OPEN ACCESS
J DATA SCI APPL. VOL. 3, NO. 1, PP. 038-047, JANUARY 2020

E-ISSN 2614-7408
DOI: 10.34818/JDSA.2020.3.37

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4.5 Decision Tree

Irne Mabarti, Annisa Aditsania*

School of Computing, Telkom University
Bandung, Indonesia

*aaditsania@telkomuniversity.ac.id

Received on 23-08-2019, revised on 20-09-2019, accepted on 11-05-2020

Abstract

Cancer is one of the leading causes of death in various countries, and the mortality rate increases each year. On the other hand, bioinformatics technology will be very beneficial for cancer predictions, one of the methods that can be considered for cancer prediction is the classification of microarray data. Microarray data is a data containing many gene expressions that describe DNA cells. Microarray data has a large dimension. In this research, the dimensional reduction method used is the Minimum Redundancy Maximum Relevance (MRMR) optimized with Genetic Algorithm (GA). The reduced data is then classified with C4.5. Two trials are conducted in this research. The first trial is using the Minimum Redundancy Maximum Relevance (MRMR) method combined with Genetic Algorithm (GA) as an optimization method and the C4.5 classification method, and the trial results in an average accuracy of 79%. While the second trial is conducted using the Genetic Algorithm (GA) method for feature selection and the C4.5 classification method. It results in an average accuracy of 78%.

Keywords: Cancer, Microarray, Minimum Redundancy Maximum Relevance (MRMR), Genetic Algorithm (GA), C4.5

I. INTRODUCTION

According to data from the WHO in 2015 there were approximately 8.8 million people died from cancer [3]. In 2018, it was estimated that about 9.6 million people died of cancer. There were 2.09 million cases of lung cancer and 1.76 million of people died from lung cancer in 2018. There were 2.09 million people suffered from breast cancer and 627 thousand of them died in 2018 [15]. From these data, we can see that there is an increase in deaths caused by cancer each year. Currently, people still detect cancer in a traditional way, by looking for the symptoms of cancer in a person. Microarray technique is one of the methods that can be used to detect cancer with a considerable accuracy.

Microarray technique is one of the methods that can help us detect cancer. Inside the microarray, there are a lot of gene expressions that can describe the DNA. Microarray data has a huge dimension. Using gene expression analysis can reduce the time to identify a person suffering from cancer compared to using traditional methods. The result of analysis can be justified under the analysis rules contained in the medical expert.

A huge dimension of microarray data can affect the accuracy of classification process. Microarray data utilization can be maximized by looking only at the influential data to determine whether a person is suffering from cancer.

In 2018 there was a study on the classification of microarray data using C4.5 Decision Tree and BPSO classification methods, with the result of 99% accuracy [16]. Therefore, the method used in this research is the Minimum Redundancy Maximum Relevance and C4.5 Decision Tree.

The classification process to detect cancer microarray data is done in 3 stages: preprocessing, dimensional reduction, and gene classification. Selection of features in the dimensional reduction utilizes a Minimum Redundancy Maximum Relevance (MRMR) optimized with Genetic Algorithm. While C4.5 Decision Tree method is used for gene classification.

The MRMR aims to reduce the number of features while still generating a good accuracy [1]. Feature selection is needed due to a very large number of features in microarrays data. Based on [17], the decision tree classification method has more accuracy than *Naïve Bayes* method, thus the decision tree can be combined with the MRMR method. Moreover, C4.5 decision tree shows promising improvement when incorporated with an optimization method [16]. So, in this study we try to use the Genetic Algorithm as the optimization method.

II. RELATED WORKS

Cancer classification using microarray data has been studied in recent years especially by bioinformatics researchers. Microarray is a DNA gene data that is used as a parameter in analyzing genes expression to map out the types of genes that causes cancer [5]. There are DNA or RNA in microarrays. It is located on a chip in a gene [8]. Microarray data has many gene expression patterns that can be classified to detect cancer [4]. This requires a level of expression of RNA obtained from different networks of microarrays, determining the level of genes that can be used as a variable expression classifier, applying rules for classification’s design of sample data, and applying failure estimation procedure [4].

Microarray is commonly used in healthcare to detect diseases, especially cancer, where abnormalities occur in the affected cells. Microarray advantages may allow researchers to analyze genes in large quantities at a time [20]. Microarray data of cancer has a fairly large dimensions, researches using the appropriate methods will therefore contribute greatly to the classification process. Here are a few studies that have been done on the cancer’s microarray data.

