Journal of Theoretical and Applied Information Technology
31st May 2021. Vol.99. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IDENTIFICATION OF THE FISH SURVIVAL RATE AND THE FISH TYPES TO LIVE IN LAKE TOBA USING MACHINE LEARNING

[1]ROMI FADILLAH RAHMAT, [2]SARAH PURNAMAWATI, [3]TIFANI ZATA LINI, [4]RIDHO FARIHA, [5]NABILA DEANNA, [6]INDRA AULIA, [7]SHARFINA FAZA

[1,2,3,4,5,6] Department of Information Technology, Universitas Sumatera Utara, Indonesia
[7] Politeknik Negeri Medan, Medan, Indonesia

E-mail: [1]romi.fadillah@usu.ac.id

ABSTRACT

Waste disposal in Lake Toba conducted by the local citizens has led to water pollution and the decreasing number of freshwater fish. Therefore, the identification of water content is mandatory to ensure the survival and cultivation of the fish in Lake Toba. We conducted two studies using the same data using machine learning methods of Long Short-Term Memory and Support Vector Machine. These studies proved that both methods were reliable in the identification process. The results of this research showed that the water contents in Lake Toba are still good enough for the freshwater fish to live and the types of fish that have to highest chance to survive in Lake Toba.

Keywords: *Identification, Freshwater Fish, Machine Learning, LSTM, SVM*

1. INTRODUCTION

Machine learning is one of the branches from Artificial Intelligence that allow data being learnt with or without supervision. A wide range of machine learning applications have been implemented in various field, such as medical, agricultural, business operation, and many more.

Over the past decades, the implementation of machine learning has been conducted and they have shown promising results. Alongside with the growing data collection, both quantity and quality, it is quite challenging to determine which method would be the most efficient to be applied for one specific case.

Several researches on classification using machine learning method have been conducted, such as a study on the identification of dental and oral diseases through the dental images using an SVM classification system with an accuracy of 94,442%. The data were tested without using the preprocessing stage and implemented three kernels, namely Polynomial, RBF, and Linear, where the test parameters used were λ (lambda), y (gamma), C (complexity), ε (epsilon) and the number of iterations. [1]

Another research was carried out to classify the epidermal area of Dengue Fever in which the data training and data testing were performed ten times, and produced an accuracy of 75%. [2].

A credit risk assessment system was developed using genetic algorithms to find the parameter values that can reach the optimum in classifying data. Various kernels and parameters of Support Vector Machine were implemented in the study. Once the testing process completed, the accuracy value will be re-tested to form an objective function, which will become a new parameter in the data classification if the performance conditions have not been met [3].

Dimpas et al. performed another research. The authors applied the long short term memory (LSTM) to identify whether news considered as a clickbait or not based on its headline. The eligible headline news for this study should be in English, Filipino, or both. This research implemented the optimizer of RMSProp with a learning rate of 0.0005 and has an accuracy rate of 91.5% [4].

Another study was conducted by Zhao et al. (2017). The study has the purpose of identifying the diagnosis of errors in industrial operational procedures and products. It also compared the system accuracy using LSTM with dynamic linear

discriminant analysis (DLDA) combined with the Support Vector Machine (SVM) algorithm, and Multilayer Perceptron (MLP). The results showed that the system accuracy using DLDA + SVM algorithm is 95.2%, while the accuracy rates of the system using MLP and LSTM are 95.8% and 98.7%, respectively [5].

Of all the studies conducted so far, there have been no studies that utilize machine learning to see the most suitable fish to live in particular area of Lake Toba. Referring to previous studies, machine learning has been proved to be one of the effective methods in handling large volumes of data and identify specific behavior patterns. This research aims at comparing two machine learning algorithms which are SVM and LSTM using the dataset of Lake Toba's water quality to examine the efficiency of these algorithms.

Lake Toba is the largest lake in Southeast Asia located in the north Sumatra archipelago Indonesia. Lake Toba is also a water ecosystem that has undergone many changes as a result of various local activities. As the largest lake in Indonesia, Lake Toba becomes the water resources with important values in the field of ecology, hydrology, and economic functions. [6].

The cultivation of freshwater fishes in Lake Toba is focused on specific fish species that are popular with local people. This leads to an imbalance habitation environment for freshwater fish and affects the economic values of fish cultivation.

