A Survey on Water Quality Analysis and Prediction

*Kavya Divakar, #Chitharanjan K

*Post Graduate Student, #Assistant Professor, SCT College of Engineering, Kerala, India,

*kavyadivakar95@gmail.com, #chitharanjan_k@yahoo.com

Abstract **The rise in population is tremendously increasing day by day by which the depletion of natural resources is putting the life at a risk. Water is one among them that is getting highly polluted and scarce due to some factors like the dumping of the waste, the release of chemicals from factories, etc. Hence, there is a need to check the quality of these water resources for each purpose. In this paper, a survey has been done about water quality analysis and their predictions. The newly emerged technologies like machine learning and deep learning is widely used to perform the analysis and prediction of various real time applications. Water quality is assessed based on the standards proposed for each parameter present in the sample. They can be efficiently utilized to accurately predict the quality of water samples. A literature survey will help us to know about the methods that have been proposed which will be useful at the current and in the future times.**

Keywords —AdaBoost, CatBoost, Gradient Boost, HistGradient Boost, Light Gradient Boost, XGBoost.

I. INTRODUCTION

The population hulk is affecting all the natural resources of earth. Freshwater management is an essential one in the current scenario. To check the quality of water samples collected, there exists a number of traditional ways. It mainly includes carrying the water sample to the nearest laboratories and to check with the help of various lab equipment. But this has many limitations. First, the cost of these equipment is high and hence the cost required to check the sample is also high. Secondly, it takes more than one day for some parameter to get checked and hence time consumption is also high.

In this perspective, there is a need for a better resolution. The role of artificial intelligence comes to play here. Technologies like machine learning, deep learning etc., help to solve these problems. Many machine learning and deep learning algorithms exist which gives proper training and accurate predictions to various problems. In the case of water quality prediction, all these algorithms can be used for proper training. Certain parameters of water samples give accurate prediction using some algorithms whereas some other algorithms fail in proper prediction. Hence, in this survey, we will discuss about water quality prediction using various machine learning and deep learning algorithms.

Under Section II, the need to evaluate water quality is discussed. Section III deals with various literature review by comparing the methods proposed by each one of them. Section IV have summarized the contents and compared the points of the existing research work.

II. WATER QUALITY EVALUATION

Water quality can be simply defined as a measure of the physical, chemical, biological, and microbiological characteristics of water. By monitoring water quality, empirical evidence is provided to make decision on health and environmental issues. During olden times, the chemicals present in the water was less and hence the quality of these water resources were good. But now in the 21st century, many chemicals are being used in our everyday lives, and that indeed make a way to water resources. The need for water quality is to ensure the purity of water that is consumed by living beings. It is not always possible to carry the water samples to nearby or far away located laboratories to check their quality. It is during this time, the need for a proper water quality monitoring system is required. A well-designed monitoring system can serve the current and future needs even if it is not foreseen.

III. LITERATURE REVIEW

The history of water quality monitoring is dated centuries back. During olden days, the methods involved to monitor water quality were time-consuming, risky and inaccurate. One of the traditional method to analyze bacterial count was Heterotrophic Plate Count(HPC). HPC estimate heterotrophic bacterial count by recording their level of growth on a non-selective nutrient containing media. But they lack accurate bacterial count and hence not a good method. Other ways to monitor water quality is by:

[1] Simple Observation of Water Sample

 DOI : 10.35291/2454-9150.2020.0492

[2] Using Portable Test Kits Available
[3] Using a Mobile Laboratory
[4] Sending Samples to Commercial Laboratory.

Since all the above mentioned methods fail at some point of time, there is a need for proper water quality monitoring. In this survey, we will discuss some methods that had been implemented for proper monitoring of water quality.

A. Towards Building a Predictive Model for Remote River Quality Monitoring for Mining Sites

Maria Regina et. al discussed about a low cost mobile electronic sensor which was developed to monitor water quality in 8 sites of baseline river water near mining sites [2]. Data collection was done once every two months for almost a year. This data is then send to mobile device via Bluetooth. A total of 150 water samples were acquired using sensors. The dataset contains information like Date, Site, Trial, pH, Dissolved Oxygen, Arsenic, Temperature, Turbidity, Salinity, Mercury and Water Quality, and the target includes Good, Not Good or Unknown based on the scores of each element present in it. The data are then processed and sent to the cloud server for remote storage.