Analysis of gene expression is more effective to help medical experts in detecting cancer, rather than using traditional methods to see a symptom or signs [1]. In 2017, Maulana Tresna Fahrudin, Iwan Sharif, Ali Ridho Barakbah [17] did a research on Data Mining Approach for Breast Cancer Patient Recovery. The study compared two methods, namely, Naïve Bayes and Decision Tree. The comparison between the two methods can be seen from their accuracy, where Naïve Bayes method has an accuracy of 92.76%, and the accuracy of Decision Tree method is 92.99%. The accuracy of the Decision Tree method is therefore higher than that of the Naïve Bayes method.

In 2018, Firda A. Ma'Ruf, et al. [2] conducted research on Analysis of Effect of Dimensional Reduction Method of Minimum Redundancy Maximum Relevance in Microarray Data Based Cancer Classification Using Support Vector Machine classifier. The study explained the use of MRMR in dimension reductions of feature selection and the use of SVM methods in its classification with linear kernel function and plynomia kernel, with the number of features used in the study amounts to 10% of the original number. This means that the accuracy of the classification is 100% and the performance of the built system is very good.

Amalya Citra Pradana, Adiwijaya, Annisa Aditsania in [16] conducted research on the Implementation of Binary Particle Swarm Optimization Algorithm (BPSO) and C4.5 Decision Tree for Cancer Detection Based Microarray Data Classification. An accuracy of 99% was obtained from the study using BPSO-C4.5. In [23], the Principal Component Analysis (PCA) reduce data dimension by forming a new subset of features so that the feature dimensions become less, therefore PCA is classified into the reduction of the extraction dimension the features.

III. PROPOSED METHOD

A. *Microarray data*

Microarray data has many gene expression patterns that can be classified to detect cancer [4]. This requires a level of expression of RNA obtained from different networks of microarrays, determining the level of genes that can be used as a variable expression classifier, applying rules for classification's design of sample data, and applying failure estimation procedure [4].

Microarray is commonly used in healthcare to detect diseases, especially cancer, where abnormalities occur in the affected cells. Microarray's advantages may allow researchers to analyze genes in large quantities at a time [20].

Microarray data of cancer has a fairly large dimensions, researches using the appropriate methods will therefore contribute greatly to the classification process. Here are a few studies that have been done on the cancer's microarray data:

B. *Normalization*

Normalization data is one of the essential process prior to data classification process. Cancer data includes data of Breast Cancer, Lung Cancer, Leukemia, Ovarian Cancer and Colon Tumors that have continuous value of data. Normalization is done on the data to change the value of data information to the 0-1 interval. The formula used in the normalization process is as follows:

$$Normalized\ Data = \frac{data - min(data)}{max(data) - min(data)} \qquad (1)$$

C. *K-Fold*

K-Fold is one of the sharing data methods between data train and data test. K-Fold method randomly divides the dataset towards K partition. In this research, the values of K used are 3 and 5. Figure 1. is an example overviewing how the K-fold works with the value of K = 3, thus the division of datasets using the K-Fold can be pictured as follows:

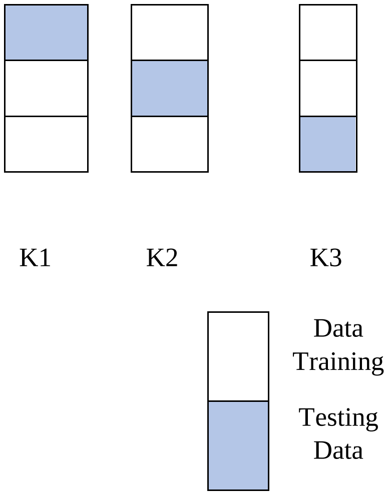

Fig. 1. Division of Dataset Using K-Fold Method with the Partition Value of K = 3

Figure 1. Visually shows how the K-Fold works with the partition value of K = 3, the dataset will then be divided into three datasets. Each dataset group of one data will be used as a data test and the rest will be used as data train.

D. Minimum Redundancy Maximum Relevance (MRMR)

Minimum Redundancy Maximum Relevance (MRMR) is used to evaluate the pre-feature with discriminatory information [14]. MRMR method tries to select a subset of features, each of which has a maximum relevance to the target class and has minimal redundancy with other features [14].

