Diagnosis of TBC Disease Using SVM and Feedforward Backpopagation

Dana Ramza Fakhma [1,*], Alamsyah [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: danaramza@students.unnes.ac.id

ARTICLE INFO

Article history
Received 10 Maret 2022
Revised 28 Maret 2022
Accepted 18 April 2022

Keywords
Tuberculosis
Classification
Feedforward Backpropagation
Support Vector Machine

ABSTRACT

Tuberculosis (TBC) is an infectious disease caused by a virus Mycobacterium tuberculosis. One of the organs is often infected by the virus Mycobacterium tuberculosis is the lungs. According to Laily et al. (2015) this disease is the second largest killer worldwide for infectious diseases after HIV/AIDS. Therefore, the level of diagnosis accuracy TBC disease needs to be improved using better methods. After the data is collected, then the data is processed in the preprocessing stage and through the normalization process so that the data range can be balanced. Furthermore, the last process is the classification process. In this classification process using two methods, namely Support Vector Machine and Feedforward Backpropagation. The two classification methods are assessed because they are simple and has a fairly precise level of accuracy, but also has a weakness in the selection of appropriate features. Based on research that has been done, using model testing with 10 executions, the accuracy results for Support Vector Machine produces an accuracy of 97.41%, while the results accuracy for Feedforward Backpropagation produces a level of accuracy by 98.51%. This shows that the Feedforward method Backpropagation is considered to improve the accuracy of diagnosis TBC disease.

1 Introduction

Tuberculosis (TBC) is still a public health problem in the world and mostly affects the productive age group, some are from socio-economic groups and have low levels of education because they don't know the disease they are experiencing and the symptoms they suffer are almost similar to other lung diseases. Every second, every time there is at least one person infected with TBC in the world. Every year, there are 8 million people with tuberculosis and there will be 3 million people with tuberculosis sufferers who die each year. 1% of the world's population will be infected with TBC within one year. One person will have the potential to transmit TBC disease every year (Fahmi, 2005). TBC is an infectious disease attack various organs or tissues of the body. Pulmonary tuberculosis is the most common form. Pulmonary TBC cases are increasing from year to year along with the increasing cases of HIV/AIDS. From the description above, it is clear that there are many similarities between the symptoms of TBC with other lung diseases. So that an error can occur early diagnosis for patients. Things like this can lead to even higher mortality rates due to early mishandling of patients.

Along with the development of technology, the completion of a problem was originally done manually, now can be done systematically through the application. Problem solving process can be done by Algorithm on an application. Algorithms are an effective method of well-defined commands to compute a function. Starting from a start condition and initial input, the instructions describe a computation when executed or processed through a limited number of sequence conditions can be well-defined and produce output (Sampurno, et al., 2018).

Artificial Neural Networks (ANN) can be problem solving in all areas field because of its ability to classify patterns, map patterns, optimize problems, and predict. JST is a branch of artificial intelligence and is a mathematical model driven by the organization and functional features of biological neural networks. Network nerves have connecting sets of artificial neurons and the process information using a connectionist form for computing. ANN has two stages, namely Backpropagation and Feedforward. Backpropagation itself is tested with one or more inputs is given as the initial state of the existing network. After input is given, the network will receive the feed and will generate a output which will then be fed back into an input. In particular, the Feedforward Neural Network (FFNN) is often affected by the number of neuron units in the hidden layer, which allows the level of the error is smaller. Support Vector Machine (SVM) is a machine technique classic learning that can help solve problems big data classification as well as assisting multi-domain applications in environments big data (Shan, 2016).

2 The Proposed Algorithm

2.1 Artificial Neural Network

Artificial Neural Network (ANN) is a branch of artificial intelligence and is a model mathematics driven by the organization and functional features of the network biological nerves. Neural networks have connecting sets of artificial neurons and process information using a connectionist form for computing. Generally, ANN is an adaptive system enhances the organization of external or internal information that travels through the network as long as learning process. The latest neural network is a nonlinear numerical data modeling tool. They usually use sophisticated models relationships between inputs and outputs or to uncover patterns in data. ANN has been applied in various applications with sufficient compliance (Nasser, 2019).

ANN is a tool for modeling non-linear processes based on information collected by a vector called the input layer, where the information is propagated layer by layer which builds the relationship between the input and the last layer called the output layer (Aulia, 2018). The intermediate or hidden layer consists of one or more units called neurons which are interconnected with the neurons of the previous and subsequent layers. The number of hidden layers and the number of neurons from each determines the topology of the network (Navares et al., 2018).

