Water quality prediction system using LSTM NN and IoT

Ann Laverene Lopez
Department of Computer Science and Engineering
Muthoot Institute of Technology and Science
Kochi, Kerala, India
17cs254@mgits.ac.in

Haripriya N A
Department of Computer Science and Engineering
Muthoot Institute of Technology and Science
Kochi, Kerala, India
17cs063@mgits.ac.in

Kavya Raveendran
Department of Computer Science and Engineering
Muthoot Institute of Technology and Science
Kochi, Kerala, India
17cs128@mgits.ac.in

Sandra Baby
Department of Computer Science and Engineering
Muthoot Institute of Technology and Science
Kochi, Kerala, India
17cs076@mgits.ac.in

Dr. Priya C V
Department of Computer Science and Engineering
Muthoot Institute of Technology and Science
Kochi, Kerala, India
priyacv@mgits.ac.in

2021 IEEE International Power and Renewable Energy Conference (IPRECON) | 978-1-6654-0137-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/IPRECON52453.2021.9640938

***Abstract*—Drinking water that is both clean and safe is critical to one's well-being. Checking the quality of water regularly can be an initial step in ensuring pure drinking water. The existing system is time consuming and monotonous manual system. Therefore, we propose a system based on Machine Learning (ML) and Internet of Things(IoT) that can measure and forecast future water quality parameters. For this, the daily water quality data was taken from the Muvattupuzha River in Kerala. Long Short-Term Memory Neural Network (LSTM NN) was used to bring out the timeseries pattern in the data. The sensors like pH sensor, turbidity sensor and total dissolved solids (TDS) sensor were used to read the current water quality parameters and this data was used to predict the future parameter values. The IoT module which includes the sensors, Arduino and NodeMCU can be installed in the water source to monitor the parameters regularly. The main benefit of this system is that users can be notified ahead of time if there is a risk of pollution, allowing them to disinfect the water before it becomes polluted.**

***Index Terms*—Drinking water, IoT, LSTM NN, machine learning, time series forecasting, water quality prediction system**

I. Introduction

Water quality and accessibility of pure drinking water is a big issue all over the world. Urbanisation, population growth, industrial wastewater discharge, excessive water use, agricultural activities, and other factors have all contributed to the decline in water quality. Contaminated water has a major impact on human health. The biggest issue we're dealing with is the diminishing supply of high-quality drinking water. City water resources, well water supplies, and freshwater sources such as reservoirs, streams, and rivers can all be contaminated. Drinking contaminated water has a variety of negative consequences for human health. About 2 billion people around the world drink water that has been polluted. Diarrhoea, cholera, dysentery, typhoid, and polio may all be spread by contaminated water. Every year, it is estimated that 485,000 people die from diarrhoea as a result of contaminated drinking water [11]. As a result, maintaining the quality of water is a critical challenge. Continuous water quality monitoring and accurate prediction of water quality are the foundation of water environment management and are critical for protecting the environment. Traditionally, water quality detection was done manually with water samples being collected and sent to laboratories for analysis. It is time-consuming and needs human resources. It also costs money. Monitoring the quality of water regularly can be the first step in preventing water pollution. A suitable machine learning model along with cost-effective IoT devices can make a water quality prediction system more accurate and reliable. The factors like pH, turbidity and total dissolved solids play a major role in determining the quality of water. In the proposed system, these factors are used to predict future water quality. Thus it helps to prevent water pollution in the future and also indicates when more preventive measures have to be taken to keep the water safe. Since the daily monitoring of water quality is done by the sensors automatically, it helps to reduce the human resource and money to be invested in monitoring water. Hence the proposed system helps to prevent users from using contaminated water as well as prevents water from getting contaminated.

II. Related Works

A water quality prediction system that uses time series data requires a suitable machine learning model to predict the quality more accurately. Many neural network models are used for water quality prediction systems. Xiu Li and Jingdong Song [1] proposed 4 different ANN-based models for water quality prediction such as ANN model, ANN-markov chain-linear model, ANN-markov chain-ANN model and ANN-markov chain-SVR model to predict the Biological Oxygen Demand(BOD) of water samples collected from Tolo Harbour. The best performance was shown by the ANN-markov chain-SVR model which gave an RMSE of 0.6174.