Relevancy of jth feature to the class(c) calculated using F-statistic showed in Eq.2

$$F\left(g_j, c\right) = \frac{\sum_{k=1}^{K} n_k\left(\bar{g}_{j,(k)} - \bar{g}_j\right) / (K-1)}{\sum_{k=1}^{K} \sum_{l=1}^{n_k}\left(g_{j,l(k)} - \bar{g}_{j,(k)}\right)^2 / (N-K)} \tag{2}$$

where

c = classification variable

n_k = number of observations belonging to the kth class

h = class for the attributes

$\bar{g}_{j,(k)}$ = average value of $\bar{g}_j$ in all samples belonging to the kth

$g_{j,l,(k)}$ = gene expression value of lth sample belonging to the kth class

Meanwhile, redundancy between feature calculated with Pearson Correlation showed in Eq.3

$$c(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1) S_x S_y} \tag{3}$$

where :

S_x, S_y : standard deviation of x, y

The data used in this research is continuous data. So, the objective function is FCQ (F-Test Correlation Quotient). Eq.4 shows FCQ equation

$$max\left\{\frac{\sum \quad F(i,h)}{\left[\frac{1}{|S|}\sum_{j\in S}|c(i,j)|\right]}\right\} \tag{4}$$

Data processing on the GA has several steps. First is the initiation - selecting individuals randomly. Second, the fitness value of individuals who have been initiated will be calculated using MRMR. Third is parent selection. On this study, the parent selection is done using a random method. The next step is crossover. On this step, the number of children that would be generated by every parent is determined. Next is mutation. This step acts to replace the missing genes from the population due to the process selection. All these processes will be done repeatedly until n iteration(generation). If stopping criteria have been fulfilled, fitness from individual in the last generation will be sorted from the largest too the smallest. Individual with highest fitness is the optimal solution obtained from Genetic Algorithm.

E. Genetic Algorithm (GA)

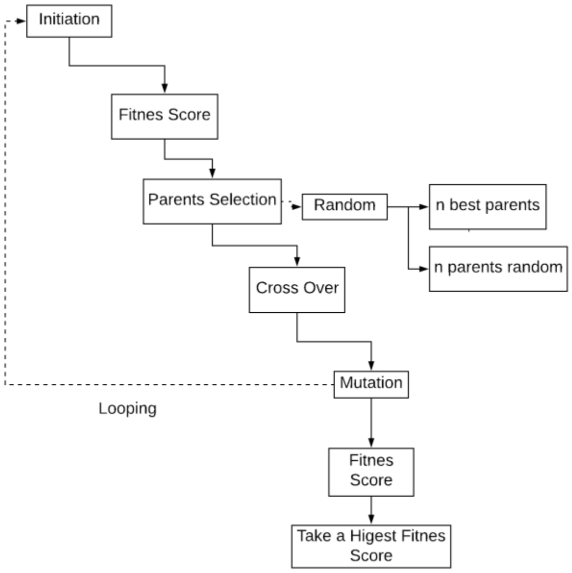

Fig. 2. Schematic Overview of methods genetic Algorithm used in research

F. C4.5 Decision Tree

The entropy of C4.5 method is used to measure the heterogeneity of sample data. Entropy is calculated using Eq.5.

$$Entropy\ (S) = -\sum_{i}^{c} p_i log\ p_i \qquad (5)$$

where:

c = The value of the target attribute (class number)

p_i = The proportion of samples in class i

Information Gain values can be calculated after the entropy value is obtained. Information Gain can be represented by a formula as follows:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v} \frac{|S_v|}{S} entropy(S_v) \quad (6)$$

where:

V = number probable values for attributes A

$|Sv|$ = Sample value for value V

S = Sample space (data) used in training

$Entropy(S)$ = is the entropy for samples with the value of V

C4.5 is the successor of ID3 method, where the Gain Ratio is used to improve the Information Gain formula:

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)} \qquad (7)$$

where :

S = Sample space (data) used in training

A = An attribute

$Gain(S,A)$ = Information gain for attributes A

$SplitInfo(S,A)$ = Split Information for attributes A

An attribute with the highest Gain Ratio value will be chosen as the test attribute in a node. With a Gain that uses value of Information Gain, this approach uses the normalization of Information Gain with Split Information. SplitInfo represents the entropy or information that has potential, with the formula:

$$SplitInfo\ (S,A) = -\sum_{i=1}^{k} \frac{S_i}{S} log_2 \frac{S_i}{S} \qquad (8)$$

where :

S = Sample space (data) used in training

A = An attribute

S_i = The number of sample of attribute-i

G. Accuracy

The performance of the system is needed to determine the assessment value of the system, as a benchmark to see whether the classification that has been done is correct. In this system, the performance is calculated using accuracy that uses True Positive, True Negative, False Positive and False Negative.

