

Hybrid methods to identify ovarian cancer from imbalanced high-dimensional microarray data

Ni Kadek Emik Sapitri, Umu Sa'adah, Nur Shofianah

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received Mar 13, 2024

Revised Nov 1, 2024

Accepted Nov 14, 2024

Keywords:

Balanced accuracy

CART

Hybrid method

Infinite feature selection

Microarray data

Ovarian cancer

ABSTRACT

Scientists have used microarray data to identify healthy people and patients with various types of cancer, including ovarian cancer. Ovarian cancer is the most dangerous of all types of cancer that attacks the female reproductive organ. The right combination of methods is needed to identify ovarian cancer from microarray data because that type of data is high-dimensional and imbalanced. This research aims to propose two hybrid methods which are a combination of infinite feature selection (IFS) as features selector with classification and regression tree (CART) as a classifier. IFS can work with two separate scenarios, namely supervised infinite feature selection (SIFS) and unsupervised infinite feature selection (UIFS). This research also compares the performance of the two hybrid methods proposed (SIFS-CART and UIFS-CART) with CART without IFS. The data used is OVA_ovary that has 10937 columns and 1545 rows. The results shows that SIFS-CART achieves maximum performance using 1000 features and UIFS-CART 5000 features. CART without IFS uses all 10935 features. The balanced accuracy results show SIFS-CART can outperform CART without IFS and UIFS-CART. Using less features to get highest balanced accuracy results, SIFS is more effective in performing feature selection on the OVA_ovary dataset compared to UIFS.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ni Kadek Emik Sapitri

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Brawijaya University

Malang City, East Java, Indonesia

Email: emikpitri@gmail.com

1. INTRODUCTION

Global Cancer Statistics 2022 stated that there were almost 20 million new cases of cancer and predicted that the annual number of new cases of cancer can reach 35 million by 2050 [1]. Some studies stated that several types of cancer, such as breast cancer [2], lung cancer [3], pancreatic cancer [4], and ovarian cancer [5] are asymptomatic in the early stages. Ovarian cancer is the most dangerous of all types of cancer that attacks the female reproductive organ system [6]. The incidence of ovarian cancer has increased significantly among young women and in low-income countries [7]. Žilovič *et al.* [8] stated that some current diagnostic tools utilized in clinical practice are not have enough sensitivity and specificity for detecting ovarian cancer in its early stages, due to the lack of early symptoms and an effective screening strategy for asymptomatic populations. Therefore, methods that can identify asymptomatic cancer, especially ovarian cancer, are urgently needed.

Scientists have used microarray data to identify healthy people and patients with various types of cancer [9], including ovarian cancer. Microarray is a famous method for identifying cancer cells [10]. Machine learning-based methods can be used to identify cancer from microarray data, for example research by Rochayani *et al.* [9]. Computational analysis with machine learning can reveal hidden patterns that not

many doctors know about [11]. However, there are two main challenges to identify cancer from microarray data using machine learning. First, microarray data contains thousands or even tens of thousands of features with a small number of observations, so it is classified as high-dimensional data [9]. In other words, microarray data has a greater number of features (columns) than the number of observations (rows). High-dimensional data can cause overfitting and have a negative impact on the accuracy of classification algorithms used in machine learning [12]. Overfitting is a phenomenon when a machine learning model performs very well on training data, but poorly on test data [13]. Second challenge is that microarray data naturally presents an unbalanced class distribution (imbalanced data) with samples of a certain class (majority class) much more than samples of another class (minority class) [14]. Imbalanced data can affect the effectiveness of machine learning models because it results in results that are biased towards the majority class [15]. Therefore, the methods that can identify ovarian cancer from microarray data need to overcome both challenges.

To overcome high-dimensional problems, one strategy is selecting variables at the preprocessing stage [9]. The variable selection stage is usually called feature selection. Feature selection technique is suitable for microarray data because a biological perspective states that only a small group of genes are associated with certain diseases [12]. Some literatures such as [16]–[19] using least absolute shrinkage and selection operator (LASSO) feature selection in various cancer classification cases from high-dimensional microarray data. LASSO originally proposed in 1996. A relatively new feature selection method named infinite feature selection (IFS) proposed in [20]. Roffo *et al.* [20] proven that IFS has better performance compared to LASSO.