ANN is one of the information processing systems is designed to imitate the workings of the human brain in solving a problem by starting the learning process through changing the weight of the synapses. ANN is able to perform past data-based activity recognition. Past data will be studied by an ANN so that it has the ability to make decisions on data that has never been studied (Alfiati, 2017).

2.2 Feedforward Neural Network

Feedforward Neural Network (FFNN) is one of the most common types of ANN. In FFNN, neurons arranged in layers and fully interconnected to form a directed graph. The FFNN layer is an input layer, a number of layers hidden, and output layers (Faris et al., 2016). An ANN with a feed-forward structure or what can be called FFNN has the characteristic of not having a learning loop where the signal travels from the input layer through the hidden layer then to the output layer (Wibowo, 2017).

FFNN is not like other algorithms, for example Genetic Algorithm, Particle Swarm Optimization, Ant Colony Optimization which are good at exploitation and exploration and can handle simultaneous adaptation in each component of FFNN. However, no single method can solve all types of problems. So, we need to improvise, adapt and build hybrid methods to optimize FFNN (Ojha et al., 2017).

The steps carried out in the feedforward process are as follows (Fernanda & Otok, 2012).

- Each input unit (x_i, i = 1, …, n) has the task of receiving the input signal x_i and spreading it to all units in the hidden layer.
- Each hidden unit (z_j, j = 1, …, p) has the task of adding up the weights obtained using the equation below.

$$Z_{inj} = V_{0j} + \sum_{i}^{n} X_i V_{ij} \; Z_{inj} = V_{0j} + \sum_{i}^{n} X_i V_{ij} \tag{1}$$

- The activation function has the task of calculating the output signal and sending it to the output layer using the equation below.

$$Z_j = f(Z_{inj})\, Z_j = f(Z_{inj}) \tag{2}$$

- Each output unit (Y_k, $k = 1, \ldots, m$) has the task of adding up the weights of the output signal using the equation below.

$$Y_{ink} = W_{ok} + \sum_{j}^{p} Z_j W_{jk}\, Y_{ink} = W_{ok} + \sum_{j}^{p} Z_j W_{jk} \tag{3}$$

- The next activation function has the task of calculating the output signal using the equation below.

$$Y_k = f(Y_{ink})\, Y_k = f(Y_{ink}) \tag{4}$$

2.3 Backpropagation Neural Network

Backpropagation Neural Network (BPNN) is an ANN training method supervised. It evaluates the error contribution of each neuron after a data set has been processed. The purpose of BPNN is to modify weights to train the neural network to map arbitrary inputs to outputs correctly. Multi-layered perceptrons can be trained using backpropagation algorithm. The aim is to study the weights for all the linkages in a multi-layered network. Minimum error function in the weight space is calculated using the gradient descent method. The resultant weight that offers the minimum error function is the solution to the learning problem (Amrutha & Remya, 2018).

BPNN is a general method of learning ANN how to complete a given task. This is a supervised learning process and is an implementation of the delta rule (Vamsidhar et al., 2010). BPNN is a supervised learning algorithm with many layers. This algorithm is also included in the type of controlled learning uses a weight adjustment pattern to achieve a minimum error value between the predicted input and the actual output. This algorithm is an excellent method for dealing with complex pattern recognition problems related to pattern recognition, prediction and identification (Wibowo, 2017).

BPNN is a well-known network that is known for its accuracy because it allows itself to learn and improve itself so that it can achieve higher accuracy. The BPNN model was developed as an extension of the perceptron learning algorithm to a multi-layered NN as shown in Figure 1.