The authors in [2] proposed a system using ML and IoT for monitoring water quality. The system consists of multi-sensors connected to NodeMCU to collect the water parameters and uses an Artificial Neural Network (ANN) algorithm to predict the quality of water. In [3] the authors introduced and summarised the machine learning-based modelling work they have been investigating for solving the three challenges-water quality prediction, data imputation and outlier detection. They studied different models like ANN, Convolutional Neural Network(CNN), RNN and concluded that all these models showed promising results in solving challenges related to water quality.

In [4], the authors used the LSTM model for water quality prediction where a comparison of an LSTM NN was made with Back Propagation Neural Network and Online Sequential Extreme Learning Machine (OSELM). LSTM NN was found to be a better model for quality prediction. Yu Jiao et al [5] proposed an LSTM model for predicting the air quality index using various parameters and the result has shown that it is a suitable model for air quality index prediction as well. An IoT-based air quality monitoring and prediction system proposed by Rutvik Mehta et al [6] also uses the LSTM model for the prediction of air quality. Improved Grey Relational Analysis was used to identify the correlation between the different parameters that affect the quality of water and the selected features were used as an input to the LSTM model for predicting the dissolved oxygen content in Tai Lake and Victoria Bay [7]. These studies help to conclude that LSTM NN is a better model for prediction which includes time-series data.

A smart water quality monitoring system based on IoT [8] uses different sensors which measure real-time pH, conductivity, temperature and turbidity of water and these parameter values are compared with WHO (World Health Organization) standards to check whether the water is drinkable or not. A system based on IoT where real-time data is collected using various sensors [9] used an ARM-based microcontroller to convert the analogue signal to a digital signal. It used a Zig-Bee module to send the data to a personal computer, then the data was analysed and notifications were sent to the users accordingly. In [6], several sensors which measure pH, temperature, turbidity, dissolved oxygen etc were used to collect the water quality data along with NodeMCU for sending the data to the database and was found to be cost-effective.

The related works discussed above used different neural network-based methods like ANN, RNN, CNN, LSTM etc for time series data prediction. From the studies it is clear that most of the methods are based on LSTM neural network as it has shown promising results in the field of time-series data.

III. System Design

A. LSTM NN

LSTM NN is a type of recurrent neural network suitable for processing and predicting events with long intervals in time series.

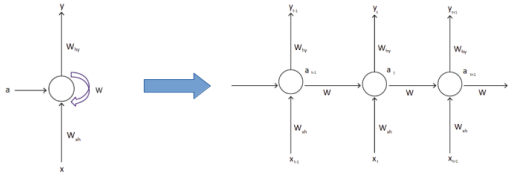

Fig. 1. Structure of RNN

Fig 1 shows the structure of RNN. It includes three layers, the input layer, the hidden layer and the output layer. W_{xh} and W_{hy} are the weights for the connection of the input layer to the hidden layer and the connection of the hidden layer to the output layer respectively. The activation of the layer is indicated by 'a'. X_i and Y_i denote the i^{th} input and i^{th} output respectively. W is the weight that is passed from one hidden layer to the next for the connection between the hidden layers. RNN uses the current information as well as the previous information for prediction. But it faces a vanishing gradient problem. To overcome this problem of RNN, LSTM NN was developed [4].

