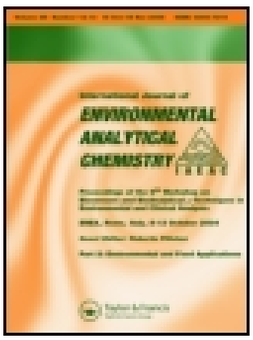

International Journal of Environmental Analytical Chemistry

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/geac20

Water quality analysis in a lake using deep learning methodology: prediction and validation

Venkata Vara Prasad D , Lokeswari Y Venkataramana , P. Senthil Kumar , Prasannamedha G , Soumya K. & Poornema A.J.

To cite this article: Venkata Vara Prasad D , Lokeswari Y Venkataramana , P. Senthil Kumar , Prasannamedha G , Soumya K. & Poornema A.J. (2020): Water quality analysis in a lake using deep learning methodology: prediction and validation, International Journal of Environmental Analytical Chemistry, DOI: 10.1080/03067319.2020.1801665

To link to this article: https://doi.org/10.1080/03067319.2020.1801665

Published online: 05 Aug 2020.

INTERNATIONAL JOURNAL OF ENVIRONMENTAL ANALYTICAL CHEMISTRY
https://doi.org/10.1080/03067319.2020.1801665

Taylor & Francis
Taylor & Francis Group

ARTICLE

Water quality analysis in a lake using deep learning methodology: prediction and validation

Venkata Vara Prasad D[a], Lokeswari Y Venkataramana[a], P. Senthil Kumar[b], Prasannamedha G[b], Soumya K.[a] and Poornema A.J.[a]

[a]Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India; [b]Department of Chemical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

ABSTRACT

Discharge of untreated waste water, municipal sewage, industrial effluents, dumping of degradable and non-degradable wastes has polluted natural water sources like river, lake, pond to a great extent, therefore, it is obligatory to test out the quality of water before consumption. With this motivation, the work explores many deep learning algorithms to estimate the Water Quality Index, which is a singular index to describe the general quality of water, and the Water Quality Class, which is a distinctive class defined on the basis of the Water Quality Index. The water samples were collected from Korattur Lake in the Chennai city. The water quality parameters such as pH, Total Dissolved Salts, turbidity, phosphate, nitrate, iron, Chemical Oxygen Demand, chloride and sodium were measured from the collected water samples. The models used for training and testing include Deep Learning models such as Artificial Neural Network, Recurrent Neural Network and Long-Short Term Memory for both binary and multi-class classification. The metrics used for evaluating the models were accuracy, precision and the execution time of the models that are used for comparing and analysing above mentioned models. From the results obtained, it was observed that LSTM yielded the highest accuracy of around 94% and also consumes the least execution time when compared with other deep learning models.

ARTICLE HISTORY
Received 24 June 2020
Accepted 6 July 2020

KEYWORDS
Water quality; prediction; artificial neural network; recurrent neural network; long short term memory; classification accuracy

1. Introduction

India is rich in water source as it is surrounded by ocean on its three sides that serves as source foe desalination treatment, enriched with perennial rivers, lakes, reservoirs and ponds. Not entire water resource is used by man only few limited quality of water is available for general purpose like domestic and industrial application because of pollution which is major crisis in India. These sources are being depleted due to anthropogenic effects that heavily contaminate water at deep levels thereby spoiling the quality of water. As days prolong, water pollution has become more serious with the rapid development of the economy and urbanisation. Hence, it is important to check the quality of water before consumption. The quality of water from the lakes has a significant importance as they are generally used for multiple purposes such as: drinking, domestic and residential water

CONTACT P. Senthil Kumar senthilkumarp@ssn.edu.in

supplies, agriculture (irrigation), other human or economic activities. Water quality plays a major role on public health and the environment because, consumption of impure water leads to many water-borne diseases like Cholera, Diarrhoea, etc. Hence, it is mandatory to check the quality of water before consumption.

The water quality depends on several interconnected parameters with a local and temporal distinction, which are subjective to the water flow rate during the year. Most of the studies related to the assessment of the water quality use several methodologies like software analysis that incorporates water flow with respects to direction, statistical analysis using statistics and its tools and machine learning process. Prediction of water quality helps in analysing the impacts caused by presence of polluted water in the chosen area of study. Here the chosen area is Korattur lake that is located in Chennai, India. Among various water quality parameters, most important one is Water Quality Index (WQI) as it incorporates statistical combination of all physio-chemical water quality parameters in analysing the effect of pollutants on water. Some of the contaminants that deteriorate water quality are heavy metals, like lead, chromium, cadmium and mercury, pesticides, emerging contaminates like pharmaceutical compounds, personal care products, aromatic amines that are dangerous for human health, since they are toxic and can be carcinogenic.

Chennai is heavily dependent on the rainfall through which lake water and reservoirs are conserved with natural source of water, due to the lack of perennial rivers within the city. Pollutants are added to the lakes mainly due to anthropogenic activities like discharge of sewage, effluents, dumping of solid waste, release of untreated waste water. As time prolongs these pollutants are transported across soil substrate. Further they are penetrated across soil zones and pollute one of the valuable water resources, ground water. Ground water is one of the important sources for drinking water, followed by lake water and reservoirs. Incidence of ground water pollution is high in areas that are congested and populated heavily as large volume of waste are concentrated and discharged in to natural zones through which hydro chemical parameters are varied. Additionally lake waters are also contaminated as they face contaminates directly through direct disposal that undergoes dispersion and dissolution in water. From lake and reservoirs water is feed for all basic and municipal activities, irrigation, and agriculture. Hence, it is important to know the quality of water before usage. The work focuses on predicting the quality of lake water in Chennai.

Many researchers have analysed the quality of water and have built various deep learning models to predict the quality of water. Some of the works related to the prediction of water quality using deep learning includes the work done by Y. khan, C. S. See et. al in developing a water quality prediction system that predicts the quality of water resources like lakes, rivers, streams and estuaries using ANN and time-series analysis. The dataset was obtained from the United States Geological Survey (USGS) of the year 2014. Major parameters like Chlorophyll, Specific Conductance, Dissolved oxygen and Turbidity were considered for evaluation. Their performance was evaluated with Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. It proved to be reliable one with high accuracy of prediction and lowest MSE and best regression value for specific conductance [1,2].