The waste disposal is one of the issues in Lake Toba and has led to the obstruction of the fish breeding and the fish population to be downsized. In fishery activities, the physical and chemical characteristics of water are the parameters that are required to bolster the success of these activities.

Several studies have been conducted on fish environment observation and both methods. Harianja et al. conducted one of the researches. The study aimed to build an application to monitor the water using Raspberry and Arduino, and it can be controlled using an android application. The tool served to replace water automatically when the water content is not suitable for the Batak fish [7].

The research performed by Harianja et al. in 2014 has been quite beneficial since it proposes an automatic water replacement tool. However, the device is limited to a small area such as an aquarium or pond. It can be expanded using a machine learning so it can directly identify the water content of the fish habitat by taking the record of changes of the water content of Lake Toba utilizing a sensor, so the fish should not be moved to aquarium or pond beforehand.

Based on the statements above, the authors proposed research to calculate the survival rate of the fish and the types of fish that are the most suitable to breed using six parameters of chemical components in Lake Toba using two methods that have been proven as effective algorithms in the identification process as comparison. Authors set the testing to scope to include 4 areas of Lake Toba: Ajibata, Haranggaol, Ambarita and Parapat II. While the objects for the fish survival test were limited to goldfish, carp, tilapia, pomfret and catfish.

MATERIALS AND METHOD

2.1 Data

The data used for this research is the water content datasets obtained from a study conducted by Rahmat et al. in 2016 [8]. The data contain various measurement parameters of water content such as dissolved oxygen (DO), pH, water temperature, air temperature, and air humidity, which were saved in .CSV file format. The data were taken from several locations in the area of Lake Toba, including Ajibata, Haranggaol, and Parapat. The format of these datasets can be seen in Figure 1.

```
1.31;4.58;30.94;34.00;34.00;6–80†6.82
0.31;1.10;30.94;50.00;34.00;6.81
0.58;2.02;30.94;50.00;34.00;6.82
0.84;2.93;30.94;51.00;34.00;6.80
1.10;3.85;30.94;50.00;34.00;6.81
```

Figure 1: Format of Datasets

Each parameter is separated by a semicolon (;). In each column, from left to right, these parameters are voltage sensors, pH, water temperature, air temperature, air humidity, and DO. The data were obtained by recording the results of water quality measurements using a sensor connected to an Arduino and some components such as GSM shield, dissolved oxygen sensor, DS18S20 waterproof temperature sensor, DHT11 air temperature, and humidity sensor and PH sensor of DFrobot SEN0161. The data were recorded two times, taken on 25 October 2016 at 2 pm - 4 pm and on 26 October 2016 at 5 am - 8 am.

2.2 General Architecture

This research designed a system by implementing two of the machine learning algorithms, called Long Short-Term Memory (LSTM) and Support Vector Machine (SVM). LSTM was designed explicitly to avoid long-term dependency problems, while SVM was applied to solve the complexity problem of linearity and non-linearity in classification and regression. LSTM algorithm is well suited to identify the water content because the obtained dataset is the time-series data and, based on the previous research, often has the highest accuracy among other algorithms in terms of identification. LSTM has a form of recurrent neural network module chain and has four connected layers, as shown in Figure 2.

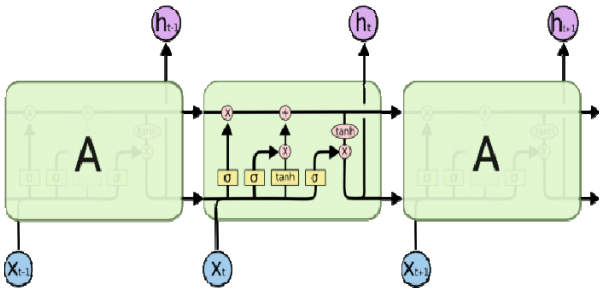

Figure 2: Structure of LSTM

SVM aims to find the best hyperplane to achieve an optimum result. SVM is not only able to separate two classes, but also sufficient to separate classes by changing it to a higher dimensional space so that the separation process occurs easier, even if the number of samples is less than the number of dimensions that will be changed. This is due to the help of the kernel trick. Kernel tricks usually work on non-linear data where the existing features are changed to higher dimensions to be able to separate the data since the trick cannot do the linear separation. The illustration of kernel tricks is shown in Figure 3.

