Scientific Journal of Informatics
Vol. 11, No. 1, Feb 2024
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Knowledge Discovery from Confusion Matrix of Pruned CART in Imbalanced Microarray Data Ovarian Cancer Classification

Ni Kadek Emik Sapitri[1*], Umu Sa'adah[2], Nur Shofianah[3]
[1, 2, 3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia

Abstract.
Purpose: The results of microarray data analysis is important in cancer diagnosis, especially in early stages asymptomatic cancers like ovarian cancer. One of the challenges in analyzing microarray data is the problem of imbalanced data. Unfortunately, research that carries out cancer classification from microarray data often ignores this challenge, so that it doesn't use appropriate evaluation metrics. It makes the results biased towards the majority class. This study uses a popular evaluation metric "accuracy" and an evaluation metric that is suitable for imbalanced data "balanced accuracy (BA)" to gain information from the confusion matrix regarding accuracy and BA values in case of ovarian cancer classification.
Methods: This study use Classification and Regression Tree (CART) as the classifier. CART optimized by pruning. CART optimal is determined from the results of CART complexity analysis and confusion matrix.
Results: The confusion matrix and CART interpretations in this research show that CART with low complexity is still able to predict majority class respondents well. However, when none of the data in the minority class was classified correctly, the accuracy value was still quite high, namely 86.97% and 88.03% respectively at the training and testing stages, while the BA value at both stages was only 50%.
Novelty: It is very important to ensure that the evaluation metrics used match the characteristics of the data being processed. This research illustrates the difference between accuracy and BA. It concluded that that classification of an imbalanced dataset without doing resampling can use BA as evaluation metric, because based on the results, BA is more fairly to both classes.

Keywords: Imbalanced microarray data, Ovarian cancer, Confusion matrix, Evaluation metrics, CART
Received January 2024 / **Revised** February 2024 / **Accepted** February 2024

INTRODUCTION

Microarray is a widely utilized technique in the identification of cancer cells, involving the examination of DNA proteins to facilitate subsequent gene analysis [1]. Scientists have used microarray data to distinguish between individuals in good health and patients with variuos types of cancer [2]. Microarray data is organized into a matrix called the gene expression matrix [1]. Gene expression from cancer cells and healthy cells was obtained using microarray technology [3]. The results of microarray data analysis can play an important role in diagnosis, prognosis and treatment planning, for example in cancer identification [4].

One way to identify cancer from microarray data is by classification. Classification is an analysis technique that is widely applied in microarray studies which aim to study differences in gene expression in various types of diseases, including cancer [5]. However, classification will face challenges when dealing with microarray data.

Microarray data naturally presents an imbalanced class distribution (imbalanced data) with samples of a certain class (majority class) far more than samples of another class (minority class) [6]. Imbalanced data can affect the effectiveness of machine learning models because it results in results that are biased towards the majority class [7]. The strategy to overcome the problem of imbalanced data is doing resampling so that the data is balanced or simply using evaluation metrics that are not sensitive to imbalanced data. One

*Corresponding author.
Email addresses: emikpitri@gmail.com (Sapitri), u.saadah@ub.ac.id (Sa'adah), nur_shofianah@ub.ac.id (Shofianah)
DOI: 10.15294/sji.v11i1.50077

evaluation metric for evaluating classification results that is insensitive to imbalanced class distribution is balanced accuracy (BA) [8].

Unfortunately, research that classifies cancer from microarray data often ignores the problem of imbalanced data. As a result, there are several studies that do not consider appropriate evaluation tools for cancer classification from microarray data, including [2], [9], [10], [11], and [12].Those five studies used accuracy as a measuring tool. In fact, accuracy can produce overly optimistic results on imbalanced data [13]. In other words, accuracy is sensitive to imbalanced data.

Accuracy and BA are both calculated using the information contained in the confusion matrix. The confusion matrix summarizes the results of the classification, namely the amount of data that has been classified correctly and that has not been classified correctly. This study aims to explore information from the confusion matrix regarding accuracy and BA values in cancer classification cases from microarray data, and also illustrates the difference between accuracy and BA.

