Journal of Soft Computing Exploration

Improve the Accuracy of C4.5 Algorithm Using Particle Swarm Optimization (PSO) Feature Selection and Bagging Technique in Breast Cancer Diagnosis

Raka Hendra Saputra[1], Budi Prasetyo[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jul 25, 2020
Revised Aug 10, 2020
Accepted Sept 3, 2020

Keywords:

Data mining
Decision Tree
Classification
C4.5
PSO
Bagging
Breast Cancer

ABSTRACT

Breast cancer is the second leading cause of death due to cancer in women currently. It has become the most common cancer in recent years. In early detection of cancer, data mining can be used to diagnose breast cancer. Data mining consists of several research models, one of which is classification. The most commonly used method in classification is the decision tree. C4.5 is an algorithm in the decision tree that is often used in the classification process. In this study, the data used was the Breast Cancer Wisconsin (Original) Data Set (1992) obtained from the UCI Machine Learning Repository. The purpose of this study was to select features that will be used and overcome class imbalances that occur, so that the performance of the C4.5 algorithm worked more optimal in the classification process. The methods used as feature selection are PSO and bagging to overcome class imbalances. Classification was tested using the confusion matrix to determine the accuracy that was generated. From the results of this study, the application of PSO as a feature selection and bagging to overcome class imbalances with the C4.5 algorithm succeeded in increasing accuracy by 5.11% with an initial accuracy of 93.43% to 98.54%.

Corresponding Author:

Raka Hendra Saputra
Computer Science Departement
Faculty of Mathematich and Natural Sciences, Universitas Negeri Semarang,
Email: rakahendrasaputra@students.unnes.ac.id

1. INTRODUCTION

Breast cancer is the second leading cause of death due to cancer in women currently. It has become the most common cancer among women in developed and developing countries in recent years [1]. Identification of breast cancer can be done manually, but this process is difficult because we must remember all the information needed for each particular situation that cause in low accuracy. Mortality from breast cancer can be reduced if it can be detected early. There are conventional methods for breast cancer detection but machine learning classifiers need to be done because they can get higher accuracy [2].
Data mining is a pattern recognition technology as well as statistical and mathematical techniques to find meaningful correlations, patterns and new trends by sorting out the data storage stacks that store large data [3].

In the medical field, data mining can be used to diagnose some diseases such as breast cancer, heart disease, diabetes, etc. [4].

Classifications in data mining are two forms of data analysis process used to extract models that describe data classes or predict future data trends. In the classification process, there are 2 phases; the first phase is training data, wherein this phase the data are studied and analyzed using classification algorithms. The model or classifier studied is presented in the form of a pattern or classification rule; the second phase is the use of models for classification, and testing data is used to estimate the accuracy generated based on classification rules [5].

The problem that often occurs is the classification has a large number of features in the dataset, but not all of them will be used. Irrelevant and redundant features can reduce performance [6]. Unnecessary features can make generalizations more difficult and increase the size of the search space which makes a major obstacle in machine learning and data mining. To maximize accuracy in classification, we can use feature selection in selecting features that will be used [7].

Feature selection is widely used to overcome irrelevant exaggerated features. Feature selection simplifies a collection of data by reducing dimensions and identifies the relevant features without reducing the prediction of accuracy [8]. Particle Swarm Optimization (PSO) is a metaheuristic optimization for feature selection because it has been proven to be competitive compared to genetic algorithms in some cases, especially in the field of optimization [4]. Metaheuristic optimization has proven to be a superior methodology for getting a good solution in a reasonable time [9]. In addition, too many available features, the dataset also often occurs data imbalances.

Data imbalance is one of the classic problems in classification in machine learning. Data imbalance has been proven to reduce the performance of machine learning algorithms [10]. Imbalance can be interpreted, for example one class (majority class) is more than the other class (minority class) [11].

Breast Cancer Wisconsin (Original) Data Set has 2 classes, namely benign written 2 in class as much as 458 (65.5%) and malignant written 4 in class as many as 241 (34.5%).

