IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.14, No.2, April 2020, pp. 189~198
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.51086

Implementation of Genetic Algorithms and Momentum Backpropagation in Classification of Subtype Cells Acute Myeloid Leukimia

Dian Mustikaningrum*[1], Retantyo Wardoyo[2]
[1]Master Program in Computer Science and Electeonics, FMIPA UGM, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]dian.mustikaningrum@mail.ugm.ac.id**, [2]rw@ugm.ac.id

Abstrak

Acute Myeloid Leukimia (AML) merupakan salah satu jenis penyakit kanker yang menyerang sel darah putih jenis myeloid. Subtipe AML M1, M2, dan M3 dipengaruhi oleh tipe sel yang sama yaitu myeloblast, sehingga untuk klasifikasi diperlukan analisis yang lebih rinci. Metode klasifikasi yang digunakan adalah Momentum Backpropagation. Dalam penerapannya, pemilihan arsitektur, learning rate, dan momentum yang optimal masih dilakukan dengan cara random trial. Hal inilah yang menjadi salah satu kekurangan Momentum Backpropagation. Penelitian ini menggunakan algoritma genetika (GA) sebagai metode optimasi untuk mendapat arsitektur, learning rate, dan momentum terbaik dari JST. Algoritma genetika adalah salah satu teknik optimasi yang meniru proses evolusi biologi.

Data set yang digunakan dalam penelitian ini adalah data fitur numerik hasil dari segmentasi citra sel darah putih yang diambil dari penelitian sebelumnya yang dilakukan oleh Nurcahya Pradana Taufik Prakisya. Berdasarkan data tersebut dilakukan proses evaluasi terhadap pemilihan parameter Momentum Backpropagation random trial dengan algoritma Momentum Backpropagation dengan Algoritma Genetika. Selanjutnya dilakukan perbandingan yang mampu memberikan hasil yang lebih akurat dengan data yang digunakan dalam penelitian ini.

Hasil penelitian menunjukkan bahwa pelatihan dan pengujian dengan optimasi algoritma genetika terhadap parameter JST menghasilkan rata-rata akurasi memorisasi sebesar 83,38% dan akurasi validasi 94,3%. Sedangkan pelatihan dan pengujian dengan momentum backpropagation random trial menghasilkan rata-rata akurasi memorisasi 76,09% dan akurasi validasi 88,22%.

Kata kunci— *Acute Myeloid Leukimia, Jaringan Syaraf Tiruan, Momentum Backpropagation, Algoritma Genetika*

Abstract

Acute Myeloid Leukimia (AML) is a type of cancer which attacks white blood cells from myeloid. AML subtypes M1, M2, and M3 are affected by the same type of cells called myeloblasts, so it needs more detailed analysis to classify.Momentum Backpropagation is used to classified. In its application, optimal selection of architecture, learning rate, and momentum is still done by random trial. This is one of the disadvantage of Momentum Backpropagation. This study uses a genetic algorithm (GA) as an optimization method to get the best architecture, learning rate, and momentum of artificial neural network. Genetic algorithms are one of the optimization techniques that emulate the process of biological evolution.

The dataset used in this study is numerical feature data resulting from the segmentation of white blood cell images taken from previous studies which has been done by Nurcahya Pradana Taufik Prakisya. Based on these data, an evaluation of the Momentum Backpropagation process was conducted the selection parameter in a random trial with the genetic algorithm. Furthermore, the comparison of accuracy values was carried out as an

Received November 1st, 2019; Revised April 20th, 2020; Accepted April 29th, 2020

alternative to the ANN learning method that was able to provide more accurate values with the data used in this study.

The results showed that training and testing with genetic algorithm optimization of ANN parameters resulted in an average memorization accuracy of 83.38% and validation accuracy of 94.3%. Whereas in other ways, training and testing with momentum backpropagation random trial resulted in an average memorization accuracy of 76.09% and validation accuracy of 88.22%.