Various sensors are used to collect parametric values.

I. The constituent elements like Arsenic, Cadmium, Copper, etc. are monitored using LIBS sensor.

II. Optical sensor is used to monitor either Copper or Zinc which is sent via Bluetooth to Electric Strip based sensor to check water quality.

III. Atmospheric sensor is used to monitor Mercury.

All these data are transmitted to the cloud which is then stored in a main server. A backup is also stored in a testing/backup server. The main server processes the collected data, and the output is sent to Mobile Applications, Display Dashboard and to the Website.

The predictive model is created by a Conditional Inference Tree. They take a set of input parameters and return the predicted outcome. For this particular dataset, the outcome was discrete and hence classification decision tree induction was more appropriate. A local server retrieves the data for analysis and model and transmit the processed data back to the server. This processed data would be sent to mobile phone of users in the form of notification.

B. Water Quality Prediction Model of a Water Diversion Project Based on the Improved Artificial Bee Colony–Backpropagation Neural Network

Siyu Chen et.al proposed an improved artificial bee colony (IABC)-back propagation neural network algorithm to predict water quality [4]. The data was collected on a daily basis for a period of two months from Jiangsu Province Hydrology and Water Resources Investigation Bureau. Seven parameters like pH, Dissolved Oxygen (DO), Permanganate Index (COD), Biochemical Oxygen Demand (BOD), Nitrate, Petroleum and Volatile Phenol. The entire study area was divided into 6 stages where they used first day's data of first stage to predict the next day's data at sixth stage.

To build Back-Propagation (BP) neural network, three parameters such as DO, COD and BOD were selected as neurons of the input layer. To train the model, water samples collected during the first 50 days were used and the remaining 16 sets of data were used for testing the results. BP neural network consists of an input layer, multiple hidden layers and a single output layer. All layers are connected to each other and not connected within the same layer. The BP neural network performs forward propagation of the signal and back propagation of the error. But it has two limitations:

- BP algorithm has ideal training for five layers, but is difficult to train in the deep structure because of the local minimum problem in the non-convex objective function.
- BP converges easily to local minimum values.

To overcome these issues, BP algorithm has to be improved. The principle of Artificial Bee Colony (ABC) algorithm is a combinational optimization algorithm which was inspired by the intelligent behaviour of honey bee swarm when foraging for food. The ABC algorithm search for the optimum connection weight values between neurons and the threshold values of each neuron. The ABC algorithm adopted the random generation method to randomly generate several individuals to form the initial group.

The ABC algorithm has two renewal processes: the updating process of the employed bees and onlookers where the former is a process of global searching and the latter is a local searching process. The popular swarm optimization (PSO) adds the present optimal solution and the global optimal solution to accelerate the convergence rate of the algorithm. The experiment had three parts:

1. First part: To compare the performance of the pure BP neural network and ABC-BP algorithm.
2. Second part: Compare the performance of the BP and IABC-BP algorithm.
3. Third part: Compare the performance of the IABC-BP and ABC-BP algorithm.

The water quality was predicted with the optimization of the connection weight and threshold value in the neural network. Thus the number of input and output node was 3 and the suitable number of neurons in the hidden layer was 7. After repeated simulation tests, the size of bee colony

 DOI : 10.35291/2454-9150.2020.0492

was found to be 100. To ensure the accuracy of the models, three indicators like Relative Error, Coefficient of Determination and Nash–Sutcliffe Efficiency (NSE) Coefficient were selected.

Four models of BP, PSO-BP, ABC-BP, and IABC-BP neural network were constructed for the data and analysed the above indicators. The relative error shows that the error of each index in the IABC-BP neural network model is almost the smallest of the four models by making it the most suitable model of the four models.

Table. 1 The Coefficient of Determination

Algorithms	**Coefficient of Determination**
BP	0.658
PSO-BP	0.918
ABC-BP	0.942
IABC-BP	0.981

Table. 2 The NSE Coefficient

Algorithms	**NSE Coefficient**
BP	0.134
PSO-BP	0.296
ABC-BP	0.541
IABC-BP	0.805

As referred from Table.1, IABC-BP is having the highest coefficient of determination as compared to other classifiers. Similarly, in Table.2, IABC-BP shows the highest value for NSE coefficient compared to other classifiers. Hence, it can be referred from both the tables, table.1 and table.2 that the IABC-BP predicting model fits better than the other three models. Overall, the results indicate that IABC-BP model is equipped with the best quality and highest reliability.