Two popular methods used in the ensemble method are Bagging and Boosting [13]. Bagging technique is superior compared to boosting when dealing with data that contains noise [14]. In addition, bagging technique is not only easy to be developed, but also strong when dealing with class imbalances if implemented correctly [15]. Bagging technique can be applied to tree-based methods to increase the value of accuracy that will be generated later [16].

A text can consist of only one word or sentence structure [2]. Information in the form of text is important information and is widely obtained from various sources such as books, newspapers, websites, or e-mail messages. Retrieval of information from text (text mining), among others, can include text or document categorization, sentiment analysis, search for more specific topics (search engines), and spam filtering [3]. Text mining is one of the techniques that can be used to do classification where, text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data [4].

The classification method itself many researchers use the Naïve Bayes Classifier where a text will be classified in machine learning based on probability [5]. Naïve Bayes Classifier is a pre-processing technology in the classification of features, which adds scalability, accuracy and efficiency which is certainly very much in the process of classifying a text. As a classification tool, Naïve Bayes Classifier is considered efficient and simple, and sensitive to feature selection [6].

The data used in this study contains hotel reviews in English so that it can be seen that the grammar used by a person is very diverse in writing the review, diversity makes the features generated through N-Gram will be very much. Therefore, here we will use N-Gram word characters with N = 1, 2, 3 to retrieve features in a review which will then be classified with the Naïve Bayes Classifier Algorithm.

It is expected that the N-Gram Naïve Bayes Classifier Algorithm in this study can be classified correctly and appropriately. So that the main purpose of this study can be fulfilled which is to know the effect of N-Gram features on Naïve Bayes Classifier for sentiment analysis of hotel reviews.

2. METHOD

Stages of data processing consist of several stages, starting from converting the dataset format which was originally .data to .csv, overcoming missing values contained in the dataset, selecting features with PSO, overcoming class imbalance by bagging, and evaluating using confusion matrix. For more details about the methods used in this study can be seen in Figure 1.