Keywords— *Acute Myeloid Leukimia (AML), Neural Network, Momentum Backpropagation, Genetic Algorithm*

1. INTRODUCTION

Acute myeloid leukimia (AML) is characterized by an increase in the number of myeloid cells in the marrow and an arrest in their maturation, frequently resulting in hematopoietic insufficiency (granulocytopenia, thrombocytopenia, or anemia), with or without leukocytosis[1]. Five-year survival rates during this period were less than 15 percent. Over the past decade, refinements in the diagnosis of subtypes of AML and advances in therapeutic approaches have improved the outlook for patients with AML. Despite these improvements, however, the survival rate among patients who are less than 65 years of age is only 40 percent. The primary reason for the outbreak of this cancer is still a mystery. Moreover, weakness, fever, tiredness or pain in joints and/or bones are also symptoms associated with AML just like other common ailments. Since the cancer is acute, it is even more important to detect it while it is in its primary stages of growing. Thus, it is very important to have a system that can detect AML accurately[2].

Artificial Neural Network (ANN) provide main features, such as : flexibility, competence, and capability to simplify and solve problems in pattern classification, function approximation, pattern matching and associative memories[3]. ANN has the aptitude for random nonlinear function approximation and information processing which other methods does not have. Different techniques are used in the past for optimal network performance for training ANNs such as backpropagation neural network (BPNN) algorithm. However, the BPNN algorithm suffers from two major drawbcaks: low convergance rate and instability. The drawbacks are caused by a risk of being trapped in a local minimum, and possibility of overshooting the minimum of the error. Another weakness found that is in the application of high levels of difficulty and complex combinations of the given criteria, namely learning speed, size, generalization ability, and resistance to data disturbances and increasing size and complexity, making artificial neural networks trapped in local optimal[4].

The combination of architectural parameters, initial weights, and initial biases greatly determines the learning ability of artificial neural networks [5]. Predictions generated by ANN so far there are no standard rules regarding how many optimal hidden layers. Each case certainly has a different number of hidden layers that will be used to get the optimal solution. In addition to some of the parameters previously mentioned, an increase in the value of learning rate can cause an increase in the speed of training in reaching the point of convergence but can also reduce the value of predictive analysis, especially in the value of precision in each class. The trial and error method is usually used in finding the highest parameter values used in learning ANN to get the highest accuracy, precision, sensitivity, and specificity[6].

Evolutionary computation is often used to train the parameter of neural network. In recent years, many improved learning algorithms have been proposed to overcome the weakness of gradient-based techniques. Genetic algorithms are widely used for optimization problems in artificial neural networks, for example research related to the diagnosis of breast cancer using genetic algorithms[7]. The same thing was done in detecting skin cancer. The combination of genetic algorithms and artificial neural networks can also provide better accuracy results than the methods used previously[8].

Genetic algorithms can be hybridized with other algorithms. For example, gradient-based methods can be used to enhance the performance of genetic algorithms. The global search capability of genetic algoithms is used to ensure a high probability of finding global optimality, whereas the derivatives or local information can be used to speed up local search. For this approach, genetic algorithms can be combined with various other algorithms to use the advantage of both algorithms[9].

2. METHODS

In this section, the proposed method is explained in detail. This Includes data descriptions, preprocessing of data, and methods in classifying subtype cells *myeloblast*, *promyelocyte*, *monoblast*, and *support* cell by using hybrid genetic algorithms backpropagation.

2.1 Subtype cells AML (Myeloblast, Promyelosit, Monoblast, and Support)

Acute Myeloid Leukemia (AML) is a type of cancer that develops rapidly which attacks blood cells and spinal cord. The method often used in AML classification was developed by French-American-British (FAB), which classifies AML nine subtypes, namely M0, M1, M2, M3, M4, M4Eo, M5, M6, M7 [1]. The characteristic of each cell can be shown in Table I.

