*ICIC Express Letters*
*Part B: Applications*
Volume **13**, Number **1**, January **2022**
 ISSN 2185-2766
pp. **89–97**

# LAKE TOBA WATER QUALITY PREDICTION USING EXTREME MACHINE LEARNING

Romi Fadillah Rahmat[1,*], Artambo Benjamin Pangaribuan[2]
Eric Suwarno[1], Sarah Purnamawati[1] and Tifani Zata Lini[3]

[1]Department of Information Technology
Universitas Sumatera Utara
Jl. Dr. T. Mansur No. 9, Kampus Padang, Bulan, Medan 20155, Sumatera Utara, Indonesia
*Corresponding author: romi.fadillah@usu.ac.id
121402071.es@gmail.com; sarah_purnamawati@usu.ac.id

[2]Department of Informatics
University of Pembangunan Nasional Veteran
Jalan RS. Fatmawati - Kecamatan Cilandak, Jakarta 12450, Indonesia
artambo@upnvj.ac.id

[3]Department of Business Information Technology
University of Twente
Enschede 7522NB, The Netherlands
tifanizatalini@student.utwente.nl

Received March 2021; accepted June 2021

Abstract. *Recent research about water quality in Lake Toba found that, based on the examination result of water sample collected at Haranggaol Horison District of Simalungun Regency, North Sumatera Province, Indonesia, the certain level of water pollution has been detected. The water quality examination work was conducted by collecting water sample from a number of locations beneath Lake Toba coast, which will be examined inside a laboratory. The result of laboratory examination for each sample will be used to measure water quality level in Lake Toba. This measurement method will increase the possibility of time and cost inefficiencies. Therefore, a method for performing water quality examination, which focused on time and cost efficiency, is required. With the development of computer technology, the implementation of microcontroller, combined with various types of sensor probes, can be done to perform water quality examination process. A research has been performed to measure water quality of Lake Toba in real time. However, the measurement data have to be processed for predicting the water quality index in Lake Toba. In this research, the water quality prediction process will be performed using extreme learning machine based on the water quality parameter measurement result data, which will return a graph showing water quality index measured according to Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status. The experiment result shows that the water quality prediction process using extreme learning machine can be done with training time ranges between 0.031 and 0.094 seconds. Also, the usage of the hard-limit function as activation function in prediction process has a better result than using the other activation functions. The measured water quality level ranges between B (good) and C (moderate) levels.*
**Keywords:** Water quality, Artificial neural network, Extreme learning machine, Prediction

1. **Introduction.** The research conducted by Haro et al. [1] shows that according to the examination performed in Haranggaol Horison District of Simalungun Regency, North Sumatera Province, the certain level of water pollution is detected in Lake Toba. The water pollution measured in the research, ranges from low level to moderate level, with

DOI: 10.24507/icicelb.13.01.89

the waste produced by residential and industrial activity within the coastal area, as the main source of water pollution in Lake Toba.

Following the advanced development of computing technology, various methods, including the usage of microcontroller combined with several sensor probes, are implemented to perform the water quality measuring process. Rahmat et al. [2] performed water quality measurement in Lake Toba using Arduino, connected with various type of sensor probes, which is updated within the finite interval. However, a method has to be implemented to perform water quality prediction, by focusing on the accuracy level and computing duration of the method.

Kasabov [3] stated that artificial neural network can be considered to perform prediction task. However, the duration of the prediction process, especially in the training process, is the main problem faced when using artificial neural network to perform prediction process. Werbos [4] and Rumelhart et al. [5] have developed backpropagation algorithm, which can be used to improve computation speed while using artificial neural network for prediction process. Despite the computing speed improvement, the backpropagation algorithm has difficulty in processing a big amount of data. Zou et al. conducted a study in water quality prediction model using stochastic theory. According to the result, the stochastic prediction model has high reliability and is feasible. However, it could be changed if there are any significant aspects or parameter which would affect the accuracy since the method is a random process regardless of any specified factors [6].