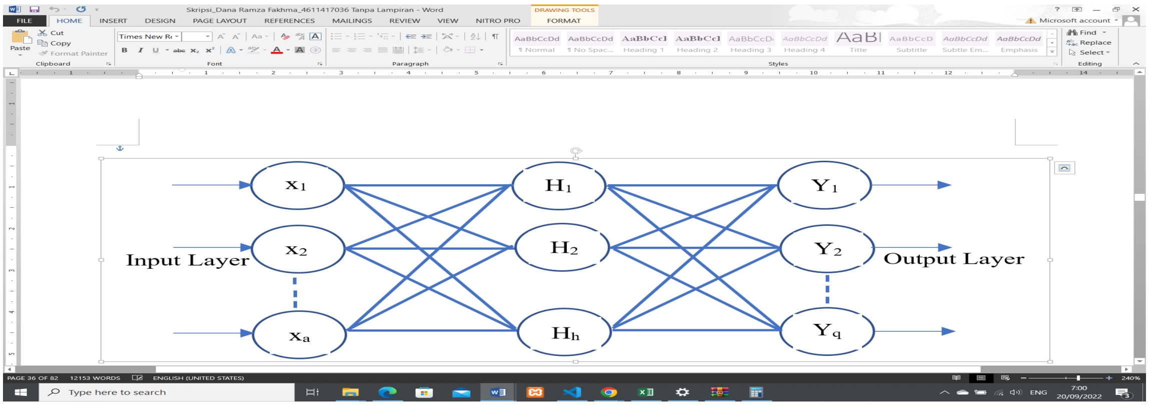
Skripsi_Dana Ramza Fakhma_4611417036 Tanpa Lampiran - Word
FILE HOME INSERT DESIGN PAGE LAYOUT REFERENCES MAILINGS REVIEW VIEW NITRO PRO FORMAT
X1
X2
Xa
H1
H2
Hh
Y1
Y2
Yq
Input Layer
Output Layer
PAGE 36 OF 82 12153 WORDS ENGLISH (UNITED STATES)
Type here to search
7:00
20/09/2022

Figure 1. Backpropagation Neural Network Architecture

The steps carried out in the backpropagation process are as follows (Fernanda & Otok, 2012).

- Each output unit (Y_k, $k = 1, \ldots, m$) has the task of receiving a target pattern based on the pattern in the input training.
- Perform the process of calculating error information using the equation below.

$$S_k = (T_k - Y_k)\, f'(Y_{ink})\; S_k = (T_k - Y_k)\, f'(Y_{ink}) \tag{5}$$

- Calculate the correction weight to update the previous *Wjk* with the equation below.

$$\nabla W_{jk} = aS_kZ_j \quad \nabla W_{jk} = aS_kZ_j \tag{6}$$

- Each hidden unit (z_j, $j = 1, \ldots, p$) has the task of adding up the input delta from the layer above it using the equation below.

$$S_{ink} = \sum_{k=1}^{m} S_k W_{jk}\; S_{ink} = \sum_{k=1}^{m} S_k W_{jk} \tag{7}$$

- Multiply the weight correction to update the previous V_{ij} with the equation below.

$$\nabla V_{ij} = aS_jX_i \quad \nabla V_{ij} = aS_jX_i \tag{8}$$

- Updating the weights and biases with the following conditions.
  a. Each unit of output (Y_k, $k = 1, \ldots, m$) has the task of updating its weights and biases using the equation below.

$$W_{jk}(baru) = W_{jk}(lama) + \nabla W_{jk} \tag{9}$$

$$W_{jk}(baru) = W_{jk}(lama) + \nabla W_{jk} \tag{10}$$

  b. Each hidden unit (z_j, $j = 1, \ldots, p$) has the task of updating its weights and biases using the equation below.

$$V_{ij}(baru) = V_{ij}(lama) + \nabla V_{ij} \tag{11}$$

$$V_{ij}(baru) = V_{ij}(lama) + \nabla V_{ij} \tag{12}$$

  c. Make testing for stop conditions.

BPNN has the ability to solve complex problems. This is possible because the network with this algorithm is trained using the supervision learning method so that it can recognize the input pattern of a data with a high level of accuracy (Fitryadi et al., 2016). BPNN is an ANN model with multilayer which is often used for estimating time series data. This method is the best method in dealing with the problem of recognizing complex patterns. Some applications that use the BPNN method are data compression, computer virus detection, object identification, and many more. In general, BPNN can be applied in various fields as a forecasting method, while in future research, applying the BPNN model to diagnose TBC cases in Indonesia with input of factors that affect TBC.

2.4 Support Vector Machine

Support Vector Mahine (SVM) is one of the classic machine learning techniques can be used to help solve big data classification problems and help multi-domain applications in a big data environment (Shan, 2016). SVM is one of the machine learning methods whose workings are based on the principle of Structural Risk Minimization (SRM) to obtain functional best separator (hyperplane) that can separate two data sets from two different classes (Rahutomo, et al., 2018).

SVM can accurately classify genes into several functional categories and make predictions to identify the function of unnotated yeast genes (Brown et al., 1999). SVM is able to classify tissue and cell types, just like the perceptron method. In addition, SVM can also be used to identify mis-labeled data. In this study, a simple kernel was applied (Furey et al., 2000). Data processing can be done with text mining techniques. To process large text data is required a machine to explore opinions, including positive or negative opinions. SVM is one of the classification algorithms that can be used for sentiment analysis. However, SVM works less well on the large-sized data. (Larasati, et al., 2019).

