JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

# Discrete Wavelet Transform (DWT) and Random Forest for Cancer Detection Based on Microarray Data Classification

Monica Triyani[1*], Adiwijaya[2], Annisa Aditsania[3]
[1,2,3] School of Computing, Telkom University
[1,2,3]Telekomunikasi Street, Terusan Buah Batu, Bandung 40257, West Java, Indonesia
*Corresponding email: monicatriyani29@gmail.com

Received 27 May 2020, Revised 16 June 2020, Accepted 21 June 2020

Abstract — Cancer is one of the leading causes of death worldwide. According to the World Health Organization (WHO), in 2018, about 9.6 million deaths caused by cancer. DNA microarray technology has played an important role in analyzing and diagnosing cancer. The accuracy resulting from the classification of Random Forests is not optimal because microarrays have large dimensional data. Therefore, it is necessary to reduce the dimensions of the Discrete Wavelet Transform (DWT) as a feature to reduce dimensions and increase accuracy in microarray data. Based on the simulation, the dimension can be reduced and improve the accuracy of classification up to 8% - 20%. DWT approximation coefficient can improve accuracy better than detailed coefficients for data on colon cancer 100%, lung cancer 100%, ovarian 100%, prostate tumor 80%, and central nervous system 83.33%.

Keywords – cancer, microarray, dimension reduction, Discrete Wavelet Transform (DWT) and Random Forest



## I. INTRODUCTION

Cancer is a leading cause of death worldwide. Cancer is caused by uncontrolled growth and spread of abnormal cells that can attack any part of the body. Usually, cancer arises from the transformation of normal cells into tumor cells that develop into malignant tumors [1]. According to the World Health Organization (WHO), in 2018, about 9.6 million deaths caused by cancer. There were 2.09 million cases of breast cancer, and 627 thousand people died due to breast cancer in 2018. An estimated 2.09 million cases of people had lung cancer and 1.76 million deaths due to breast cancer in 2018[2].

In recent years, DNA microarray technology has played an important role in analyzing and diagnosing cancer [3][21]. Before knowing the DNA microarray technology, cancer detection still uses the traditional way to look at the symptoms of cancer disease. DNA microarray technology developed by Patrick O. Brown, Joseph DeRisi, and David Botstein allows researchers to collect large amounts of gene expression at simultaneously and be able to analyze changes in gene expression patterns under certain conditions [4]. Gene expression is used to determine the type of cancer cells, and the level of gene expression in the human body can be measured through DNA microarrays experiments [18]. Analysis of gene expression can convince medical experts whether a patient has cancer or not compared to the traditional way.

Microarray with large data dimensions results in not optimal accuracy of the classification process [5]. This problem affects system performance and computing time. Therefore, it is necessary to reduce the microarray data 's dimensions to increase the accuracy value and avoid overfitting the classification.

In the previous research, [8] with using dimension reduction and classification Random Forest for ten datasets in three conditions got an accuracy of 94.18% for Leukemia on the condition (1), 96.20% for lymphoma in the condition (2) and 83.71% for Adenocarcinoma of the condition (3). In 2018[6], there was a research on the classification of microarray data using the Discrete Wavelet Transform (DWT) and Naïve Bayes, with an accuracy of 98.4126% for

ovarian, 78.95% for colon and 83.33% for lung. So, by using dimension reduction in Discrete Transform (DWT), the accuracy obtained is better than without dimension reduction [22][23]

In 2018, Adiwijaya et al. [3] conducted research on Dimension Reduction Using Principal Component Analysis for Cancer Detection based on Microarray Data Classification. In this research, it is explained that PCA is used as dimension reduction, and two classification methods (SVM and LMBP) are used as a comparison. The comparison can be seen from the results of accuracy, using PCA and SVM methods produces an accuracy of 94.98% while using PCA and LMBP achieves an accuracy of 96.07%. The accuracy of the LMPB method is better than the SVM method because the LMBP method can generalize new data using the model obtained in the testing process better than SVM on microarray data.

Aydadenta and Adiwijaya [9] conducted a classification using the Random Forest algorithm with clustering combined with the relief method feature selection. Accuracy results obtained are, 85.87% for colon cancer, 98.9% for lung cancer, and 88.97% for prostate tumor. Damayana [17] classified skin cancer using the K-nearest Neighbor (KNN) method, and the extraction of DWT features as dimension reduction. The process of this research consists of image input, preprocessing, DWT feature extraction, and KNN classification process. The accuracy obtained is 76%.