C. Back to the future: IoT to improve aquaculture: Real-Time Monitoring and Algorithmic Prediction of Water Parameters for Aquaculture Needs

Maxime Lafont et.al described the feedback of the use of IoT for the real-time monitoring of water quality and its development in prediction systems in an aquaculture [11]. For this, they have developed a device for real-time monitoring of water quality using physico-chemical parameters. They deployed three devices in three kinds of farms: an oyster and mussel farm, a fish farm and a spirulina farm where:

- Oyster/mussel and fish farms-Mediterranean Sea(No electricity and no internet).
- Spirulina farm- Greenhouse in France(Have electricity but no internet).

The monitoring device is made up of an electronic card, rechargeable batteries, solar panel, LoRaWan module and antenna, waterproof sealed box and water quality sensors. The sensors include Temperature, Salinity (derived from conductivity, turbidity, dissolved oxygen).

The sensor measure was made every 20 minutes and sent to the gateway in an encrypted LoRaWan packets and then send to LoRa Server for the storage. For the aquaculturists to observe and visualise data, they have developed a real-time monitoring interface. Also, when any of the parameters exceeds a certain critical value, automatically, the user receives a message and an email to inform him.

D. Deep Neural Network for Marine Water Quality Classification with the Consideration of Coastal Current Circulation Effect

Carlin et.al performed a classification performance among machine learning models using Deep Neural Network [8]. The data was collected from the Hong Kong Environmental Protection Department(EPD) with 24 marine water quality parameters. The target is the concentration level of red tide. Four time lagged measurements like Chlorophyll-a, Total Nitrogen, Total Phosphorus and Dissolved Oxygen are selected to model the occurrence of red tide. Two network structures with 3-12 layers are used to study the two factors to improve the modelling performance:

- Increase in the number of layers.
- Increase in the number of nodes within a layer.

One network structure was constructed with an increasing number of nodes in each successive layer, and hence called Funnel-type structure. The other network structure was constructed using a fixed number of nodes in hidden layer, and hence called 3-node structure.Non-linear hyperbolic tangent activation function is applied for at-most 4 layers before the output layer. Linear identity function is used in the rest of the additional hidden layers and in the output layer, Soft-max activation function is used. Red tide occurrence is labelled as the positive class. To compare the performance of these models, k fold Cross-Validation is implemented with k=5. The empirical analysis is carried out using SAS v9.4 and SAS Enterprise Miner v14.1. Neural networks with 11 layers (funnel-type architecture) had good performance regarding Average Profit, Misclassification Rate and ROC Value whereas it had poor performance for the average profit under 3-node architecture.

E. Design of River Water Quality Assessment and Prediction Algorithm

Sheng Cao et. al proposed a water quality assessment model using the Least Squares Support Vector Sorter (LS-SVC) [14] to analyse the water quality level. The samples were collected from National Surface Water Monitoring Centre. LS-SVC model was used because it is an improved version

 DOI : 10.35291/2454-9150.2020.0492

of SVC which replaced the slack variable with the square of the training error to reduce the workload. LS-SVM based adaptive Particle Swarm Optimization (PSO) algorithm was used to increase the accuracy of prediction.

Fuzzy information granulation method has to be combined with the LS-SVR algorithm and for that the information has to be divided into independent windows first. The window size has to be properly selected to maintain the original information and to effectively simplify the time series data. Next, each window has to be blurred to generate fuzzy information particles.

The trained model was used to evaluate the samples collected from the river. The first 150 groups, 153 groups and 156 groups of data was processed for fuzzy information grain processing, and the LS-SVR regression model was used to predict the water quality.

The accuracy of this model in predicting the training sample was 97% and that of the testing sample was 94%. Hence, the improved PSO algorithm improved the results' accuracy by 1.5% and also the convergence speed was significantly faster when compared to basic PSO algorithm.