3 Method

Before in this research, diagnosis of TBC disease performed by applying SVM and Feedforward Backpropagation as classification method. The desired results in this research are the accuracy of the proposed method. The flowchart of this research method can be seen in Figure 2.

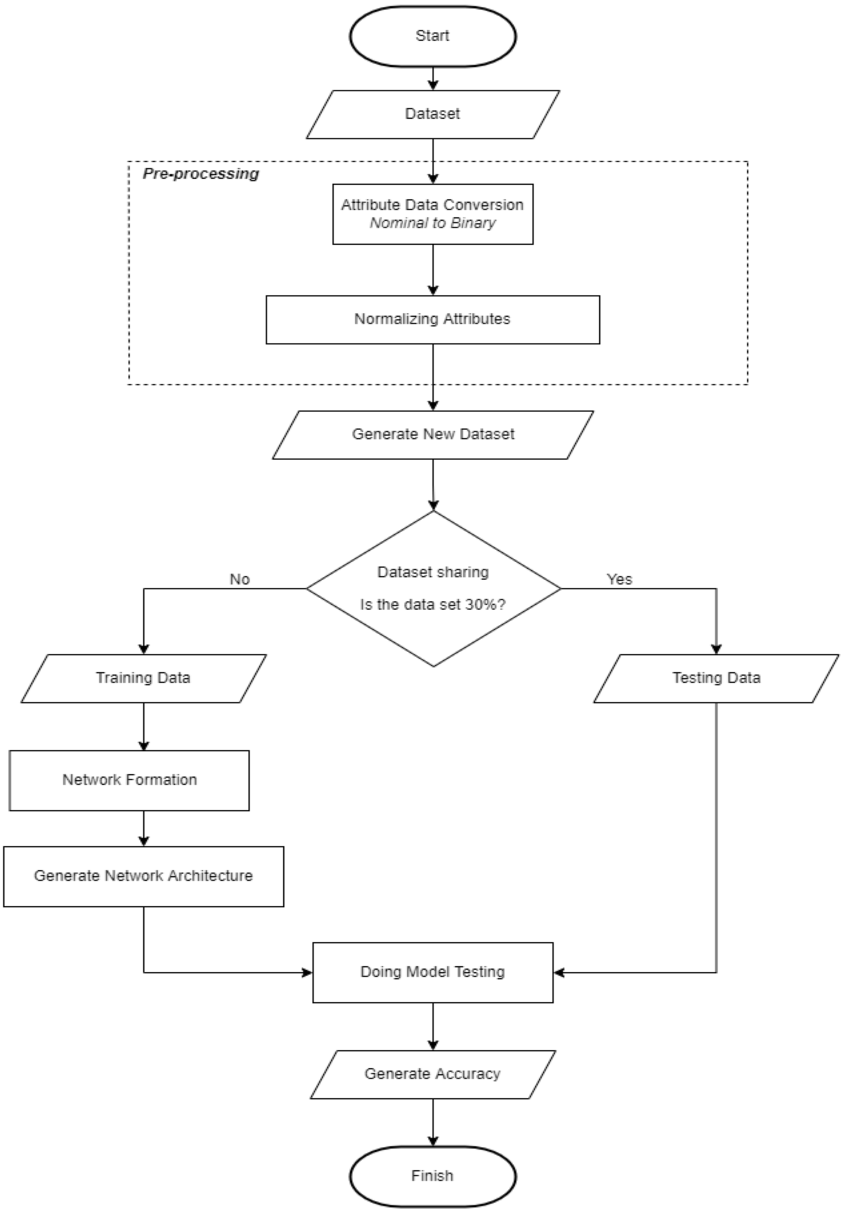

Figure 2. Figure of research steps

This research begins by preparing the dataset which carried out in a preprocessing stage. In this case, the dataset only goes through a step, which is attribute normalization. The research on analyzing TBC disease using SVM and Feedforward Backpropagation is carried out in several stages. The first stage is to prepare the database, which is collected from Kaggle. The dataset is split into 70% training data and 30% test data. The second step is to normalize the attribute of the dataset, the x_i data in the range of [-1–1] which has z_{max} as the maximum value and z_{min} as the minimum value from the data attributes.