Score, which can also be referred to as accuracy is the assessment value of a system when doing classification. True Positive (TP) is the system's success value if the system gave a positive label and original label is positive. True Negative (TN) is the system's success value if the system gave a negative label and original label is negative. False Positive (FP) is the system's success value if the system gave a positive label and the original label is negative. False Negative (FN) is the system's success value if the system gave a negative label and original labeling is positive.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

IV. RESEARCH METHOD

A. *Datasets*

There are 5 data used in this research, namely Leukemia Cancer, Breast Cancer, Colon Tumors, Lung Cancer and Ovarian Cancer. Table 1 shows Microarray Data used in this research [21]

Table 1. Data Research Specifications

Name Data	Amount of data	Dimension Data	Positive Class	Negative Class
Leukemia	72	7129	AML	ALL
Breast Cancer	97	24 481	Relapse	Non-relapse
Colon tumors	62	2000	Positive	Negative
Lung Cancer	181	12533	Mesothelioma	ADCA
Ovarian Cancer	253	15154	Cancer	Normal

B. *MRMR GA and C4.5 Scenario System*

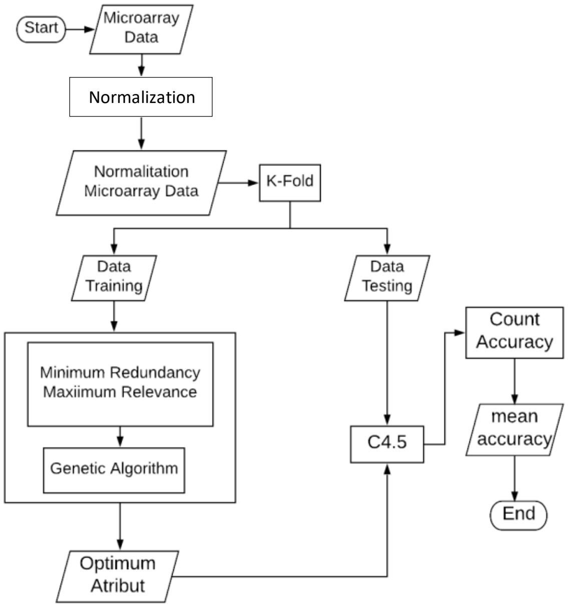

Fig. 3. Research System's Schematic Design of Implementation of Cancer's Microarray Data Classification Using MRMR GA and C4.5 Decision Tree

C. GA and C4.5 Scenario System

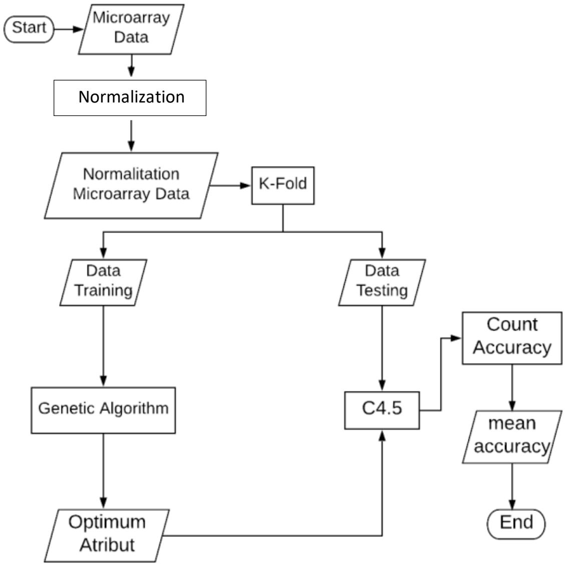

Fig. 4. Research System's Schematic Design of Implementation of Cancer's Microarray Data Classification Using MRMR GA and C4.5 Decision Tree

V. RESULTS AND DISCUSSION

Two processes of dimension reduction is used in this research. Fig.5 shows accuracy comparison between MRMR optimized by GA and GA.