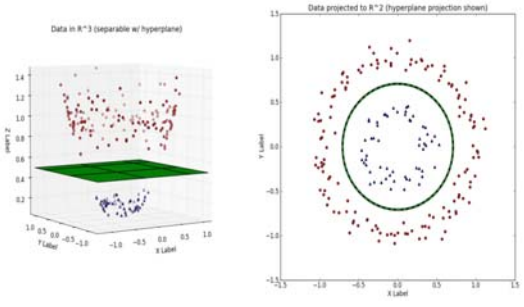

Figure 3: Illustration of Kernel Tricks in Support Vector Machine

The system was built to have two features, which are to identify the water contents in Lake Toba using Long Term Short Memory and to detect the types of fish to be more likely to survive in the lake. The research has two similar general architectures where LSTM identify the water contents and calculate the accuracy of the system when compared to the actual data, while SVM served to generate the best fish to live in Lake Toba. The general architectures are shown in Figure 4.

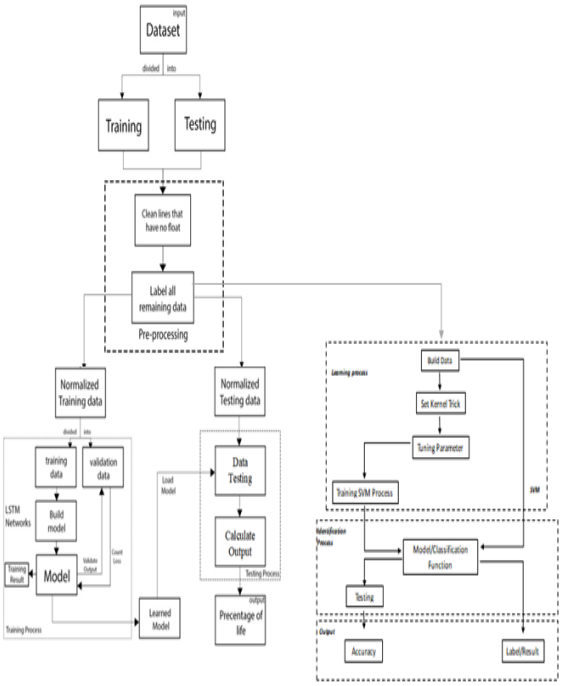

Figure 4: General Architecture

2.2.1 Input

The input is a dataset that has been divided into two parts. For the training process, the data were amounted to 11993 lines, consisted of the data combination obtained from the location of Ajibata and Parapat, while for the testing process, there are 7634 data which were collected from several locations that will be processed in the preprocessing,

2.2.2 Pre-processing

Once the data were acquired, the next step will be the preprocessing stage consisting of cleaning and labeling processes.

Journal of Theoretical and Applied Information Technology
31st May 2021. Vol.99. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The cleaning process aims to cleanse the data from the non-float characters, numbers, or symbols and remove the unnecessary columns in the training process such as the data from voltage sensor, air temperature sensor, and air humidity sensor. The process of data cleaning can be seen in Figure 5.

This process begins with inputting the dataset and initializing each row. If the value of the selected row has a non-float value or number, then the line will be deleted. The process will be repeated until the last line. Then the columns that are not required, such as the voltage sensor, air temperature sensor, and air humidity sensor, will be removed.

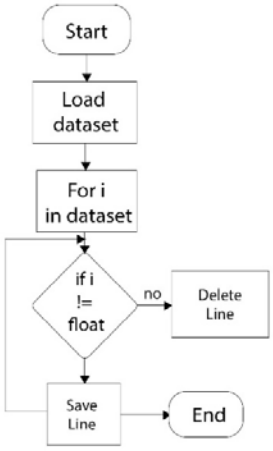

Figure 5: Data cleaning

The data were separated by a comma that states the column differences. These parameters include the level of oxidation-reduction potential, the acidity level of the water, water temperature, humidity, air temperature, and the level of dissolved oxygen. Details of the range of parameter values of each fish can be seen in Table 1.