IFS is a graph-based feature selection method. IFS can work in two scenarios, supervised infinite feature selection (SIFS) and unsupervised infinite feature selection (UIFS). In [20], both SIFS and UIFS were tested on 11 datasets, 5 of which were high-dimensional microarray data. The classification results on 5 microarray data shows that both SIFS and UIFS can make the classifier produce higher classification accuracy compared to the same classifier that uses other feature selection methods. However, Roffo *et al.* [20] has two weaknesses. First, it ignored the problem of imbalanced data. Second, it has not presented classification results at the training and testing stages, and has not presented a comparison of classification results on a classifier without IFS. As a result, the effectiveness of SIFS and UIFS as a feature selection is not yet known.

Some 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 [9], [20]–[23]. Those five studies used accuracy as a measuring tool. In fact, accuracy can produce overly optimistic results on imbalanced data [24]. In other words, accuracy is sensitive to imbalanced data. One tool for evaluating classification results that is insensitive to unbalanced class distribution is balanced accuracy (BA) [25]. Sapitri *et al.* [26] concluded that BA can perform more fairly to both classes in imbalanced microarray data. In other words, to overcome imbalanced data challenges, BA can be used as an evaluation metric.

This research proposes hybrid methods which are constructed from two machine learning methods: a feature selection method and a classifier. IFS was chosen as the feature selection method. Both IFS scenarios (SIFS and UIFS) are separately used to overcome high-dimensional problems. IFS was chosen because it is still rarely used but has potential to produce good feature selection results based on [20]. The classifier method chosen is a decision tree algorithm named classification and regression tree (CART) that was originally proposed in 1984. Even though CART is an old method, it is superior in terms of interpretation [17] and can handle outlier data [27]. Therefore, this research proposed two hybrid methods called SIFS-CART and UIFS-CART. BA is used as an evaluation metric for imbalanced microarray data. So, this research fixed two research weaknesses in [20] specifically in the case of ovarian cancer classification from microarray data. It is hoped that this research can be used as a reference regarding the purposes of hybrid methods SIFS-CART and UIFS-CART, and also concerning the effectiveness of the two IFS scenarios as feature selection in CART especially in the case of ovarian cancer classification from high-dimensional imbalanced microarray data.

2. PROPOSED METHODS

There are two hybrid methods proposed by this research, namely SIFS-CART and UIFS-CART. In general, how the hybrid methods work is summarized in Figure 1. In our hybrid methods, the data that has gone through the preprocessing stage is used as an input for IFS in the feature selection stage. The IFS method, both UIFS and SIFS, produces output in the form of a list of features that have weights and have been ranked based on these weights. Some features that have certain ranks are chosen. It formed new data that has smaller dimensions (fewer feature columns) because the number of features used is reduced by IFS in the feature selection stage. The new data that only contains a group of features with a certain rank is then used as input for CART. So, our hybrid methods works by using the output from IFS as an input for CART.

