Lake Toba Water Quality Prediction Using *Extreme Learning Machine*

Eric Suwarno, Romi Fadillah Rahmat, Maya Silvi Lydia
Faculty of Computer Science and Information Technology
University of Sumatera Utara
Medan, Indonesia
E-mail: 121402071.es@gmail.com | romi.fadillah@usu.ac.id | maya2@usu.ac.id

***Abstract*—Recent research about water quality done 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 is done 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 type of sensor probes, can be done to perform water quality examination process. A research has been done for performing real time water quality measurement in Lake Toba. However, the measurement data have to be processed for predicting water quality index measured 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 Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air. 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 hard-limit function as activation function in prediction process will return better result than using the other functions as activation function. The measured water quality level ranges between B (good) to C (moderate) level.**

Keywords—water quality prediction, artificial neural network, extreme learning machine

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

Water quality measurement in Lake Toba is performed by collecting water sample at several locations within the coast of Lake Toba. Each water samples 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.

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] states 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 of the computing speed improvement, the backpropagation algorithm has difficulty in processing a big amount of data.

Extreme learning machine [6] is one of the algorithm that can be used to improve training speed of artificial neural network. By randomizing input weights and bias, and using Moore-Penrose matrix invers 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.

In this research, the data obtained from measurement process done by Rahmat *et al*. [2] will be processed using extreme learning, to predict the water quality level in Lake Toba calculated by each data rows. The calculation method of water quality index used in this research is based on Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 tentang Pedoman Penentuan Status Mutu Air [7]. The final result

of this research is a graph showing water quality index obtained from prediction process using extreme learning machine.

II. PROBLEM IDENTIFICATION

Water quality measurement is performed in Lake Toba by obtaining water samples in several locations within coastal area of the lake, which will be examined inside a laboratory. Because of the possibility of time and cost inefficiency caused from this measurement method, a method to perform water quality prediction, which focusing on time and cost efficiency while performing measurement process, is required.

III. METHODOLOGY

This section will describe the data source utilized in this research, along with the general architecture of the research. Subsection A will describe the data utilized in this research, while Subsection B will describe the steps performed in this research, as shown by the general architecture.

A. 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 location within the coast of Lake Toba. The parameters used in the research is dissolved oxygen (DO), acidity level (pH), oxidation reduction potential level (ORP), water temperature, humidity level, and surface temperature. The measurement process resulted in several 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. The details of the data utilized in this research is described by Table I.

TABLE I. DETAILS OF THE UTILIZED DATA [2]

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	Parapat	2446	1452	872	580
DATA parapat resume.txt					
DATA samosir.txt	Ambarita	6129	3113	1869	1244
DATA samosir resume.txt					

B. General Architecture

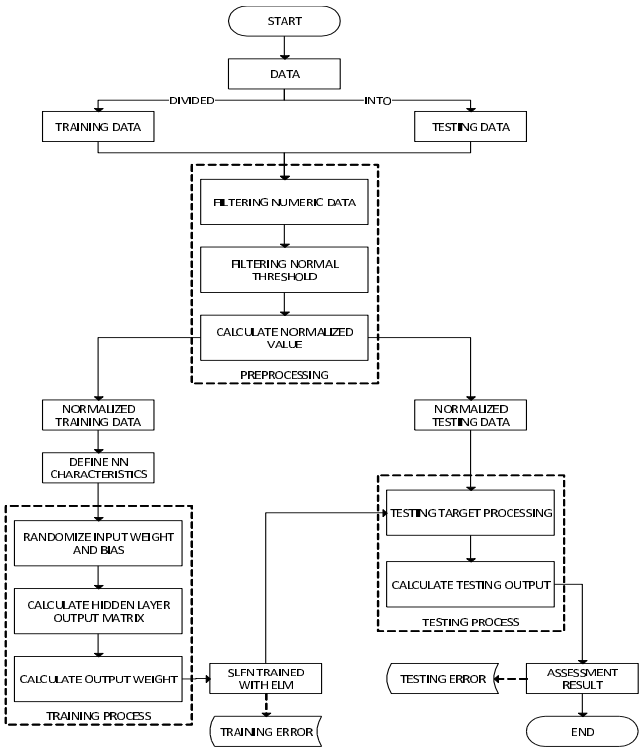

Fig 1. General Architecture

Figure 1 describes the general architecture implemented in this research. Each data will be processed by using the steps below:

- *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 rows 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):

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

  where 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.
- *Determine neural network characteristics*: Before the training process is performed, two of the neural network characteristics, namely the number of hidden neuron and the activation function of each neuron, has to be determined. In this research, sine, sigmoid, cosine,

and hard-limit function, will be used as the activation function.

- *Training*. 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 β 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.
- *Testing*. 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 rows.

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

$$RMSE = \sqrt{\frac{\sum(Y_i - Y_i)^2}{N}} \tag{3}$$

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

C. Water quality index (WQI) calculation method

The method of water quality index calculation is based on Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air. The calculation method of water quality index is done by comparing the value of each parameter measurement with minimum standard value of the parameter. Each parameters that does not meet the minimum standard value will decrease the water quality index, with the amount of decrease explained by Table II.