Table 1: Threshold values of water content for freshwater fish

Name	pH	Temperature (°C)	DO (mg/L)
Tilapia	7 – 8	25 – 30	3 – 5
Iridescent Shark	6 – 8,5	25 – 30	3 – 7
Pomfret	7 – 8	25 – 30	≥ 3
Carp	6 – 8	20 – 32	7 – 8
Goldfish	6,5 – 8,5	25 – 30	≥ 5

The next stage is the labeling process to determine the output based on the features of pH, temperature, and DO. Labeling on these data is referred to a study by Bhatnagar et al. in 2013 [9], where there are several criteria for the water content to be suitable for the freshwater fish. These requirements are shown in Table 2.

Table 2: Criteria for the water content of freshwater fish

No.	Parameter	Tolerable levels	Optimal levels	Poor levels
1	pH	7 - 9.5	6.5 - 9	<4 or> 11
2	DO (mg / L)	3-5	5	<4 or> 8
3	Water temperature (°C)	15-35	20-30	<12 or> 35

The formula to determine the label is explained in the syntax in Figure 6.

```
if (n >= nmin and n <= nmax)
    y = 1
else if ((n >= Xnmin and n < nmin) or (n >
nmax and n <= Xnmax))
    y = 0.5
else if (n < Xnmin or n > Xnmax)
    y = 0
else
    y = 0

z = y1 + y2 + y3

if (z == 3)
    zout = 2   #optimal levels
else if (z >= 2 and z < 3)
    zout = 1   #tolerable levels
else:
    zout = 0   #poor level
```

Figure 6: Syntax for Labeling Process

Based on Figure 6, n is the value of pH, temperature, or DO contained in a single row of data, n_{min} is the minimum point of optimal value, and n_{max} is the maximum point of optimal value. Xn_{min} is the minimum tolerance value, and Xn_{max} is the maximum tolerance value, while y is the output of the calculation. If the output is optimal, it would be denoted as 1. If it is a tolerance value, it will be given a value of 0.5, and if the output is undefined, it will be marked as 0. If the n value is less or greater than the tolerance value, the water quality will be considered as poor. z is the output of the sum value of pH value (y^1), water temperature value (y^2), and DO (y^3), which was calculated for each row of the data. After the value of z was obtain, the rules can be set. If the value of z_{out} is 3, then it will be labeled as "2", if the value < 3 or >=

2 then it will be marked as "1" and others will be labeled as "0".

After the process completion, the result will be saved into a .CSV file containing columns of pH value, water temperature values, DO, and the labels. The result of this process can be seen in Figure 7.

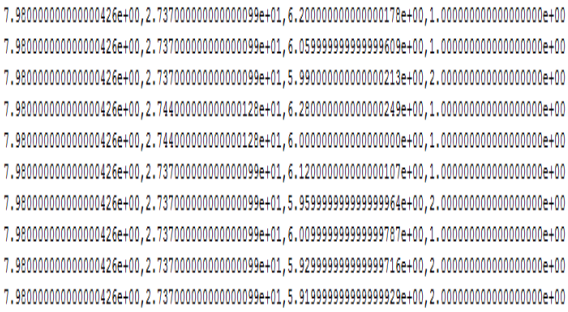

Figure 7: Result of Labeling Process

After the preprocessing stage completed, the data will be trained and will be carried out for the identification process using the results of the training process. Both LSTM and SVM were using the labeled data as the input to conduct the next stages.

2.2.3 Identification Process

The system allows the users to choose which features the users would like to access. For water contents identification, the users can choose the method of LSTM to acquire the contents of Lake Toba water and the accuracy of the result, while SVM provides the users with the list of fish that are suitable to live in certain areas of Lake Toba.

2.2.3.1 Long Short Term Memory (LSTM)

The identification process of the viability of fish in Lake Toba consisted of two steps, which were training and testing processes. For the training, the dataset will be divided into two parts with a proportion of 80:20. 80% of the total data will be used in the testing, while the rest for validation. Then, the model will be created based on the parameters.

Determining the number of nodes on the neural network, especially in the hidden layer, is an important thing to do before starting the training because the hidden layer has an important role in the calculation of the final result of the neural network. In this research, the modeling process applied LSTM, which has three hidden layers. The user can select the activation function on each hidden layer before the training process started. There are four types of activation functions, namely softmax, sigmoid, tanh, and Relu.