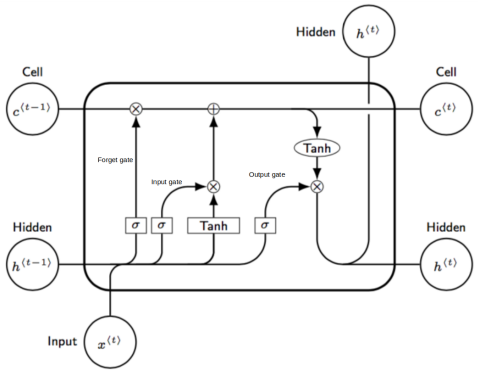

Fig. 2. LSTM

The neurons of LSTM NN called the memory block (Fig 2) have a structure that maintains a constant error flow indicating that it is better to deal with time-series data. The LSTM is made up of cell states and gates. Throughout the sequence's processing, the cell state can convey important information. As

a result, information from earlier time steps can make its way to later time steps, lessening the short-term memory effects. Information is added or withdrawn from the cell state via gates as the cell state travels. The gates determine which information about the cell state is permitted. LSTM includes 3 gates called the input gate, forget gate and output gate. The forget gate is the one that decides what information should be thrown away or kept. Information from the previous hidden state and current input is passed through the corresponding activation function. The input gate is used to update the cell state and the output gate decides what the next hidden state should be. Thus, the LSTM uses the present information, past output and memory to produce new output and adjusts its memory.

B. Internet of Things (IoT)

Internet of Things is a sort of "universal global neural network" in the cloud which connects various sensors and devices to the internet [12]. Various traditional fields of embedded systems, wireless sensor networks, control systems etc contribute to enabling the Internet of things. The communication of various sensors and devices can be done directly, hands-free, remotely over the Internet. Sensors play an important role in the automation of any application. It measures and processes the collected data for detecting changes in physical things. If a change occurs in any physical condition for which the sensor is made, it produces a measurable response [13]. There are different types of sensors. One among them is the water quality sensor which measures the quality of water by monitoring the ion. This type of sensor includes a pH sensor, turbidity sensor, dissolved oxygen sensor etc which monitor the water quality in the IoT environment.

C. Water quality prediction using LSTM NN and IoT

The proposed system checks the water quality and warns the user before the water gets contaminated. As many parameters can determine the purity of water, some parameters like pH, turbidity and total dissolved solids are taken into account. The dataset used here was collected from Kerala Water Authority which contains daily data of water quality parameters from 1 January 2016 to 31 December 2020. The LSTM model is used for prediction as it is known for giving better performance with high accuracy. A comparison of LSTM with ANN is also performed which uses the same test set.

The different phases involved in training and testing of the model includes:

- Data Collection and preparation: The dataset was collected from Kerala Water Authority. It was then converted to a comma-separated values (.csv) file format. The data is then stored into a DataFrame and the date column was set as an index.
- Data Pre-processing: The required dataset has been taken and only three parameters, i.e. turbidity, TDS and pH was used for computational purposes. The dataset was split into training and testing set in the ratio 7:3. Min-Max normalization has been applied on the dataset to normalize the dataset into a range of 0 and 1.
- Creating the models: Three separate models based on LSTM NN were created for the prediction of each parameter. Each model has one input layer, one hidden layer and one output layer.
- Compiling the models: Before training, we compiled the models using the loss function and the optimizer. Adam(Adaptive Moment Estimation) has been used as the optimizer and Mean Squared Error as the loss function.
- Training the models: 70% of the dataset was used for training and to improve the performance, the models were trained with different numbers of epochs and neurons in the hidden layers to identify which could give better performance.
- Model Evaluation: RMSE and MSE of the LSTM NN model and a similar ANN model were used to identify the best model for time series prediction. LSTM has shown better performance compared to ANN. RMSE and MSE of the models with different numbers of epochs and neurons in the hidden layer were calculated to identify which can give better performance.
- Testing: 30% of the dataset was used for the testing of the three LSTM models.

The system uses 3 LSTM NN's which individually learns the time-series pattern of these water quality parameters from the historical training data. Each LSTM in the proposed system consists of one input layer, one hidden layer and one output layer. The number of neurons in the hidden layer is determined by experimenting with different values.