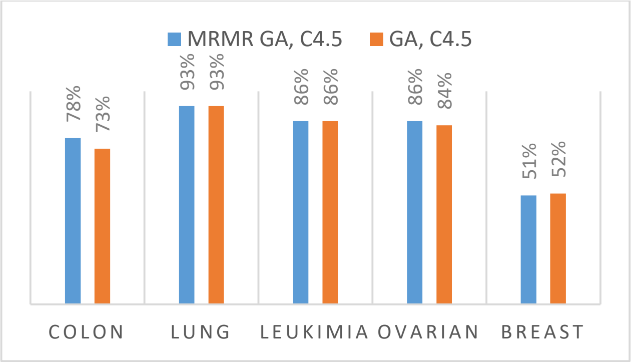

Fig. 5. Accuracy Comparison

We compare the accuracy of the study's results between MRMR GA and C4.5 method with GA and C4.5 method. In Figure 5, Colon dataset has a mean accuracy of 79% for MRMR GA and C4.5, and 78% accuracy for GA and C4.5. The highest accuracy is on Lung Dataset and the lowest accuracy is on Breast Cancer Dataset.

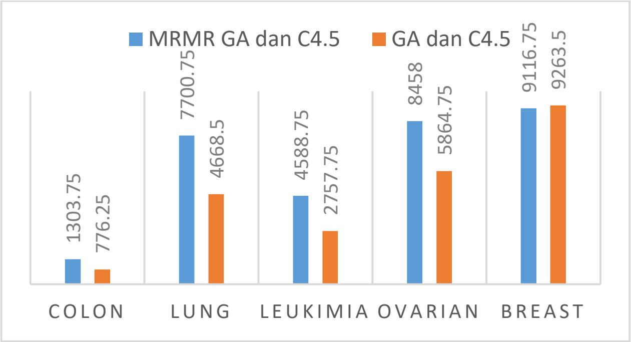

Fig. 6. Dimension Comparison Bar Diagram

In Figure 6, dimensions of cancer microarray data processing using GA and C4.5 is smaller compared to using MRMR GA and C4.5. But Breast Cancer has a larger dimensions. This happens because the GA method has parent selection step that makes Breast Cancer dataset significantly larger.

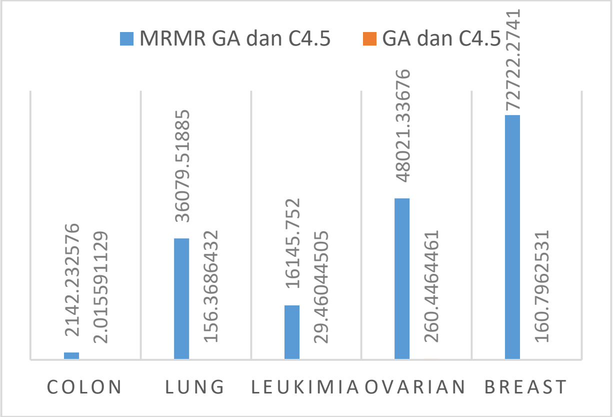

Fig. 7. Execution Time Comparison Bar Diagram

In Figure 7, Execution Time Comparison Bar Diagram. The execution time using MRMR GA and C4.5 is 9.37 hours, and the execution time using GA and C4.5 is 0.0338 hours. MRMR GA and C4.5 method have a significantly longer execution time compared to GA and C4.5.

VI. CONCLUSION

It can be concluded that the MRMR GA and C4.5 method can be implemented in the cancer microarray data. The results of the MRMR GA and C4.5 method on cancer data can be categorized as good. A comparison between MRMR, GA and GA C4.5 and C4.5 methods shows accuracy differences. Higher accuracy is produced by the MRMR GA and C4.5 method.

Utilization of MRMR GA and C4.5 method is more appropriate for the Lung Cancer data because it produces a high accuracy. While the data that has the lowest accuracy namely Breast Cancer have information that does not support data processing using MRMR GA and C4.5 method.

If the microarray data is processed only by using the C4.5 classification method, then the accuracy of the result is smaller compared to using MRMR GA and C4.5 or C4.5 method and the GA. Thus, the dimension reduction method is very influential on the accuracy of a system built for cancer microarray data classification.

The more methods used in building a system, will affect the execution time. The system's execution time using MRMR GA and C4.5 is longer than the system built using GA method and C4.5. So that further research can be done using schemes that require less computing time. In the future, it can also be done by using other dimension reduction, optimization and classification methods.