The first step of this phase is data training. At this step, the initial value of weights and biases will be set randomly; then, it will be calculated using equation 1 to get the ft value. After that, we look for the it value using equations 2 and the value of O_t using equation 3.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

Where it is an input gate, ft is the forget gate, and O_t is the output gate. σ a sigmoid function. W_i is the weight values of the input gate, W_f is the weight values of the forget gate, and W_o is the weight value of the output gate. bi is a bias of the input gate, bf is a bias of the forget gate, and b_o is a bias of the output gate. h_{t-1} is the hidden state on the network memory cell, which is an output value before order t, and x_t is the input of order t.

The calculation results will be validated in each batch with the validation dataset to see whether the given output is in accordance with the actual value. The validation process will produce an error called loss to be used as a parameter to determine the given weight of the next batch. The result of this process is a neural network model called the learned model, which has been completed the training and will be reloaded while performing the testing process. In addition, the total loss and accuracy will also be printed at the end of each training process.

The next step is data testing. In this process, the learned model that has been generated in the previous training process will be reloaded and calculated the output value based on the parameters given during the training process. The testing process was conducted to determine the effectiveness of Long Short-Term Memory in the identification process.

2.2.3.2 Support Vector Machine

For the identification of fish in Lake Toba, a kernel was determined to process all the labeled data. Not all data can be separated linearly; SVM is only capable of separating the data linearly. SVM needs element support to allow SVM to separate data non-linearly; one of them is by adding kernel function. By adding the kernel function in SVM, the data will be mapped into a higher vector so that the hyperplane can be constructed.

Each kernel has a different tuning parameter due to the difference in the work principle. The values of C, which is a regularization parameter, were 1, 10, 100, and 1000. The default values of r, γ, and d are [1, 10,100, and 1000], [0.01, 0.1, 1, 10], and [2, 4] respectively.

The learning process in SVM aims to find the best hyperplane. Furthermore, based on the labeled data, the kernel selection, and tuning parameters, the data will be trained, and SVM will learn based on the input data and determine the classification according to the specified label. If the data dimensions are difficult to classify, then the shape will be changed to a higher dimension to facilitate the search of the best hyperplane. The amount of training data will be deducted from the number of test data. Then, the pattern will be formed to be implemented in the later stage.

The classification process will be calculated from each support vector in the database. Then for each vector in the training will be compared with the vector in the test. After the pattern is formed, the test will be conducted based on the desired amount of test data, which were previously taken from the training data before the learning process. Then, the test will be reviewed whether the test results match the values on the label so that the accuracy rate of the system can be determined.

The prediction process will be conducted using raw data. It filtered the data to match the necessity as described above but only limited to the input data. This data will be labeled because the main purpose of the prediction is to see whether the model can correctly classify and identify the data as expected output. The expected output is for the system to determine which fish can live in Lake Toba as per the input data and parameters.

2.2.4 Output

The output of this process is in the form of efficiency rate of two compared method in processing the data of the viability level of freshwater fish in Lake Toba in the form of percentage diagrams as well as the content of the water, whether it is optimal, can be tolerated, or bad; and also which fish are suitable to live in Lake Toba.

2. RESULT AND DISCUSSION

The tests were conducted when the system had been established. The tests have the purpose of finding whether the system is well-functioned in identifying the viability level of the fish, the most suitable fish to live in Lake Toba, as well as in finding the best parameters especially optimizer, activation function, and loss function – for LSTM, and the best kernel – for SVM; also to see whether the optional parameters can affect the identification results.

3.1 Test Result of Long Short-Term Memory

The first test of LSTM identification was conducted by combining each optimizer and the available activation function to find the best accuracy rate with the lowest error. The testing process was performed 24 times by changing the optimizers with a variety of activation functions, with selections of the optimizer such as SGD, RMSProp, Adam, Adamax, Adgrad, and Adelta and activation functions such as Softmax, Tanh, Relu, and Sigmoid. Moreover, the other parameters that cannot be changed are the epoch value of 100, the learning rate of 0,001, the batch size of 64, the 32 neurons, and the loss function of Categorical Cross-entropy. After the first test was completed, the best five combinations of optimizer and activation function will be tested using different loss functions. Two types of the loss function can be used for testing, namely, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The purpose of this test is to see if different function loss may impact the outcome of accuracy and error. Another test was performed to compare the results using three parameters of pH, water temperature, and DO. These are the main parameters to measure the survival rate of the fish with the help of another six parameters, namely voltage sensor, pH, water temperature, air temperature, air humidity, and DO. This test aims to see whether combining the unnecessary parameters can affect the accuracy or the error of the final result.