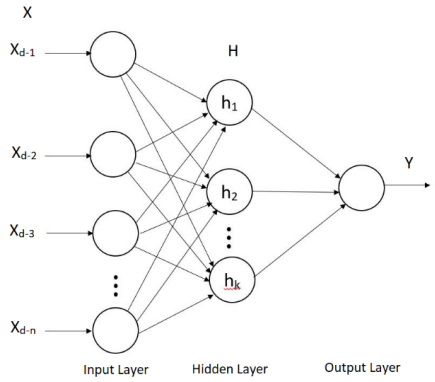

Fig. 3. Water quality prediction system based on LSTM NN

In the Fig 3, X denotes the input vector, $X = (x_{d-1}, x_{d-2}, ...x_{d-n})$ with n as the time step. $H = (h_1, h_2, ...h_k)$ denotes the hidden layer neurons and Y represents the output vector.

To evaluate the performance of the model, RMSE is used as it indicates the difference between the predicted value and the actual value. Equation (1) shows how the RMSE of the model is calculated. Here n denotes the number of data points and e is the error which is the difference between the predicted and actual value.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(e)^2} \tag{1}$$

The overall system architecture is shown in Fig 4.

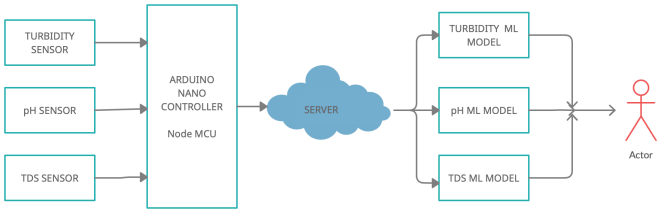

Fig. 4. Architecture of Proposed System

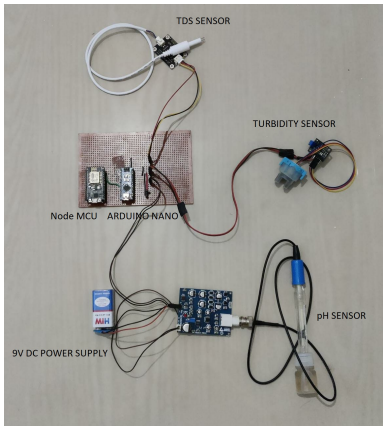

Fig. 5. Experimental setup of IoT module

As shown in Fig 5, the IoT module consists of a pH sensor, turbidity sensor, TDS sensor, Arduino nano and NodeMCU. The sensors are connected to an Arduino nano microcontroller which uses the ATmega328P. The Arduino Nano has 30 male I/O headers in a dip-30 arrangement that can be programmed with the Arduino software integrated development environment (IDE). The sensors are of analog type and they are connected to the analog pin of the Arduino board. Fig 6 depicts the flow chart of the working procedure of the controller and sensors.

The data read by the sensors are sent to the database with the help of NodeMCU which is connected to the Arduino nano micro controller. There is a serial communication between both. The NodeMCU contains a firmware that uses ESP8266 Wi-Fi SoC and it is used to send data. This IoT module is installed in the water source to monitor the water quality regularly.

TABLE I
DRINKING WATER SPECIFICATION(INDIAN STANDARD).

Sl No	Parameters	Unit	Acceptable Limit	Permissible limit
1	pH		6.5-8.5	No relaxation
2	Turbidity	NTU	1	5
3	TDS	mg/litre	500	2000

The current water quality parameter value read by the sensor along with previous 6 days values which are stored in the