The type of cancer chosen as a case study in this research is ovarian cancer. Ovarian cancer is the most dangerous of all types of cancer that attacks the female reproductive organ system [14]. In 2020, Harsono reported that only 20% of ovarian cancer was diagnosed at stage 1 (early) when the disease is limited to the ovaries. In fact, 90% of patients in the early stages respond well to existing therapy [15]. In addition, the American Cancer Society estimates that there will be 19710 new cases of ovarian cancer in 2023 and 13270 people are expected to die from this disease [16] and it is estimated that the death rate due to ovarian cancer in 2040 will significantly increase [15].

The classifier used in this research is Classification and Regression Tree (CART). CART is a decision tree algorithm proposed by Breiman [17]. CART was chosen considering that the purpose of classification is not only to obtain high accuracy, but also to seek new knowledge from the classification results [5]. CART has the advantage of being interpretable. Research [2] uses CART as a breast cancer data classifier. The interpretation results obtained indicate that the genes selected to build CART are closely related to breast cancer growth based on several related studies in the medical field. Apart from that, CART has advantages compared to other decision tree algorithms such as Iterative Dichotomizer 3 (ID3) and C4.5 because CART is able to handle data outliers [18].

Tuan et al.'s research [19] stated that cost-complexity pruning (ccp) plays a role in reducing the overfitting phenomenon in tree-based algorithms. Overfitting is a phenomenon when a model performs very well on training data, but poorly on test data [20]. The amount of ccp in CART can be set with the ccp_alpha parameter. Therefore, CART in this study was optimized by pruning by setting the ccp_alpha parameter in the decision tree formed.

It is hoped that this research can be used as a reference regarding confusion matrices and evaluation tools related to classification cases from microarray data that have an imbalanced class distribution. It is hoped that the interpretation of CART in this study will make it easier for readers to understand the relationship between CART complexity and classification results.

METHODS

Before explain the data analysis steps, we give some brief reviews about CART, confusion matrix, and evaluation metrics.

Classification and Regression Tree (CART)

The predictor variables in microarray data all contain continuous data. This explanation will focus on the CART algorithm for predictors containing continuous data. The splitting criterion for creating tree branches used in the CART algorithm is Gini impurity. The main components of a decision tree model are nodes and branches, and the most important steps in building a model are splitting, stopping, and pruning [21]. There are three types of nodes, including root nodes, inner nodes or internal nodes, and leaf nodes.

The following explains the CART algorithm which is a summary of [17] and [22]. Let D be a node, $K = 1,2, \dots, k$ represents the number of classes, and p_K is the proportion of class K observations in node D. The Gini impurity of D which is denoted $Gini(D)$ is expressed by the following equation:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2. \quad (1)$$

After getting the Gini impurity from D, the total Gini impurity value is calculated if D is partitioned binary by a sorter s into D_{left} and the D_{right} expressed by the weighted average. The formula for calculating total Gini impurity is listed in Equation (2).

$$Gini_s(D) = \frac{|D_{left}|}{|D|} Gini(D_{left}) + \frac{|D_{right}|}{|D|} Gini(D_{right}), \quad (2)$$

where $|D_{left}|$ states many observations on the left node, $|D_{right}|$ states many observations on the right node, and $|D|$ is the number of observations. Counting $Gini_s(D)$ carried out for each candidate split point on each variable. The best dividing point is the one that has lowest $Gini_s(D)$.

The CART algorithm for classification of predictor variables containing continuous data is briefly presented as follows:

Input: predictor variables $X_j = (x_{1j}, x_{2j}, x_{3j}, \dots, x_{nj})^T$ and response variable (Y), where $j = 1,2,3, \dots, m$.

- For all j:
  1. Select X_j to perform the split.
  2. Sort the observation values on X_j from smallest to largest. Let the sorted values written as $\{x_{sj} | s = 1,2,3, \dots, n\}$.
  3. For all x_{sj} where $s \neq n$:
     a. Calculate the midpoint $md_s = x_{sj} + \frac{(x_{(s+1)j} - x_{sj})}{2}$.
     b. Set the md_s as a candidate split point.
     c. Create the left node and the right node.
     d. Calculate the Gini impurity of the left and right nodes using Equation (1).
     e. Calculate the total Gini impurity of md_s using Equation (2).
  4. Set the midpoint md_s that has the smallest total Gini impurity as the splitting point.
- Set the predictor variable that has the smallest total Gini impurity as the root node.
- Exclude selected predictor variable from list of predictor variables.
- Repeat the process for remaining predictor variables until all leaves are formed.