Another study about water quality prediction using neural networks was conducted by Zhou et al. The study implemented LSTM as the prediction method. The highlight of the research was its training time which took only about 30 minutes and was considered to improve the structure of the neurons for shorter training time [7].

Extreme learning machine [8] is one of the algorithms that can be used to improve training speed of artificial neural network. By randomizing input weights and bias, and using Moore-Penrose matrix inverse technique, extreme learning machine, which is performed inside a single hidden layer feedforward neural network, can perform prediction task with higher accuracy level and faster computation speed.

Compared to the recent research [6] and [7], this study took into account of specific parameters to ensure a high accuracy rate and it took shorter time to process the training data when compared to previous research. The calculation method of water quality index used in this research is based on Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status [9]. The final result of this research is a graph showing water quality index obtained from prediction process using extreme learning machine.

Water quality measurement in Lake Toba is performed by collecting water sample at several locations within the coast of Lake Toba. Each water sample collected in this process will be examined inside a laboratory, by which the water quality level measured in Lake Toba will be determined. By using this method of measurement, there will be the possibility of time and cost inefficiencies occurred from performing the measurement process. Therefore, a method to perform water quality prediction, which has the ability to reduce the time and cost usage, is required.

## 2. Methodology.

This section will describe the data source utilized in this research, along with the general architecture of the research.

### 2.1. Data source.

The data utilized in this research is the measurement result from the research conducted by Rahmat et al. [2], which is done in several locations within the coast of Lake Toba. The parameters used in the research were dissolved oxygen (DO), acidity level (pH), oxidation reduction potential level (ORP), water temperature, humidity level, and surface temperature. The measurement process resulted in several

Table 1. Details of the utilized data

| Dataset file name | Location | Number of data rows | | | |
|---|---|---|---|---|---|
| | | ***Initial*** | ***Post-filtering*** | ***Training*** | ***Testing*** |
| DATA ajibata.txt | Ajibata | 2203 | 2112 | 1268 | 844 |
| DATA Haranggaol.txt | Haranggaol | 6374 | 3532 | 2120 | 1412 |
| DATA parapat.txt<br>DATA parapat resume.txt | Parapat | 2446 | 1452 | 872 | 580 |
| DATA samosir.txt<br>DATA samosir resume.txt | Ambarita | 6129 | 3113 | 1869 | 1244 |