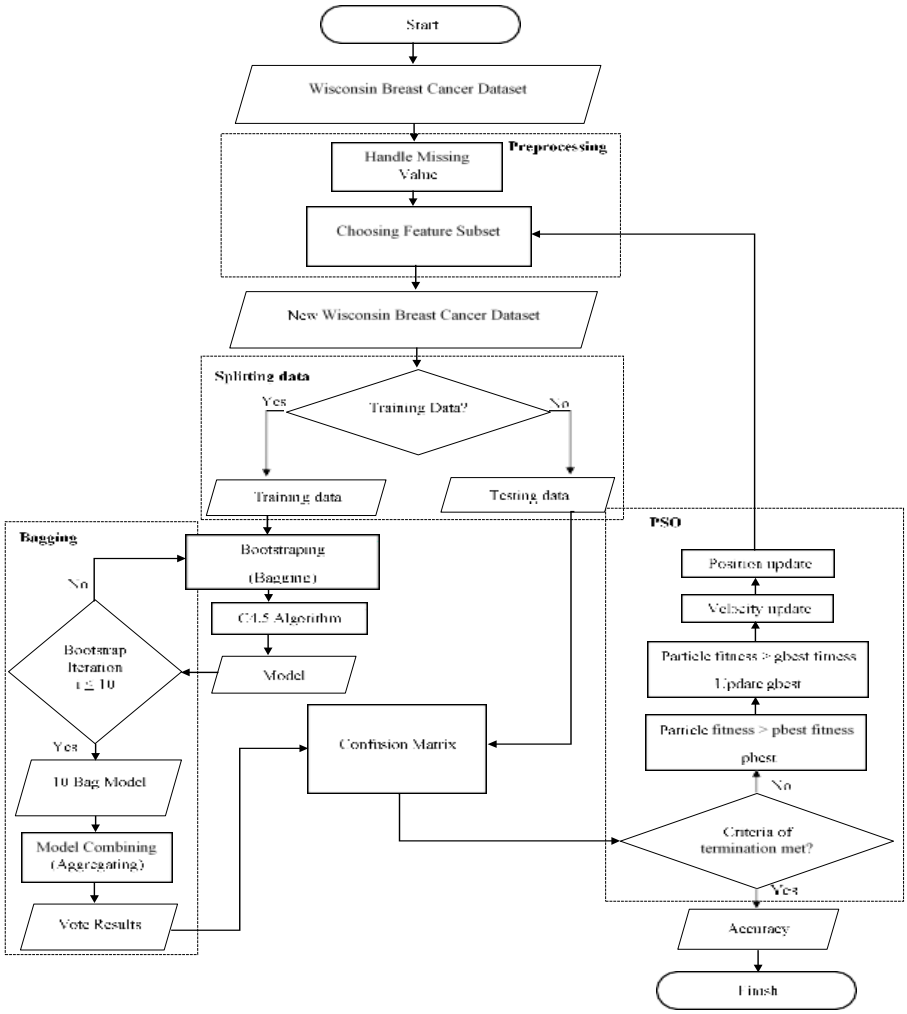

Figure 1. C4.5 algorithm using pso and bagging technique

2.1 Handling Missing Values

The dataset used in this study experienced a missing value of 16 data. This was known with the help of WEKA tools as shown in Figure 2.

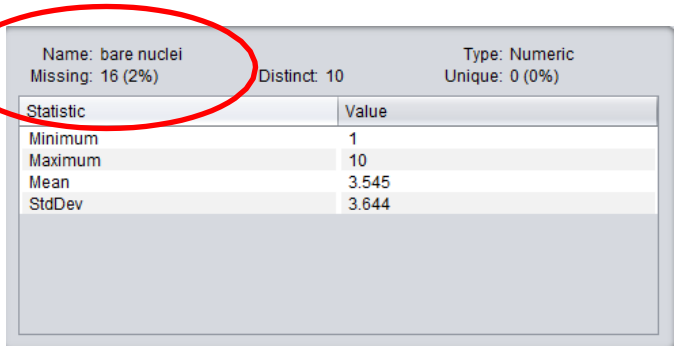

Figure 2 Missing Value on the Dataset

The missing value was in the bare nuclei attribute which meant as much as 16 data in the bare nuclei attribute was not filled. In this study, 16 data that experienced missing values were overcome by imputation methods, so as not to interfere with the classification process. So that, the amount of data in the dataset was reduced, which was originally 699 data to 683 data.

2.2 Particle Swarm Optimization (PSO)

PSO was selected as the best features available in the Breast Cancer Wisconsin (Original) Data Set. The best features were the features selected to be used in the next process. The PSO steps in selecting the best features are shown in Figure 3.

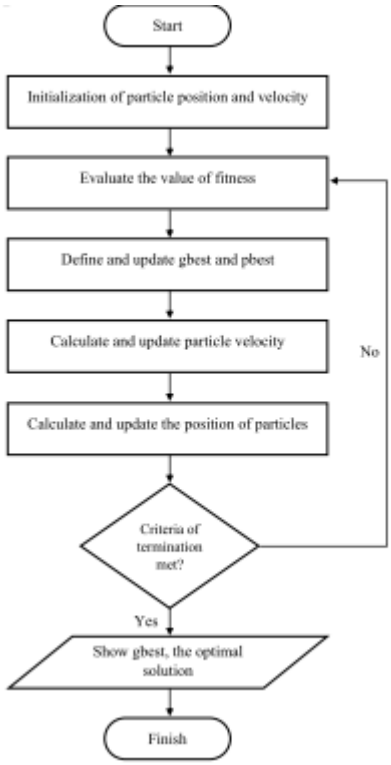

Figure 3. Flowchart PSO

Based on Figure 3, the PSO steps can be seen more clearly as follows.