Muharemi et al. [3] developed a water quality prediction system that deals with time series data. The data were collected from a public water company located in Germany.

Evaluation was carried out using machine learning and deep learning models such as logistic regression, linear discriminant analysis, and support vector machines (SVM), artificial neural network (ANN), deep neural network (DNN), recurrent neural network (RNN) and long short-term memory (LSTM). The parameters like Time, Turbidity, pH, Electrical Conductivity, Water Temperature, Chloride (cl), Redox Chlorine dioxide and flow rate were chosen [3]. Liu et al. [4] developed a water quality prediction system using the deep learning model LSTM. The dataset was collected from the Guazhou Water Source of the Yangtze River in Yangzhou from January 2016 to June 2018. The water quality parameters such as Water Temperature, pH, Dissolved Oxygen, Conductivity, Turbidity, Chemical Oxygen Demand (CODMn), NH3-N were chosen [4]. Wang et al. [5] designed a time series water quality prediction system by using LSTM Neural Network to predict the water quality. The data were collected from Taihu Lake which was measured monthly from 2000 to 2006. The water quality parameters such as DO and total phosphorus were chosen [5]. Solanki et al. [6] predicted the water quality of a reservoir. The data were collected from Chaskaman reservoir by considering the parameters such as pH, dissolved oxygen and turbidity. Deep learning model ANN was to predict the water quality whose performance was evaluated with Mean squared error and Mean absolute error [6]. Xu and Liu [7] designed a water quality prediction model to predict the water quality of intensive freshwater pearl breeding ponds in Duchang County in China's Jiangxi Province using wavelet neural networks. By considering parameters like Dissolved oxygen, pH, Temperature, Air humidity, Wind speed and solar radiation levels results were evaluated. Result showed that the developed model is superior to the BP neural network and the Elman neural network [7]. Haghiab et al. [8] predicted the quality of Tireh River located in the southwest of Iran using artificial neural network (ANN), group method of data handling (GMDH) and support vector machine (SVM). Parameters such as DO, COD, BOD, EC, pH, Temperature, K, Na, Mg were considered [8]. But all these works had some drawbacks like consideration of insufficient parameters of water quality, inappropriate accuracy, inefficient to handle the multi-dimensional and imbalanced datasets. The water quality indexes have not been considered.

Hence, the work involves the deep learning models such as ANN, RNN and LSTM to meet the requirements that the previously used models failed to achieve. The parameters considered are: pH, Total dissolved salts (TDS), turbidity, phosphate, nitrate, iron, COD, chloride and sodium. The objective of the current work is to overcome the drawbacks of the existing models by using the models that can handle large datasets which are complex and of nonlinear type which are well suited for making predictions on time series data and also when the number of parameters considered is large. It also works well with unstructured and semi structured data. The output obtained is more informative than any other algorithms. Based on these predicted values, the accuracy of the deep learning models are analysed and compared. The quality of -the water in the next 5 years is also known [9–14]. In precise, the objective of this work is to predict the quality of water in Chennai region. The study area chosen is Korattur lake which is located in north of Chennai-Arakkonam railway line. It is one of the largest lakes in thewestern part of city. It is a chain of three lakes comprising of Ambattur lake, Madhavaram lake and Korattur lake [15]. The deep learning models were used to predict the accuracy of water quality, precision and execution time.

2. Materials and methods

Figure 2 depicts the design of the water quality prediction system. The first step is the data pre-processing, which involves cleaning the data and feature selection. Data cleaning is the process of removing missing or inappropriate records from the dataset. Feature selection refers to selecting the most essential features which contribute most to the prediction variable (Class). The next step is determining water quality index and assigning classes to the data by considering the value of the water quality index. The WHO guidelines for drinking water are used for determining the ground truth for drinkable and non-drinkable water. The data are thus classified into binary class (Good and Bad water samples) and multi-class (Excellent, Good, Average, Bad and Poor water samples). Once the classes have been assigned for all the data, the input dataset is divided into training and testing sets. Among the available data, 80% of the data is used for training and the remaining 20% is used for testing. The models used for training include the deep learning models such as ANN, RNN and LSTM. The models are then tested to find the accuracy, precision and time taken. Based on the outcomes obtained from deep learning models, the results are analysed to find the best suitable model for our system. The model with highest accuracy and less execution time is then used for predicting the quality of the water.

2.1. Data pre-processing

Data processing includes four reasonable steps that are applied in the prediction. Steps include dataset collection, dataset description, data splitting and water quality index. All the four steps help in water quality assessment of the chosen location.

2.1.1. Dataset collection

Korattur Lake is one of the largest lakes in the Chennai city. It is spread over 990 acres and is located to the north of Chennai. It has been a major source of drinking water for about 18 years. The dataset for our work was collected from the Korattur Lake. The dataset consists of water data for over 10 consecutive years (2010 to 2019). The dataset consists of about 5,000 records and consists of 9 parameters such as PH, TDS, turbidity, phosphate, nitrate, iron, COD, chloride and sodium. Figure 1 represents the satellite view of Korattur Lake.

2.1.2. Dataset description

The data set consists of 5000 records in both binary class as well as multi-class classification consisting of 9 parameters as shown in Table 1. The desirable range of parameters for drinking water is shown in Table 2.

2.1.3. Data splitting

Before training the deep learning model it is necessary to divide the data into training and testing sets. After splitting the data, the model is trained and tested with certain parts of the data to measure the accuracy of the model's performance. The data was split as a fraction of 4:1 for training and testing, respectively. Thus, of the total of 5000 samples, 4000 samples were used for training and 1000 samples for testing.