Table I. Cell Characteristics

Subtype	**General Name (% case)**	**Description**
M1	*Acute myeloblastic leukimia without maturation* (15-20%)	Auer-Rods that are seen in the indentation of the nucleus, morphologically are still undifferentiated
M2	*Acute myeloblastic leukimia with maturation* (25-30%)	There are blast cells with clear basophilic cytoplasm with pale perinucleus (Golgi zone), sometimes translocation can be seen
M3	*Acute promyelocytic leukimia* (5-10%)	There is an atypical granular promyelocyte with cytoplasm filled with Auer-rods

2.2 Data Description

Sources of data in this study using research data on previous researchers [10]. The data used is in the form of discrete data from segmented image features. The data used in this study has six features consisting of area, edge area or perimeter, roundness, nucleus ratio, mean, and standard deviation.

Table II. Information of these datasets

Subtype Cells				**Total**
Myeloblast	**Promyelosit**	**Monoblast**	**Support**	
201	6	0	17	**224**
161	20	19	25	**225**
17	101	10	157	**285**
379	**127**	**29**	**199**	**734**

2.3 Preprocessing Data (*Data Normalization*)

The classification data used in this study is numerical data consisting of six feature parameters, namely area area, edge area, cell roundness, nucleus ratio, mean, and standard deviation. The data that has been obtained then passes the normalization process, this is necessary because the extraction results have a variety of values. The range of feature values can be described : area and edge area have a range of integer (integer) values, The roundness and nucleus ratio has a range of real number values between 0 to 1, mean and standard deviations are real numbers with a range of values ranging from - to 255. All feature data will be normalized, before entering the training process, use (1)

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where:

x_i = The value to-*i* before normalized
$x_i{}'$ = The value to-*i* after normalized
x_{min} = The minimum value of the data
x_{max} = The maximum value of the data

2.4 Backpropagation

The very general nature of backpropagation training method means that a backpropagation net (a multilayer, feedforward net trained by backpropagation) can be used to solve problems in many areas[11]. In tis study Backpropagation model can be grouped into three layers, namely, input layer, hidden layer and output layer, Fig 1 represents the architecture.

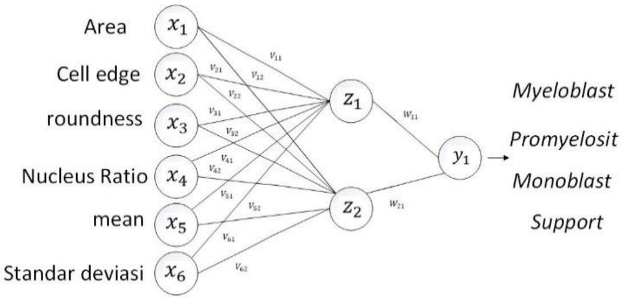

Figure 1 Represents the architecture

ANN modeling is divided into two stages (i.e. training and testing). The first part of the modeling is the training stage in which formulation of the initial structure of ANN is executed. Subsequently, validation is stage to ensure teh accuracy of the final model. In addition, the datasets are distributed using *k-fold cross validation, k*=3.

2.5 Genetic Algorithm

There are many advantages of genetic algorithms over traditional optimization algorithms[9]. GA is a method for solving both constrained and unconstrained optimization problems. The key concept of GA mechanism bases on natural selection, the process that drives biological evolution. The method begins with a set of individuals, called an initial population. GA repeatedly modifies a population of individual solutions. At each generation, GA selects individuals from the current population to be parents and uses them to produce the children for the next generation. Over successive generation, the population evolves to an optimal solution.

2.5.1 Initialization

Populations contain a group of individuals as a solution to the problems represented on chromosomes. Each chromosome contains genes which are information from individuals. The length of the chromosome is designed not to be fixed because the determination of the search area for the solution are learning rate, momentum, the number of hidden layers and the number of neurons. The chromosome representation in this study uses a real number encoding representation scheme where each hyperparameter in ANN: learning rate, momentum, the number of hidden layers, and the number of neurons in each hidden layer into genes.