Unlike previous research [8][20], this research will use DWT dimension reduction as feature extraction. and classification Random Forest to improve the accuracy of the other microarray data.

## II. RESEARCH METHOD

### A. *Microarray Data*

There are five microarray data used in this research, namely Colon Cancer, Lung Cancer, Ovarian, Central Nervous System, and Prostate Tumor. The data obtained from Kent Ridge Biomedical Data Set Repository (http://leo.ugr.es/elvira/DBCRepository/).

Table 1. Microarray Data

| Dataset | Number of Records | Number of Features | Class | Sample |
|---|---|---|---|---|
| Colon Cancer | 62 | 2000 | 2 | 62 (22 positives, 40 negative) |
| Lung Cancer | 181 | 12533 | 2 | 181 (31 Mesothelioma, 150 ADCA) |
| Ovarian | 253 | 15154 | 2 | 253 (91 normal, 162 cancer) |
| Central Nervous System | 60 | 7129 | 2 | 60 (21 Class1, 39 Class0) |
| Prostate Tumor | 136 | 12600 | 2 | 136 (77 Tumor, 59 Normal) |

Table 1 shows the specification of the microarray data used in this research. For each cancer data, the number of records, features, and sample are different.

### B. *General Scheme*

This research aims to detect or diagnose cancer based on microarray data. Microarray Data has large dimensions so that the resulting accuracy is not optimal. Therefore, dimension reduction and classification processes are needed. The scheme of the system can be seen in Fig. 1.

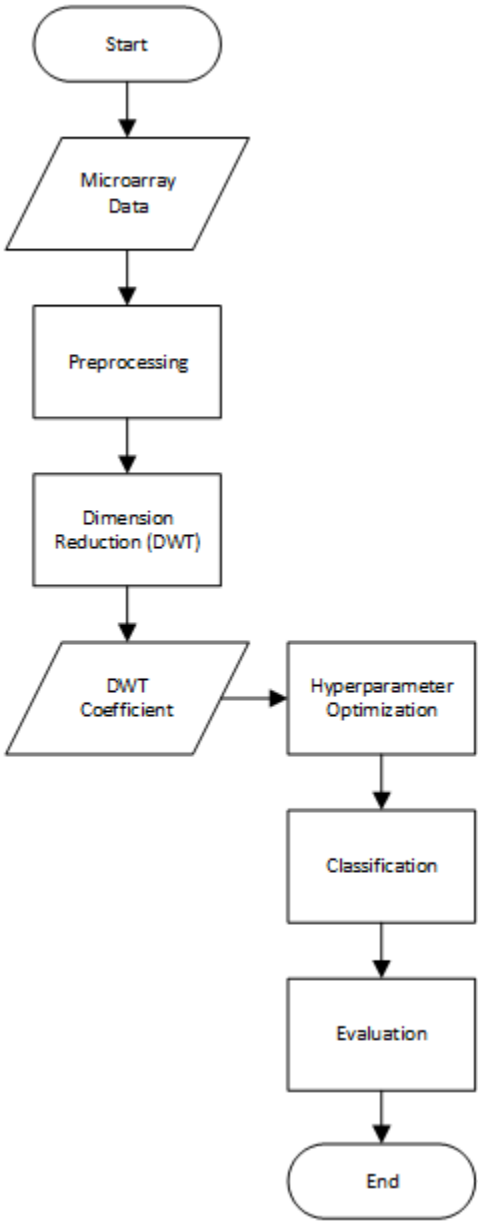


Fig. 1. General Scheme

a) Preprocessing

Preprocessing in this research consists of two processes. The first is to split the data into training data and data testing. The second process is normalization. The normalization process changes data values into intervals 0 to 1 using the MinMax Scaler algorithm. Normalization is used so that the range of values among the data is not too much different. Normalization is calculated using Equation (1) [18]:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where:

$X'$ The new value of the feature in the normalization domain

$X$ Value of feature before normalization process

$X_{min}$ The lowest value of the feature in normalization data

$X_{max}$ The highest value of the feature in normalization data

b) Dimension Reduction

Microarray Data has an enormous dimension and the complexity of data because it contains more features than samples. According to [11], this will cause complex problems in the classification process, often called the curse of dimensionality. A process that is often used is the dimension reduction process [7] to solve the problem. Dimension reduction is used to reduce complexity in the data microarray. Dimension reduction is of two types namely feature selection and feature extraction. Feature selection is to choose some features that are considered essential to speed up data processing by reducing dimensions and avoiding overfitting of the classifier.