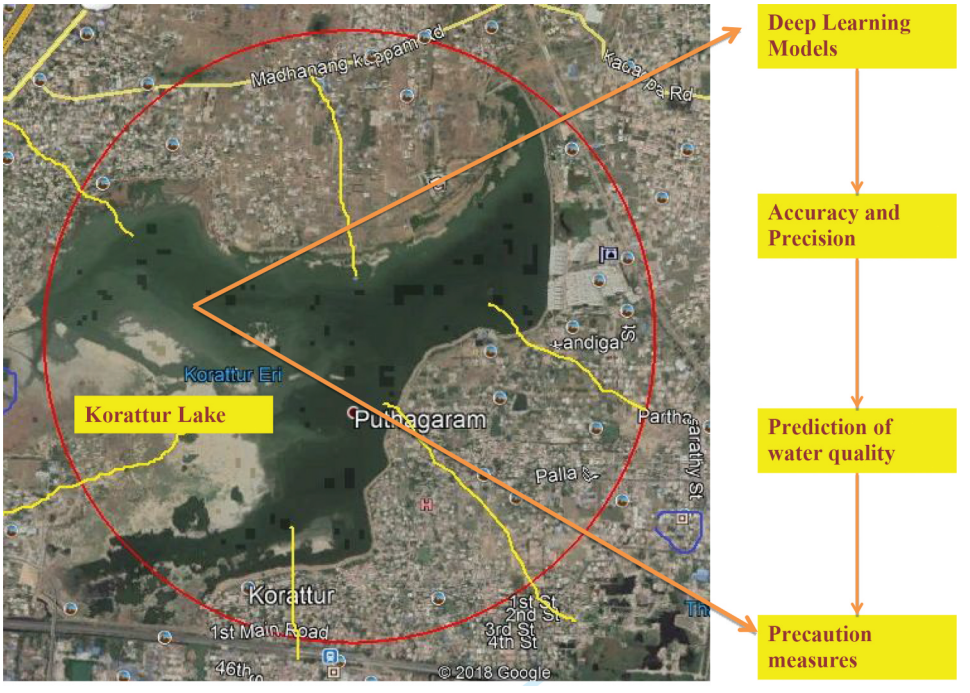

Figure 1. Satellite view of Korattur Lake, Chennai, India.

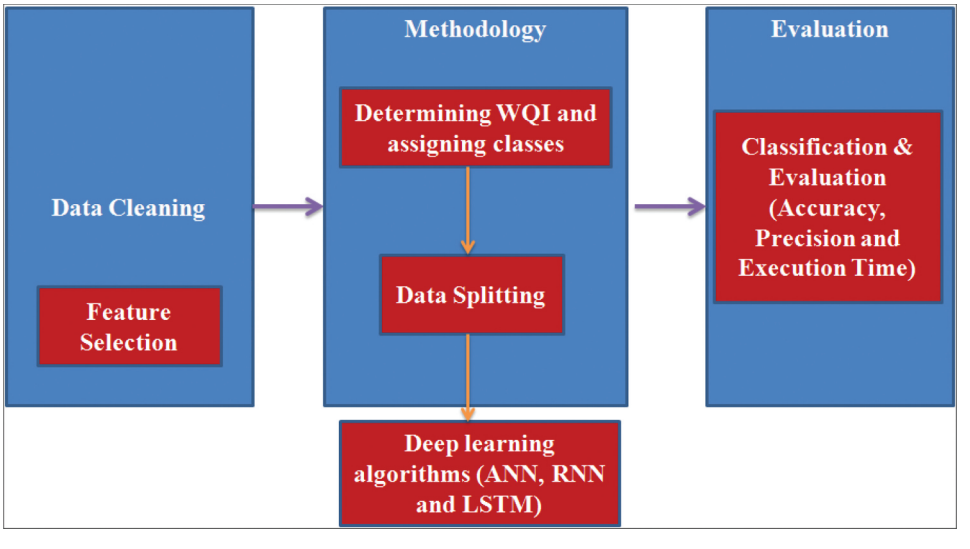

Figure 2. Water Quality Prediction System.

2.1.4. Water quality index

The water quality index is calculated based on the nine parameters such as pH, TDS, turbidity, phosphate, nitrate, iron, COD, chloride and sodium that can provide a simple indicator of water quality. The weights are assigned to each parameter based on the

Table 1. Dataset description.

Dataset	No. of Records	No. of Parameters	No. of Classes	Class Distribution
Korattur Lake	5000	9	2	Drinkable – 4325 Non Drinkable – 675
	5000	9	5	Excellent – 649 Very good – 1831 Good – 1450 Poor – 620 Very Poor – 450

Table 2. Desirable range of drinking water.

S.No	Parameters	Suitable range	Reference
1	pH	6.5–8.5	[23–28]
2	Phosphate	0.005–0.5 mg/L	
3	Total Dissolved Solids	300–600 mg/L	
4	Turbidity	<5 NTU	
5	Nitrate	<10 mg/L	
6	Iron	0.3 mg/L	
7	Chlorides	4 mg/L	
8	Sodium	< 20 mg	
9	COD	3–6 mg/L	

highest difference between min value and max value of that parameter [16]. After assigning weights for each and every parameter, the quality rating scale is found by using the formula (1) and (2)

$$Q_i = \left(V_i - V_{i0}/S_i - V_{i0}\right) * 100 \tag{1}$$

Where,

V_i stands for estimated value of n^{th} parameter

S_i is the desirable or permissible range

V_{i0} ideal value of n^{th} parameter in pure water. All ideal values are taken as zero for drinking water except pH = 7.0.

Then, the water quality index is found by

$$WQI = \Sigma\left(W_i * Q_i\right)/\Sigma_{i=1}{}^{n}\mathbf{W_i} \tag{2}$$

Where,

W_i is the weight allocated to each parameter. Based on the WQI, the classes are classified as shown in Table 3.

Table 3. Quality of water-based on WQI.

