International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 1, February 2021, pp. 728~735
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i1.pp728-735

A hybrid feature selection on AIRS method for identifying breast cancer diseases

Achmad Ridok[1], Nashi Widodo[2], Wayan Firdaus Mahmudy[3], Muhaimin Rifa'i[4]
[1,3]Faculty of Computer Science (FILKOM), Brawijaya University, Indonesia
[2,4]Department of Biology, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia

Article Info

Article history:

Received Aug 29, 2018
Revised Jul 27, 2020
Accepted Aug 25, 2020

Keywords:

AIRS
Breast cancer
FCBF
Hybrid feature selection
IG

ABSTRACT

Breast cancer may cause a death due to the late diagnosis. A cheap and accurate tool for early detection of this disease is essential to prevent fatal incidence. In general, the cheap and less invasive method to diagnose the disease could be done by biopsy using fine needle aspirates from breast tissue. However, rapid and accurate identification of the cancer cell pattern from the cell biopsy is still challenging task. This diagnostic tool can be developed using machine learning as a classification problem. The performance of the classifier depends on the interrelationship between sample sizes, some features, and classifier complexity. Thus, the removal of some irrelevant features may increase classification accuracy. In this study, a new hybrid feature selection fast correlation based feature (FCBF) and information gain (IG) was used to select features on identifying breast cancer using AIRS algorithm. The results of 10 times the crossing (CF) of our validation on various AIRS seeds indicate that the proposed method can achieve the best performance with accuracy =0.9797 and AUC=0.9777 at k=6 and seed=50.

Corresponding Author:

Achmad Ridok
Faculty of Computer Science (FILKOM),
Brawijaya University,
8 Veteran Road Malang, Indonesia.
Email: acridokb@ub.ac.id

1. INTRODUCTION

Breast cancer is the most famous female killer between the ages of 35 and 54 years old and the most common cancer in women [1]. The cheap diagnostic tool for early detection of this disease is essential to prevent fatal incidence. This diagnostic tool can be developed using machine learning to identify cancer cell profile from cell biopsy. Machine learning is one of the sub-fields of artificial intelligence that has been applied to automatically identify the hidden patterns through learning or classify from experience of previous data [2].

There are several classification methods that have been developed, such as artificial neural networks (ANN), K-nearest neighbor (KNN), support vector machine (SVM) and AIRS. However, one of the advantages of AIRS compared to other methods is that the exact settings do not need to be known beforehand but are determined by yourself. With this capability, this method is considered a smart supervised classification [3]. In addition, AIRS is one of these techniques that has been used successfully in the problem of medical classification [4].

The performance of a classifier depends on the interrelationship between sample sizes, some features, and classifier complexity. Classification results would be better if using more training datasets with fewer attributes. Thus, the removal of some irrelevant features may increase classification accuracy [5-8]. Several feature selection algorithms have been used to improve the performance of the AIRS algorithm, such as C4.5 [9], principal component analysis (PCA) [10], correlation-based feature (CFS) and FCBF [11].

Besides, Utilization of feature weighting algorithms has been performed by some researchers such as the weighting of fuzzy preprocesses and information gains (IG) [12-15]. Besides that, the performance of the classification algorithm also is determined by the parameters of each algorithm. The effect of setting parameters on the performance of the AIRS algorithm for some data such as Iris, Ionosphere, Diabetes and Sonar has been conducted [2]. This work has reported that different seed assignments have a significant effect on accuracy. Meanwhile, the effect of setting parameters of AIRS for identifying breast cancer datasets has not been conducted. Therefore, this study proposes a combined method of selecting FCBF and weighting features IG on the AIRS classifier algorithm.

2. RELATED RESEARCH

In this section, some researches are explained related to efforts of improving the performance of AIRS classification algorithms including feature selection and feature weighting.