In contrast, feature extraction is projects data into features that are few but still reflect the original data [5]. The dimension reduction process used in this research is the Discrete Wavelet Transform (DWT) method as feature extraction. The dimension reduction process uses the Discrete Wavelet Transform (DWT).

Discrete Wavelet Transform (DWT) is a feature extraction method that processes the signals for generating genes to be treated [14]. In this research, the microarray feature plays the signal input in the dimension reduction process in DWT. DWT used is Daubechies Level 4. DWT decomposition performs a signal division process into two parts, namely highpass and lowpass filters. The highpass filter is used to analyze the portion of the signal that has a high frequency (large scale) while the lowpass filter used to analyze signals that have a low frequency (small scale) [12]. The feature will be convoluted using a highpass filter, then downsampling to produce a detailed coefficient (cD1). Features that are convoluted using lowpass filters will also be downsampling to produce an approximation coefficient (cA1) that describes the signal's identity. Downsampling causes the length of the coefficient to be about half the initial features-the output of DWT coefficients in the form of approximation and detail. The decomposition process can be calculated using Equation (2) and (3) [13].

$$c_{j+1}(k) = \sum_m h(m - 2k)\, c_j(m) \quad (2)$$

$$d_{j+1}(k) = \sum_m h_1(m - 2k)\, c_j(m) \quad (3)$$

where :

$c_j$ = coefficients approximation

$d_j$ = coefficients detail

$m$ = initial features

$h(m - 2k)$ = lowpass filter

$h_1(m - 2k)$ = highpass filter

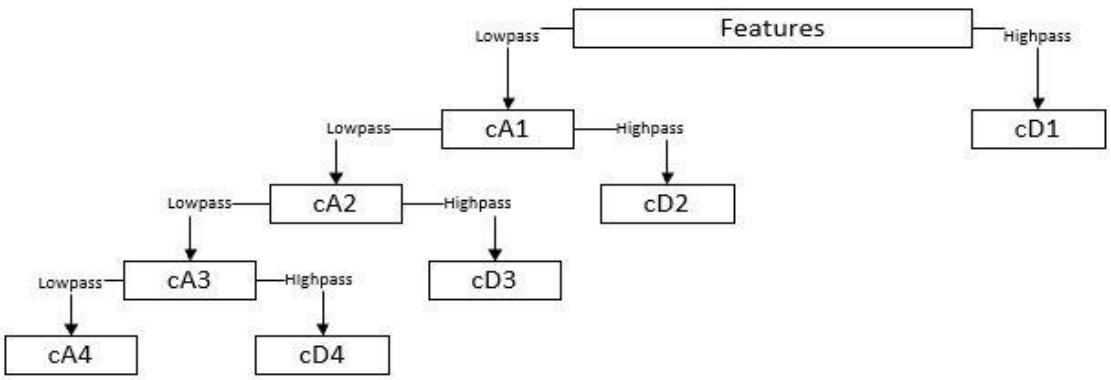


Fig. 2. Decomposition using DWT level 4

c) Hyperparameter Optimization

Hyperparameter optimization is used to find the best solution for the model being built. There are two parameter optimization methods used in this research, namely Random Search and Grid Search. Random search is a technique with a sampling search on parameter combinations, while Grid Search is a Brute Force algorithm that searches by trying every possible parameter combination [15]. Random Search has a slightly faster computing time than Grid Search but does not guarantee to get the best results. Therefore, in this research, both methods are to get the best optimization parameters.

The parameters used to build the Random Forest classification model are n_estimator, max_features, min_samples_split, and min_samples_leaf in Table 2. In this research, Random Search will search randomly to find the best optimization parameters from the combination of existing parameter values. Experiments carried out in Random Search are as many as 50 iterations. The best optimization parameter values obtained from Random Search will be reused in the Grid Search method to further improve the best parameters by passing an experiment of 5 iterations. However, the parameter n_estimator will be searched again for more details on the Grid Search method.

d) Classification

Microarray data classification is a bioinformatics science that has been widely studied and used to analyze cancer [4]. Grouping data can classify cancer detection that has been determined [7].