TABLE II. WATER QUALITY CALCULATION METHOD [8]

Number of parameter	Value type	Parameter type		
		Physical	Chemical	Biological
Below 10	Minimum	-1	-2	-3
	Maximum	-1	-2	-3
	Average	-2	-4	-6
10 and above	Minimum	-2	-4	-6
	Maximum	-2	-4	-6
	Average	-6	-12	-18

After calculating water quality index, the quality class of a water resource is determined by Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 tentang Pedoman Penentuan Status Mutu Air, as shown by Table III.

TABLE III. 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 is described by Table IV. Every measurement result that does not meet the standard value, will cause the decrease of water quality index.

TABLE IV. STANDARD VALUE OF WATER QUALITY PARAMETERS ([8], [10])

Parameter	Unit	Standard value	Minimum value	Maximum value
Dissolved oxygen	mg/L	≥ 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	deg C	Deviation 3	20.0	37.0
Surface temperature	deg C		20.0	37.0
Humidity	%	-	0.0	100.0

IV. IMPLEMENTATION AND EXPERIMENT

This section will describe the results obtained from experiments performed in this research. The main focus of the result, which will be described in this section, are training error and testing error obtained in the experiments, and prediction result obtained from each datasets.

A. Prediction accuracy comparison

The best training error obtained in each experiments using sigmoid function as activation function, with different amount of hidden neurons is shown by Table V. Each value is shown inside the table three digits behind comma 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 of the improvement of accuracy level, the testing accuracy will be lower while using higher amount of hidden neuron.

TABLE V. BEST TRAINING ERROR BASED ON HIDDEN NEURON AMOUNT USING SIGMOID FUNCTION AS ACTIVATION FUNCTION

Data set	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

The comparison of best testing error obtained in the experiment between the activation function is shown by Table VI. Each value shown in the table is approximated by three digits behind comma sign. The number of hidden neuron implemented using sigmoid, sine, and cosine function implemented in this experiment is 15 neurons, while the number of hidden neuron implemented using hard-limit function is 20 neuron. The experiment result shows that among the activation functions implemented in the experiment, the best testing error can be obtained while using hard-limit function as activation function in training and testing process.

TABLE VI. BEST TESTING ERROR BASED ON ACTIVATION FUNCTION

Data set	Act. function	Best testingerror			
		Training error (RMSE)	Testing error (RMSE)	Training duration (s)	Testing duration (s)
Ajibata	Sigmoid	0.161	0.103	0.047	0.047
	Sine	0.153	0.077	0.047	0.031
	Cosine	0.163	0.076	0.047	0.047
	Hard-limit	$2.2x10^{-14}$	$4.2x10^{-14}$	0.047	0.047
Ambarita	Sigmoid	1.270	3.440	0.047	0.031
	Sine	1.239	3.436	0.047	0.047
	Cosine	1.247	3.440	0.047	0.031
	Hard-limit	0.978	1.538	0.062	0.078
Haranggaol	Sigmoid	1.507	0.661	0.094	0.062
	Sine	1.494	0.639	0.062	0.094
	Cosine	1.352	0.654	0.078	0.078
	Hard-limit	1.672	0.577	0.109	0.047
Parapat	Sigmoid	0.487	0.622	0.078	0.062
	Sine	0.464	0.559	0.078	0.047
	Cosine	0.513	0.574	0.062	0.031
	Hard-limit	0.687	0.251	0.047	0.031

B. 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 by Figure 2 represents the water quality index predicted by extreme learning machine, shown by the black lines in the graph, based on the measurement record of Lake Toba in Ajibata. The experiment is 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 III, 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.

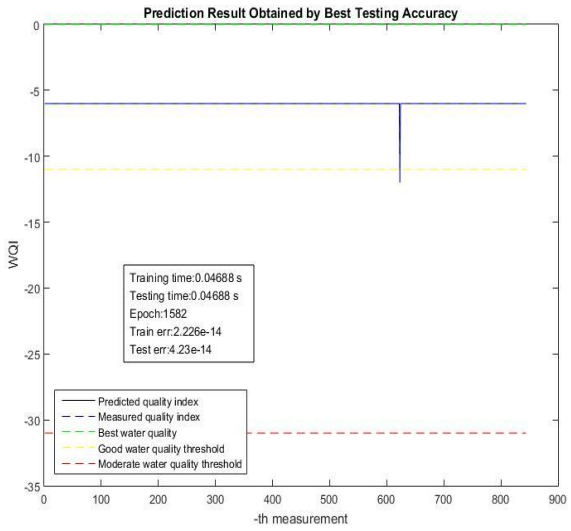

Fig 2. Prediction result using Ajibata measurement record

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 result is shown by black lines in the graph, meanwhile blue striped lines in the graph represents expected water quality index. 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.