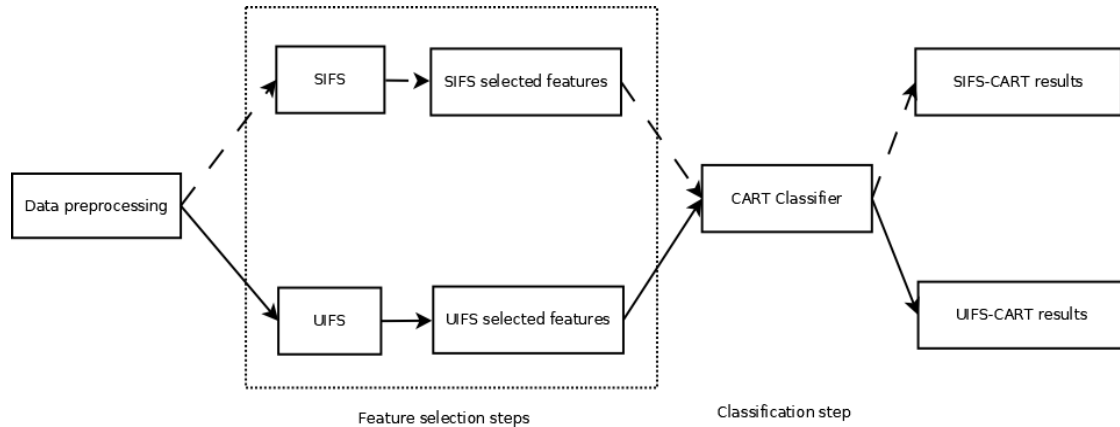


Figure 1. Proposed methods

3. METHOD

This research tested three methods according to Table 1. There is one single method (CART) and two hybrid methods (SIFS-CART and UIFS-CART). The single method used to show the effectiveness of the two IFS scenarios as feature selection in CART in the case of ovarian cancer classification from high-dimensional imbalanced microarray data.

Table 1. The methods

No.	Method	Feature selection	Feature selection optimization	Classifier	Classifier optimization
1	CART	-	-	CART	minimum cost-complexity pruning
2	SIFS-CART	SIFS	5-fold cross validation	CART	minimum cost-complexity pruning
3	UIFS-CART	UIFS	5-fold cross validation	CART	minimum cost-complexity pruning

Based on Table 1, different optimization techniques are applied in IFS and CART. The full descriptions of IFS algorithms and formulas can be seen on [20]. Based on Roffo *et al.* [20], both SIFS and UIFS transforms the data (table form) into a complete graph and calculate a graph matrix called weighted adjacency matrix. Both scenarios have different formulas to calculate the matrix. Then, the matrix processed algebraically to get a column matrix that contains the weights of each feature. SIFS requires 3 parameters ($\alpha_1, \alpha_2, \alpha_3$) and UIFS requires a parameter α to produce a weighted adjacency matrix. This research optimizes these parameters using k-fold cross validation [28] with $k = 5$ or 5-fold cross validation for short. Furthermore, CART optimization is carried out by pruning using the minimum cost-complexity pruning technique [29].

The data used in this research is the OVA_ovary dataset which is an open access microarray data, available on OpenML website as initiated by [30]. The OVA_ovary data file has ".arff" format or attribute-relation file format which consists of metadata and dataset. The extracted dataset consists of 10937 columns and 1545 observations/samples (rows). These columns include 1 ID_REF column (sample ID), 10935 gene columns (feature columns), and 1 Tissue (target) column listed in Table 2.

Table 2. OVA ovary dataset

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

The "other" values in the target column indicate not ovarian cancer tissue samples and "ovary" values indicate ovarian cancer tissue samples. The OVA_ovary dataset is unbalanced data. The "other" class is the majority class containing 1347 samples, while the "ovary" class only contains 198 samples.

The data analysis process was conducted using the Python programming language. The software used is JupyterLab version 3.6.3. The hardware used is a laptop with 13th generation Intel Core i7 processor

specifications and has dual channel 8 GB RAM (16 GB RAM in total). In general, the data analysis steps carried out consisted of four stages: data preprocessing, feature selection (only for hybrid methods), classification, and evaluation. Figures 2 and 3 illustrate the difference analysis steps in single method and hybrid methods. Every stage is described as follows.

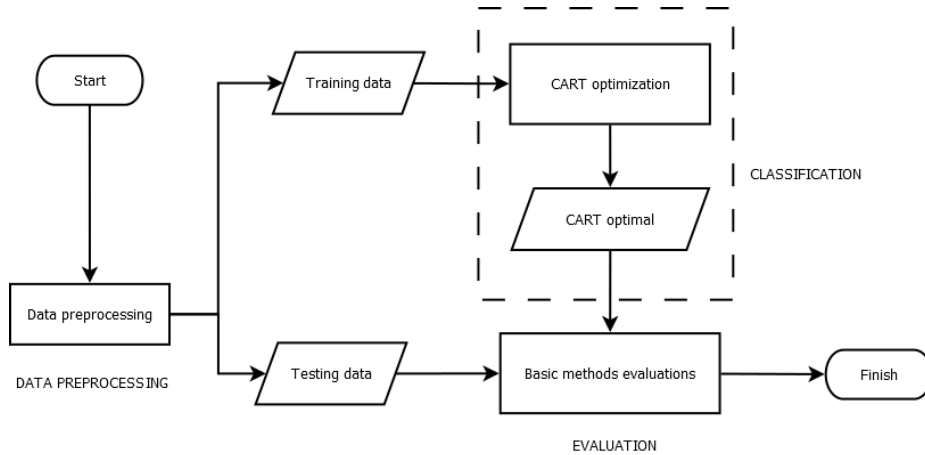


Figure 2. Flowchart of single method (CART)