Table 1. Parameter Optimization

| Parameter | Value | Information |
|---|---|---|
| n_estimator | 200, 400, 600, 800, 1000 | Number of trees in the random forest |
| max_features | auto / log2 | Maximum number of features |
| min_samples_split | 2, 5, 10 | The minimum required a number of samples to split an internal node. |
| min_samples_leaf | 1, 2, 4 | The minimum required number of samples to be at a leaf node. |
| bootstrap | True / False | Method for sampling data points |

The classification process is done after the dimension reduction process. At this stage, the input data in the form of reduced data dimensions will be processed to diagnose whether a person has cancer or not. The classification process in this research uses the Random Forest method.

Random Forest algorithm that uses the ensemble decision tree method is a classification method that produces more than one model so that Random Forest consists of more than one tree [10]. Each decision tree is constructed using a random vector. The Random vector used in the tree building process is to select the arbitrary value 'X' as many as the input attribute X, which will be shared at each node in the decision tree formed. Parameters to set the power of random forest algorithms lies in the selection of X values and the number of trees to be formed [9][16].

Random forest is a combination of several decision trees. The decision tree used in this study is the Gini Index. Gini Index is one of the methods used to determine the best breaking point. The general formula of the Gini Index can be seen in equation (4). $D$ is the node, and $P_i$ is the probability $D$ is in class $C_i$, and m is the class. If the data is broken down to attribute A into two subsets $D1$ and $D2$, then the Gini index equation is as follows (5)[19].

$$Gini\ (D) = 1 - \sum_{i=1}^{m} P_i{}^2 \quad (4)$$

$$Gini_A(D) = \frac{|D1|}{|D|} Gini(D1) + \frac{|D2|}{|D|} Gini\,(D2) \quad (5)$$

e) Evaluation

Evaluation of a system is needed to determine whether the classification is done correctly or not. In this research, confusion matrix that records the value of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), is used [24][25]. Positive data is categorized as cancer, while negative data is data that is classified as non-cancerous. Table 3 is a confusion matrix.

Table 2. Confusion Matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| **Actual Positive** | True Positive (TP) | False Negative (FN) |
| **Actual Negative** | False Positive (FP) | True Negative (TN) |

True Positive (TP) was the amount of data when data showed positive predictive cancer, and the actual data showed positive for cancer. True Negative (TN) was the amount of data when data showed negative predictive cancer (non-

cancerous), and the actual data showed negative cancer (non-cancerous). False Positive (FP) was the amount of data when the data showed positive predictive cancer, and the actual data showed negative cancer (non-cancerous). False Negative (FN) was the amount of data when data showed negative predictive cancer (non-cancerous), and the actual data showed positive for cancer. Furthermore, the calculated values of accuracy, precision, and recall using Equation (6) (7) and (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

## III. RESULT

The testing is carried out on five cancer datasets; Colon Cancer, Lung Cancer, Ovarian, Prostate Tumor, and Central Nervous System. There are three scenarios in this research, the classification without dimension reduction, classification with DWT coefficients approximation, and classification with DWT coefficients detail. Each scenario performs experiments ± 10 attempts to obtain the best result. After that, an evaluation is carried out. Here are the test results of the three scenarios.

In Table 4, there are three scenarios carried out at five cancer datasets. The best results are obtained in the lung for classification without reduction, and ovarian data is 100%. While for the classification with DWT coefficients approximation, the best results are obtained in colon, lung, and ovarian data with accuracy up to 100% and for the classification with DWT coefficients detail the best result obtained in the lung data with an accuracy of 100%. Lung and ovarian data obtain constant accuracy in the classification without reduction or classification with the DWT approximation coefficient of 100%. While prostate data obtains constant accuracy in all three classification scenarios is 85.71%. Of the three classification scenarios, the best result for five cancer datasets obtained in the classification with DWT coefficients approximation with accuracy results in 100% for colon cancer, lung cancer 100%, ovarian 100%, prostate tumor 85.71%, and central nervous system 83.33%. So, by using dimension reduction in Discrete Wavelet Transform (DWT), the accuracy obtained is better than without dimension reduction. A smaller number of features can produce better accuracy for some data such as colon cancer and the central nervous system. Some other data, such as lung cancer, ovarian, and prostate tumors, have constant accuracy. However, in the classification with DWT coefficient detail for ovarian data has a slight decrease in accuracy after reduced dimensions. Classification with DWT coefficient approximation is more capable of producing better accuracy because the generated features are the best.