REFERENCES

[1] Adiwijaya., U.N. Wisesty., E. Lisnawati, A., A. Aditsania., & D.S. Kusumo, (2018). "Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification". Journal of Computer Science 14(11).

[2] Ma'Ruf, FA, Adiwijaya, A., & Wisesty, UN 2018. Influence Analysis of Dimensional Reduction Method of Minimum Redundancy Maximum Relevance In Microarray Data Based Cancer Classification Using Support Vector Machine classifier. Journal of Physic: Conference Series, Volume 1192, Conference 1. [1].

[3] World Health Organization, the American Cancer Society. 2015. Cancer facts & Tableures 2015. [accesed 12 October 2018]

[4] Braga-Neto, Ulisses & R Dougherty, Edward. 2004. Is cross-validation is valid for small-sample microarray classification ?. Bioinformatics. 20. 374-380

[5] Liu, X., Krishnan, A., & Mondry, A. 2005. An entropy-based gene selection method for cancer classification using microarray data. BMC bioinformatics, 6 [1], 76.

[6] Nurviarelda, R., Adiwijaya, A., & Rohmawati, AA 2018. Microarray Data Classification Using Discrete Wavelet Transform And Naive Bayes Classification. eProceedings of Engineering, 5 [1].

[7] Ramadhani, PT, Wisesty, UN, and Aditsania, A. 2017. Cancer Detection Microarray Data Classification is based on using the Functional Link Neural Network with Genetic Algorithm Selection feature. Indonesian Journal on Computing (Indo-JC), 2 [2], 11-22.

[8] Singh, RK, & Sivabalakrishnan, M. 2015. Feature selection of gene expression classification of data for cancer: a review. Procedia Computer Science, 50, 52-57.

[9] Kresno, SB 2011. Micro-RNA and its Implications on Cancer. Indonesian Journal of Cancer, 5 [3].

[10] Fodor, IK 2002. A survey of dimension reduction techniques (No. UCRL-ID-148 494). Lawrence Livermore National Lab., CA (US).

[11] Li, L., & Li, H., 2004. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 20 [18], 3406-3412.

[12] Saeys, Y., Inza, I., & Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. bioinformatics, 23 [19], 2507-2517.

[13] Materka, A., & Strzelecki, M. 1998. Texture analysis methods-a review. Technical university of lodz, institute of electronics, COST B11 report, Brussels, 9-11.

[14] Li, Z., Zhou, X. Dai, Z., & Zou, X. 2010. Classification of G-protein coupled receptors based on support vector machine with minimum redundancy maximum relevance and genetic algorithms. BMC bioinformatics, 11 [1], 325.

[15] World Health Organization. [on line]http://www.who.int/en/news-room/fact-sheets/detail/cancer/ [Accesed 12 October 2018]

[16] Adiwijaya. (2018). "Deteksi Kanker Berdasarkan Klasifikasi Microarray Data". Media Informatika Budidarma.

[17] Fahrudin, TM, Sharif, I., & Barakbah, AR 2017. Data Mining Approach for Breast Cancer Patient Recovery. EMITTER International Journal of Engineering Technology, 5 [1], 36-71.

[18] Budhi, RK 2008. Application of Genetic Algorithm for Scheduling Optimization Event Class. Transformatika Journal, 6 [1], 1-9.

[19] Haupt, RL, & Ellen Haupt, S. 2004. Practical genetic algorithms.

[20] Singh, RK, & Sivabalakrishnan, M. 2015. Feature selection of gene expression classification of data for cancer: a review. Procedia Computer Science, 50, 52-57.

[21] Manik, A., Adiwijaya, A., & Utama, D. Q. (2019). "Classification of Electrocardiogram Signals using Principal Component Analysis and Levenberg Marquardt Backpropagation for Detection Ventricular Tachyarrhythmia". Journal of Data Science and Its Applications, 2(1), 78-87.

[22] Purbolaksono, MD, Widiastuti, KC, Mubarok, MS, & Ma'Ruf, FA 2018, March. Implementation of mutual information and Bayes theorem for classification of microarray data. In the Journal of Physics: Conference Series [Vol. 971, No. 1, p. 012 011]. IOP Publishing.

[23] Astuti, Astuti., Adiwijaya. (2019). "Principal Component Analysisi Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis". Jurnal Media Informatika Budidarma, Vol 3, No 2.