Based on the testing result, the system obtained some combinations of optimizer and activation, which have the highest accuracy and the fastest execution time, and the system indicated that the difference in optimizer and activation function could affect the accuracy level. The testing result is shown in Table 3.

Table 3: The accuracy rates of the combination of optimizer and activation functions

	SGD	RMS Prop	Adg rad	Ada m	Adam ax	Adel ta
Softmax	77%	97%	76%	76%	97%	76%
Tanh	77%	95%	76%	99%	98%	75%
Relu	76%	99%	77%	99%	99%	76%
Sigmoid	76%	99%	76%	99%	99%	75%

The next test was to examine the top five combinations of optimizer and activation function with the highest accuracy and the fastest execution time. It can be seen in Table 4 using different Loss Functions.

Table 4: The test result using six parameters

	Adam-Relu	Adam-Tanh	Adamax-Relu
Accuracy	98%	98%	99%
Loss	6%	7%	4%
Time	49.82 seconds	50.21 seconds	49.91 seconds

Another test was performed to see whether other parameters can affect the accuracy and error. The test was conducted using the best three combinations of optimizer and activation function with the highest accuracy and smallest error, namely Relu-Adam, Adam-Tanh, and Adamax-Relu, with the loss function of Categorical Cross-entropy. The result of this test is shown in Table 5.

After all the tests were completed, the last step is to identify the survival rate of fish in the specified area of Lake Toba using the three combinations of optimizer and activation functions, namely, Relu-Adam, Adam-Tanh and Adamax-Relu with the loss function of Categorical Cross-entropy.

Table 5: The test result using different loss function

Optimizer	Activation	Mean squared error	Mean absolute error	categorical
Rmsprop	Relu	Accuracy: 99% Loss: 0%	Accuracy: 75% Loss: 16%	Accuracy: 99% Loss: 2%
Adam	Tanh	Accuracy: 99% Loss: 0%	Accuracy: 77% Loss: 16%	Accuracy: 100% Loss: 1%
Adam	Relu	Accuracy: 99% loss: 1%	Accuracy: 75% Loss: 16%	Accuracy: 99% Loss: 2%
Adamax	Relu	Accuracy: 99% loss: 1%	Accuracy: 78% Loss: 15%	Accuracy: 99% Loss: 1%
Adamax	Sigmoid	Accuracy: 99% Loss: 0%	Accuracy: 76% Loss: 16%	Accuracy: 99% Loss: 2%

3.2 Test Result of Support Vector Machine

The testing process using SVM was conducted several times on the dataset using all kernels and various tuning parameters, which can be seen in Table 6.

Table 6: Accuracy of kernel test

Test-	Kernel	C	coef	gam-ma	Deg-ree	Accuracy
1	*Linear*	1	A			84,122%
2	*linear*	1000				88,015%
3	*rbf*	1		0.1		97,018%
4	*rbf*	1		1		99,223%
5	*rbf*	10		0.1		98,955%
6	*rbf*	10		10		99,866%
7	*sigmoid*	1	1	1		54,259%
8	*polynomial*	1	10	0.1	2	97,732%
9	*polynomial*	1	10	1	2	98,789%
10	*Polynomial*	1	100	0.1	4	99,49%

Confusion Matrix is a method using a matrix table that can provide an assessment of the classification performance based on true or false

objects and contains actual and predictive information in the classification system. The accuracy of the classification using RBF kernel, tuning values of parameters C = 10, and gamma of 10 is at 99.866%, which is the maximum accuracy value. In multilabel data, the model of the matrix table can be seen in Table 7.

Table 7: Multiclass matrix

		Actual				
		None	BGMNP	B	G	M
Predict	None	34	0	0	0	0
	BGMNP	0	938	0	0	0
	B	0	0	28	0	0
	G	0	0	0	0	0
	M	0	0	0	0	0

Actual is the original value of the labeling results in the form of a certain value. While predict is of the prediction result of the labeled value from training data. After drawing a diagonal line on the table, cells that are not affected by the line are called false, while the ones affected by straight lines are called true values. An empty column heading indicates no fish. BGMNP represents all fish that can live in Lake Toba with B, G, and M stand for Tilapia, Pomfret, and Goldfish, respectively.