Water Quality Index Level	Water Quality Status	Reference
0–25	Excellent	[16,29]
25–50	Good	
50–75	Poor	
76–100	Very Poor	
> 100	Unfit for drinking	

2.2. Deep learning algorithms

The deep learning algorithms such as Artificial Neural Network (ANN), Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) are used for training and testing. The dataset is given as input for the deep learning models. The deep learning models are applied for both binary class as well as multi-class data. The parameters considered are: pH, Total dissolved salts (TDS), turbidity, phosphate, nitrate, iron, COD, chloride and sodium. The output obtained from various algorithms are compared and analysed to find the best suitable deep learning algorithm. The various deep learning algorithms are discussed below:

2.3. Artificial neural network

An Artificial Neural Network is an information processing technique that is intended to simulate the behaviour of the human brain. The ANN is composed of neurons where the processing takes place. ANN includes a large number of connected processing units that work together to process information. Figure 3 shows the structure of ANN. A neural network contains the following three layers:

2.3.1. Input layer

The purpose of the input layer is to receive as input the values of the attributes for each observation. Usually, the number of neurons in an input layer is equal to the number of attributes. Since our dataset contains 9 attributes, the input consists of 9 neurons, one for each input attribute. The input layer communicates to one or more hidden layers.

2.3.2. Hidden layer

In the hidden layer, the actual processing is done via a system of weighted 'connections'. There may be one or more hidden layers and the hidden layers can have any number of

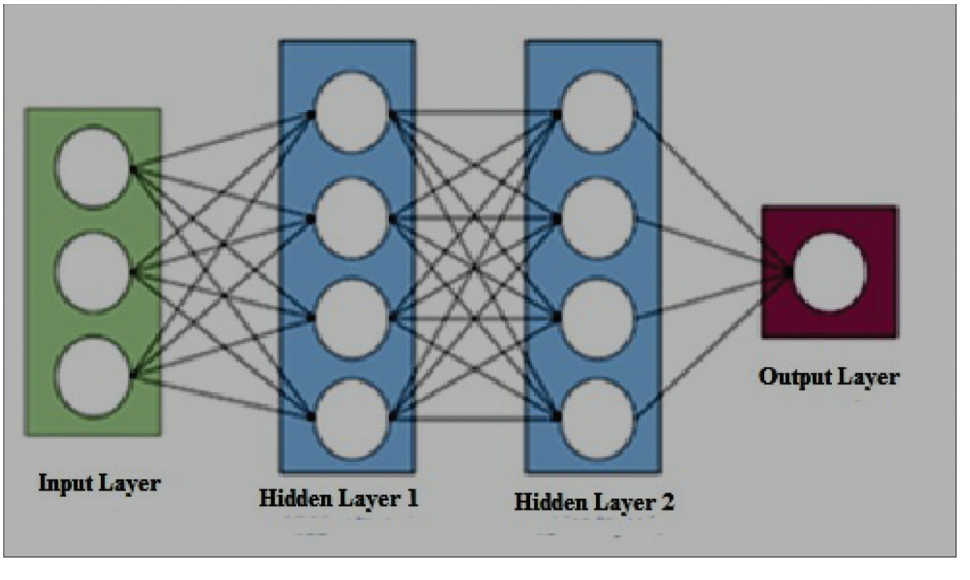

Figure 3. Structure of ANN.

neurons. The values entering a hidden node multiplied by weights, a set of predetermined numbers stored in the program. The weighted inputs are then added to produce a single number. The output of the last hidden layer is given as input to the output layer. We have used two hidden layers which consist of 15 and 9 neurons, respectively.

2.3.3. Output layer

The hidden layers then link to an output layer. It returns an output value that corresponds to the prediction of the response variable (Class). The number of neurons in the output layer corresponds to the number of output classes. Thus, in binary classification, the output layer has 1 neuron. For multi-class classification, the output layer consists of 3 neurons [17,18].

2.4. Recurrent neural network

Recurrent Neural Network (RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. RNN have a 'memory' which remembers all information about what has been calculated. Figure 4 shows the structure of RNN.

The RNN consists of an input layer, one hidden layer and an output layer. RNN also consists of an Embedding layer with masking, a recurrent layer which uses LSTM and a dropout layer. Input layer consists of 9 neurons since there are 9 attributes in the dataset. The embedding layer is used to compress the input features to an input

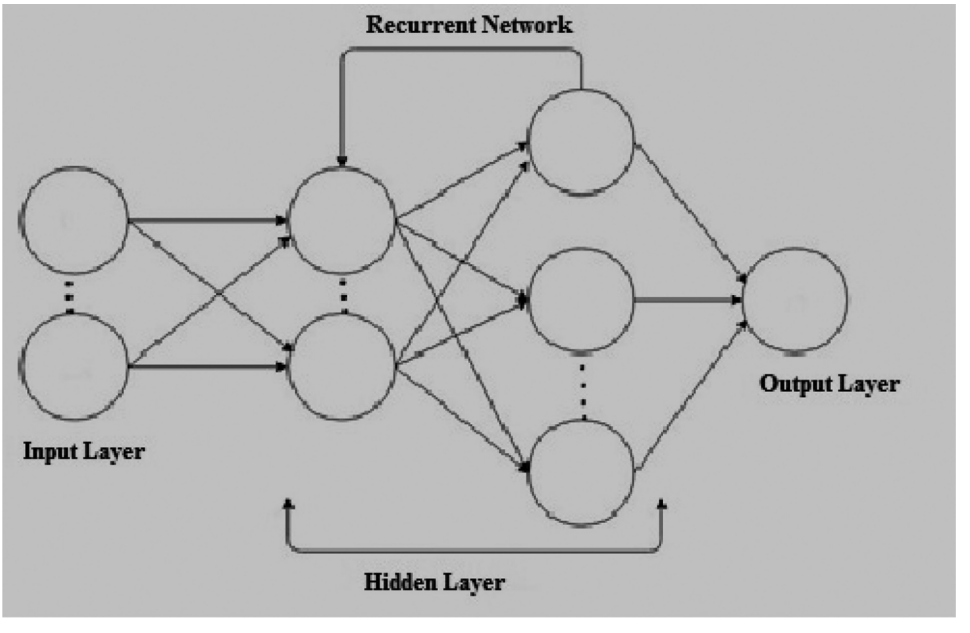

Figure 4. Structure of RNN.

vector. The maximum size of the input feature and the embedding size are given as input to this layer. Masking is used to skip the timestamps in the input data. The recurrent layer is the key layer of RNN. It calculates the gradient each time and recursively sends the value of the gradient to the hidden layers to reduce the value. The weights are adjusted each time to reduce the value of the gradient. Sometimes, there is a chance that the gradient can diminish to zero while back propagation. This problem is called the vanishing gradient problem. So, LSTM is used in the recurrent layer to overcome this problem and to increase the accuracy. Hidden layer can have any number of neurons. The values entering a hidden node multiplied by weights, a set of predetermined numbers stored in the program. The weighted inputs are then added to produce a single number. The output of the hidden layer is given as input to the output layer. Dropouts are added between the hidden layers to reduce the problem of over fitting in the neural network. Output layer returns an output value that corresponds to the prediction of the response variable (Class). The number of neurons in the output layer corresponds to the number of output classes. Thus, in binary classification, the output layer has 1 neuron. For multi-class classification, the output layer consists of 3 neurons [19].