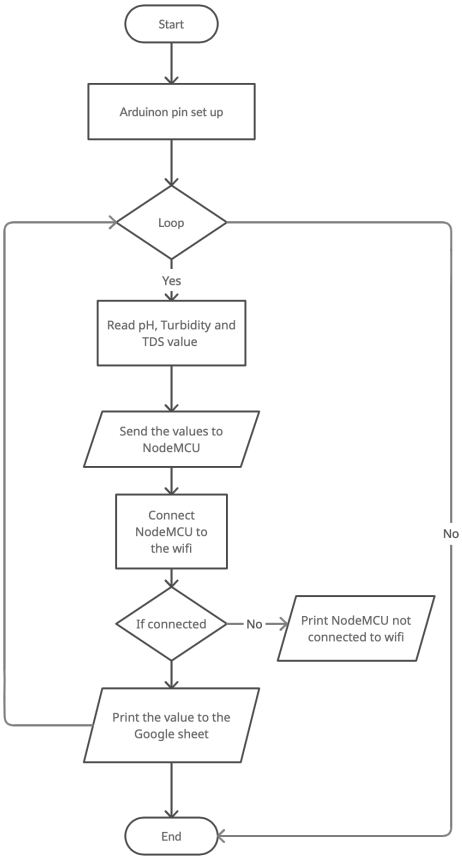

Fig. 6. Flow chart of working procedure of IoT module

database are given to the pre-trained LSTM models. Each model predicts the corresponding parameter values for the next week. The predicted values of different parameters are compared with their standard values as per the Indian Standard (Table I). If these predicted values are not within the standard range, an alert in the form of a mail is provided indicating the chance of water contamination in the future.

IV. RESULTS AND DISCUSSIONS

The daily water quality data provided by the Kerala Water Authority is the dataset used in the proposed system. It contains real-time observed data of river water collected between January 2016 and December 2020. It consists of about 1826 data in total. From this, 1295 data were used for training and the remaining 531 data for testing. It includes data of water quality parameters like pH, turbidity and TDS. Here the turbidity was measured in NTU and TDS in mg/L. Table II depicts the first 5 samples of the dataset.

Table III, shows the minimum and maximum value of each parameter in the training and testing data during the 3 main seasons(indicated in terms of months in Table III) in Kerala. In the dataset used for training and testing all the parameters are showing very little variation in the months between March to May(as shown in Table III). This indicates that during the summer season the river water is not on the verge of pollution.

TABLE II
DATASET

Date	pH	Turbidity	TDS
01/01/2016	7.2	2.9	30.2
02/01/2016	7.3	1.3	30.4
03/01/2016	7.1	1.7	30.3
04/01/2016	7.3	2.6	30.7
05/01/2016	7.1	1.5	30.1

The months between October and February is the time of the northeast monsoon season in Kerala. The turbidity value has gone out of range in this season for both training and testing data. pH and TDS are within the limit.During the period of southwest monsoon in Kerala, there will be heavy rain and all the rivers will be filled with water. That is between the months of June and September, the pH value is on the verge of pollution range. Also, the turbidity has gone out of range indicating that water has been polluted. Thus we can conclude that the river water is mainly getting polluted in the rainy season and the parameter most affected is turbidity as it goes out of range during this season.

TABLE III
RANGE OF PH,TURBIDITY AND TDS DURING DIFFERENT SEASONS

Training Data			
Season Months	**pH**	**Turbidity**	**TDS**
Oct-Feb	7.1-7.7	0.4-26.2	21.9-42.5
Mar-May	7.1-7.8	0.4-4.6	30.7-40.4
Jun-Sep	6.6-7.6	0.3-46.6	19.7-40.1
Testing Data			
Season Months	**pH**	**Turbidity**	**TDS**
Oct-Feb	7.1-7.6	0.3-26.2	22.3-42.5
Mar-May	7.1-7.6	0.4-4.6	30.7-40.4
Jun-Sep	6.7-7.6	0.3-46.6	19.7-40.1

The below figures[7-9]shows the variations in pH, turbidity and TDS throughout different years.

Fig. 7. Changes of pH throughout the year

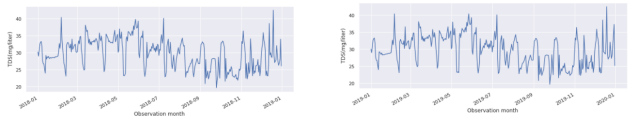

Fig. 8. Changes of TDS throughout the year

As shown in the graphs the pH has very small variations throughout the year. Similarly, TDS also shows slight variations. But turbidity is showing large variation especially in

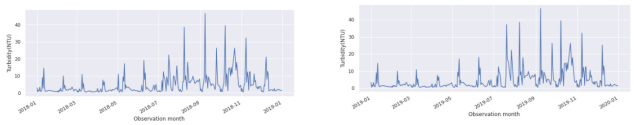

Fig. 9. Changes of turbidity throughout the year

months like June, July etc. This helps us to conclude that the turbidity of the river water is out of the standard range during rainy seasons.