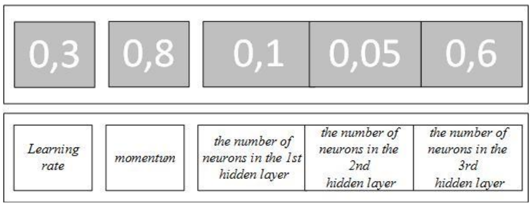
0,3
0,8
0,1
0,05
0,6
Learning rate
momentum
the number of neurons in the 1st hidden layer
the number of neurons in the 2nd hidden layer
the number of neurons in the 3rd hidden layer

Figure 2. Structure of GA Chromosomes

The individual shown in Figure 2, an individual will have a hidden layer structure of 1 to 3 with neurons in each layer having a different number with a range of 0 to 100 units. Chromosomes will be divided into three parts, 1 gene for learning rate, 1 gene for momentum, and 1-3 non-fix genes that indicate the number of neuron in each hidden layer. In the third part of the chromosome the denormalization process will be carried out to get the number of neurons in each hidden layer. Calculation of denormalization on chromosome part 3 can use the formula (2).

$$\boldsymbol{x' = x_i \text{ x } 100} \qquad (2)$$

where :

x' = normalization data

x_i = Chromosomes *i*

2.5.2 Fitness Function or Evaluation Function

The objective function of genetic algorithm in this study is to minimize errors. Fitness values state how an individual can be the solution of the problem that is defined. This fitness function can be used to see which individual is producing the smallest error value. Fitness function calculations can be obtained from the average root value of errors in the system or commonly called MSE (Mean Square Error). Optimal value is obtained by getting the smallest MSE value so that the greater the value of fitness. For optimal value problems, the fitness function shown (3).

$$Fitness = \frac{1}{MSE} \qquad (3)$$

Based on the results of the fitness function: learning rate, momentum, the number of hidden layers, and the number of neurons in each hidden layer will be obtained that corresponds to the individual from the calculation process.

In our framework based on GA, we have employed the following algorithm/methods Roulette Wheel Selection, Whole Arithmatic Crossover, flip mutation and Elitism.

2.5.3 Termination criteria

In this enhanced GA Backpropagation approach, the learning /evolutionary process is terminated, if it meets the condition that number of fitness evaluation reaches its maximum count.

2.6 Process

In this study will be constructed model of architecture based on these data, an evaluation of the Momentum Backpropagation process was conducted the selection of a random trial parameter with the genetic algorithm. Genetic algorithm method used to analyze and classify *myeloblast*, *promyelocyte*, *monoblast*, and *support* cell to determine the number of hidden layers, number of neurons, learning rate, and momentum in backpropagation neural networks. The expected output is to get the best-classified result value with the smallest MSE value. The best-classified value results can be obtained by testing the amount of data, as well as testing on the genetic parameters: population size, crossover probability, and mutation probability.