3.3 Identification Result

The result will be obtained after the identification process was completed. It is in the form of the percentage of feasibility rate of the freshwater fish in the area. The result can be seen in Figure 8.

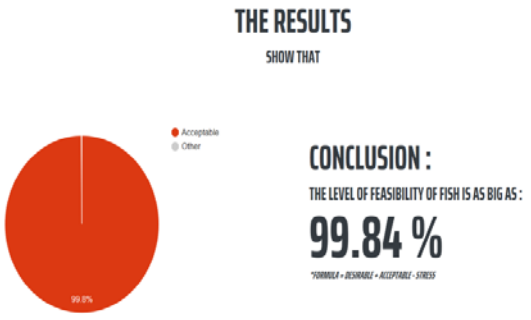

Figure 8: Identification Result of feasibility level of fish in Lake Toba

Based on the results of the first test, as shown in Table 2, it can be seen that the difference in the optimizer and activation functions greatly affects the accuracy rate. That is because the difference in calculation techniques of each optimizer and activation function, also the compatibility of the format data on these calculations so that the selection of optimizer and activation functions must be adjusted to the datasets and the methods.

Based on Table 3, it can be seen that the Loss Function can also affect accuracy, especially Mean Absolute Error. It decreases the accuracy rate. However, the loss functions of Categorical Cross-entropy and Mean Squared Error do not have a significant impact on the accuracy rate, only a slightly different loss value. From the testing results, we can see that the unfit loss function will lead to the decreasing accuracy rate of the neural network. As an example, Mean Absolute Error is commonly used in research for a prediction and is not a suitable classification. Therefore, it is mandatory to carefully choose the loss function because every issue requires a different loss function.

Based on the test results using the six parameters, it showed that the unnecessary parameters could affect the results of training, especially on the loss percentage. The loss percentages of the combinations of Adam-Relu, Adam-Tanh, Adamax-Relu were increasing from 2% to 6%, 1% to 7%, and 1% to 4%, respectively. This loss significantly affects the identification result.

As for the test results of Support Vector Machine, it showed that the algorithm achieved the objective of the study. In Table 6, it can be seen that from 1000 rows, the test can be classified well based on the comparison of actual values and predict values by drawing a diagonal straight line where the true value is 1000, while the False value is 0. The most suitable fish to live in Lake Toba in sequence are Pomfret, Goldfish, Carp, Tilapia, and Iridescent Shark. The result can be seen in Figure 9.

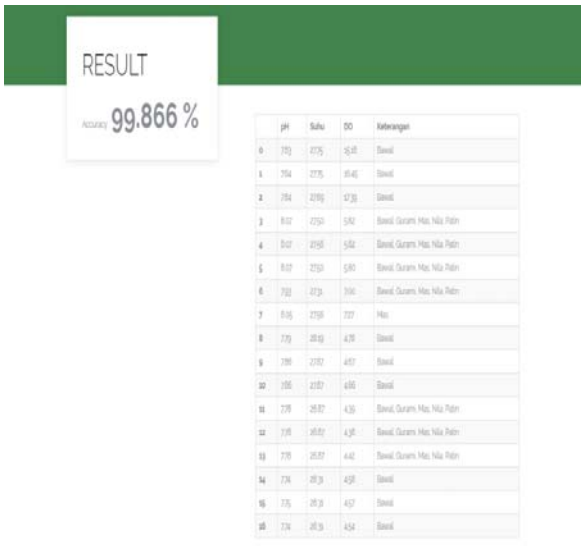

Figure 9: Identification Result of The Most Suitable Fish to Inhabit in Lake Toba.