Output: maximum tree.

To reduce the maximum complexity of the tree, pruning is carried out. The decision tree obtained after the pruning process is called the optimal decision tree or in this case optimal CART.

Confusion matrix and evaluation metrics

The classifier estimates the class of each data sample, groups it into labels in the target class, so that at the end of the classification procedure each sample falls into one of four cases [13]. The four cases are summarized in the confusion matrix. An example of the confusion matrix display can be seen in Table 1.

Table 1. Confusion matrix

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

In the confusion matrix, TP and TN represent data that has been classified correctly or in other words corresponds to the original data, while FP and FN represent data that has not been classified correctly. In the case of cancer classification, the meaning can be explained as follows:

1. TP states the event when data on people with cancer is predicted correctly;
2. TN states the event when data from people who do not have cancer are predicted correctly;
3. FP states an incident when data on a person who does not have cancer, but the classification results instead state that the person has cancer;
4. FN states that the data occurs when a person has cancer, but the classification results state that the person does not have cancer.

The TP, TN, FP, and FN values can be used to calculate evaluation metrics. There are various types of evaluation metrics for classification cases. Some of them are accuracy, TP rate, FP rate, precision, recall, Matthews correlation coefficient [23], and balanced accuracy [24]. The evaluation metrics used in this research are accuracy and balanced accuracy (BA). Accuracy can be formulated in Equation (3) [13] and BA can be formulated in Equation (4) [24].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3)$$

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \quad (4)$$

Data analysis steps

This research carries out cancer classification from microarray data with Pruned CART, namely CART that is optimized by pruning the decision tree. The stages of data analysis are summarized in Figure 1. The data analysis steps can be described as follows:

1. Download the dataset. The data used in this research is the OVA_ovary dataset. This data is microarray data about ovarian cancer [25] and can be downloaded via the OpenML website with the link: https://www.openml.org/search?type=data&status=active&id=1166.
2. Data preprocessing, including data extraction, deleting irrelevant columns, adjusting data types, data scaling with minmax normalization [26], and data splitting with a proportion of training data and test data of 80%:20%.
3. Classification with CART.
4. CART optimization by pruning the tree according to the ccp_alpha value.
5. Print the results of the confusion matrix for the training and testing stages for each ccp_alpha values.
6. Evaluate the model based on accuracy and BA formula.

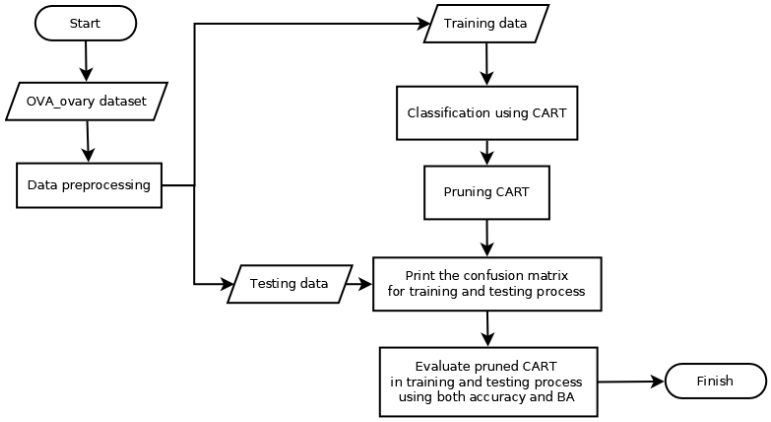


Figure 1. Flowchart of pruned CART modelling

This research uses Python programming language starting from the data preprocessing stage to the final stage. The software used is JupyterLab version 3.6.3. The hardware used is a laptop with 13th generation Intel Core i7 processor and has dual-channel 8GB RAM (total RAM 16GB).

CART is formed with the 'DecisionTreeClassifier' package from the 'sklearn.tree' library. This library can form decision tree using various algorithms. To use CART algorithm that using Gini impurity as splitting criteria, we set the parameters criterion = 'gini' and splitter = 'best'. The cost-complexity pruning parameter (in Python it called ccp_alpha) is searched with 'cost_complexity_pruning_path' in the 'DecisionTreeClassifier' which uses the minimum cost-complexity pruning technique.