2.1. Utilization of features selection in AIRS

Polat *et al.* introduced the FS-AIRS method to diagnose breast cancer [9]. In this research, the C4.5 decision tree algorithm was used to select features that reduce from 9 features to 6 features. Evaluation with 10-CV obtained accuracy of 98.51 at k=1. In 2007, Polat and Salih used PCA-AIRS for predicting hepatitis disease that achieved accuracy 94.12 at k=1. Utilization of PCA in this study has reduced the number of features from 19 to 5 [16]. The same method has been used for lung cancer detection that reduced the number of features from 57 features to 4 features and achieved 100% accuracy at k KNN=1 T [10]. Katsis *et al*. have used (CFS) on AIRS in order to detect early breast cancer. The investigation result at k values of 1 to 7 showed the best results at k=3 in AIRS and AIRS+CFS methods. Nevertheless, the best results between them occur on AIRS without CFS. Besides, this study has shown that the accuracy of the AIRS algorithm is better than other comparative methods, such as SVM and C4.5 [17]. Ridok *et al*., [11] have combined FCBF as a features selection on AIRS for classifying breast cancer datasets. This method achieved accuracy 100% at k=1 to 30.

2.2. The weighting of the preprocess

All features of training dataset are weighted using a specified algorithm before they are used for training in classification. Polat *et al*. have introduced the AIRS hybrid method and the weighting of preprocessed using fuzzy for diagnosing thyroid disease. The accuracy of this method reached 85 at k=2 [18]. Shamshirband *et al*. [19] uses a combination of fuzzy logic and AIRS to diagnose Tuberculosis. The features were normalized through a fuzzy rule based on a labeling system. The labeled features were categorized into normal classes and tuberculosis using AIRS. The validation results using a 10-CV showed 99.14% accuracy on the learning rate of 0.8. Several studies related to the use of IG-AIRS for weighting of features have been conducted. Kodaz *et al*. [12] introduced the utilization of information gain (IG) on AIRS algorithm to identify atherosclerosis disease with success rate 99.09%. With the sam method, Kodaz *et al.* [13] have used this method to identify Thyroid disease with accuracy 95.90. Kara *et al*., also applied this method to the classification of microorganisms with a success rate of 92.35 [14].

2.3. Dataset

This study used Wisconsin breast cancer datasets collected by Dr. W.H. Wolberg from fine-needle aspirates from human breast cancer tissue (WBCD) (https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin). The WBCD consists of 683 samples that divided into two group, i.e., 444 for benign, and 239 for malignant. Each of samples consists of nine features as follows: (1) Clump thickness (x1); (2) Uniformity of cell size (x2); (3) Uniformity of cell shape (x3); (4) Marginal adhesion (x4); (5) Single epithelial cell size (x5); (6) Bare nucleoli (x6); (7) Bland chromatin (x7); (8) Normal nucleoli (x8); and (9) Mitoses (x9).

2.4. Method

As elucidated in the introduction that classification performance can be enhanced by selection and weighting of features, therefore this study proposes a combination of FCBF and IG as a hybrid feature selection algorithm to improve AIRS classification algorithm's performance in identifying breast cancer disease. The proposed method can be illustrated as in Figure 1. In the first step, the FCBF algorithm reduces the feature of the dataset with 9 features into a dataset with 6 features. In the second step, IG algorithm is used to determine the weight of each new dataset feature which will be used for a weighting factor when calculating the distance between features on the internal AIRS. In the third step, the new dataset is divided into two parts, namely as a training dataset and the other as a test dataset using 10-fold cross validation as outlined in 4.2. The fourth step, The AIRS algorithm generated cell memory as a result of the learning

process of each training data under the condition of the specified parameters as depicted in Figure 2. The fifth step, KNN determines the label of each training data based on k voting majority. In the last step, accuracy is calculated based on the number of similarities between the label results of the classification and the origin label of each training dataset. The third step until the last step is repeated 10 times according to the combination of dataset pairs and training data as the result of 10-fold cross-validation.

According to Watkin *et al*. [2], as mentioned in the introduction, changes in seed cells affect the accuracy of AIRS classification. Therefore, the proposed method was evaluated on the variation of seed from 10 to 100 to find out the best performance of classification in terms of accuracy and AUC. In some previous studies, investigations were rarely conducted on KNN with different k values. Therefore, the study investigated the results of the KNN classification at k=5, k=6 and k=7.