Step 1: Initialization of particle position (*xt*), weight of inertia (*w*) = 0.72, and acceleration coefficients (c_1 and c_2) = 0.7. Initialization of particle velocity (*vt*) = 0. Number of particles = 50 and iterations performed = 100.

Step 2: Calculate and evaluate the fitness value of each particle using the C4.5 algorithm.

Step 3: Determine the *pbest* value of each particle based on the accuracy value produced by C4.5. Determine *gbest* value based on the highest *pbest* value.

Step 4: Calculate particle of velocity and position using Equations 1 and 2.

$$v_{id}^{t+1} = w \times v_{id}^{r} + r_{1i} \times (p_{id} - x_{id}^{t}) + c_2 \times r_{2i} \times (p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (2)$$

Step 5: Determine the optimal criteria. Determination of particle probability 0 or 1 based on the speed value using the sigmoid function in Equation 3.

$$x(t+1) = \begin{cases} 1 & if \quad rand < s(v(t+1)) \\ 0 & if \quad not \end{cases} \quad (3)$$

The value rand () is a random number that is uniformly distributed between 0 and 1. The *S*( ) function is a sigmoid function calculated using Equation 4.

$$s(v_{ij}(t+1)) = \frac{1}{1+e^{-v_{id}^{t+1}}} \quad (4)$$

Step 6: Displays *gbest* and optimal solution in the form of selected features that will be used.

2.3 Bagging technique

Bagging is a method that combines bootstrapping and aggregating. Bootstrap samples are obtained by changing the number of elements or resampling the same number of elements as the original dataset [21]. The bagging process was done in training data, while the steps are as in Figure 4.

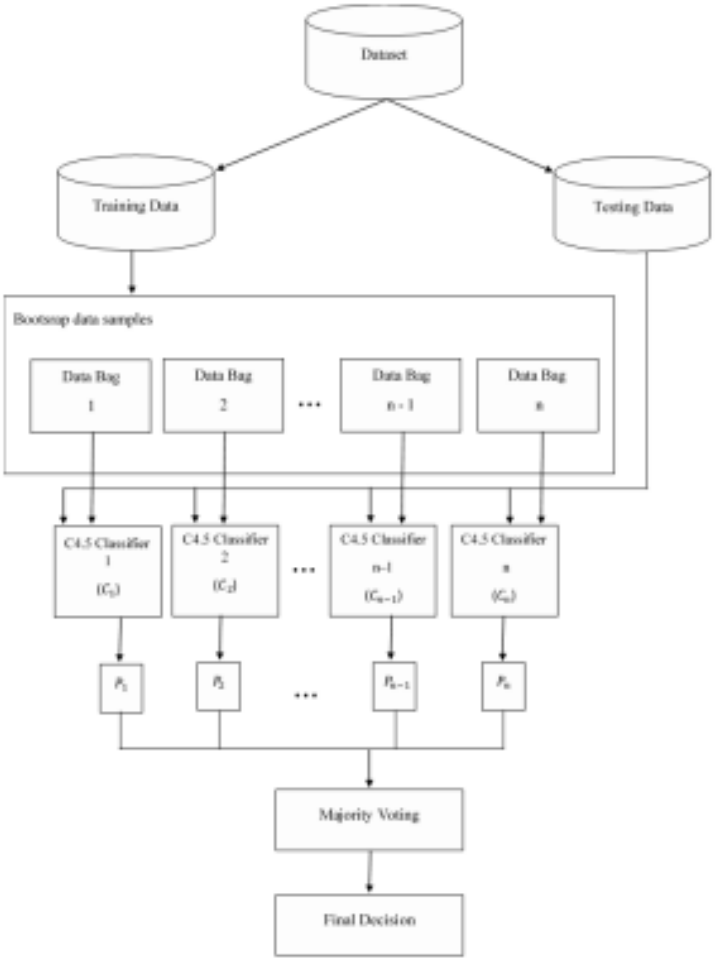

Figure 4. The concept of bagging process

Based on Figure 4, it can be seen more clearly bagging steps are as follows.