RESULTS AND DISCUSSIONS

Result of data preprocessing

The OVA_ovary dataset file has the format ".arff" or Attribute-Relation File Format. The file consists of metadata and dataset. After going through the extraction process with JupyterLab, information was obtained

that the available dataset consisted of 10937 columns and 1545 observations (rows). These columns include 1 ID_REF column, 10935 gene columns, and 1 Tissue (target) column according to Table 2.

Table 2. OVA_ovary dataset after data extraction process

ID_REF	1007_s_at	121_at	...	AFFX-ThrX-M_at	Tissue
117704	3196.7	3844.8	...	1094.5	Other
301664	3532.6	397.9	...	612.1	Other
203673	5109.7	563.7	...	1578.4	Other
⋮	⋮	⋮	⋮	⋮	⋮
277715	7334.8	660.9	...	588	Ovary
179866	4225.5	1125.5	...	1306.2	Ovary

There are 10935 genes that act as independent variables with the Tissue column act as a dependent variable consisting of 2 classes. Tissue contains 2 classes, namely 'Other' (not ovarian cancer sufferers) which is the majority class and 'Ovary' (ovarian cancer sufferers which are a minority class. OVA_ovary dataset after going through data preprocessing stage has a value range of [0, 1].

This research use 'train_test_split' from 'sklearn.model_selection' library to randomize the data splitting process. The proportion of training data and testing data of this research is 80%:20% as in [27] which use same dataset. After splitting, there are 1236 rows on training data and 309 rows on testing data. The composition of 'Ovary' and 'Other' classes before and after splitting process can be seen on Table 3.

Table 3. Class distribution of OVA_ovary dataset

	Ovary	Other
Original data (before splitting)	198	1347
Training data	161	1075
Testing data	37	272

Insights from the CARTs and confusion matrix elements

Table 4 summarizes the data contained in the confusion matrix in the training and testing stage of the OVA_ovary dataset. In Table 4, iteration 0 represents the maximum tree. The tree continues to be pruned at each iteration until only the root node is left at the 18^{th} iteration.

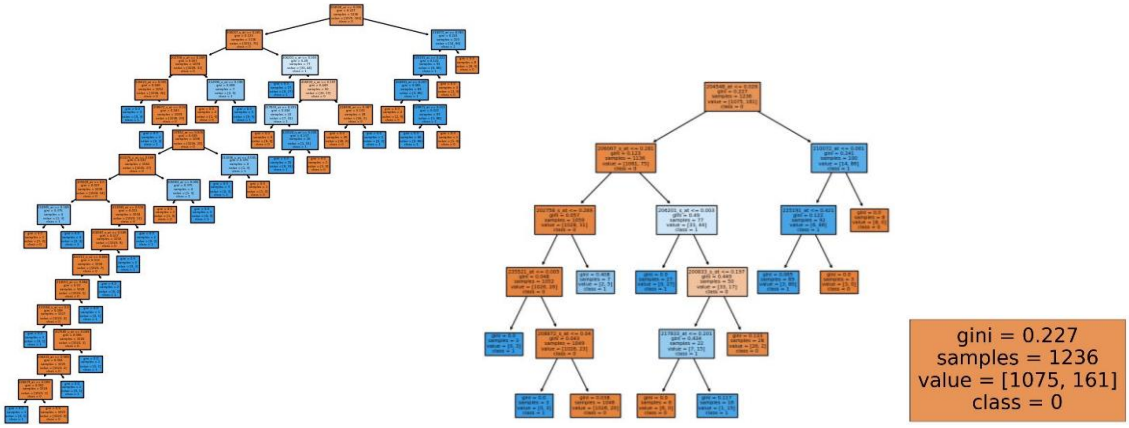
The ccp_alpha value $= 0$ meaning maximum tree, or in other words the CART has maximum number of nodes (Figure 2a). As the ccp_alpha increase, there are more nodes that have been pruned. So, lower ccp_alpha produce more complex CART.

TP values in the training and testing stage tend to decrease as the ccp_alpha value increases. This means that the more complex the CART form, the more data on ovarian cancer patients can be detected. In iteration 0 of the testing stage, it can be seen that the FN and FP results are still quite high. In fact, in the training process both have a value of zero. This indicates that the maximum tree is not able to generalize the data well at the testing stage. So, the optimization process needs to be carried out.