Three separate predictive models based on LSTM NN were established using Keras which is a deep learning framework for the three parameters pH, turbidity and TDS. Each model predicts the future value of each parameter individually. The dataset was divided in the ratio of 7:3 for training and testing. To improve the performance of the LSTM NN for the data, the model was tested with a different number of epochs(No.E) and hidden layers(No.N) and is recorded in Table IV. The RMSE reduces as the number of neurons in the hidden layer rises, as shown in Table IV. After a certain point, as the number of neurons in the hidden layer increases, RMSE rises as well.

TABLE IV
LSTM NN WITH DIFFERENT NUMBER OF EPOCHS AND NEURONS IN HIDDEN LAYER.

No.E	No.N	pH		Turbidity		TDS	
		MSE	RMSE	MSE	RMSE	MSE	RMSE
25	10	0.018	0.136	0.007	0.084	0.013	0.117
50	10	0.018	0.136	0.007	0.084	0.012	0.111
25	25	0.018	0.136	0.007	0.084	0.014	0.120
50	25	0.018	0.136	0.006	0.083	0.012	0.111
25	50	0.018	0.137	0.006	0.083	0.012	0.111
50	50	0.018	0.137	0.006	0.083	0.011	0.110
25	75	0.018	0.137	0.006	0.083	0.012	0.111
50	75	0.018	0.137	0.006	0.083	0.012	0.111

The LSTM NN model was compared with the ANN by checking the accuracy of both the models on the same test set. The RMSE of LSTM NN for each parameter was found to be less than that of the RMSE of ANN for each parameter as shown in Table V. The result has shown that LSTM NN is a better model for water quality prediction which uses time series data.

TABLE V
COMPARISON OF LSTM NN AND ANN

	RMSE of LSTM NN	RMSE of ANN
pH	0.136	0.138
Turbidity	0.083	0.090
TDS	0.110	0.118

Table VI shows the RMSE values of different parameters for different samples of the test set for LSTM and ANN. The 151 samples indicate the data from October to February, 92 samples indicate the data from March to May and 122 samples

show the data from June to September of the year 2020. As shown in the Table VI, the RMSE of pH for LSTM which is 0.136 is smaller than that of ANN which is 0.138. Even though LSTM has lower RMSE, the difference in the RMSE is very less when compared to ANN. But for turbidity and TDS the difference in the RMSE value for LSTM and ANN is 0.007 and 0.008 respectively which is larger than that for pH. Thus LSTM has shown better performance as it has a lower RMSE value compared to ANN for all parameters, especially turbidity and TDS. Also among the three LSTM models for each parameter, the LSTM for turbidity is having an RMSE of 0.083 which is less compared to the LSTM model for pH and TDS. From this, we can conclude that LSTM for turbidity is best among the three models.

TABLE VI
RMSE VALUES OF DIFFERENT PARAMETERS FOR DIFFERENT SAMPLES

Test Data	pH		Turbidity		TDS	
	LSTM	ANN	LSTM	ANN	LSTM	ANN
151 (Oct-Feb)	0.142	0.143	0.016	0.020	0.081	0.091
92 (Mar-May)	0.140	0.141	0.122	0.124	0.102	0.103
122 (June-Sep)	0.107	0.109	0.081	0.081	0.133	0.140

The below figures[10-12] depicts the prediction of LSTM NN and ANN on various parameters. It helps to conclude that LSTM NN has a better performance.