2.5. Long short term memory

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. But in LSTM, the output from each node is given as input to itself for better performance and accuracy as shown in Figure 5. It is used to tackle the problem of long-term dependencies of RNN in which the RNN cannot predict

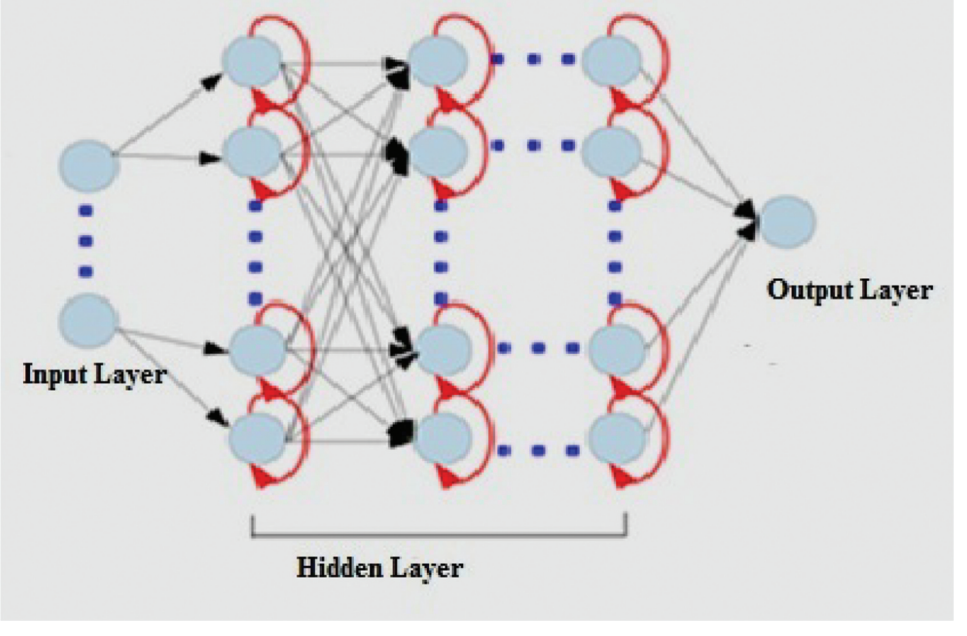

Figure 5. Structure of LSTM.

the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting and classifying on the basis of time series data.

The LSTM consists of an input layer, a LSTM layer and an output layer. LSTM also consists of an Embedding layer, a Conv1d layer, a max-pooling layer and a dropout layer. Input layer consists of 9 neurons since there are 9 attributes in the dataset. The embedding layer is used to compress the input features to an input vector. The maximum size of the input feature and the embedding size is given as input to this layer. Conv1D layer is used to map the input features to the output variable. Max-pooling layer is used to extract import features from the given attribute. LSTM layer is the key layer of LSTM where the actual processing takes place. That is, the output from each node is given as input to itself. The LSTM output size is usually determined according to the input data; we have set 70 as the LSTM output size. Dropouts are added between the hidden layers to reduce the problem of over fitting in the neural network. Output layer returns an output value that corresponds to the prediction of the response variable (Class). The number of neurons in the output layer corresponds to the number of output classes. Thus, in binary classification, the output layer has 1 neuron. For multi-class classification, the output layer consists of 3 neurons. The confusion matrix is obtained as output from all the deep learning models. From the confusion matrix accuracy and precision are calculated which are discussed in following section [19,20].

3. Results and discussions

3.1. Model evaluation metrics

3.1.1. Confusion matrix

A confusion matrix is a map between the actual and the predicted class label. The confusion matrix is shown in Table 4.

3.1.2. Accuracy

Accuracy is the measure of samples which correctly predicted in percentage. It is the important metric used to evaluate the performance of the model. It is the ratio of number of correct predictions by total no. of predictions [21]. The formula for accuracy is given by:

$$(TN + TP)/(TN + TP + FN + FP) \tag{3}$$

Table 4. Confusion matrix.

Class Names	Class 0 (Actual)	Class 1 (Actual)	Expansion
Class 0 (Predicted)	TN	FP	True Negative (TN): Samples which are negative, and are predicted to be negative. (i.e.) Non-drinkable water is correctly predicted as non-drinkable. False Positive (FP): Samples which are negative, but are predicted positive. (i.e.) Non-drinkable water predicted wrongly as drinkable.
Class 1 (Predicted)	FN	TP	True Positive (TP): Samples which are positive, and are predicted to be positive. (i.e.) Drinkable water correctly predicted as drinkable. False Negative (FN): Samples which are positive, but are predicted negative. (i.e.) Drinkable water predicted wrongly as non-drinkable.

3.1.3. Precision

Precision is a useful measure of success of prediction when the classes are very imbalanced. Precision refers to the fraction of the results which are pertinent. That is, precision is the rate of the number of samples correctly predicted as drinkable (Class 1) out of all the samples classified as drinkable (Class 1) by the model [21,22]. The formula for precision is given by:

$$TP/(TP + FP) \tag{4}$$

3.1.4. Execution time

The execution time refers to the time taken by the model for training and testing. It is measured in seconds. It includes the time taken for training and testing the model.

3.2. Accuracy

On comparing the accuracy of the deep learning models as shown in Figure 6, it is seen that the accuracy of binary classification and multiclass classification is almost the same. Out of all the algorithms, the LSTM algorithm was found to have the highest accuracy of 92% for binary classification and 94% for multiclass classification.