Step 1: Perform the bootstrap process on the training data, dividing the data according to the specified number of bags. In this study 100 bags were used.
Step 2: Classify each bag using the C4.5 classifier to get the model.
Step 3: Next, the model obtained was tested using testing data.
Step 4: Each bag produces accuracy, then vote on all results.
Step 5: The results of the final decision were the results based on majority voting.

2.4 C4.5 algorithm

C4.5 algorithm is an algorithm that is widely used in classifications to make decisions because it can produce decision trees that are easy to interpret and understand, it has an acceptable level of accuracy, and are efficient for dealing with discrete and numerical attributes [22]. Stage C4.5 was conducted on the training data in conducting the classification process can be seen in Figure 5.

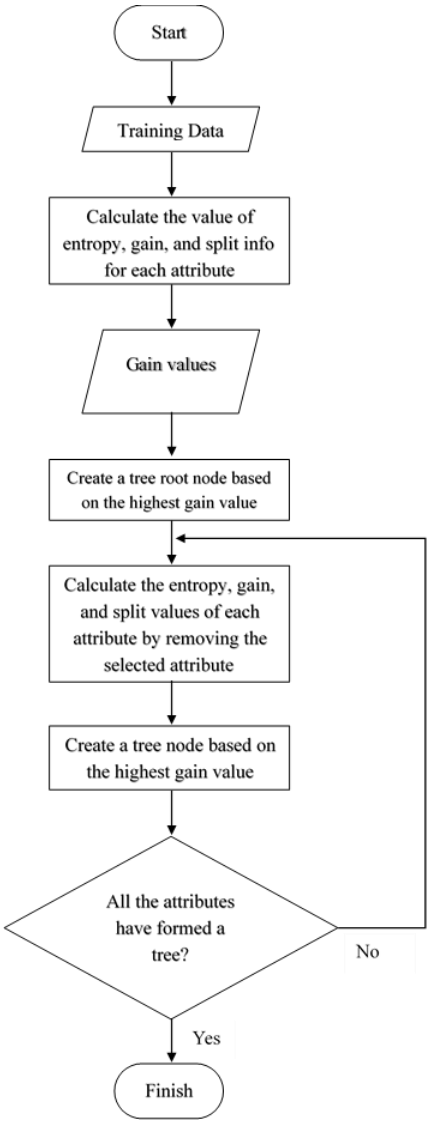

Figure 5. C4.5 Algorithm

Based on Figure 5, it can be seen more clearly the steps of the C4.5 algorithm are as follows.

Step 1: Calculate the entropy of each attribute with Equation 5.

$$Entropy\ (S) = \sum_{i=1}^{n} p_i \times \log_2 p_i \tag{5}$$

Step 2: Calculate the gain info for each attribute by Equation 6.

$$Info\ Gain\ (S.A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{6}$$

Step 3: Calculate the split info for each attribute using Equation 7.

$$Split\ Info\ (S.A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \tag{7}$$

Step 4: Calculate the gain of each attribute by Equation 8.

$$Gain\ (A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{8}$$

Step 5: Calculate the gain of each attribute by Equation 8.
Step 6: Repeat steps 1 to 4 to determine the branch by removing the selected attributes.
Step 7: Create a branch based on the highest gain value.
Step 8: Continue to repeat the process of determining branches until all attributes form a tree.

2.5 Evaluate with the Confusion Matrix

The evaluation stage was carried out at the end of the research process. This stage is useful for testing the model and calculating the resulting accuracy. In this study the evaluation was carried out with a confusion matrix. The steps are as follows.

Step 1: Enter the test results in the confusion matrix table as seen in Table 1.

Table 1. Testing the confusion matrix

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

Step 2: Calulate the accuracy value, determine the highest accuracy with Equation 9 .

$$Accuracy = \frac{TP+TN}{P+n} \times 100\% \quad (9)$$

Step 3: State the conclusions from the accuracy results obtained.