F. Ground Water Quality Prediction using Machine Learning Algorithms in R

S. Vijay et.al discussed about the characteristics of ground water quality in some towns of Vellore district in Tamil Nadu [9]. They had chosen Arcot, Ranipet, and Walljah pet towns for the study. The water samples were collected from 30 locations of different bore wells of these towns which are exclusively used for drinking purpose.

The water parameters such as pH, Chloride, Sulphate, Nitrate, Carbonate, TDS, EC, Bicarbonate, metal ions and trace elements were observed. The target of the observations were High water contamination and Low water contamination. RStudio was used as the tool to analyze the data and observe the result. Machine learning algorithms like C5.0, Naive Bayes and Random Forest were used to classify the water quality data.

Table. 3 Performance Comparison of Models

Algorithms	Accuracy	Kappa	Specificity	Sensitivity
Naive Bayes	100%	1.0	1.0	1.0
Random Forest	100%	1.0	1.0	1.0
C5.0	93%	1.0	1.0	1.0

The results as shown in Table.3 indicates that the classifiers such as Naive Bayes and Random Forest had 100% accuracy along with other measures like Kappa, Specificity and Sensitivity to be 1 whereas, C5.0 had an accuracy of 93% with other measures of Kappa, Specificity and Sensitivity to be 1. Hence Naive Bayes and Random Forest classifiers were good in classifying data.

G. Evaluation of Groundwater Quality at Eloor, Ernakulam District, Kerala Using GIS

Fehmida Fatima et.al proposed the ground water quality in Eloor in Ernakulam district of Kerala by using Geographical Information System (GIS) [10] based method. Eloor, being a river island is formed between two distributaries of Periyar river. 30 water samples were collected during the period January 2018 to February 2018 from open wells at Eloor. Parameters such as pH, TDS, Temperature, Chloride, Sulphate, Iron, Nitrate, Hardness, COD and Lead were analyzed. The coordinates of these samples were found using handheld GPS.

Geostatistical methods were used for mapping spatial variation of parameters. Kriging is one such method, which can not only predict values at unmeasured locations, but also determine the uncertainity in prediction. Before using this method, exploratory data analysis to determine the suitability of data was done. Normal distribution of the data was evaluated using Histograms and Q-Q plots. If they are not normal, logarithmic and arcsine transformations were done to normalize the data.

3 types of models such as spherical, exponential and Gaussian were used to predict parameters at unmeasured locations. The best fitting model will be selected based on Root Mean Square Error (RMSE). The model with low RMSE will have more accurate prediction and will be used to generate the thematic map. All three models were found to fit some parameters like:

- Spherical Model: TDS, Nitrate, Hardness and Iron.
- Exponential Model: pH.
- Gaussian Model: Temperature, Chloride and Sulphate.

The results indicated that the north eastern parts of Eloor had high acidity whereas it decreased when moved towards the western and southern regions. High temperature were shown in the central, southern and eastern parts. TDS and sulphate were shown to be within the limit in the whole area, but still had high value in the eastern region. Chloride, Iron and Hardness concentration exceeded the limits in almost all parts of Eloor. Nitrate concentration was high in the north-western parts. Overall, groundwater of Eloor in northern and eastern parts were found to be polluted and unfit for human consumption.

H. Predictive Analysis of Water Quality Parameters using Deep Learning

Archana Solanki et.al proposed a predictive analytics of water quality data of Krishna river in India on WEKA tool [1]. The data include the parameters like Dissolved Oxygen,

 DOI : 10.35291/2454-9150.2020.0492

pH and Turbidity. Since they had no target labels, clustering technique was applied to create three clusters based on seasons like Winter, Summer and Monsoon. Two unsupervised learning methods called Denoising Auto-Encoder and Deep Belief Network were used to develop the model.

The stacked denoising auto-encoder learned one layer at a time in which each layer was trained as a denoising auto-encoder by minimizing the reconstruction error. Once all the layers got trained, it goes to the next stage called fine-tuning. Fine-tuning was done to minimize the prediction error on a supervised task. To perform this, at first, logistic regression layer was added on the top of the network and then the entire network was trained as in Multi Layer Perceptron(MLP). Both the auto-encoder and sigmoid layers of MLP share parameters and also the results of intermediate layers of MLP is fed as input to the auto-encoder.