3.3. Precision

Precision may also be defined as the number of samples predicted correctly as drinkable water out of all the samples predicted as drinkable. On comparing the calculated precision of all the models as shown in Figure 7, it is seen that the precision is almost the same for binary classification and multiclass classification. Out of all the three algorithms, the

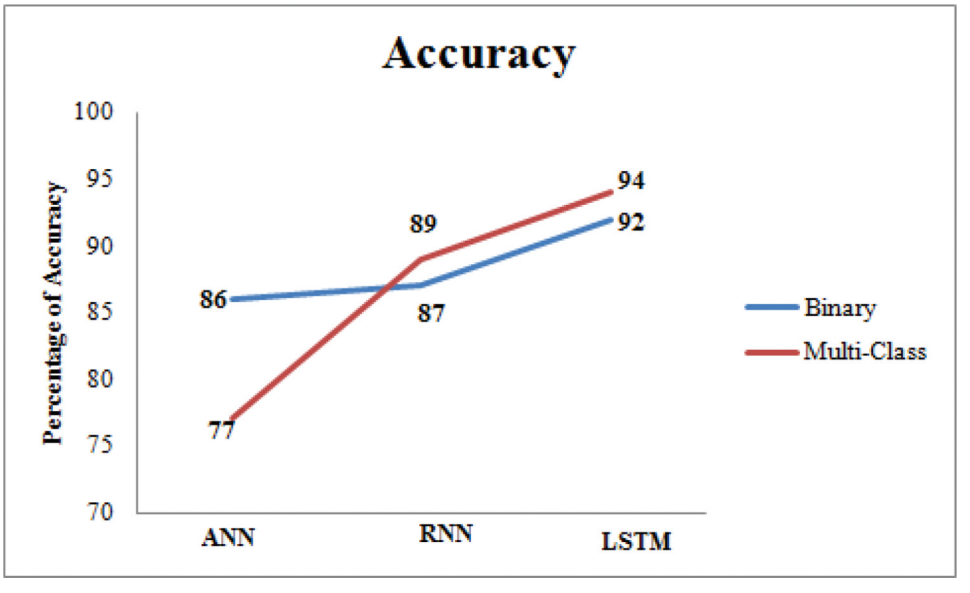

Figure 6. Accuracy of deep learning models.

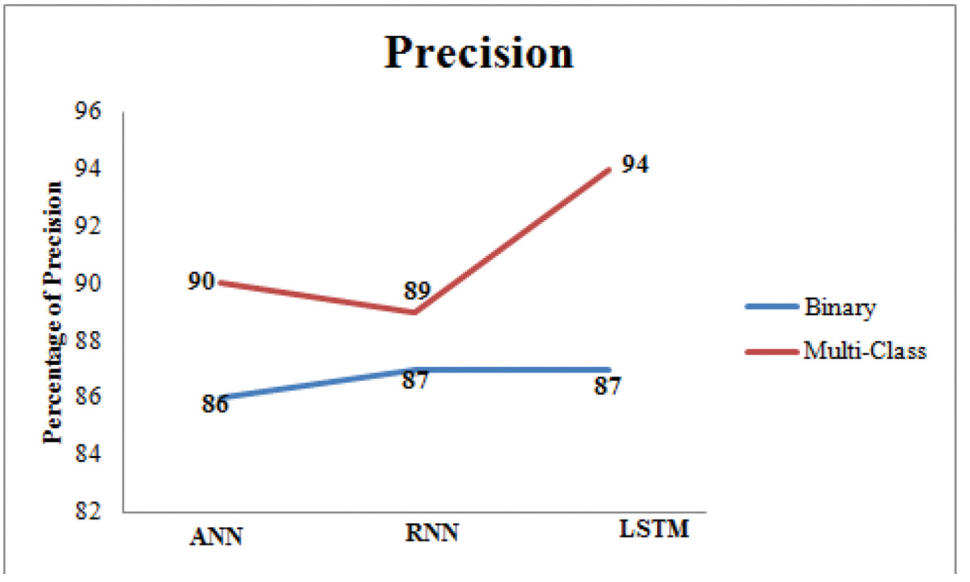

Figure 7. Precision of deep learning models.

LSTM has the highest precision of 87% for the binary classification of data and 94% for multi-class classification.

3.4. Execution time

The execution time refers to the time taken by the model for training and testing. The execution time of each algorithm is calculated in seconds. Table 5 shows that the execution time for both binary classification and multi-class classification is almost the same. LSTM has the least execution time of around 29 seconds for binary classification and 35 seconds for multi-class classification. The analysis and comparison of all the algorithms showed that among the deep learning models, LSTM has the highest accuracy of about 92% and also has the highest precision. The time taken for LSTM is the least. So the Deep learning model LSTM is considered to be the best suitable model for our dataset and hence is used for prediction of water quality for the next three years.

The current work is compared with the work done by F. Muharemi, D. Logofatu, et al. [3]. The authors in previous work have employed deep learning algorithms such as ANN, DNN, RNN and LSTM for prediction of binary data. Their ANN structure contains two hidden layers with two neurons each and RNN contains 10 hidden layers RNN, with a single linear output layer. They considered the parameters such as Time, Turbidity, pH, Electrical Conductivity, Water Temperature, Chloride (cl), Redox Chlorine dioxide and

Table 5. Execution time (in seconds) of the deep learning models.

Algorithm	Binary Classification	Multi-Class Classification
ANN	51 seconds	57 seconds
RNN	57 seconds	55 seconds
LSTM	29 seconds	35 seconds

flow rate to predict the quality of a lake in Germany. They evaluated their work by using the metrics such as F1-score and precision. Out of the algorithms they employed, DNN produced better results for their prediction. The precision of DNN was found to be 92%.