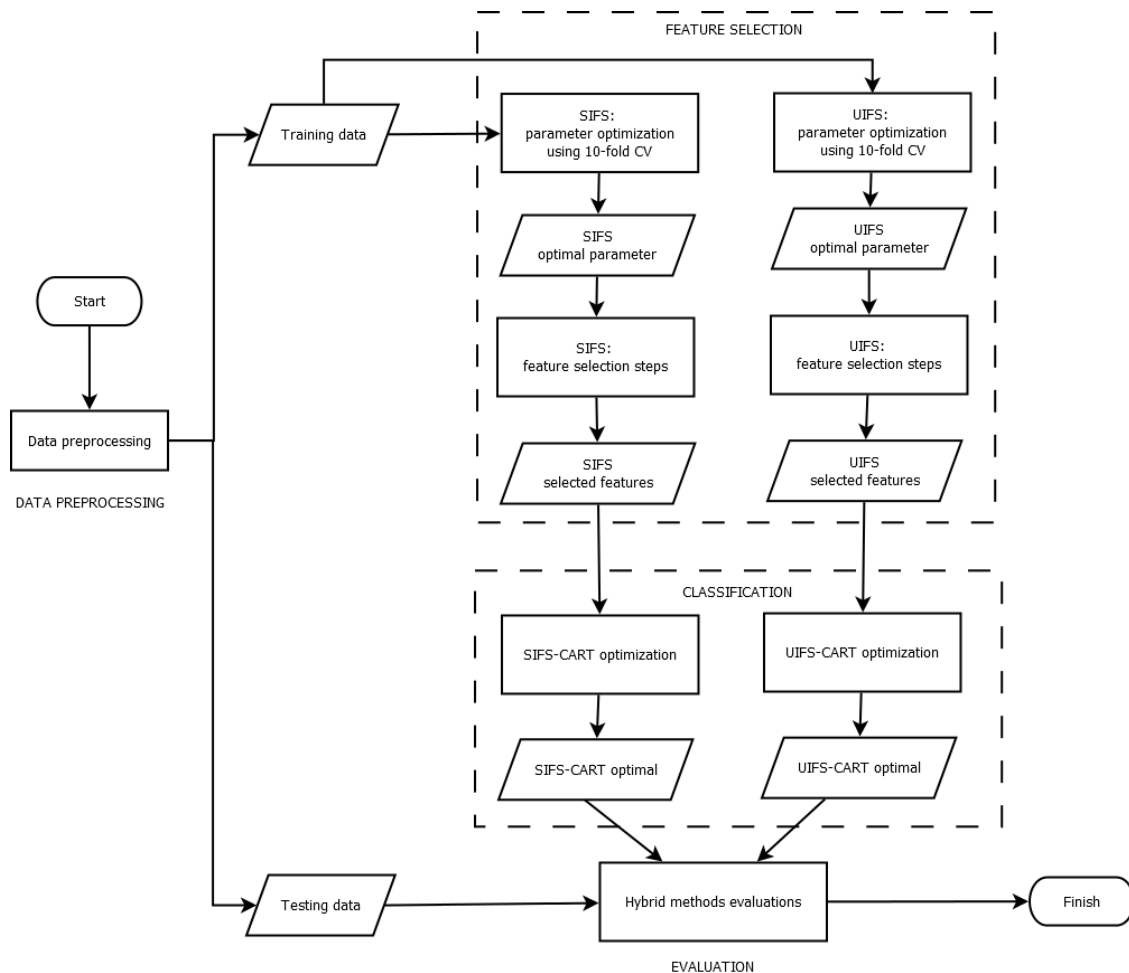


Figure 3. Flowchart of hybrid methods (SIFS-CART and UIFS-CART)