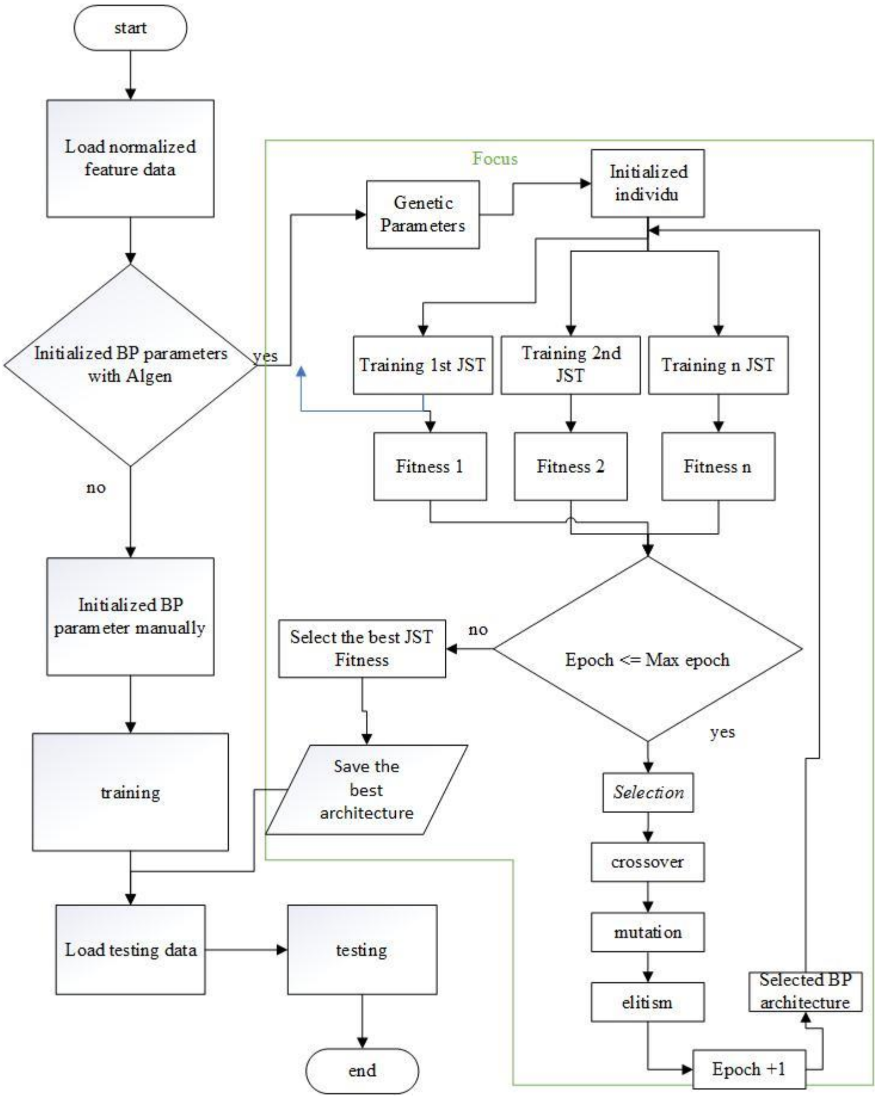

Figure 3 Flowchart diagram for *Myeloblast, Promyelosit, Monoblast, and Support* with Genetic Algorithm

The first step is to specify the input data to be used, then preprocessing the data using the Min-Max (0 – 1) normalization. Training data and normalised test data for easy calculation. The next step is to determine the type of kernel and the value of the parameters to be used, then the process of initializing parameters the user is given the choice to determine the architecture manually or with genetic algorithms. This process is the focus of research where the results of these two methods will be compared to be able to see the difference in terms of accuracy with the aim of seeing whether this method can provide better learning ability than standard methods without GA modification. After sequential process of selecting parameters with GA, population initialization is done by arousing individuals with real representation chromosomes. After that, the calculation of fitness values by looking at the accuracy of the training process ANN which represents each individual in the population that has been raised. Individuals who have met maximum epoch will be stored for later use in training data that will be compared with momentum backpropagation with the generation of parameters in a random trial without GA. The process of loading test data, entering testing data that will be used in the testing process. Broadly, the system design plot is depicted in the flowchart in Figure 3.

3. RESULTS AND DISCUSSION

The results are derived from experiments that have been performed on each genetic parameters. The test parameter is attempted based on the constraints of the parameter value that has been defined. Based on this test, there will be a pair of the best parameter values of each fitness based on the smallest MSE value. There are several tests conducted on the system including testing the learning process of genetic algorithms which includes the influence of genetic parameters on fitness values and comparative analyzes with backpropagation momentum without GA of the resulting accuracy.