In the current work, a long-term time series data of Korattur Lake in Chennai were considered with the most important parameters that contribute to the water quality such as pH, Total Dissolved Salts (TDS), turbidity, phosphate, nitrate, iron, Chemical Oxygen Demand (COD), chloride and sodium. The deep learning algorithms considered are ANN, RNN and LSTM. For ANN we have used two hidden layers which consist of 15 and 9 neurons and RNN consists of an input layer, one hidden layer and an output layer also consists of an embedding layer with masking, a recurrent layer which uses LSTM and a dropout layer. We have done both binary and multi class classification and in multi-class we have five different classes including Excellent, Good, Average, Bad and Poor. The LSTM algorithm was found to have the highest accuracy and precision of 94% compared to all the other algorithms. LSTM can keep track of a large number of patterns from a long series time data by giving weightage to the values which are passed deciding their level of importance so that it can classify process and predict time series data producing highly accurate results. Figure 8 shows the comparison of precision of both the works. It is seen that the LSTM model used in the current work has a higher precision than the DNN model used by F. Muharemi, D. Logofatu, et. al [3].

3.5. Future forecast

The methodology of analysing water quality of natural resources through machine learning and deep learning process can be extended with multiple water quality parameters along with its interaction with soil strata. Every pollutant has its own nature of transportation in water and its related zones with respect to prevailing environment.

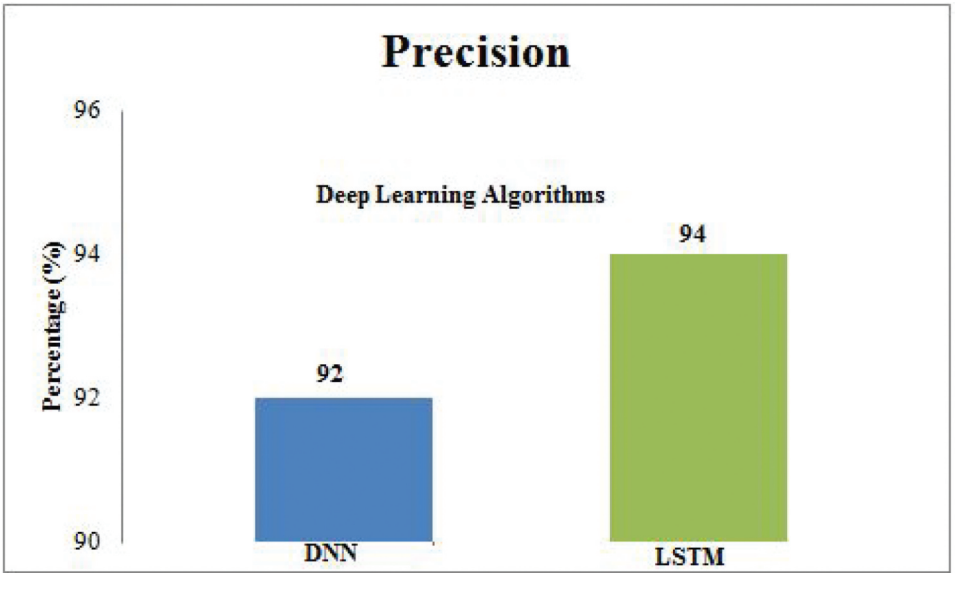

Figure 8. Comparison of precision.

Hence, water quality could be analysed before and after monsoon period as during post monsoon period dispersion of pollutants will be high hence transportation will be huge in the surrounding area. This is a easy and economical way of analysing water quality which could be adopted with maximum level of hydrological and hydrochemical data.

4. Conclusion

The models used for Training and Testing include Deep Learning models such as ANN, RNN and LSTM for binary classification and multi-class classification. The deep learning models produced an average accuracy of around 93%. Out of the three deep learning algorithms, LSTM was found to be the best suitable algorithm for our work since it produces the highest accuracy of 94%. LSTM also has the highest precision and consumes the least time for execution compared to all the other deep learning algorithms. So, the model can be used for prediction of water quality for the next 3 years. The water quality of the Korattur Lake for the next 3 years (2020–2022) was predicted using LSTM by collecting the current samples of water from the lake. The attributes of the water were given as input to the model and the model predicted the quality of water to be 'Not drinkable' since the sewage water has recently been drained into the lake water and its water quality is very poor. The Korattur Lake is under water treatment currently. So, testing the water after the treatment will give accurate results for the next 3 years. Certain preventions could be adopted in order to prevent contamination of lake which would conserve resource at its own cost based on the prediction done in the work and the results given by deep learning algorithm. The impacts of due to dispersion of pollutants in lake is shown in Figure 9.

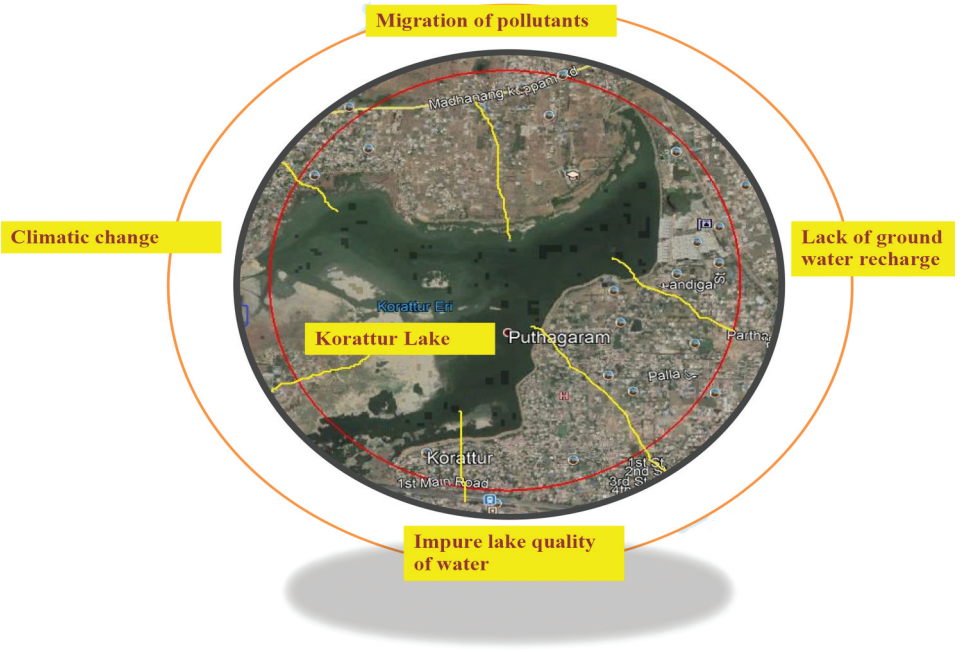

Figure 9. Impacts of pollutants on lake water and its hydrology.