Deep Belief Network (DBN), being a deep neural network with many hidden layers are connected to each other, but the individual units are not connected. DBN was constructed using Restricted Boltzmann Machine(RBM). Greedy layer unsupervised training was applied to DBNs with RBMs as the building blocks for each layer. The predictive model was built using continuous deep belief network that use continum of decimals instead of binary data. After the training, feed forward neural network was applied to get the prediction and the results thus obtained are compared with MLP and Linear Regression.

The results indicated that the parameter turbidity had high variability among all the three parameters. Also, the deep learning approaches were able to handle the under and over fitting of predictions as compared to MLP and Linear regression.

I. Internet of Things Enabled Real Time Water Quality Monitoring System

S. Geetha et.al presented a survey work conducted based on the technologies used in water quality monitoring and have also proposed an in-pipe [7] water quality monitoring system. This model was used to test water samples and also an alert mechanism to detect any deviation in water quality parameters. Water quality sensors were used to monitor real-time parameter readings of Conductivity, Turbidity, Water Level and pH. A single chip microcontroller TI CC3200 with an in-built Wi-Fi module and ARM Cortex M4 core was used which could be connected to the nearest Wi-Fi hotspot for internet connectivity. The software part was programmed using ENERGIA IDE.

The sensor readings were collected by microcontroller which was then sent to Ubidots cloud. In this cloud, threshold was set for each parameter based on the standards proposed by WHO. Ubidots features included a real-time dashboard to analyze data and also to create events based on measured parameter values. If any of these parameters' threshold cross a limit, a message would be sent to user's mobile. Finally, a mobile application was developed to display the sensor readings which could be possibly checked by water authorities and end-users**.**

J. Water Quality Monitoring: From Conventional to Emerging Technologies

Umair Ahmed et.al has done a survey regarding water quality monitoring system from conventional methods [13] to the newly emerged technologies like IoT, Machine Learning, etc. Also they have proposed a low cost, water quality monitoring system utilising IoT and Machine Learning. The proposed architecture is an advanced version of the existing water monitoring system along with the detection of anomalous events. This architecture consists of various modules like Sensing module, Coordinator module, Data processing and Analysis module and the Storage and Core Analytics module. The details of each are given below:

- Sensing Module: The sensing module consists of four sensors to monitor Temperature, pH, Turbidity, and Total Suspended Solids.
- Coordinator Module: Act as a coordinator between the sensing module and the data processing and analysis module. It uses Arduino microcontroller to take the sensor readings and transmits them through a ZigBee transceiver to the on-site computer.
- Data Processing and Analysis Module: This module comprises of web services, data preprocessing service, storage service, and the data analysis service. Once the sensor data is received, they are locally stored in a MySql database using storage service and also useful data is filtered using data preprocessing service. Since a large amount of data is stored, it is analysed to find the useful information.
- Storage and Core Analytics Module: At first, they ensure long-term storage of water parameter readings and second, they predict water quality using machine learning techniques and detect anomalous events. Once the data preprocessed stage, it is transferred to the cloud using the REST web service. The detection of anomalous events is also done in this module. Two machine learning models are deployed: one of them is Artificial Neural Network(ANN) and the other one is K-means clustering. The water quality data is classified into three clusters: Class 0, Class 1 and Class 2. Normal data points which are fit to drink are assigned to Class 1, those data points fit for other purposes are assigned to Class 2 and those data which is totally unfit is assigned to Class 0. Any new points which clear the

 DOI : 10.35291/2454-9150.2020.0492

anomalous detection is only allowed to join a cluster based on its peculiarity.

- Application Dashboard: The dashboard is used to visualise water quality data on desktop, web and mobile platforms. The dashboard which is developed in Java visualises the information in the form of graphs and heat maps. The web dashboard has two instances: one which is deployed on the local desktop computer on site and the other instance is deployed on the internet and uses the REST web service. Also, the android application can also access the data through the same web service.

IV. CONCLUSION

Water is one of the precious resources that is highly necessary and inevitable for the existence of all living beings on this planet. Due to its high demand and scarce availability, there is a need to check the proper quality of these water resources. Water quality monitoring deals with the change in variability of parameters present in the water bodies. A lot of methods have been proposed since centuries back for the assessment of water quality. Each method dealt with a unique pattern to assess the quality and to conclude its result. But, most of these traditional methods failed at some point of time because it could not meet the needs as required.