The graph in Fig. 3 shows the precision results for each cancer data from three test scenarios. Precision is the ratio of a person's positive predictive cancer to the overall positive, predictable outcome. Colon cancer data get a precision of 100% on the classification with DWT coefficients approximation, which means the ratio of people who are predicted to be positively affected by colon cancer from the entire test data is significant. In comparison, for lung cancer and ovarian data, obtain 100% precision in all three scenarios. Prostate tumor data also obtains the same precision in all three scenarios by 94%. While the central nervous system data obtain a precision of 33.33% in the classification with the DWT coefficients approximation, which means the ratio of someone positively affected by the central nervous system from the overall positive central nervous system test data is very small. For classification without dimension reduction and classification with DWT coefficients, detail produce 0% precision in the central nervous system data which means the person's ratio is positively affected by the central nervous system from the overall prediction results of the positive central nervous system is absent.

The graph in Fig. 4 presents the results of recall or sensitivity to each cancer data from three testing scenarios. The recall is the ratio of a person's prediction of positive cancer compared to the positive overall data that is . Colon and lung cancer data obtained 100% recall in all three classification scenarios, which means the ratio of positive people predicting colon and lung cancer from all the data is significant. In comparison, ovarian data received 100% recall in the classification without reduction and classification with approximation DWT coefficient. Prostate tumor data obtained a recall of 85% in all three classification scenarios, which means the ratio of predictive positive people affected by prostate tumors from the overall data is only 85%. While the central nervous system data only obtained a recall in the classification with a DWT coefficient approximation of 100%.

Based on the research conducted, if the value of precision is small and large recall is caused by the value of TP (True Positive), and FP (False Positive) value is large, whereas if the value of recall is small and the precision is large due to the value of TP (True Positive) and the value of FN (False Negative) large. The precision and recall values will be the same if there are no classifications. That is shown from ± 10 experiments with the same precision and recall, all of which have 100% precision and recall and also 100% accuracy.

Table 3. Test Results of Three Classification Scenarios

| Dataset | Number of Initial Features | Number of Reduction Features | Accuracy(%) | | |
|---|---|---|---|---|---|
| | | | Without Reduction | Approximation Coefficient | Detail Coefficient |
| Colon Cancer | 2000 | 1003 | 84% | 100% | 92% |
| Lung Cancer | 12533 | 6270 | 100% | 100% | 100% |
| Ovarian | 15154 | 7580 | 100% | 100% | 98% |
| Prostate Tumor | 12600 | 6303 | 85.71% | 85.71% | 85.71% |
| Central Nervous System | 7129 | 3569 | 75% | 83.33% | 75% |