Table 3. The confusion matrix values of CART on training and testing stages

Iteration	ccp_alpha	Training results				Testing results			
		TP	TN	FP	FN	TP	TN	FP	FN
0	0	161	1075	0	0	28	251	21	9
1	0.001213592	161	1072	3	0	29	250	22	8
2	0.00151699	161	1071	4	0	29	250	22	8
3	0.001599524	161	1070	5	0	29	250	22	8
4	0.001610253	156	1070	5	5	29	251	21	8
5	0.002311604	156	1068	7	5	29	251	21	8
6	0.003005086	154	1068	7	7	28	251	21	9
7	0.003091215	154	1066	9	7	28	251	21	9
8	0.003186245	148	1066	9	13	28	252	20	9
9	0.003497633	139	1069	6	22	25	262	10	12
10	0.004384821	139	1066	9	22	27	262	10	10
11	0.00464387	133	1066	9	28	27	263	9	10
12	0.005350411	128	1068	7	33	25	264	8	12
13	0.006205869	128	1062	13	33	25	263	9	12
14	0.007427395	113	1069	6	48	25	267	5	12

15	0.010406641	113	1061	14	48	25	265	7	12
16	0.012357836	130	1028	47	31	26	260	12	11
17	0.034140344	86	1061	14	75	22	266	6	15
18	0.093753886	0	1075	0	161	0	272	0	37



(a) (b) (c)

Figure 2. CART illustration on 0 iteration (a), 9th iteration (b), and last iteration (c) training stage

Figure 2 presents an illustration of CART in several iterations of the training stage to understand the changes in tree complexity produced in each iteration. In Figure 2, 'gini' represents the total Gini impurity at that node. 'Samples' represents the number of samples at that node. Because there are 1236 rows of training data, the 'sample' at the root node (Figure 2c) has a value of 1236. Furthermore, 'value' contains the number of samples in each class, and 'class' states the dominant class at that node. At the root node, the dominant class is class 0 because the initial data is imbalanced and class 0 is the majority class. Based on Figure 2, information can be obtained that the greater the ccp_alpha value, the more branches will be cut. As a result, the number of nodes in CART decreases.

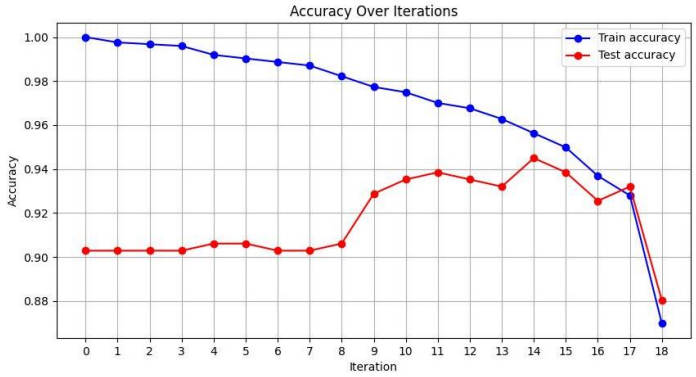


Figure 3. The changes of accuracy over iterations

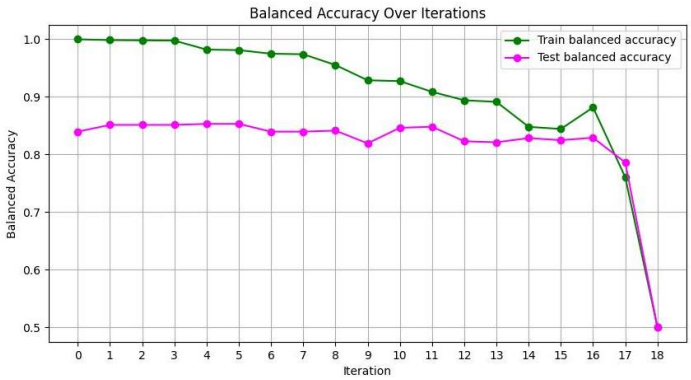


Figure 4. The changes of BA over iterations

Based on Table 4, the TN values are quite fluctuating, but in the last two iterations the values are high. Thus, CART with low complexity is still able to predict 'respondents who do not suffer from ovarian cancer well. However, it should be noted that the class 'not suffering from ovarian cancer' or class 0 (Other), is the majority class.