3. CONCLUSION

Based on the test results using Long Short Term Memory to identify the feasibility level of the freshwater fish in Lake Toba based on the water content, it can be concluded that the test result showed that the survival rate of the freshwater fish in Haranggaol region is 99.84%, while the water content of Lake Toba contains 0.08% of bad water content. From the identification result, it can be concluded Lake Toba has good water content but not quite optimal for the freshwater fish in Lake Toba. The best optimizers and activation functions in LSTM model to achieve the best accuracy are Adamax and Adam for the optimizers and Relu and Tanh for the activation functions. In the identification process, long short term memory can conduct training and testing processes in less than one minute with 100 epoch. Unsuitable loss function may decrease the accuracy of the training process, so the loss function has to be chosen in accordance with the type of problems. Adding optional parameters may result in a higher error and decreasing the system accuracy of the identification process.

As for Support Vector Machine, the algorithm can identify the fish to inhabit Lake Toba well. The selection of kernel and tuning parameters affect the accuracy rate of the system, as well as minimizing the error value. If the kernel is exempted, the generalization of the model will be decreased, forcing the computation process to increase, while tuning parameters assist the system in separating the classes/groups, which affects the accuracy. The choice of the kernel that can achieve maximum accuracy rate is RBF kernel with tuning parameters of C at 10 and gamma of 10. The best accuracy level is at 99.866% because RBF or Gaussian can map the data to form a new vector space with infinite dimensions. So the higher the dimensions of a vector space, the more it is likely to find the best hyperplane to separate the data. It was also supported by the confusion matrix, where the number of false obtained is 0. Based on the result, the types of fish that have a higher survival rate are Pomfret and Goldfish.

In conclusion, both methods performed greatly in predicting the survival rate of the local fish to cultivate in Lake Toba. Comparative analysis of the classifiers shows that SVM outperformed LSTM with a slightly better accuracy at 99.86% that was 0.02% higher than the identification using Long Short-Term Memory. This indicates that machine learning could be implemented in predicting a pattern if given the right parameters and the right combination of activation function-optimizer and kernel-turning parameter.

REFRENCES:

[1] J. Manurung, “Parameter Optimation of Support Vector Machine using Genetic Algorithm in Credit Risk Assessment, 2017.

[2] A. M. Puspitasari, et al., “Dental and Mouth Classification using Support Vector Machine,” 2018.

[3] H. Situmorang, “Classification of Dengue Fever Epidemic Areas Using the Support Vector Machine Method in Medan City,” 2015.

[4] P. K. Dimpas, et al., “Filipino and English Clickbait Detection Using a Long Short Term Memory Recurrent Neural Network,” IEEE, 2014.

[5] H. Zhao, et al., “Sequential Fault Diagnosis Based On LSTM Network,” IEEE. 2017

[6] H. Tjahjo, et al., “Potential of Fisheries Resources V at Lake Toba, North Sumatra,” 1998.

[7] R. Harianja, et al., “Smart Monitoring Apps for Salvaging Neolissochillus Thienemanni Sumateranus (Batak Heritage) from Extinction,” IEEE, 2017, pp. 1-5..

[8] R. F. Rahmat, et al., “Real-Time Monitoring System for Water Pollution in Lake Toba,” International Conference on Informatics and Computing (ICIC),” 2016.

[9] A. Bhatnagar, P. Devi, "Water quality guidelines for the management of pond fish culture," International Journal of Environmental Sciences, 2013, p. 14.

[10] A. O. Pusphita, et al., "Application of Support Vector Machine (SVM) Classification Method in Primary School Accreditation Data in Magelan Regency," 2014.

[11] T. A. Barus, "The Biodiversity of Lake Toba Ecosystem and its Conservation. Inauguration of the Position of Professor Still in the Field of Limnology Science. Medan," Thesis of Universitas Sumatera Utara. 2007.

[12] R. F. Rahmat, et al., "Prediction of the Water Quality in Danau Toba," IEEE, 2017.

[13] Khairunnisa, "Regional Suitability Analysis for Aquaculture of Floating Cage Fish in Girsang Sipangan Bolon of Lake Toba," 2014.

[14] D. D. Haro, "Impact of Activity of Floating Net Cage Cultivation on Water Quality of Lake Toba in Haranggaol Horison District, Simalungun Regency, North Sumatra," 2014.

[15] T. Ginting , et al,. "RSS-Based Machine Learning for Localization using Cell-Id Global System For Mobile Communication (GSM)," 2015.

[16] R. A. Swief, et al. "Support vector machines (SVM) based short term electricity load-price forecasting," IEEE Bucharest, 2009, pp. 1 – 5.