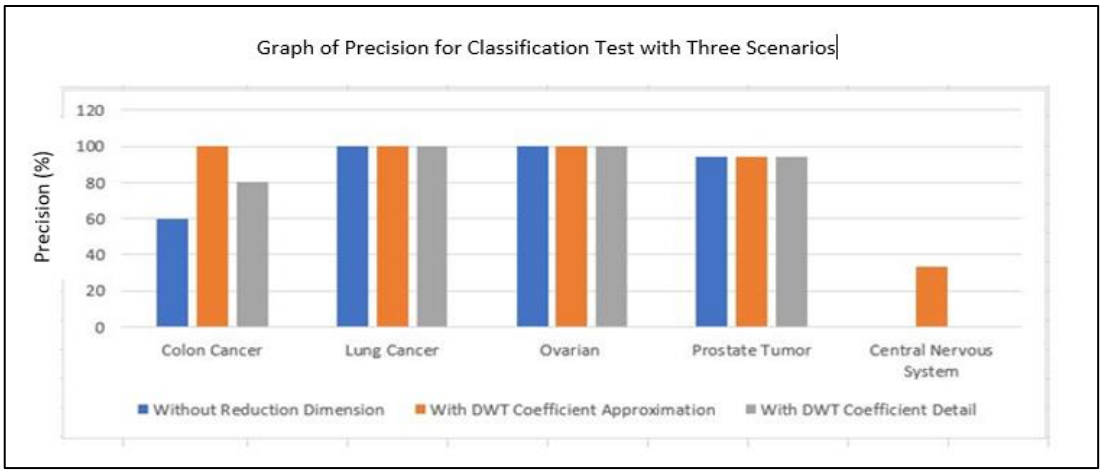


Fig. 3. Precision for Classification Test with Three Scenarios

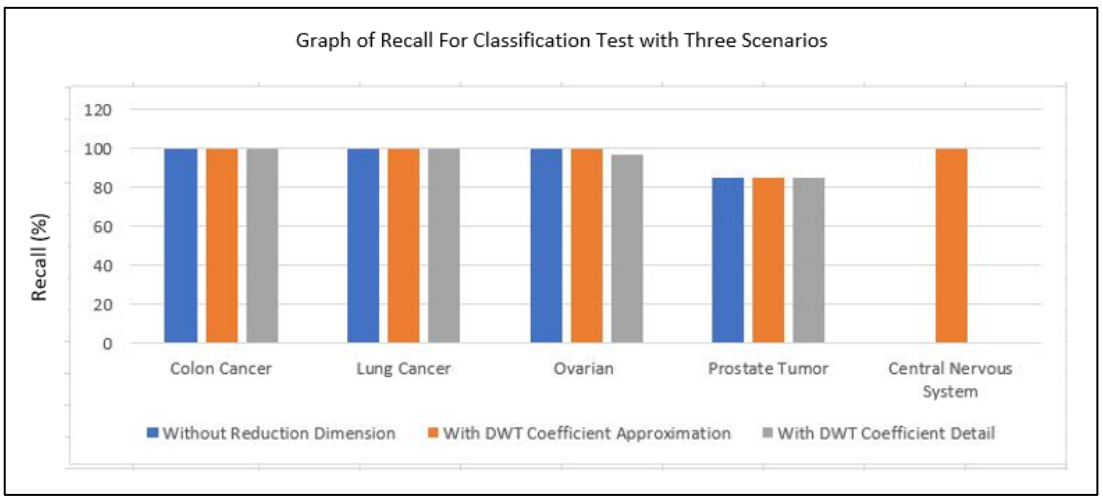


Fig. 4. Recall for Classification Test with Three Scenarios

## IV. DISCUSSION

From the result, we gather some insights from this research. From the three classification scenarios carried out, classification without reduction, classification with DWT coefficient approximation, and classification with DWT coefficient detail, the best accuracy performance is obtained by the classification with DWT coefficient approximation. Before dimension reduction, colon cancer data obtained 84% accuracy and after reduced dimensions increased to 100%. The same case occurred in the central nervous system, which obtained accuracy from 75% to 83.33%. Taking the best features of the approximation coefficient causes the classification with the DWT approximation coefficient can increase better accuracy for all data. The approximation coefficient is a filtered feature with low frequency able to store information. Reducing the dimensions of the Discrete Wavelet Transform (DWT) is very influential in improving accuracy with a smaller number of features than without dimension reduction.

## V. CONCLUSION

Based on the research results that have been obtained, the conclusion drawn from this research is that the classification using Random Forest with the reduction of the dimensions of the Discrete Wavelet Transform (DWT) can produce the best accuracy, reaching 100%. Reduction of dimensional Discrete Wavelet Transform (DWT) can improve the accuracy of classification up to 8% - 20%. Before dimension reduction, colon cancer data obtain 84% accuracy. After reducing dimensions, it increases to 100%, central nervous system before being reduced gets an accuracy of 75% and increase to 83.33%, prostate tumor having a constant accuracy of 85.71%, and in the same case for lung and ovarian also has constant accuracy of 100% by making different attributes. The accuracy obtained is influenced by the value of the parameters contained in the optimization parameter. Besides being influenced by the hyperparameter optimization values, increasing accuracy is affected by taking the best features in the dimension reduction process. The three classification scenarios carried out produces different accuracy. However, the classification with the DWT coefficient approximation is better. It can improve accuracy better than the classification with DWT coefficient detail and the classification without reduction.

## ACKNOWLEDGMENT

The authors would like to thank Telkom University have supported this research and publish the paper.

## REFERENCES

[1] American Cancer Society, “Surveillance Research,” p. 5, 2019.

[2] World Health Organization, “Cancer Factsheets,” *World Health Organization*, 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed: 19-Sep-2019].

[3] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, “Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification,” *J. Comput. Sci.*, vol. 14, no. 11, pp. 1521–1530, 2018.

[4] W. Yip, S. B. Amin, and C. Li, *Handbook of Statistical Bioinformatics*. 2011.