To choose optimal ccp_alpha values, this research considering the results of confusion matrix and preventing overfitting. Overfitting is a phenomenon when a model performs very well on training data, but poorly on test data [20]. This phenomenon can be seen on the difference between training and testing results.

In the last two iterations, TP values was low. Even though the highest accuracy and BA values in the testing stage fell on the 17th iteration, this research did not choose the CART in that iteration as the optimal CART. With the same perspective, CART in the 18th iteration was also not chosen as the optimal CART, even though the respective differences in accuracy and BA values at the training stage and the testing stage were the lowest. Since the second lowest difference of accuracy is on 17th iteration, this study chose CART with the difference in BA values at the training stage and the testing stage is the second lowest. In other word, the optimal CART is the CART in the 15th iteration and the optimal ccp_alpha value is 0.010406641.

Figure 3 and Figure 4 show the changes in accuracy and BA values over iterations. Illustrations of the optimal CART can be seen in Figure 5. The first line at each root node and terminal nodes in Figure 5 contains the gene feature used as the splitting criterion.

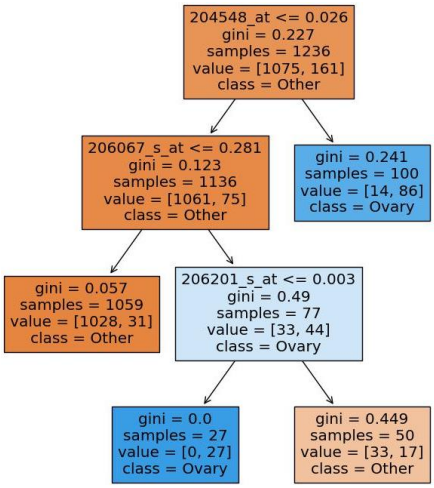


Figure 5. Optimal CART

Insights from the evaluation metrics

Chicco and Jurman in [13] state that accuracy can produce overly optimistic results on imbalanced data. It can be said that accuracy is sensitive to imbalanced data. This study illustrates this by using two evaluation metrics, namely accuracy and BA in ovarian cancer classification cases from the OVA_ovary dataset.

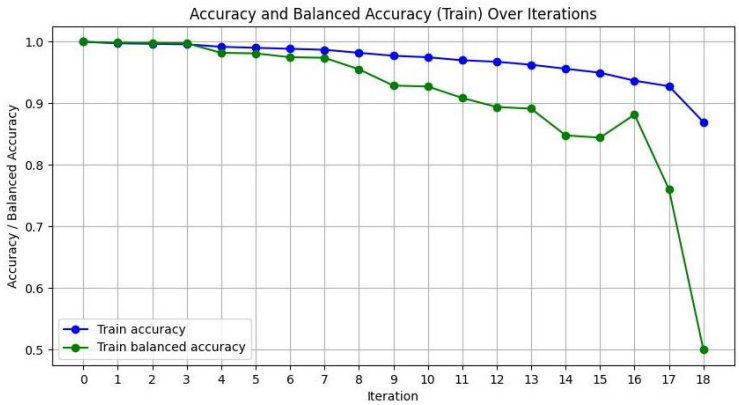


Figure 6. The changes in accuracy and BA values at training stage

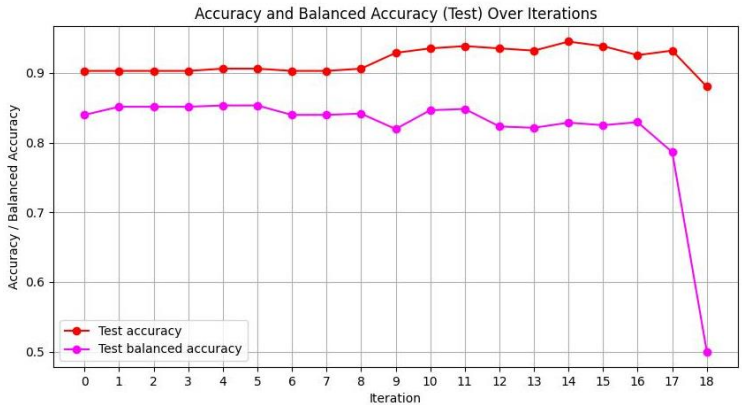


Figure 7. The changes in accuracy and BA values at testing stage

Figure 6 and Figure 7 present the differences in the results of two evaluation metrics at the training and testing stages. When maximum CART works too well on the training data (iteration 0), the accuracy results obtained are the same as BA because all the data can be classified correctly. Referring to Table 4, Figure 6, and Figure 7, differences in accuracy and BA values begin to appear when at least one observation that is not classified correctly.