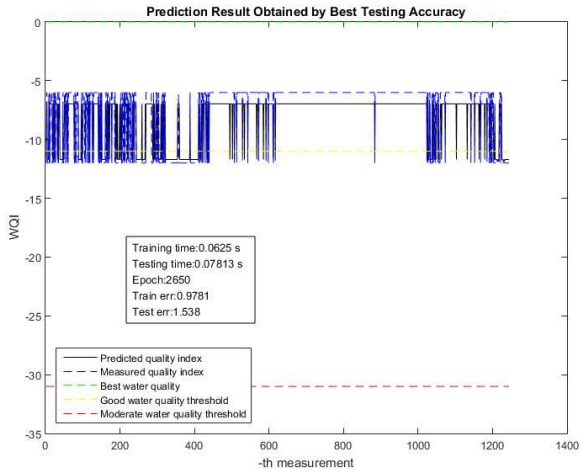

Fig 3. Prediction result using Ambarita measurement record

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. The black line in the graph represents predicted water quality index, while the striped blue line in the graph represents expected water quality index. 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 are in good to moderate level according to Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air.

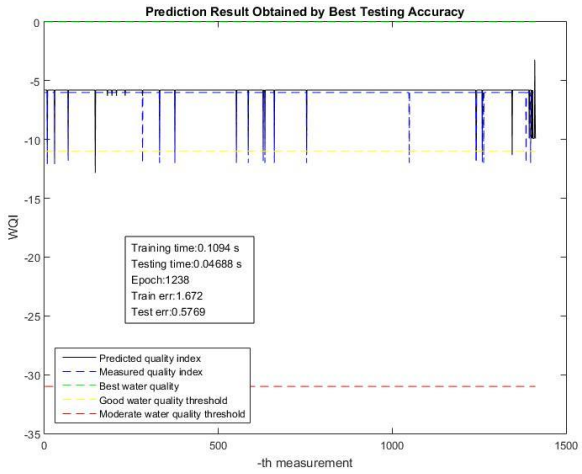

Fig 4. Prediction result using Haranggaol measurement record

Meanwile, 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 is shown by black line in the graph, and the expected water quality index is shown by blue striped line in the graph. 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.

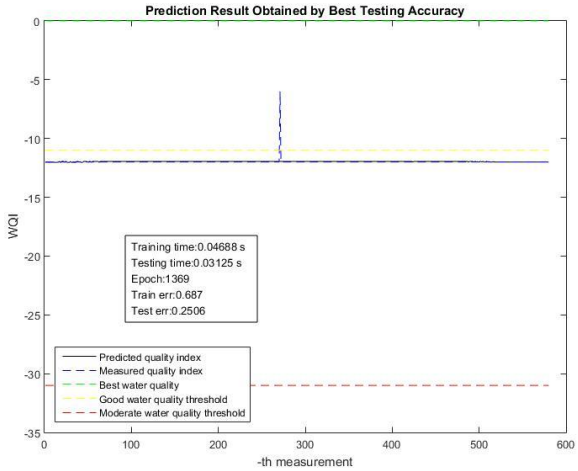

Fig 5. Prediction result using Parapat measurement record

V. CONCLUSION

In this research, extreme learning machine is implemented to predict water quality of Lake Toba, by using water quality parameter measurement data obtained 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, the predicted water quality level in Lake Toba according to Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air ranges between good to moderate level.

For the future researches, the addition of various water quality parameters, for example, total dissolved solids (TDS), ammonia level, and the other parameters, are recommended. The further addition of measurement data is also recommended to improve the results obtained from this research. The various neural network architecture, 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, "Kondisi Kualitas Air Danau Toba di Kecamatan Haranggaol Horison Kabupaten Simalungun Sumatera Utara," *AQUACOASTMARINE*, vol. 1, no. 1, 2013.

[2] R. F. Rahmat, M. F. Syahputra, M. S. Lydia, and others, "Real time monitoring system for water pollution in Lake Toba," in *Informatics and Computing (ICIC), International Conference on*, 2016, pp. 383–388.

[3] N. Kasabov, *Evolving Connectionist Systems*, 1st ed. New York: Springer, 2007.

[4] P. W. Werbos, "Beyond regression: new tools for prediction and analysis in the behavioral science," Harvard University.

[5] D. E. Rumelhart, G. E. Hinton, andR. J. Williams, "Learning representations by back-propagating errors," *Cogn. Model.*, vol. 5, no. 3, p. 1, 1988.

[6] 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.

[7] Menteri Negara Lingkungan Hidup, *Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air*. Jakarta: Sekretariat Negara, 2003.

[8] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," Mar.2015.

[9] G.-B. Huang and Q.-Y.Zhu, "Extreme Learning Machines," 2004. [Online]. Available: http://www.ntu.edu.sg/home/egbhuang/elm_random_hidden_nodes.html. [Accessed: 27-Feb-2017].

[10] 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," in *2012 IEEE Sensors*, 2012, pp. 1–4.

[11] Yoan Miche, A. Sorjamaa, P. Bas, 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, Jan.2010.

[12] S. H. Dinius, "DESIGN OF AN INDEX OF WATER QUALITY," *JAWRA J. Am. Water Resour. Assoc.*, vol. 23, no. 5, pp. 833–843, Oct.1987.