3. RESULT AND DISCUSSION

This research was conducted using tools, namely the Python 3 programming language, scikit-learn library, and Pyswarms documentation. While the material used was the Breast Cancer Wisconsin (Original) Data Set obtained from the UCI Machine Learning Repository. The tools and materials in this study were public so they can be accessed and used by anyone who will conduct or prove the validity of the research conducted by previous researchers. PSO feature selection was done to get selected features that will be used for the classification process. The dataset which previously had 9 attributes after being processed by PSO left 8 selected features that will be used in the next process. This can optimize the performance of the C4.5 algorithm in classifying the dataset. The results of the selected features can be seen in Figure 6. A total of 8 selected features when used in the classification process with the C4.5 algorithm produce an accuracy of 95.62%.

Figure 6. Selected Features by PSO

Bagging was done to overcome the class imbalance that occurs in the dataset used. Bagging was done in training data by dividing the data into 100 bags randomly, the total data of the whole bag was the same as the total training data. Bagging will produce the best bag of 100 bags then the results will be processed by C4.5 algorithm to do the classification. 1 bag with the highest accuracy will be used as the final decision in the classification process. In this study, data from 1 bag selected when processed by the C4.5 algorithm produced an accuracy of 97.81%.

This study recorded every accuracy that results from the classification process that has been done. The results can be seen in Table 2

Table 2. Results of each method used

Algorithm	Accuracy
C4.5	93,43%
C4.5 + PSO	95,62%
C4.5 + Bagging	97,81%
C4.5 + PSO + Bagging	98,54%

Based on Table 2, it was known that there was an increase in each method used. C4.5 algorithm without using PSO and bagging produced an accuracy of 93.43%. C4.5 algorithm with PSO without using bagging produced an accuracy of 95.62%. C4.5 algorithm with bagging without using PSO produced an accuracy of 97.81%. And the purpose method which in this case was an algorithm with PSO and bagging produces an accuracy of 98.54%. So it can be concluded that there was an increase in accuracy of 5.11% when comparing the C4.5 algorithm without PSO and bagging with the purpose method in this study.

When the method used in this study was compared with previous studies, it can be seen that the accuracy produced in this study was 98.54 which shows better than some previous studies using the Breast Cancer Wisconsin (Original) Data Set as in Table 3. Akay [23] in his research showed that the distribution of training and testing data respectively 80% and 20% is the most optimal when used for classification of breast cancer. Lavanya & Rani [24], in her study showed the application of bagging to decision trees which in this case was CART produces an accuracy of 97.85% . Muslim MA et al., [4] in his research succeeded in increasing accuracy by 0.88% by using PSO as a feature selection on the C4.5 algorithm [4]. Shrivas & Singh [25] in his research showed C4.5 using the distribution of training and testing data respectively 80% and 20% for the classification of breast cancer resulting in an accuracy of 92.857% [25].

Table 3. Comparison of research accuracy

Method	Accuracy
Akay	97,91%
Lavanya & Rani	97,85%
Muslim *et al*	96,49%
Shrivas & Singh	92,857%
The Purpose Method	98,54%

4. CONCLUSION

Based on the results of research and discussion related to C4.5 algorithm using Particle Swarm Optimization (PSO) feature selection and bagging technique in breast cancer diagnosis, it can be concluded that PSO was used from a number of features in the dataset. In this case, the feature can be referred to an attribute. The dataset originally had 9 attributes and 1 class became 8 attributes and 1 class after PSO was applied. Bagging was used to overcome class imbalances that occur in the dataset. Bagging produced the best bag to be used in the classification process of the C4.5 algorithm in order to make its performance more optimal. Accuracy results obtained when applied PSO and bagging on the C4.5 algorithm were 98.54%. While, C4.5 without PSO and bagging produced an accuracy of 93.43%. So, it can be seen an increase of 5.11% based on the comparison of the resulting accuracy. This showed that PSO and bagging had an important role in optimizing

REFERENCES

[1] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, "Diagnosis of breast cancer using decision tree data mining technique, " *International Journal of Computer Applications*, vol. 98, no. 10. 2014.