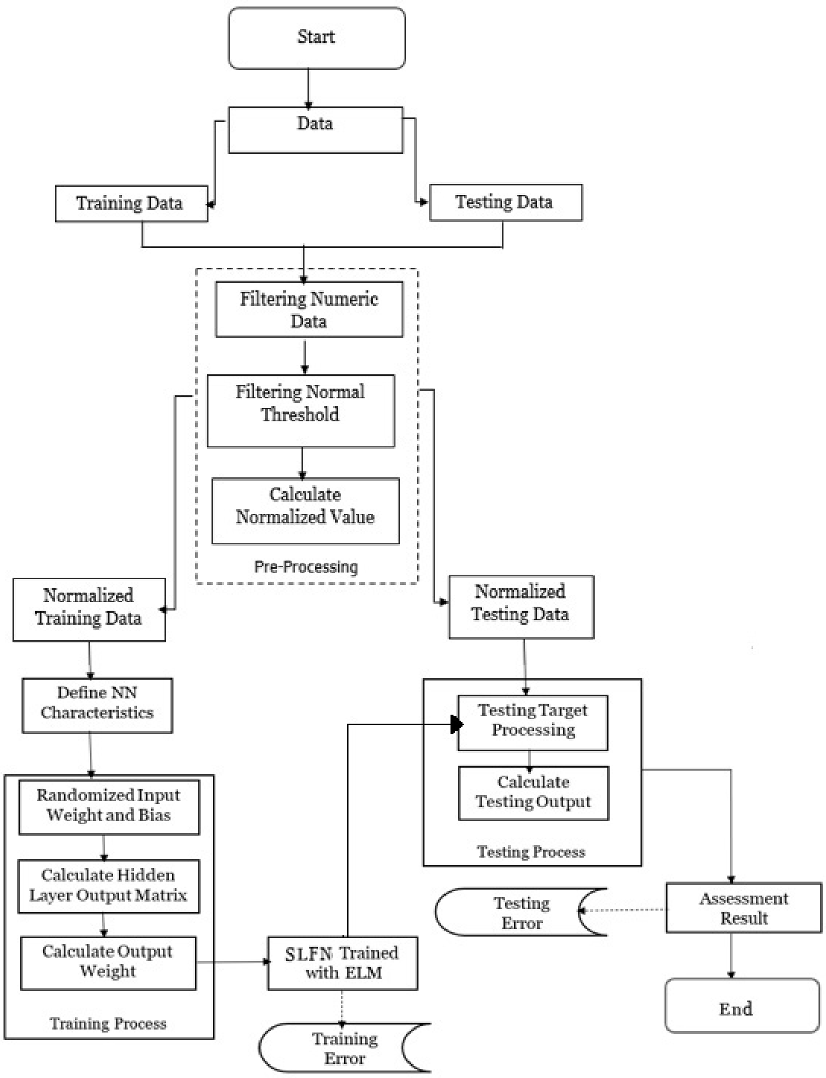


Figure 1. General architecture

sets of data, which is named according to the measurement location, while each data file contains certain rows of data, along with the measurement result.

2.2. **General architecture.** Figure 1 describes the general architecture implemented in this research.

The first step is preprocessing. Each dataset utilized in this research will be preprocessed in order to enable extreme learning machine to perform training and testing process. First, the filtering process will be performed by each row to ensure that each parameter contains a valid value. Normalization method utilized in this process is min-max normalization

[8], where the normalized value of each parameter is calculated by using (1):

$$A' = \left(\frac{A - A_{\min}}{A_{\max} - A_{\min}}\right) * (D - C) + C \tag{1}$$

Based on Equation (1), $A'$ is normalized value of parameter data with value of $A$ with the range of $[C, D]$, $A_{\min}$ is the minimum value of $A$, and $A_{\max}$ refers to the maximum value of $A$. The result of the preprocessing step is training datasets and testing datasets, which can be processed by extreme learning machine.

The next phase would be determining neural network characteristics: Before the training process is performed, two of the neural network characteristics, namely the number of hidden neurons and the activation function of each neuron, have to be determined. In this research, sine, sigmoid, cosine, and hard-limit function, will be used as the activation function.

The process continued to training stage. In this step, extreme learning machine will be performed for training process, which is done by performing three steps [9], described as below.

- Input weight and bias randomization. In this step, input weights of the vector connecting input neurons and hidden neurons, and bias value of hidden neurons, will be obtained by randomization process. The result of randomization process will be used for calculating the output of hidden layer.
- Hidden layer output calculation. In this step, the input weight received from input neurons, along with bias value, will be calculated according to the activation function determined in the previous process, which will be stored in the matrix $H$.
- Output value calculation. In this step, the output value will be calculated by using (2):

  $$\beta = H^{\dagger}T \tag{2}$$

  where $\beta$ refers to the matrix containing output value, $H^{\dagger}$ refers to Moore-Penrose inverse matrix of the matrix $H$, and $T$ refers to the output target of each data rows. These steps will produce a neural network that is ready to perform water quality prediction.

After training process, the data will go through the testing process. In this step, the neural network produced in training process using extreme learning machine will be utilized to perform water quality prediction process. The output target will be processed before the prediction process is started in this step. The final result of this step is a graph showing prediction result using extreme learning, shown by water quality index, compared to the measured water quality index, for each data row.

In training and testing process, the accuracy level of each process will be calculated for every iteration in each experiment. The accuracy level is measured by using root mean square error (RMSE), which is done by using (3):

$$RMSE = \sqrt{\frac{\sum\left(Y_i - \bar{Y}_i\right)^2}{N}} \tag{3}$$

where $Y_i$ refers to the predicted value, $\bar{Y}_i$ refers to the expected value, $i = [1, 2, 3, \ldots, N]$, and $N$ refers to the amount of data rows.