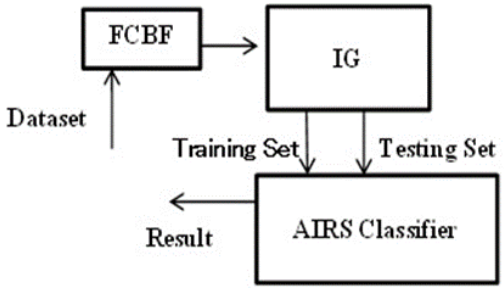


Figure 1. AIRS classification based on IG and FCBF feature selection

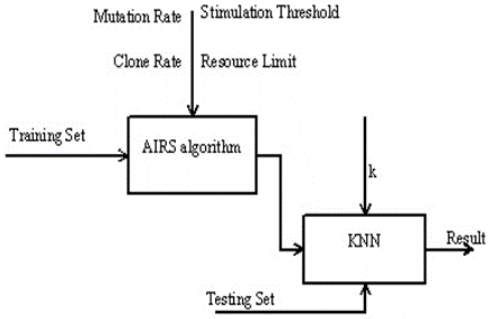


Figure 2. AIRS classification based

2.5. Classification performance measurement

Measurement of classification accuracy can be evaluated by calculating four components, such as the number of correctly recognized class instances (true positive), the number of correctly recognized instances that are not included in the class (true negative), and the wrong example but assigned to the class (false positives) or who is not recognized as a class instance (false negatives). All of these components can be shown as a confusion matrix shown in Table 1 for the case of binary classification. The row of the table represents predicted label, meanwhile, the actual label is represented by a column of the table. From the table, some commonly used metrics for measuring classification performance can be generated as shown in Table 2. The performance of the proposed method was evaluated only used three metrics namely accuracy, error rate and AUC. In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. Otherwise, misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated. For two-class problem, AUC is one of the popular ranking type metrics which reflects the overall ranking performance of a classifier [20, 21].

Table 1. Confusion matrix

	Actual +label	Actual label
Predicted +label	True + (TP)	False - (FN)
Predicted -label	False + (FP)	True - (TN)

Tabel 2. Metrics [20, 22]

Metrics	Formula
Accuracy (acc)	$\frac{tp+tn}{tp+fp+tn+fn}$
Error rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$
Sensitivity (Recall)	$\frac{tp}{tp+fn}$
Specifity (sp)	$\frac{tn}{tn+fp}$
Precision (P)	$\frac{tp}{tp+fp}$
AUC	$\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$

3. METHOD

3.1. System development environment

The system was developed using the Ruby programming language with JRuby 1.7.3 on NetBeans IDE 8.1. The system runs on a Windows 7 Professional 32-bit environment on a Laptop with an Intel (R) Core (T M) i3-2310M processor and 4GB of RAM.

3.2. 10-fold cross-validation

Model validation for machine learning algorithms should ensure that data were transformed to the model properly and the model represents the system with an acceptable accuracy [23]. Therefore, to minimize the bias associated with the random sampling used during the training phase, the system was evaluated using n-fold cross-validation. In this approach, the instances are randomly divided into n equal stratified subsets. At each iteration, n-1 subsets are merged to form the training set, and the remaining set is used as the testing set which classification accuracy of the algorithm is measured on it. This process is repeated n times, choosing a different subset as the test set each time. Therefore, all data instances have been used n-1 time for training and once for testing. The final predictive performance is computed over all folds in the usual manner. This study used tenfold cross-validation for evaluation purposes.

3.3. Experimental framework

To obtain the best performance of the proposed method, it was evaluated in various conditions k of KNN between k=5, k=6 and k=7 on each number of seed as mentioned in 2.4. The results of this evaluation will be compared with three methods as illustrated in Table 3.