3.1 The influence of genetic parameters on fitness values

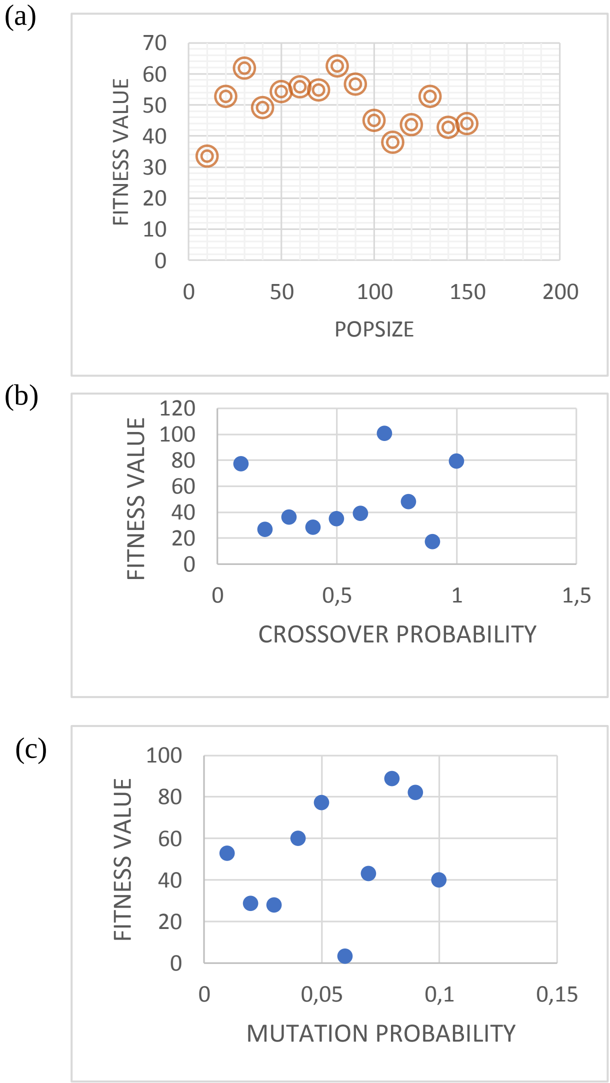

Figure 4 Effect of parameter changes (a) *Popsize*; (b) *Crossover Probabilitiy*; (c) *Mutation Probability*

Figure 4 (a) is a plot test result of popsize population size on the system on its fitness value. The first experiment was started by taking population size = 10, Pc = 0.5, and Pm = 0.01. The next experiment is by increasing the population size but the other parameter values are fixed.. In this test, the best fitness value 62.5 which can be achieved in experiments with a population size of 80. Figure 4 (b) is a plot graph of the effect of crossover probability change. As seen in the test results shows that the best fitness value occurs when the Pc value of 0.7 with a large fitness value of 100.4. The experiments were carried out using the best population size in the previous observations and the initial Pc value. Then for the next PC is determined by the user. In this discussion the influence of Pc value on fitness value will be observed. The best population size that has been obtained previously is 80. The first experiment used was a population size of 80, Pc 0.1, and Pm 0.01 followed by experiments with other Pc values increase constantly. Figure 4 (c) is a plot graph that regulated by using the parameters of the previous observations namely the population of 80 and Pc 0.7. Furthermore, for the next Pm value the user will be determined. The first experiment used a Pm value of 0.01, a population of 80, and a Pc of 0.01. Subsequent experiments will add regular Pm values with a fixed population and Pc.

3.2 Comparative Analyzes by Resulting Accuracy

The results obtained in 10 calibration of backpropagation momentum training data testing produce an average accuracy of 76.09%. These results are obtained by randomizing parameters by looking at the possibility of the best parameters in the previous discussion. The highest results in testing using this training data obtained with an accuracy of 84.06%.

In addition using testing data, experiments were conducted using the k-fold cross validation method with k = 3, this method did the test three times in accordance with the number of folds. The test results combined to calculate the value of the confusion matrix and its predictive analysis. Details of the test three times per algorithm using the k-fold cross validation method concluded that the average of GA testing produced higher results than the momentum of backpropagation without GA. The average of GA testing obtained an average value 94.3%, while the momentum of backpropagation testing without GA obtained an average value of 88.22%.