[2] A. Gupta, and B. N. Kaushik, "Feature selection from biological database for breast cancer prediction and detection using machine learning classifier, " *J. Artif. Intell*, vo. 11, pp. 55-64, 2018.

[3] D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc. 2004.

[4] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, "Optimization of C4. 5 algorithm-based particle swarm optimization for breast cancer diagnosis, " *Journal of Physics: Conference Series*, vol. 983, no. 1, 2018.

[5] D. Singh, N. Choudhary, and J. Samota, "Analysis of data mining classification with decision tree technique, " *Global Journal of Computer Science and Technology*, vol. 13, pp. 1-5, 2013.

[6] B. Xue, M. Zhang, and W. N Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach, " *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1656-1671. 2012.

[7] I. A. Gheyas, and L. S. Smith, "Feature subset selection in large dimensionality domains, " *Pattern recognition*, vol. 43, no. 1, pp. 5-13. 2010.

[8] M. H. Aghdam, S. Heidari, "Feature selection using particle swarm optimization in text categorization,

" *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, pp. 231-238. 2015
[9] S.C. Yusta, "Different metaheuristic strategies to solve the feature selection problem, " *Pattern Recognition*, vol. 30, no. 5, pp. 525-534. 2009.
[10] T. W. Cenggoro, "Deep learning for imbalance data classification using class expert generative adversarial network, " *Procedia Computer Science*, vol. 135, pp. 60- 67. 2018
[11] N. Rout, D. Mishra, and M.K. Mallick, "Handling imbalanced data: a survey, " *International Proceedings on Advances in Soft Computing, Intelligent Systems and Application*s. Singapura, 2018, pp. 431-443.
[12] B. W. Yap, K . A . Rani, H.A.A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah,. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).* Singapura: Springer.
[13] D. Opitz and R. Maclin, " Popular ensemble methods: an empirical study, " *Journal of Artificial Intelligence*, vol. 11, pp. 169-198. 1999.
[14] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, A. "Comparing boosting and bagging techniques with noisy and imbalanced data, " *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol .41, no. 3, pp. 552-568, 2010.
[15] W. Feng, W. Huang, W, and J. Ren, "Class imbalance ensemble learning based on the margin theory, " *Applied Sciences*, vol. 8, no. 5, pp. 815. 2018.
[16] C. D. Sutton, "Classification and regression trees, bagging, and boosting." *Handbook of statistics*, vol. 24, pp. 303-329. 2005
[17] M. Bramer, Principles of data mining, London: Springer. 2007
[18] Y. Yang, and W. Chen, "Taiga: performance optimization of the C4. 5 decision tree construction algorithm". *Tsinghua Science and Technology*, vol. 21, no. 4, pp. 415-425, 2016.
[19] B. Boukenze, H. Mousannif, and A. Haqiq, "Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease, " *Int. Journal of Database Managment systems*, vol. 8, no. 30, pp. 1-9, 2016.
[20] K. R. Pradeep, and N. C. Naveen, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics, " *Procedia computer science*, vol. 132, pp. 412-420, 2018.
[21] E. Alfaro, M. Gámez, and N. Garcia, "Adabag: and package for classification with boosting and bagging, ". *Journal of Statistical Software*, vol. 54, no. 2, pp. 1- 35, 2013.
[22] S.J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making, " *Journal of Biomedical Informatics*, vol. 78, pp. 144-155, 2018.
[23] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis, " *Expert systems with applications*, vol. 36, no. 2, pp. 3240-3247. 2009.
[24] D. Lavanya, and K. U. Rani, "Ensemble decision tree classifier for breast cancer data, " *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, pp. 17, 2012.
[25] A.K. Shrivas, and A. Singh, "Classification of breast cancer diseases using data mining techniques, ". *International Journal of Engineering Science Invention*, vol. 5, no. 12, pp. 62-65, 2016.