3.1. Data preprocessing

Data preprocessing carried out in this research including: i) OVA_ovary data extraction from “.arff” format to “.xlsx” format, ii) data cleaning by deleting irrelevant columns (ID_REF column), iii) adjusting data types to make all columns contain numerical data, iv) data scaling with minmax normalization as in [31], and v) data splitting to training data and testing data using 80%:20% proportion as in [32]. For the single model (CART), the training data goes through the classification stage as in Figure 2. Meanwhile, for the hybrid methods, the training data then goes through the feature selection stage as in Figure 3.

3.2. Feature selection using IFS

Both IFS scenarios require parameters. This research tested several combinations of parameter values according to Tables 3 and 4. Optimization of these parameters was carried out using 5-fold cross validation. The 5-fold cross validation divides training data into five folds. The five folds are processed in five iterations. In each iteration, one-fold acts as testing data and the rest as training data in turn. Best parameter is the parameter combination that has best performance averaged from every iteration.

Based on Roffo *et al.* [20], SIFS and UIFS do not perform feature selection directly, but rather sort the features based on the weights obtained. Thus, the number of selected features needs to be specified manually. This research tested 10 selected feature sizes, namely 10, 50, 100, 150, 200, 250, 500, 1000, 2500, and 5000 features with higher rank. Then, the selected features are used as an input for the CART classifier in classification stage.

Table 3. SIFS parameters values

Combination	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0.1	0.1	0.8
5	0.1	0.8	0.1
6	0.8	0.1	0.1
7	0.1	0.45	0.45
8	0.45	0.1	0.45
9	0.45	0.45	0.1
10	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Table 4. UIFS parameter values

Combination	α
1	0
2	0.1
3	0.2
4	0.3
5	0.4
6	0.5
7	0.6
8	0.7
9	0.8
10	0.9
11	1

3.3. Classification using CART

All methods in Table 2 use CART as a classifier. The difference between the three lies in the number of features used. The first method (CART) uses all the features (10935 features). The second and third methods use features with varying sizes according to the 10 feature sizes tested.

In this research, CART is created using the 'DecisionTreeClassifier' package from the 'sklearn.tree' library by setting the parameters criterion = 'gini' and splitter = 'best'. To reduce the 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 called optimal CART.

Minimum cost-complexity pruning [29] is an algorithm used to prune decision trees with the aim of preventing overfitting. This algorithm requires a parameter $ccp \geq 0$, which is called the cost-complexity parameter to measure the complexity of a decision tree. The ccp parameter which in Python is called ccp_alpha is searched using the 'cost_complexity_pruning_path' package in the 'DecisionTreeClassifier'. It automatically produces some ccp_alpha values and creates a list of pruned CART based on those ccp_alpha values. Pruned CART that fulfill three criteria: i) produce a fairly high BA value at the testing

stage; ii) has a low difference in BA scores at the training and testing stages; and iii) the interpretation of CART is not too simple (it does not only consist of root nodes), is chosen as optimal CART.

3.4. Evaluation

This stage evaluates the performance of CART, SIFS-CART, and UIFS-CART based on evaluation metrics, model complexity, and the runtimes. The evaluation metrics used in this research is BA. BA can be formulated in (1) [33].

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

BA calculated based on classification results. 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 [24]. The four cases are true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

4. RESULTS AND DISCUSSION

4.1. Results

OVA_ovary dataset after going through the preprocessing stage has a value range of [0, 1]. There are 1236 rows of training data and 309 rows of testing data. For the single model (CART), training data which has 10935 feature columns and 1236 samples rows used in classification stage. In other words, CART used 10935 features. After that, the optimal CART obtained goes through an evaluation stage. The optimal CART model achieved BA values on training and testing 84.44% and 82.5% respectively.

For hybrid methods, training data is used in the feature selection stage before the classification stage. To do feature selection using IFS, it is needed to optimize the parameters first. The parameters in SIFS and UIFS that were tested are in accordance with Tables 3 and 4. The optimal parameter values in this study are the parameters that produce the highest average BA in 5-fold cross validation. BA is used as a benchmark because it can apply more fairly to imbalanced data [26].