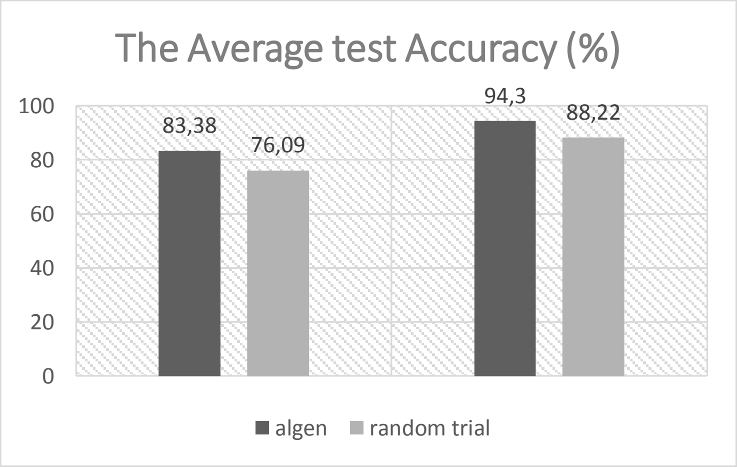

Figure 5. The average test accuracy (%)

4. CONCLUSIONS

Based on the research that has been done, it can be concluded that genetic algorithm as an alternative method of learning momentum backpropagation is able to provide cell prediction results that are closer to the actual value compared to momentum backpropagation with parameters obtained by random trial. This is evidenced by the acquisition of ten times the calibration test results with this pattern with an average memorization accuracy of 83.38% and a three-calibration test for the validation accuracy of 94.30%. These results indicate a higher accuracy compared to ANN without algen optimization with an average memorization accuracy of 76.09% and validation accuracy of 88.22%.

The scheme of combining ANN with Algen in the data used in this study can be an alternative learning that is able to produce ANN hyperparameter without random trial. Hyperparameter optimized in this study includes learning rate, momentum, and the number of hidden layers and each neuron of each hidden layer.

REFERENCES

[1] A. Lowenberg, Bob, Downing, James R., Burnett, “Acute Myeloid Leukimia,” *N. Engl. J. Med.*, vol. 341, no. 14, 1999.

[2] P. Kumar, “Automatic Detection of Acute Myeloid Leukemia from Microscopic Blood Smear Image,” *Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1803–1807, 2017.

[3] N. M. Nawi, A. Khan, and M. Z. Rehman, “A New Back-Propagation Neural Network Optimized with Cuckoo Search Algorithm.”

[4] C. Luo, Yan, Hou, Yuanbin, Liu, Gaiye, Tang and Hangying, “Transformer Fault Diagnosis Method Based on QIA Optimization Backpropagation Neural Network,” *Inf. Technol. Networing, Electron. Autom. Control Conf.*, pp. 1623–1626, 2017.

[5] R. Suhendra, C.D., Wardoyo, “Penentuan Arsitektur Jaringan Syaraf Tiruan Backpropagation (Bobot Awal dan Bias Awal) Menggunakan Algoritma Genetika,” vol. 9, no. 1, pp. 77–78, 2015.

[6] A. Harjoko *et al.*, “Classification of acute myeloid leukemia subtypes M1 , M2 and M3 using active contour without edge segmentation and momentum backpropagation artificial neural network,” *MATEC Web Conf.*, vol. 01041, pp. 1–6, 2018.

[7] N. Kdhimayoob and A. A. Hussein, “Breast Cancer Diagnosis Using Genetic Algorithm for Training Feed Forward Back Propagation,” *Int. Conf. New Trends Inf. Commun. Technol. Appl.*, no. March, pp. 7–9, 2017.

[8] R. B. Aswin, J. A. Jaleel, and S. Salim, “Hybrid Genetic Algorithm - Artificial Neural Network Classifier for Skin Cancer Detection,” *Int. Conf. Control. Instrumentation, Commun. Comput. Technol.*, pp. 1304–1309, 2014.

[9] X. Yang, *Nature-Inspired Optimization Algorithms*. Elsevier Inc, 2014.

[10] T. Prakisya, “Klasifikasi Subtipe Penyakit Acute Myeloid Leukimia M1, M2 dan M3 dengan Segmentasi Active Contour Without Edge dan Momentum Backpropagation Artifial Neural Network,” Gadjah Mada, 2017.

[11] L. Fausett, *Fundamentals of Neural Network: Architectures, Algorithm, and Application*. New Yersey: Prentice-Hall, 1994.