[5] W. Astuti and A. Adiwijaya, “Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis,” *J. Media Inform. Budidarma*, vol. 3, no. 2, pp. 72–77, 2019.

[6] R. Nurviarelda, A. A. Rohmawati, F. Informatika, U. Telkom, F. Informatika, and U. Telkom, “Klasifikasi Data Microarray Menggunakan Discrete Wavelet Transform Dan Naive Bayes Classification “, vol. 5, no. 1, pp. 1536–1540, 2018.

[7] Adiwijaya, “Deteksi Kanker Berdasarkan Klasifikasi Microarray Data,” *Media Inform. Budidarma*, vol. 2, no. 4, pp. 181–186, 2018.

[8] K. Moorthy and M. S. Mohammad, “Random forest for gene selection and microarray data classification,” no. July, 2013.

[9] H. Aydadenta and Adiwijaya, “A clustering approach for feature selection in microarray data classification using random forest,” *J. Inf. Process. Syst.*, vol. 14, no. 5, pp. 1167–1175, 2018.

[10] L. Breiman, “Random Forest Draft,” pp. 1–33, 2001.

[11] D. H. Mazumder and R. Veilumuthu, “An Enhanced Gene Selection Methodology for Effective Microarray Cancer Data Classification,” *Int. J. Simul. Syst. Sci. Technol.*, pp. 1–7, 2018.

[12] Khadijah and H. S., “Klasifikasi Data Microarray Menggunakan Discrete Wavelet Transform dan Extreme Learning Machine,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 9, no. 1, pp. 33–42, 2015.

[13] Y. Liu, “Detect key gene information in classification of microarray data,” *EURASIP J. Adv. Signal Process.*, vol. 2008, 2008.

[14] J. Bennet, C. A. Ganaprakasam, and K. Arputharaj. “A Discrete Wavelet based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis”. Anna University, Department of Computer Science and Engineering, 2014

[15] P. Liashchynskyi, “Grid Searh, Random Search, Genetic Algorithm : A Big Comparison for NAS”, Cornell University, 2019.

[16] M.D. Purbolaksono, K. C. Widiastuti, Adiwijaya, M. S. Mubarok, and F. A. Ma’ruf. Implementation of mutual information and bayes theorem for classification microarray data. In Journal of Physics: Conference Series, vol. 971, no. 1, p. 012011. IOP Publishing, 2018.

[17] I. Damayana, R. D. Atmaja, and H. Fauzi, "Menggunakan Wevelet Transform Detection of Skin Cancer Melanoma Based on Digital Image," *Deteksi Kanker Kulit Melanoma Berbas. Pengolah. Citra Menggunakan Wevelet Transform*, vol. 3, no. 3, pp. 4718–4723, 2016.

[18] Ma'ruf, Firda Aminy, and Untari Novia Wisesty. "Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier." In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012011. IOP Publishing, 2019.

[19] M. Yusa, Ema Utami and Emha T.Luthi. "Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes." In Journal of Buana Informatika, vol. 7, no. 4, 2016.

[20] Effendy, V., Adiwijaya, and Baizal, Z.A., 2014, May. Handling imbalanced data in customer Effendy, Veronikha, and ZK Abdurahman Baizal. "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest." 2014 2nd International Conference on Information and Communication Technology (ICoICT). IEEE, 2014.

[21] Mabarti, I., Aditsania, A., "Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4.5 Decision Tree". Journal of Data Science and Its Applications, 3(1), 2020.

[22] 1. Rohmawati A., Adiwijaya, 2017. A Daubechies Wavelet Transformation to Optimize Modeling Calibration of Active Compound on Drug Plants. In 5th International Conference on Information and Communication Technology (ICoICT). Pp.1-4. IEE

[23] Adiwijaya, Maharani, M., Dewi, B.K., Yulianto, F.A. and Purnama, B., 2013. digital image compression using graph coloring quantization based on wavelet-SVD. In Journal of Physics: Conference Series (Vol. 423, No. 1, p. 012019). IOP Publishing.

[24] Daeli, N.O.F, Adiwijaya. Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 2020

[25] Purnomoputra, Riko Bintang, Adiwijaya Adiwijaya, and Untari Novia Wisesty. "Sentiment Analysis of Movie Review using Naïve Bayes Method with Gini Index Feature Selection." Journal of Data Science and Its Applications 2(2) pp. 85-94. 2019