2.3. **Water quality index (WQI) calculation method.** The method of water quality index calculation is based on Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status. The calculation method of water quality index was conducted by comparing the value of each parameter measurement with minimum standard value of the parameter. Each parameter that does not meet the minimum standard value will decrease the water quality index.

After calculating water quality index, the quality class of a water resource is determined by Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status, as shown by Table 2.

Table 2. Water quality class [8]

| Water quality class | Water quality type | Index range | |
|---|---|---|---|
| | | Maximum | Minimum |
| A | Very good | 0 | 0 |
| B | Good | 0 | −11 |
| C | Moderate | −11 | −31 |
| D | Poor | −31 | – |

With the parameters used in this research, the standard value, minimum value, and maximum value of each parameter are described by Table 3. Every measurement result that does not meet the standard value, will cause the decrease of water quality index.

Table 3. Standard value of water quality parameters [8,10]

| Parameter | Unit | Standard value | Minimum value | Maximum value |
|---|---|---|---|---|
| Dissolved oxygen | mg/L | $\geq 6.0$ | 0.0 | 18.0 |
| Acidity (pH) | – | 6-9 | 0.1 | 14.0 |
| ORP | | +650 – +800 | −2,000.00 | +2,000.00 |
| Water temperature | °C | Deviation 3 | 20.0 | 37.0 |
| Surface temperature | °C | | 20.0 | 37.0 |
| Humidity | % | – | 0.0 | 100.0 |

3. **Results and Discussion.** This section will describe the results obtained from experiments performed in this research.

3.1. **Prediction accuracy comparison.** The best training error obtained in each experiment using sigmoid function as activation function, with different amount of hidden neurons is shown by Table 4. Each value is shown inside the table three digits behind period sign. From the results mentioned by the table, it is known that the increased amount of hidden neurons used in prediction process will result in higher accuracy level, which is represented by the low value of RMSE. Despite the improvement of accuracy level, the testing accuracy will be lower while using higher amount of hidden neuron.

Table 4. Best training error based on hidden neuron amount using sigmoid function as activation function

| Dataset | Hidden neuron amount | Best training error | | | |
|---|---|---|---|---|---|
| | | Training error (RMSE) | Testing error (RMSE) | Training duration (s) | Testing duration (s) |
| Ajibata | 15 | 0.122 | 1.097 | 0.047 | 0.031 |
| | 30 | 0.053 | 5.113 | 0.047 | 0.031 |
| Ambarita | 15 | 1.207 | 3.658 | 0.078 | 0.047 |
| | 30 | 1.149 | 3.511 | 0.062 | 0.047 |
| Haranggaol | 15 | 1.153 | 1.412 | 0.062 | 0.094 |
| | 30 | 1.078 | 5.602 | 0.078 | 0.062 |
| Parapat | 15 | 0.377 | 3.058 | 0.062 | 0.047 |
| | 30 | 0.173 | 3.433 | 0.062 | 0.062 |

3.2. **Prediction result.** The result of water quality prediction process shows that based on the measurement record collected in several locations within Lake Toba coast, the observed water quality ranges between good to moderate. The details of prediction result will be described for each location.

The graph shown in Figure 2 represents the water quality index predicted by extreme learning machine based on the measurement record collected in Ajibata. The experiment was performed to obtain the graph using hard-limit function as activation function and 20 hidden neurons. Based on the graph, it is known that the water quality index observed in the region of Ajibata ranges between −6 and −12. Based on the water quality class described in Table 2, it is also known that the water quality level in Lake Toba observed in Ajibata ranges between B to C class, which explains that the water resource of Lake Toba observed in Ajibata region is in good to moderate level of quality.