Prediction is an important thing used to know the future events by recognizing the patterns of events in the past. While knowing the events that will happen, it makes everyone better prepare everything, both for human life and their property (Hikmawati, et al., 2017).

4 Results and Discussion

In this study, the SVM and Feedforward Backpropagation methods were used to compare the process in diagnosing TBC starting from testing, modeling, system design, and discussing system results. A more complete explanation regarding the research results is as follows.

4.1 Parameter Testing Result

Parameter test results obtained parameter values in the method used. The results of data classification are carried out by dividing the data into two parts, 70% of the data is used for training data and 30% of the data is used for test data. The total data obtained were 404 training data and 173 test data. So that the tests carried out will be divided into two, it is testing the SVM parameters and the Feedforward Backpropagation parameters as follows.

4.1.1 Support Vector Machine Parameter Test Results

The results of the SVM parameter test using the best cost (*C*) parameter using a Linear Kernel. The parameter initialization search process can be seen in Table 1.

Table 1. *C* Parameter Test Results

Kernel	Parameter *C*	Accuracy (%)
Linear	0.001	48.88
Linear	0.01	48.88
Linear	0.1	98.26

The results of testing the initialization parameter C using training data get the best results using a linear kernel and C = 0.1 with an accuracy of 98.26%. To calculate the accuracy using the Confusion Matrix shown in Table 2.

Table 2. Confussion Matrix SVM

		Prediction	
		Positive	Negative
Actual	Positive	197	0
	Negative	7	199

$$\text{Accuracy} = \frac{197+199}{197+199+7+0}\ x\ 100\%\ \frac{197+199}{197+199+7+0}\ x\ 100\% = 98.26\%$$

4.1.2 Feedforward Backpropagation Parameter Test Results

The results of testing the Feedforward Backpropagation parameters tested include the number of hidden layers, the number of iterations (epochs), and the learning rate value to determine the results of data processing using training data. The results of this test produce Mean Square Error (MSE) and Accuracy (%) values which are the reference values to determine the best parameters in this study. There are three parameters tested in this study, such as.

Table 3. Feedforward Backpropagation Parameter Test Results

Input/ Hidden layer	Learning rate	Epoch	Mean Square Error	Accuracy (%)
1/25	0.2	40	0.00043	95.41
1/20	0.005	50	0.00089	34.92
1/10	0.01	30	0.00066	54.44
1/5	0.05	20	0.00063	56.39
1/10	0.2	40	0.22886	95.69
1/20	0.05	10	0.00068	47.28
1/15	0.005	20	0.25365	38.79
1/20	0.01	30	0.00062	51.62
1/5	0.1	50	0.00067	58.35
1/25	0.2	40	0.00051	98.51

4.1.3 Result of Determination of Support Vector Machine Model

The results of the parameter testing of the SVM method get a fairly good accuracy result, which is 98.26% by using the parameter *C* = 0.1. Therefore, the process of determining the model will use the parameter C = 0.1 and use the Confusion Matrix calculation. The results of determining the SVM model will be shown in Table 4.

Table 4. Result of Determination of Support Vector Machine Model

		Prediction	
		Positive	Negative
Actual	Positive	66	0
	Negative	3	47

$$Accuracy = \frac{66+47}{66+47+3+0}\ x\ 100\% \frac{66+47}{66+47+3+0}\ x\ 100\% = 97.41\%$$

4.1.4 Result of Determination of Feedforward Backpropagation Model

The results of determining the Feedforward Backpropagation model are carried out using the best results from the parameter testing that has been done previously. The values of these parameters are shown in Table 5.

Table 5. Feedforward Backpropagation Parameters

Parameter	**Nilai**
Input layer	1
Hidden layer	25
Learning rate	0.2
Epoch	40

The training results from the calculation process using the Feedforward Backpropagation method and the parameters are shown in Table 6.

Table 6. Data Training Results

Hidden layer	Epoch	Learning rate	MSE	Accuracy (%)
25	40	0.2	0.00064	58.35
25	40	0.2	0.00068	58.56
25	40	0.2	0.00047	63.77

Then the model obtained from the training results is tested using test data, the test results from the test data are shown in Table 7.

Table 7. Data Testing Results

Hidden layer	Epoch	Learning rate	MSE	Accuracy (%)
25	40	0,2	0,00043	95,41
25	40	0,2	0,22886	95,69
25	40	0,2	0,00051	98,51

The research that has been conducted, applies the SVM and Feedforward Backpropagation methods to get the results of a comparison of the level of accuracy in the diagnosis of TBC disease. This study resulted in the workings and performance of the method, as well as to find out which one is better in diagnosing TBC disease. In this study, a comparison of the accuracy of the SVM and Feedforward Backpropagation methods will be carried out as seen from the data testing process.