Conservation of resource would help nearby public and municipalities in performing its domestic activities without any interruptions at any cause. Precautions like

- Proper channelling of sewage drains would prevent contamination.
- Dumping and disposing of solid waste in dump yard should be encouraged.
- Encouraging public in carrying domestic activities nearby lake should be avoided.
- Proper conserving of rain water would help in proper recharge of groundwater and lake water.
- Afforestation should be encouraged for conserving water and soil fertility.
- Treatment of sewage using novel methodologies may destroy all with held contaminants thereby reducing the toxicity after discharge in natural streams.
- Discharge of untreated effluents should be prohibited.

The work can be extended by including some more classes to the data and the data set can be trained using hybrid models of machine learning and deep learning. Since deep learning algorithms can handle large amounts of data, the hybrid model might be more efficient as it can produce high accuracy as well as can handle large data sets. This is another way of predicting water quality that excludes interaction of soil with water. It is very well known that hydrological contaminants get transported across water and reach sea water that further contaminates marine system. Hence, it is very well understood that once a part of water resource is contaminated every part is destructed. These deep learning models could be easily adopted in predicting the water quality across world for proper conservation of resources. But consideration of interaction of water contaminates with soil will give much more idea on nature of contamination in the chosen zone which would be considered in future work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Y. Khan and C.S. See, IEEE Long Island Systems, Applications and Technology Conference (LISAT) IEEE, New York, 2016, 1.
[2] A. Kistan, V. Kanchana and A.T. Ansari., Int. J. Sci. Res. (IJSER) **4** (5), 3469 (2013).
[3] F. Muharemi, D. Logofătu and F. Leon, JIT **3**, 294 (2019).
[4] P. Liu, J. Wang, A.K. Sangaiah, Y. Xie and X. Yin, Sustainability **11**, 2058 (2019). doi:10.3390/su11072058.
[5] Y. Wang, J. Zhou, K. Chen, Y. Wang and L. Liu, 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) IEEE, NanJing, JiangSu, China, 2017, 1.
[6] S.H. Agrawal and K. Khare, Int. J. Comput. Appl. **125**, 975 (2015).
[7] L. Xu and S. Liu, Math. Comput. Model **58**, 807 (2013). doi:10.1016/j.mcm.2012.12.023.
[8] A.H. Haghiabi, A.H. Nasrolahi and A. Parsaie, Water Qual. Res. J **53**, 3 (2018). doi:10.2166/wqrj.2018.025.
[9] National Water Quality Monitoring Programme, *Fifth Monitoring Report (2005–2006)* (Pakistan Council of Research in Water Resources (PCRWR), Islamabad, Pakistan, 2007).
[10] S. Mehmood, A. Ahmad, A. Ahmed, N. Khalid and T. Javed, Sci Rep **2**, 1 (2013).

[11] A. Azizullah, M.N.K. Khattak, P. Richter and D.P. Häder, Environ. Int. **37**, 479 (2011). doi:10.1016/j.envint.2010.10.007.
[12] N.M. Gazzaz, M.K. Yusoff, A.Z. Aris, H. Juahir and M.F. Ramli, Mar. Pollut. Bull. **64**, 2409 (2012). doi:10.1016/j.marpolbul.2012.08.005.
[13] https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-ravaged-by-raw-sewage-korattur-lake-now-lies-encroached/articleshow/74328391.cms (accessed May 11, 2020).
[14] https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-deadline-over-government-seeks-1-month-to-plan-korattur-lake-clean-up/articleshow/74372292.cms (accessed May 12, 2020).
[15] P. Deepa, R. Raveen, P. Venkatesan, S. Arivoli and T. Samuel, Int. J. Chem. Stud. **4** (3), 116 (2016).
[16] I.N. Balan, M. Shivakumar and P.D.M. Kumar., J. Chron. Young Scient. **3**(2), 146 (2012). doi:10.4103/2229-5186.98688.
[17] M.H. Hassoun, *Fundamentals of Artificial Neural Networks* (MIT Press, Cambridge, MA, 1995).
[18] https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (accessed May 12, 2020).
[19] https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e (accessed May 11, 2020).
[20] S. Hochreiter and J. Schmidhuber, Neural Comput. **9**, 1735 (1997). doi:10.1162/neco.1997.9.8.1735.
[21] M. Sokolova, N. Japkowicz and S. Szpakowicz, Australasian joint conference on artificial intelligence, Springer, Berlin, Heidelberg, 2006, 1015.
[22] E.G. Goutte, European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2005, 345.
[23] World Health Organization. Guidelines for drinking-water quality. World Health Organization, 1, (1993).
[24] World Health Organization. Water quality and health-review of turbidity: information for regulators and water suppliers (No. WHO/FWC/WSH/17.01). World Health Organization. (2017).
[25] A. Colter and R.L. Mahler. *Iron in Drinking Water* (University of Idaho, Moscow, Idaho, 2006).
[26] O. Fadiran, S.C. Dlamini and A. Mavuso, Bull. Chem. Soc. Ethiop. **22** (2008). doi:10.4314/bcse.v22i2.61286
[27] https://water-research.net/index.php/chlorides-and-salts-in-water-future-problem-for-groundwater-users (accessed May 12, 2020).
[28] D. W. Advisory, *Consumer Acceptability Advice and Health Effects Analysis on Sodium. US Environmental Protection Agency Office of Water (4304T)* (Health and Ecological Criteria Division, Washington, DC, 2003), p. 20460.
[29] A.K. Chaurasia, H.K. Pandey, S.K. Tiwari, R. Prakash, P. Pandey and A. Ram, J. Geol. Soc. India **92**, 76 (2018). doi:10.1007/s12594-018-0955-1.