Table 3. Experimental framework

	seed10	…	Seed90
AIRS		…	
FCBF+AIRS		…	
IG+AIRS		…	
FCBF+IG+AIRS		…	

Each method was evaluated using the 10-fold cross validation scheme. In this scheme, each method is evaluated by ten repetitions. In each repetition, the dataset was divided into 10 sections, 1 part was used as test data while the remainder as training data. Determination of the test data section is determined based on the repeat number. If the repeat number is 1, then part 1 is taken as test data and the remainder as training data as well as soon. One loop represents the training process that produced mc. At this stage, the training time and the number of cells of each mc would be calculated. The classification process be performed by finding the majority voting label of each training data for k=5, k=6, and k=7. The classification results of each k were calculated accuracy, error rate, and AUC through the calculation result from the confusion matrix. At the end of each k-fold process be calculated the average of each calculation.

3.4. Setting parameters

As illustrated in Figure 2, the AIRS algorithm has some predetermined parameters as in Table 4 of column 1. Clone rate is used to adjust the rate of cloning, mutation rate to regulate mutation of mutations, thresh stimulation is used as the average stimming limit value of a cell in the pool at the time of cloning and mutation, maximum resource denotes the maximum amount of resources during cloning and mutation. Meanwhile, the FCBF algorithm requires 1 parameter, i.e. δ. In this case set=v0.1 as recommended by Yu and Liu [24]. While on the IG algorithm there is one parameter that needs to be set namely step=4.

Table 4. Setting parameters

AIRS	FCBF	IG
clone_rate = 100 mutate_rate = 0.2 stim_thresh = 0.5 max_res = 1000	δ = 0,1	Step=4

4. RESULTS AND DISCUSSIONS

The experimental results and discussion are discussed in this section. Comparison of average accuracy results between proposed method and another method on variation seeds can be pointed out as

shown in Figure 3(a). From this figure, the accuracy of the proposed method outperforms the other three methods on seed=50. As a consequence, the proposed method has the most average error as shown in Figure 3(b).

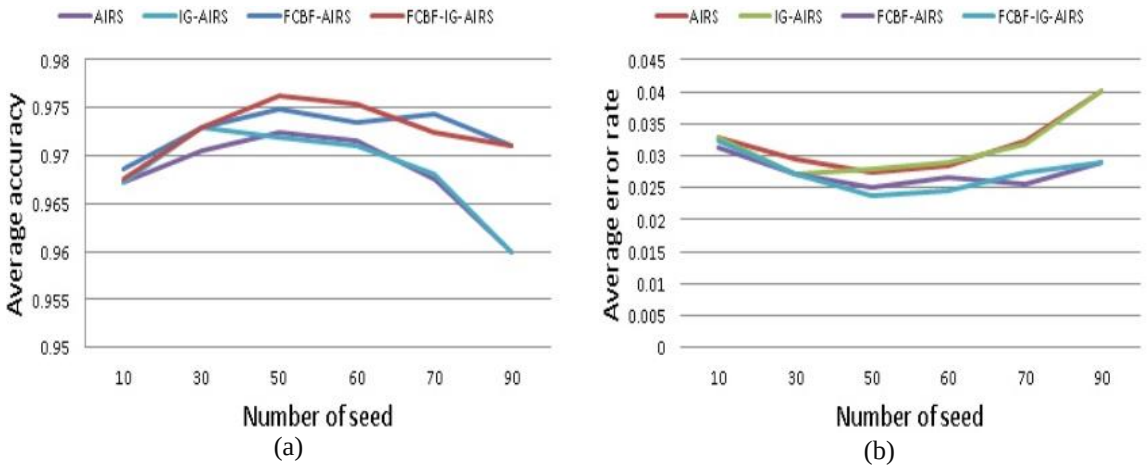


Figure 3. (a) Average accuracy and (b) Average error rate

The further analysis we used AUC to identify the best classification method for this study. AUC is one of the favorite ranking type metrics which can be used for comparing learning algorithms [25]. The value of AUC reflects the overall ranking performance of a classifier. The AUC was proven theoretically and empirically better than the accuracy metric [26, 27] for evaluating the classifier performance and discriminating an optimal solution during the classification training. The proposed method still outperforms compare to the other three methods as shown in Figure 4