The most striking difference is seen in 18th iteration. In the 18th iteration, none of the cancer positive patients were classified correctly (refer to the TP value = 0 in Table 4). However, the accuracy value is still quite high, namely 86.97% at the training stage and 88.03% at the testing stage, while the BA value is only 50%. Based on Equation (4), BA treats positive patients class (Ovary) and negative patients class (Other) equally. The 50% on BA value means that only values in one class classified correctly. In this case, all values on 'Other' class classified correctly (refer to TN values that has same values of training and testing data in Other class in Table 3), but all data in the 'Ovary' class not classified correctly (all TP values classified as FN). This is in accordance with the statement in [13] which states that the accuracy results are too optimistic, so that the results biased on the majority class (class 0). In contrast, BA can worked more fairly. When all the majority classes were classified correctly, but none of the minority classes were classified correctly, the BA score was only 50%. Therefore, it is very important to ensure that the evaluation metrics used are appropriate to the characteristics of the data being processed.

Discussion

The opportunities for further research are still wide open because the OVA_ovary dataset has not been widely used in research related to ovarian cancer classification. Several things can be done, for example analyzing differences in BA results from CART when using different data scaling techniques, using feature selection methods (like Least Absolute Shrinkage and Selection Operator (LASSO) [28] or others) before doing classification, or using resampling techniques such as oversampling and undersampling. In other perspective, it is also interesting to analyze the differences between accuracy and BA values from many imbalanced datasets, just like the simulation conducted in [5].

CONCLUSION

The results of the confusion matrix show that the more complex the CART form, the more data on ovarian cancer patients (minority class) that are successfully detected. However, this is an indication of overfitting because the maximum tree is not able to generalize the data well at the testing stage so optimization needs to be carried out. On the other hand, CART with low complexity is still able to predict respondents who do not suffer from ovarian cancer (majority class) well. In this research, the optimal CART is the CART in the 15th iteration. The accuracy value is 94.98% at the training stage and 93.85% at the testing stage. The BA value was 84.44% at the training stage and 82.49% at the testing stage.

This research shows that accuracy and BA produce different values when not all data classified correctly. The most striking difference was seen in the 18th iteration when none of the data in the minority class was classified correctly, but the accuracy value was still quite high, namely 86.97% at the training stage and 88.03% at the testing stage, while the BA value was only 50%. The accuracy value is quite high because

the majority of all classes are classified correctly. So, the classification of an imbalanced dataset without doing resampling can use BA as evaluation metric because BA can do more justice to both classes.

REFERENCES

[1] H. Almazrua and H. Alshamlan, "A Comprehensive Survey of Recent Hybrid Feature Selection Methods in Cancer Microarray Gene Expression Data," *IEEE Access*, vol. 10, pp. 71427–71449, 2022, doi: 10.1109/ACCESS.2022.3185226.

[2] M. Y. Rochayani, U. Sa'adah, and A. B. Astuti, "Two-stage Gene Selection and Classification for a High-Dimensional Microarray Data," *J. Online Inform.*, vol. 5, no. 1, pp. 9–18, 2020, doi: 10.15575/join.v5i1.569.

[3] E. Lotfi and A. Keshavarz, "Gene expression microarray classification using PCA–BEL," *Comput. Biol. Med.*, vol. 54, pp. 180–187, 2014, doi: 10.1016/j.compbiomed.2014.09.008.

[4] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method," *Artif. Intell. Med.*, vol. 123, p. 102228, 2022, doi: doi.org/10.1016/j.artmed.2021.102228.