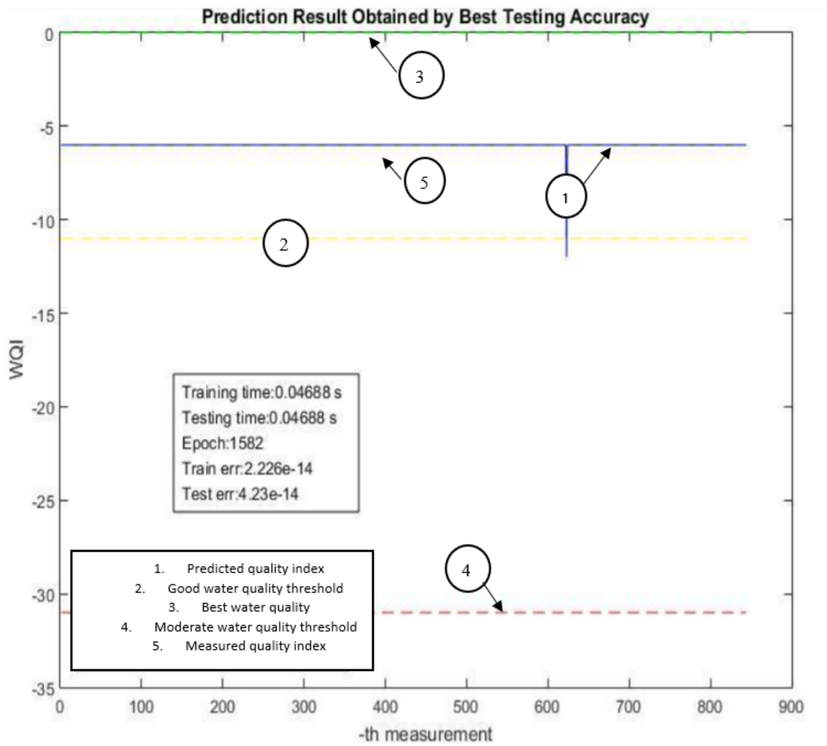


Figure 2. Prediction result for water quality in Ajibata region

The graph shown by Figure 3 explains water quality index predicted by extreme learning machine based on the measurement record collected in Ambarita. The prediction results and the expected results of the water quality index are shown in the graph. The experiment is performed to obtain prediction graph using hard-limit function as activation and 20 hidden neurons. According to the graph, it is known that the water quality index observed in Ambarita region ranges between −6 to −12. Therefore, the water resource condition of Lake Toba observed in Ambarita are in good to moderate level.

The graph shown by Figure 4 describes the prediction result obtained by using measurement records collected in Haranggaol region. The experiment is performed to obtain the graph by using hard-limit function as activation function and 20 hidden neurons. According to the graph, the water quality index predicted by measurement record collected in Haranggaol ranges from −6 to −12. Therefore, the water resource of Lake Toba observed in Haranggaol region is in good to moderate level according to Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status.