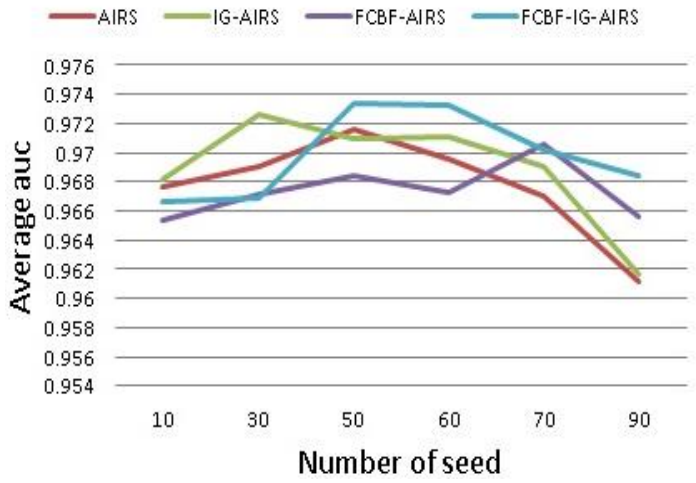
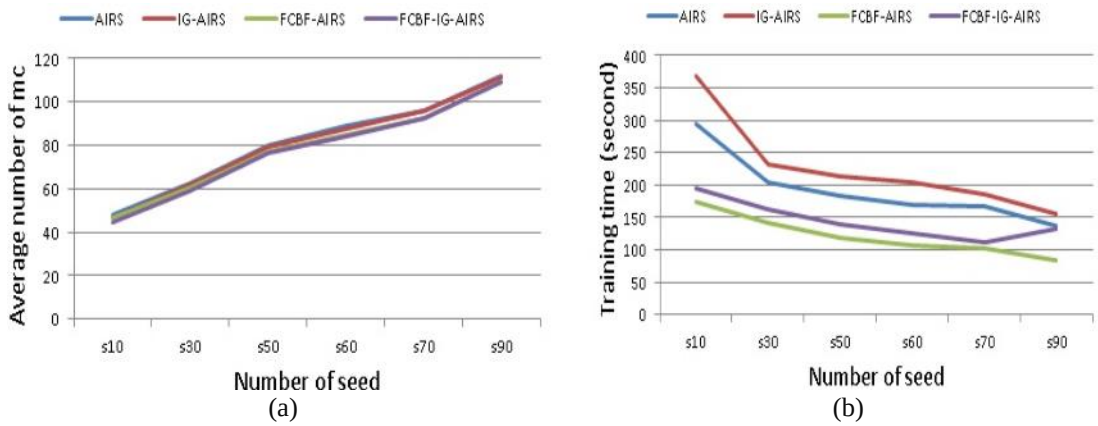


