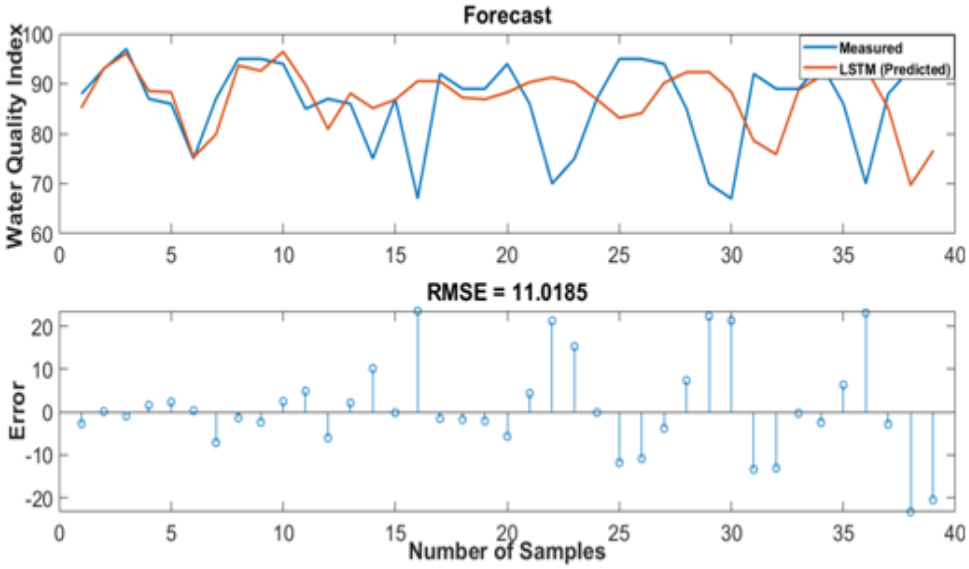

Figure 8: Comparisons of the forecast with observed values

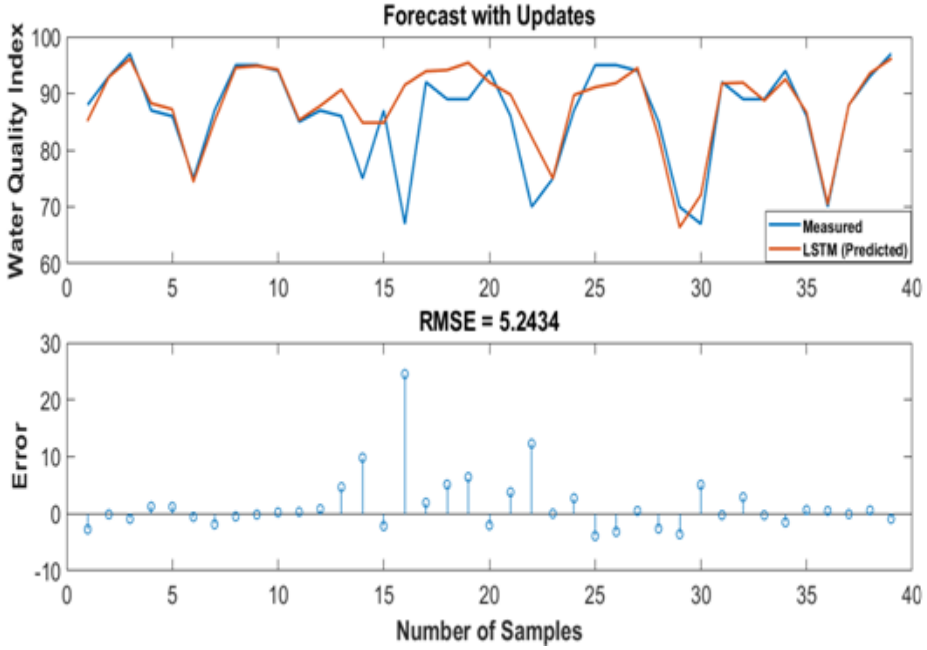

Figure 9: Comparisons of the forecast with the observed values

V. CONCLUSION

Water quality index prediction is indeed very important for scientist and engineers to guide the administration of water assets as well as for the counteractive action of water contamination. In light of the successive attributes of water quality markers, this paper proposes a more accurate prediction strategy based on LSTM for water quality forecast. The paper proposes new LSTM model to predict the value of water quality index which is the one of the main indicators of water quality assessment during road construction project and any other applicable areas. The LSTM model is prepared using dataset of water quality parameters which are provided by the Kaggle online database which monitors the stations and maintains the database and instantaneous values. The datasets contain the water quality parameters information like temperature, DO, pH and turbidity. The result obtained shows that the LSTM model can

accurately predict the water quality index with better than that of RNN based on the $R^2$and RMSE, Considering about the disadvantage of a long preparing cycle or long training cycles experienced using LSTM structure, an increasingly successful memory block can be structured in future further research work.

REFERENCES

[1] O. Elijah et al., "A concept paper on smart river monitoring system for sustainability in river," Int. J. Integr. Eng., vol. 10, no. 7, pp. 130–139, 2018, doi: 10.30880/ijie.2018.10.07.012.

[2] A. N. Prasad, K. A. Mamun, F. R. Islam, and H. Haqva, "Smart water quality monitoring system," in 2nd Asia-Pacific World Congress on Computer Science and Engineering, APWC on CSE 2015, 2016, pp. 1–6, doi: 10.1109/APWCCSE.2015.7476234.