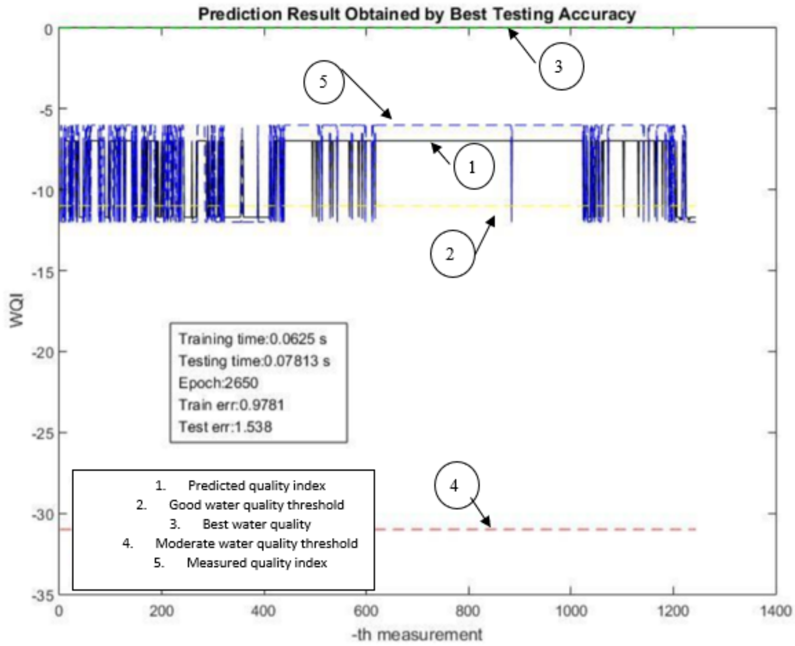


FIGURE 3. Prediction result for water quality in Ambarita region

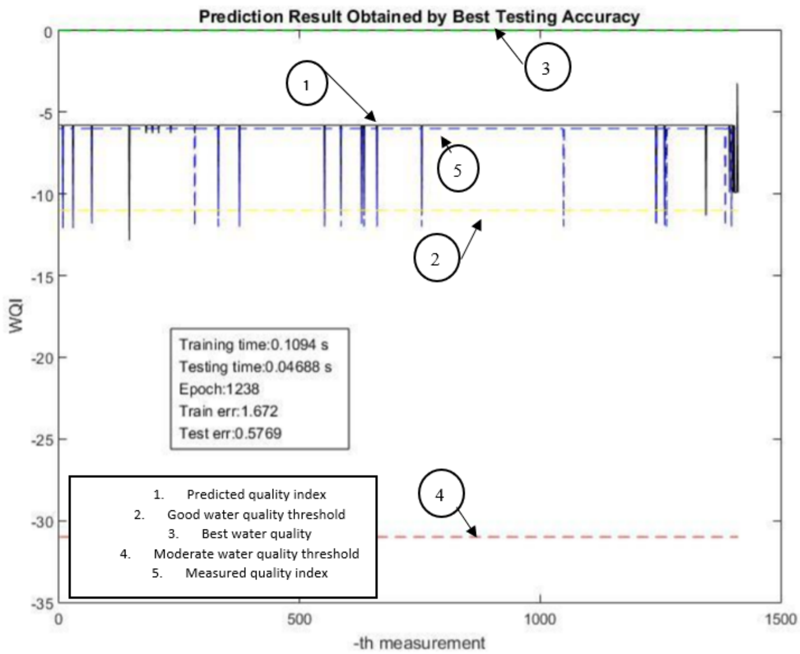


FIGURE 4. Prediction result for water quality in Haranggaol area

Meanwhile, the graph shown by Figure 5 describes water quality index predicted by extreme learning machine based on the measurement record collected in Parapat region. The prediction result shows that water quality index of water resources of Lake Toba observed in Parapat region varies from $-6$ to $-12$. Based on this result, it is known that the water resource of Lake Toba in Parapat region varies between good to moderate condition.

4. **Conclusion.** In this research, extreme learning machine is implemented to predict water quality of Lake Toba, by using water quality parameter measurement data obtained