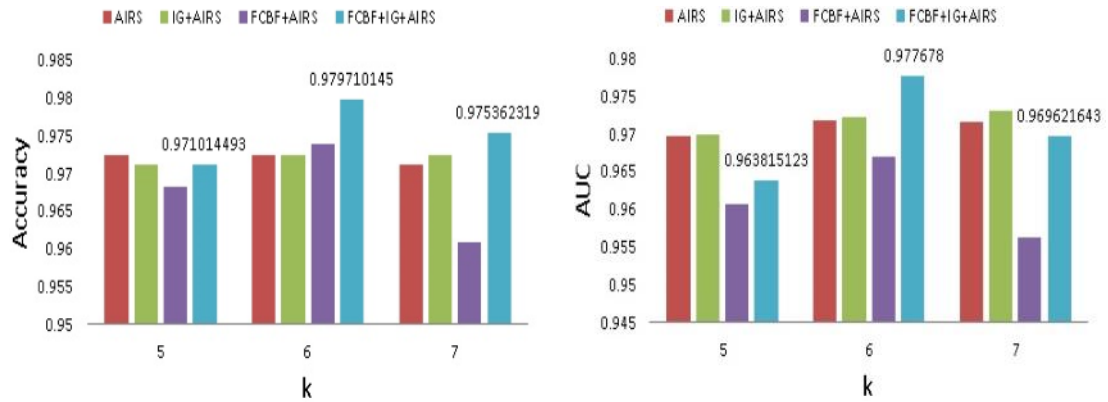
Figure 4. Comparison of average AUC

The average numbers of mc generated for all methods are linear with the number of seeds specified as shown in Figure 5(a). Conversely, the determination of the number of seeds in all methods is inversely proportional to the time for training as pointed out in Figure 5(b). From Figures 3 and 4 it can be seen that the highest performance is achieved when seed=50. Furthermore, a deeper investigation on seed=50 was performed to determine the value of k which produces the most optimal performance. The highest performance in terms of accuracy and AUC is achieved when k=5 with accurac=0.9797 and AUC=0.9777 as shown in Figures 6(a) and 6(b) respectively.



(a) (b)

Figure 5. (a) Average numbers of mc and (b) average training time



(a) (b)

Figure 6. (a) The best accuracy and (b) the best AUC

5. CONCLUSION

This research has tried to use a new hybrid feature selection by combining FCBF and IG for identifying breast cancer disease using AIRS algorithm. Experimentation on different seed with 10-fold cross-validation test yields the best performance at k=6 and seed=50 with accuracy=0.9797 and AUC=0.9777.

REFERENCES

[1] M. Ghoncheh, Z. Pournamdar, and H. Salehiniya, "Incidence and Mortality and Epidemiology of Breast Cancer in the World," *Asian Pacific Journal of Cancer Prevention (APJCP)*, vol. 17, pp. 43-46, 2016.

[2] A. Watkins, J. Timmis, and L. C. Boggess, "Artificial Immune Recognition System (AIRS): AN Immune Inspired Supervised Machine Learning Algorithm," *Genetic Programming and Evolvable Machines*, vol. 5, no. 3, pp. 291-317, 2004.

[3] J. Brownlee, "Clever Algorithm -Nature-Inspired Programming Recipes*," First Edit. Lulu Enterprises*, 2011.

[4] M. A. Chikh, M. Saidi, and N. Settouti, "Diagnosis of Diabetes Diseases Using An Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest neighbor," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2721-2729, 2012.

[5] T. M. Mohamed, "Efficient breast cancer detection using sequential feature selection techniques," *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, pp. 458-464, 2015.

[6] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," *Data Classif. Algorithms Appl.*, pp. 37-64, 2014.

[7] H. Omara, *M. Lazaar, and Y. Tabii, "Effect of Feature Selection on Gene Expression Datasets Classification Accurac," International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3194-3203, 2018.

[8] O. Cigdem and H. Demirel, "Performance analysis of different classification algorithms using different feature selection methods on Parkinson's disease detection," *Journal of Neuroscience Methods*, vol. 309, pp. 81-90, 2018.

[9] K. Polat, S. Sahan, H. Kodaz, and S. Günes, "A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)," *International Conference on Natural Computation*, vol. 3611, pp. 830-838, 2005.