The first method used is the SVM method. This method will calculate the accuracy value, then the result of the largest accuracy value will be selected. From the results of experiments carried out using the parameter $C = 0.1$, it produces an accuracy value of 97.41%. Feedforward Backpropagation method each execution produces a different accuracy value. After doing several experiments, the highest accuracy value is 98.51%. Comparison of the results of the accuracy of each method can be seen in Table 8.

Table 8. Comparison of The Results of The Accuracy of Each Method

Method	Accuracy (%)
Support Vector Machine	97.41
Feedforward Backpropagation	98.51

Research conducted by researchers has the advantage that the Feedforward Backpropagation method can produce a better level of accuracy than the SVM method. However, this study also has a drawback, namely the Feedforward Backpropagation method produces a low accuracy value, if using parameters other than those specified. Therefore, further development is needed so that data processing in each method can run as much as possible.

5 Conclusion

The application of the Feedforward Backpropagation method can be demonstrated starting from four stages, it called the data collection stage, the pre-processing stage, the system design stage, and the system testing stage. The data retrieval stage is in the form of searching and sorting TBC datasets from open source in the form of Kaggle, then the pre-processing stage, which is normalizing the data into binary numbers. Furthermore, the system design stage is displayed using the Flask Framework to display an attractive and user-friendly system display. The last stage is testing the system using predetermined parameters in order to produce maximum accuracy in diagnosing TBC disease. The highest accuracy results are obtained from the application of the SVM method to get 97.41% using the C = 1 parameter and the Feedforward Backpropagation method 98.51% using the input neuron parameter = 1, hidden layer = 25, learning rate = 0.2, and epoch = 40. Thus, it is found that the Feedforward Backpropagation method can improve the results of better accuracy than the SVM method in the diagnosis of tuberculosis.

References

Alfiati, W. (2017). PERAMALAN PENJUALAN PIPA DI PT. CIKAL TIRTA SARANA SURAKARTA DENGAN MENGGUNAKAN ALOGARITMA ARTIFICIAL NEURAL NETWORK (Doctoral dissertation, STMIK Sinar Nusantara Surakarta).

Amrutha, J., & Ajai, A. R. (2018, May). Performance analysis of backpropagation algorithm of artificial neural networks in verilog. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1547-1550). IEEE.

Fahmi, A. U. (2005). Manajemen penyakit berbasis wilayah. *Jakarta: Penerbit Buku Kompas*.

Fernanda, J. W., & Otok, B. W. (2012). Boosting neural network dan boosting cart pada klasifikasi diabetes militus tipe II. *Jurnal Matematika*, *2*(2), 33-49.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906-914.

Hikmawati, Z. S., Arifudin, R., & Alamsyah, A. (2017). Prediction the Number of Dengue Hemorrhagic Fever Patients Using Fuzzy Tsukamoto Method at Public Health Service of Purbalingga. *Scientific Journal of Informatics*, *4*(2), 115-124.

Larasati, U. I., Muslim, M. A., Arifudin, R., & Alamsyah, A. (2019). Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis. *Scientific Journal of Informatics*, *6*(1), 138-149.

Nasser, I. M., & Abu-Naser, S. S. (2019). Lung cancer detection using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS)*, *3*(3), 17-23.

Navares, R., Díaz, J., Linares, C., & Aznarte, J. L. (2018). Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. *Stochastic environmental research and risk assessment*, *32*(10), 2849-2859.

Ojha, V. K., Abraham, A., & Snášel, V. (2017). Metaheuristic design of feedforward neural networks: A review of two decades of research. *Engineering Applications of Artificial Intelligence*, *60*, 97-116.

Rahutomo, F., Saputra, P. Y., & Fidyawan, M. A. (2018). Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine. *Jurnal Informatika Polinema*, *4*(2), 93-93.

Shan, Y., Zhang, Y., Zhuo, X., Li, X., Peng, J., & Fang, W. (2016). Matrix metalloproteinase-9 plays a role in protecting zebrafish from lethal infection with Listeria monocytogenes by enhancing macrophage migration. *Fish & Shellfish Immunology*, *54*, 179-187.

Wibowo. (2017). Manajemen Kinerja. Edisi Kelima. Depok: PT. Raja Grafindo Persada.