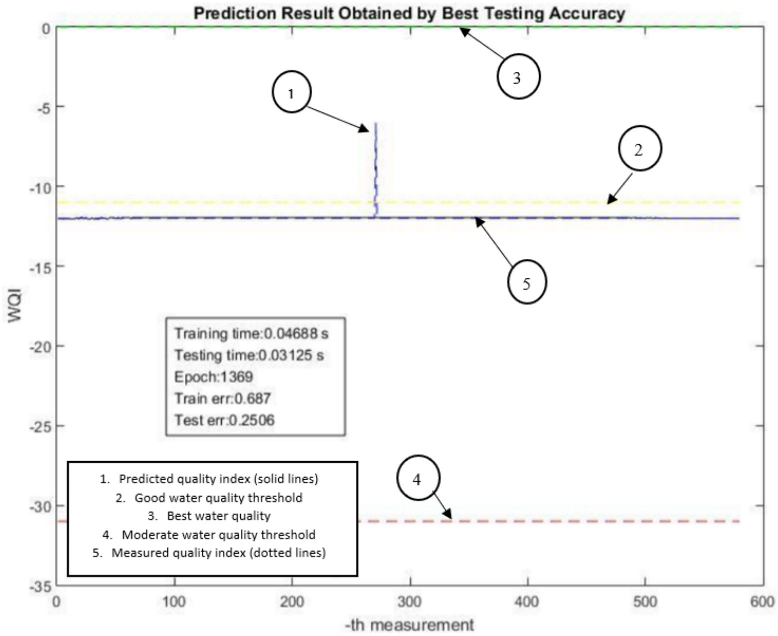


FIGURE 5. Prediction result for water quality in Parapat area

by Rahmat et al. [2]. The experiment result shows that extreme learning machine can be implemented to perform water quality prediction of Lake Toba with high accuracy, which is represented by low RMSE value, along with fast computation speed. The experiment result also shows that the best training error and testing error obtained from the implementation of extreme learning machine can be achieved by using hard-limit function as activation function in each experiment. Generally, for the predicted water quality level in Lake Toba according to Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status, the ranges were between good to moderate level.

For the future research, the addition of various water quality parameters, for example, total dissolved solids (TDS), and ammonia level, are recommended. The further addition of measurement data is also recommended to improve the results obtained from this research. The various neural network architectures, along with the improved version of extreme learning machine, for example, OP-ELM [11], can also be implemented to be compared by the result of this research. Finally, various water quality index calculation method, such as Oregon water quality index (OWQI) [12], can also be implemented.

## REFERENCES

[1] D. D. Haro, Y. Djayus and Z. A. Harahap, Water quality condition of Lake Toba at Haranggaol Horison District – Simalungun Regency North Sumatra, *AQUACOASTMARINE*, vol.1, no.1, 2013.
[2] R. F. Rahmat, M. F. Syahputra, M. S. Lydia et al., Real time monitoring system for water pollution in Lake Toba, *2016 International Conference on Informatics and Computing (ICIC)*, pp.383-388, 2016.
[3] N. Kasabov, *Evolving Connectionist Systems*, Springer, New York, 2007.
[4] P. W. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*, Harvard University, 1974.
[5] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors, *Nature*, vol.323, pp.533-536, 1986.
[6] K. Zou et al., Establishment of water quality prediction model in Pingdingshan region based on stochastic theory – Markov process, *Journal of Physics: Conference Series*, 2020.
[7] J. Zhou, Y. Wang, F. Xiao, Y. Wang and L. Sun, Water quality prediction method based on IGRA and LSTM, *Water*, vol.10, DOI: 10.3390/w10091148, 2018.

[8] G. B. Huang, Q. Y. Zhu and C.-K. Siew, Extreme learning machine: Theory and applications, *Neurocomputing*, vol.70, no.1, pp.489-501, 2006.
[9] Indonesian Minister for the Environment, *Decree of the Minister for the Environment Number 115 of 2003 concerning Guidelines for Determining Water Quality Status*, State Secretariat, Jakarta, 2003.
[10] S. G. K. Patro and K. K. Sahu, *Normalization: A Preprocessing Stage*, 2015, https://arxiv.org/abs/1503.06462, Accessed on 23-Jan-2017.
[11] T. P. Lambrou, C. G. Panayiotou and C. C. Anastasiou, A low-cost system for real time monitoring and assessment of potable water quality at consumer sites, *2012 IEEE Sensors*, pp.1-4, 2012.
[12] A. M. Yoan, P. B. Sorjamaa, O. Simula, C. Jutten and A. Lendasse, OP-ELM: Optimally pruned extreme learning machine, *IEEE Trans. Neural Networks*, vol.21, no.1, pp.158-162, 2010.
[13] S. H. Dinius, Design of an index of water quality, *JAWRA J. Am. Water Resour. Assoc.*, vol.23, no.5, pp.833-843, 1987.