Newly emerged technologies in Artificial Intelligence such as machine learning and deep learning are used in most of the real time applications to produce an effective monitoring system. From the survey we have conducted, we came to know that, the machine and deep learning algorithms could produce an effective water quality monitoring system that trains the collected samples to verify the permissible limit suggested for each parameter according to the water quality standards proposed so far. By comparing the performance, we understood that the deep learning neural networks are more effective in training a large number of samples by calculating the gradient descent and updating on its own to make the correct predictions. This in turn reveals that the future is in the hands of artificial intelligence.

REFERENCES

[1] Archana Solanki, Himanshu Agrawal and Kanchan Khare, "Predictive Analysis of Water Quality Parameters using Deep Learning," *International Journal of Computer Applications*, vol.125, No.9, September 2015.

[2] Maria Regina Justina E. Estuar, Emilyn Q. Espiritu, Erwin Enriquez, Carlos Oppus, Andrei D. Coronel,Maria Leonora Guico and Jose Claro Monje, "Towards Building a Predictive Model for Remote River Quality Monitoring For Mining Sites", *TENCON 2015 IEEE Region 10 Conference*, 07 January 2016.

[3] A. N. Prasad, K. A. Mamun, F. R. Islam and H. Haqva, "Smart Water Quality Monitoring System", *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), IEEE*, 23 May 2016.

[4] Siyu Chen, Guohua Fang, Xianfeng Huang and Yuhong Zhang, "Water Quality Prediction Model of a Water Diversion Project Based on the Improved Artificial Bee Colony–Backpropagation Neural Network", *www.mdpi.com/journal/water*, 18 June 2018.

[5] Mr. Riyaj K. Mulla and Mr. Shrikant M. Bhosale, "Water Quality Analysis and Simulation of Panchaganga River Using Matlab", *International Journal of Engineering Sciences & Research Technology*, pp. 613-620 (8), August 2016.

[6] P. Varalakshmi, S. Vandhana and S. Vishali, "Prediction of Water Quality using Naive Bayesian Algorithm", *2016 IEEE Eighth International Conference on Advanced Computing (IcoAC)*, 19 June 2017.

[7] S. Geetha and S. Gowthami, "Internet of Things Enabled Real Time Water Quality Monitoring System",*Springer Open Access Smart Water*, 2017.

[8] Carlin C.F. Chu, S.C. Yuen and Y.K. Wong, "Deep Neural Network for Marine Water Quality Classification with the Consideration of Coastal Current Circulation Effect", *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE,21 June 2018.

[9] S.Vijay and Dr.K.Kamaraj, "Ground Water Quality Prediction using Machine Learning Algorithms in R", *International Journal of Research and Analytical Reviews*, vol.6, Issue 1, March 01 2019.

[10] Fehmida Fatima S and Bindu A G, "Evaluation of Groundwater Quality at Eloor, Ernakulam District, Kerala Using GIS", *International Journal of Engineering and Advanced Technology (IJEAT)*, vol.8, Issue- 4C, December 2018.

[11] Maxime Lafont, Samuel Dupont, Philippe Cousin, Ambre Vallauri and Charlotte Dupont, "Back to the future: IoT to Improve Aquaculture: Real-Time Monitoring and Algorithmic Prediction of Water Parameters for Aquaculture Needs", *2019 Global IoT Summit (GIoTS), IEEE*, 22 July 2019.

[12] C.Ashwini, Uday Pratap Singh, Ekta Pawar and Shristi, "Water Quality Monitoring Using Machine Learning And IoT", *International Journal of Scientific & Technology Research*, vol.8, Issue.10,October 2019.

[13] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Sadaf Mumtaz and Ali Mustafa Qamar, "Water Quality Monitoring: From Conventional to Emerging Technologies", *International Water Association Publications (IWA Publishing)*, vol.20,Issue.1,01 February 2020.

[14] Sheng Cao and Shucheng Wang, "Design of River Water Quality Assessment and Prediction Algorithm", *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), IEEE*,26 March 2020.

 DOI : 10.35291/2454-9150.2020.0492