[10] K. Polat and S. Güneş, "Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm," *Expert Systems with Applications*, vol. 34, no. 1, pp. 773-779, 2008.
[11] A. Ridok, F. M. Wayan, and M. Rifai, "An Improved Artificial Immune Recognition System with Fast Correlation Based Filter (FCBF) for Feature Selection," *2017 Fourth International Conference on Image Information Processing (ICIIP)*, Shimla, pp. 1-6, 2017.
[12] H. Kodaz, S. Kara, F. Latioğlu, S. Güneş, F. Latioğlu, and S. Güneş, "A new hybrid classifier system: Information gain-based artificial immune recognition system," *Experimental Techniques*, vol. 31, no. 6, pp. 36-43, 2007.
[13] H. Kodaz, S. Özşen, A. Arslan, and S. Güneş, "Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease," *Expert Systems with Applications*, vol. 36, no. 2, part 2, pp. 3086-3092, 2009.
[14] S. Kara, B. H. Aksebzeci, H. Kodaz, S. Güneş, E. Kaya, and H. Özbilge, "Medical application of information gain-based artificial immune recognition system (IG-AIRS): Classification of microorganism species," *Expert Systems with Applications*, vol. 36, no. 3, part 1, pp. 5168-5172, 2009.
[15] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 4, pp. 1880-1888, 2016.
[16] K. Polat and S. Güneş, "A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS," *Computer Methods and Programs in Biomedicine*, vol. 88, no. 2, pp. 164-174, 2007.
[17] C. D. Katsis, I. Gkogkou, C. A. Papadopoulos, Y. Goletsis, and P. V. Boufounou, "Using Artificial Immune Recognition Systems in Order to Detect Early Breast Cancer," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 5, no. 2, pp. 34-40, 2013.
[18] K. Polat, S. Şahan, and S. Güneş, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1141-1147, 2007.
[19] S. Shamshirband et al., "Tuberculosis disease diagnosis using artificial immune recognition system," *International Journal of Medical Sciences*, vol. 11, no. 5, pp. 508-514, 2014.
[20] M. Hossin and M. N. Sulaiman, "Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 1-11, 2015.
[21] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299-310, 2005.
[22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.
[23] C. Catal, "Performance evaluation metrics for software fault prediction studies," *Acta Polytechnica Hungarica*, vol. 9, no. 4, pp. 193-206, 2012.
[24] L. Yu and H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution," *Proceedings of the twentieth international conference on machine learning*, Washington DC, pp. 1-8, 2003.
[25] A. Rakotomamonjy, "Optimizing Area Under Roc Curve with SVMs," *ROCAI*, pp. 71-80, 2004.
[26] C. X. Ling, J. Huang, and H. Zhang, "AUC: A better measure than accuracy in comparing learning algorithms," *Conference of the Canadian Society for Computational Studies of Intelligence*, vol. 2671, pp. 329-341, 2003.
[27] C. X. Ling, J. Huang, and H. Zhang, "AUC: A statistically consistent and more discriminating measure than accuracy," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 519-524, 2003.

BIOGRAPHIES OF AUTHORS

Achmad Ridok is a doctoral student at the department of Biologi, Faculty of Mathematics and Natural Science, Brawijaya University. He obtained his Bachelor degree in Matematics from department of Mathematic, Faculty of Mathematics and Natural Science, Brawijaya University. He received his M.Kom. in Computer Science from the University of Indonesia. His research interests are machine learning, artificial intelligence and biominformatics.

Widodo, Ph.D is professor at Biology Department, Brawijaya University Indonesia. He now active research on elucidation the molecular mechanism of JAMU (Indonesia traditional medicine) by using Metabolomic, Metagenomic and Bioinformatic approach. Since 2019, He was appointed by Directorate general of higher education for conduct world class research to elucidate the complexity of JAMU. Also, he active as board member of International Association for Comparative Medicine (IACM). He got several awards such as, young investigator from Tissue Culture Association-japan (2007), Best original paper from nestle council research-Japan (2008), Young scientist from Kalbe-Ristek Dikti-Indonesia (2012) and Best teacher from Brawijaya University (2014). He has experience as visiting Professor at Ristumeikan University-Japan (20017-2019).

Wayan Firdaus Mahmudy obtained bachelor degree in mathematics from Universitas Brawijaya, Indonesia, master degree in information technology from Institut Teknologi Sepuluh November (ITS), Indonesia, and completed his Ph.D. in Manufacturing Engineering at University of South Australia. He is a Lecturer at Department of Computer Science, Universitas Brawijaya. His research interests include optimization of combinatorial problems and machine learning.

Muhaimin Rifa'i, Ph.D is a Professor of Immunology in Department of Biology, Brawijaya University, Indonesia. He is now actively conducting research using herbal medicine. He is a scientist who is active in conducting research and scientific publications in various international journals including journals that have a high reputation. Currently he is conducting research related to cancer, diabetes mellitus, and atherosclerosis. As a researcher he also gives lectures at the undergraduate, masters, and doctoral degree. After holding the position as head of the Physiology Laboratory, he is currently become the head of Biology Department, Brawijaya University, Indonesia.