[5] M. Y. Rochayani, U. Sa'adah, and A. B. Astuti, "Simulation Study of Imbalanced Classification on High-Dimensional Gene Expression Data," *Sci. J. Informatics*, vol. 10, no. 1, pp. 45–54, 2023, doi: 10.15294/sji.v10i1.40589.

[6] Y. He, J. Zhou, Y. Lin, and T. Zhu, "A class imbalance-aware Relief algorithm for the classi fi cation of tumors using microarray gene expression data," *Comput. Biol. Chem.*, vol. 80, pp. 121–127, 2019, doi: 10.1016/j.compbiolchem.2019.03.017.

[7] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, "Evolutionary Machine Learning: A Survey," *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–35, 2021, doi: 10.1145/3467477.

[8] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," *arXiv*, pp. 1–17, 2020.

[9] N. A. Al-thanoon, O. S. Qasim, and Z. Y. Algamal, "Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification," *Comput. Biol. Med.*, vol. 103, pp. 262–268, 2018, doi: 10.1016/j.compbiomed.2018.10.034.

[10] T. N. Nuklianggraita, Adiwijaya, and A. Aditsania, "On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier," *Infotel*, vol. 12, no. 3, pp. 89–96, 2020, doi: https://doi.org/10.20895/infotel.v12i3.485 89.

[11] A. M. Alharthi, M. H. Lee, and Z. Y. Algamal, "Gene selection and classification of microarray gene expression data based on a new adaptive L1 -norm elastic net penalty," *Informatics Med. Unlocked*, vol. 24, p. 100622, 2021, doi: doi.org/10.1016/j.imu.2021.100622.

[12] G. Roffo, S. Melzi, U. Castellani, A. Vinciarelli, and M. Cristani, "Infinite Feature Selection: A Graph-based Feature Filtering Approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4396–4410, 2021, doi: 10.1109/TPAMI.2020.3002843.

[13] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, pp. 1–13, 2020.

[14] C. Slatnik and E. Duff, "Ovarian cancer: Ensuring early diagnosis," *Nurse Pract.*, vol. 40, no. 9, pp. 47–54, 2015, doi: 10.1097/01.NPR.0000450742.00077.a2.

[15] A. B. Harsono, "Kanker Ovarium: 'The Silent Killer,'" *Indones. J. Obstet. Gynecol. Sci.*, vol. 3, no. 1, pp. 1–6, 2020.

[16] National Cancer Institute, "SEER Cancer Stat Facts: Ovarian Cancer," 2022.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall, 1984.

[18] S. Singh and P. Gupta, "Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A Survey," *Int. J. Adv. Inf. Sci. Technol.*, vol. 27, no. 27, pp. 97–103, 2014.

[19] N. M. Tuan, huynh T. K. Chi, and N. Van Hop, "A Hybrid Machine Learning Approach in Predicting E-Commerce Supply Chain Risks," in *14th International Conference on Knowledge and Systems Engineering (KSE)*, 2022, pp. 1–6. doi: doi: 10.1109/KSE56063.2022.9953787.

[20] G. Kunapuli, *Ensemble Methods for Machine Learning*, 6th ed. New York: Manning Publications, 2022.

[21] Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015, doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215044.

[22] U. Saadah, M. Y. Rochayani, D. W. Lestari, and D. A. Lusia, "Pohon Keputusan," in *Kupas Tuntas*

Algoritma Data Mining dan Implementasinya Menggunakan R, Malang: UBPress, 2021, pp. 143–168.
[23] Ž. Đ. Vujović, "Classification Model Evaluation Metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 1–8, 2021, doi: 10.14569/IJACSA.2021.0120670.
[24] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, pp. 1–22, 2021.
[25] G. Stiglic and P. Kokol, "Stability of Ranked Gene Lists in Large Microarray Analysis Studies," *J. Biomed. Biotechnol.*, vol. 2010, 2010.
[26] X. Tang, S. X. D. Tan, and H. Chen, "SVM Based Intrusion Detection Using Nonlinear Scaling Scheme," in *4th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2018, pp. 1–4. doi: doi: 10.1109/ICSICT.2018.8565736.
[27] U. Sa'adah, M. Y. Rochayani, and A. B. Astuti, "Knowledge discovery from gene expression dataset using bagging lasso decision tree," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 2, pp. 1151–1159, 2021, doi: 10.11591/ijeecs.
[28] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.