[3] O. Elijah et al., "Application of UAV and Low Power Wide Area Communication Technology for Monitoring of River Water Quality," in 2018 2nd International Conference on Smart Sensors and Application, ICSSA 2018, 2018, pp. 105–110, doi: 10.1109/ICSSA.2018.8535994.

[4] R. Rosly et al., "The Study on the Accuracy of Classifiers for Water Quality Application," Int. J. u- e- Serv. Sci. Technol., vol. 8, no. 3, pp. 145–154, 2015, doi: 10.14257/ijunesst.2015.8.3.13.

[5] T. K. Anyachebelu, "Prediction of a Water Quality Index using Online Sensor Data," University of Bedfordshire, 2019.

[6] P. Boccadoro, V. Daniele, P. Di Gennaro, D. Lofù, and P. Tedeschi, "Water Quality Prediction on a Sigfox-compliant IoT Device: The Road Ahead of Water," pp. 1–13, 2020, [Online]. Available: http://arxiv.org/abs/2007.13436.

[7] W. C. Wong, E. Chee, J. Li, and X. Wang, "Recurrent Neural Network-Based Model Predictive Control for Continuous Pharmaceutical Manufacturing," 2018, doi: 10.3390/math6110242.

[8] P. Boccadoro, I. Student, V. Daniele, P. Di Gennaro, and D. L. Ieee, "Water Quality Prediction on a Sigfox-compliant IoT Device : The Road Ahead of WaterS," arXiv Prepr. arXiv2007.13436, no. July, 2020.

[9] Z. Jianfeng, Y. Zhu, X. Zhang, M. Ye, and J. Yang, "Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas," J. Hydrol., vol. 561, no. April, pp. 918–929, 2018, doi: 10.1016/j.jhydrol.2018.04.065.

[10] P. K. Kashyap, S. Kumar, A. Jaiswal, M. Prasad, and A. H. Gandomi, "Towards Precision Agriculture: IoT-Enabled Intelligent Irrigation Systems Using Deep Learning Neural Network," IEEE Sens. J., vol. 21, no. 16, pp. 17479–17491, 2021, doi: 10.1109/JSEN.2021.3069266.

[11] Otuoze AO, Mustafa MW, Sofimieari IE, Dobi AM, Sule AH, Abioye AE, et al. Electricity theft detection framework based on universal prediction algorithm. Indones J Electr Eng Comput Sci 2019;15. doi:10.11591/ijeecs.v15.i2.pp758-768