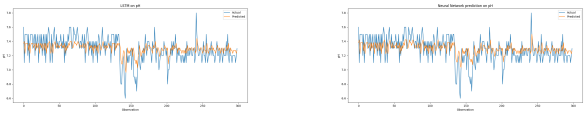

Fig. 10. Prediction on pH

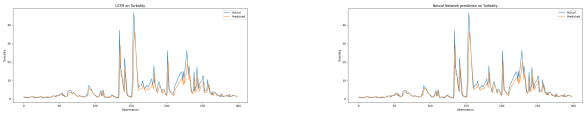

Fig. 11. Prediction on Turbidity

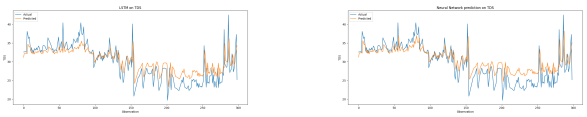

Fig. 12. Prediction on TDS

V. CONCLUSION

Water is an important element for the existence of life and good quality water is essential for human health. Nowadays there is a lack of availability of pure drinking water. Water quality prediction has a huge significance for the management of water resources. Predicting the quality of water in advance can help us to avoid water pollution. The proposed water quality prediction system based on LSTM NN and IoT predicts the future water quality by considering the sequential characteristics of water quality parameters. The model is trained with the data from the Muvattupuzha river obtained from Kerala Water Authority. Several sensors like pH, turbidity and TDS sensors were used to collect real-time data from the water source. Based on the data received from the sensors the model predicts the future value of different water quality parameters. If the predicted value of each parameter is out of the standard range, a warning is provided regarding the contamination of water in future. The model was compared with ANN and the result shows that the performance and accuracy of LSTM NN are higher. This system helps to prevent water from contamination and is cost-effective. Since it is automated, it also helps in checking the water quality automatically, thus reducing the manpower. The system can be improved by improving the accuracy of LSTM NN and by including more sensors to collect more parameters. Biosensors can also be used for better water quality prediction.

REFERENCES

[1] X. Li and J. Song, "A New ANN-Markov Chain Methodology for Water Quality Prediction," International Joint Conference on Neural Networks, pp. 12-17 July, 2015

[2] M. Mukta, S. Islam, S. D. Barman, A. W. Reza and M. S. Hossain Khan, "Iot based Smart Water Quality Monitoring System," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 669-673, doi: 10.1109/CCOMS.2019.8821742.

[3] Zhang, Yifan Thorburn, Peter Vilas, Maria Fitch, Peter. (2019). Machine learning approaches to improve and predict water quality data. 10.36334/modsim.2019.D5.zhangYiF.

[4] Y. Wang, J. Zhou, K. Chen, Y. Wang and L. Liu, "Water quality prediction method based on LSTM neural network," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, pp. 1-5, doi: 10.1109/ISKE.2017.8258814.

[5] Y. Jiao, Z. Wang and Y. Zhang, "Prediction of Air Quality Index Based on LSTM," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 17-20, doi: 10.1109/ITAIC.2019.8785602.

[6] T. W. Ayele and R. Mehta, "Air pollution monitoring and prediction using IoT," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 1741-1745, doi: 10.1109/ICICCT.2018.8473272.

[7] Zhou, Jian Wang, Yuanyuan Xiao, Fu Wang, Yunyun Sun, Lijuan. (2018). Water quality prediction method based on IGRA and LSTM. Water. 10. 1148. 10.3390/w10091148.

[8] http://www.ijstr.org/final-print/oct2019/Water-Quality-Monitoring-Using-MachineLearning-And-Iot.pdf

[9] S. R. Khaire and R. M. Wahul, "Water Quality Data Analysis and Monitoring System in IoT Environment," 2018 3rd International Conference on Contemporary Computing and Informatics (IC3I), Gurgaon, India, 2018, pp. 326-330, doi: 10.1109/IC3I44769.2018.9007289.

[10] https://www.omicsonline.org/open-access/public-health-hazards-due-to-unsafedrinking-water-2167-7719-1000138-101933.html

[11] https://www.who.int/news-room/fact-sheets/detail/drinking-water

[12] P. V. Dudhe, N. V. Kadam, R. M. Hushangabade and M. S. Deshmukh, "Internet of Things (IOT): An overview and its applications," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 2650-2653, doi: 10.1109/ICECDS.2017.8389935.

[13] D. Sehrawat and N. S. Gill, "Smart Sensors: Analysis of Different Types of IoT Sensors," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 523-528, doi: 10.1109/ICOEI.2019.8862778.