The average BA value in each experiment is summarized in Table 5. Based on Table 5, experiment number 8 in SIFS has the highest average BA. As a result, the SIFS parameters used for the feature selection stage are $\alpha_1 = 0.45$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.45$. Meanwhile, the UIFS parameter used for the feature selection stage is $\alpha = 1$ from the last experiment.

Table 5. Average BA results on SIFS and UIFS

SIFS					UIFS		
No.	α_1	α_2	α_3	Average BA	No.	α	Average BA
1	1	0	0	0.4938	1	0	0.5002
2	0	1	0	0.5094	2	0.1	0.5132
3	0	0	1	0.5002	3	0.2	0.5021
4	0.1	0.1	0.8	0.4972	4	0.3	0.5055
5	0.1	0.8	0.1	0.5025	5	0.4	0.5140
6	0.8	0.1	0.1	0.5018	6	0.5	0.4839
7	0.1	0.45	0.45	0.4870	7	0.6	0.4937
8	0.45	0.1	0.45	0.5203	8	0.7	0.5101
9	0.45	0.45	0.1	0.4931	9	0.8	0.5075
10	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.4915	10	0.9	0.5081
					11	1	0.5189

In the feature selection stage, all training data and the optimal parameter obtained are used to calculate the feature weights. The features ranked based on the weights obtained. Then, the features selected based on the rank and the selected feature sizes used. This research tested 10 selected feature sizes, namely 10, 50, 100, 150, 200, 250, 500, 1000, 2500, and 5000 features.

The training and testing data whose columns are adjusted according to the results of selected feature sizes are used in the classification stage. The BA results of SIFS-CART and UIFS-CART for each selected feature size in the classification stage are shown in Figures 4 and 5. The first dot in the line plots indicate the BA results when the hybrid methods used 10 features and so on.

Figure 4 shows that BA values are volatile. It is related to the different numbers of TP, TN, FP, and FN obtained from every number of selected features. For example, SIFS-CART that use 50 features achieved less TP value than 10 features. It makes the BA values smaller and the line plot goes down. In other words, the more features used in SIFS-CART does not guarantee an increase in BA value.

From the trend of BA values in SIFS-CART, the highest BA values at the testing stage occurred when using 1000 features. This model produces best classification results (highest TP and TN values, lowest FP and FN values) compared to the others. The BA values at the training and testing stages of SIFS-CART with 1000 features are 85.65% and 83.23% respectively.

Due to the different scenarios in IFS, Figure 5 shows a slightly different trend from Figure 4. In Figure 5, the BA value is not volatile, but tends to increase as the number of selected features increase. Thus, in UIFS-CART, the BA values in the testing stage reach the highest value at a feature size of 5000. The BA values in the training and testing stages of UIFS-CART with 5000 features are 77.50% and 75.74% respectively.

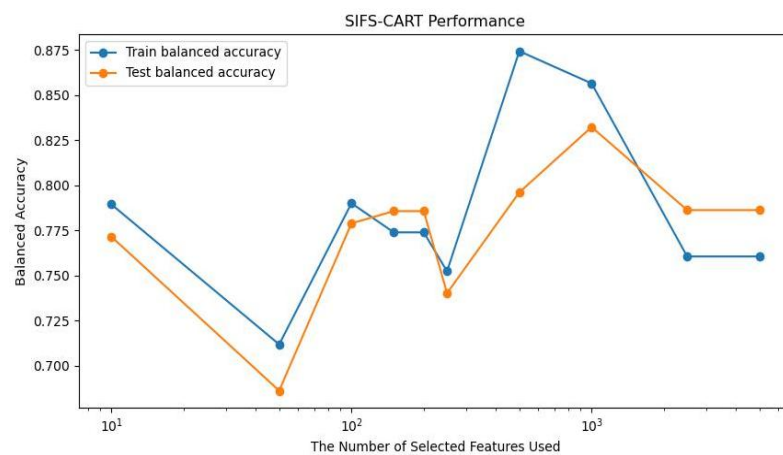


Figure 4. SIFS-CART performance on different numbers of selected features

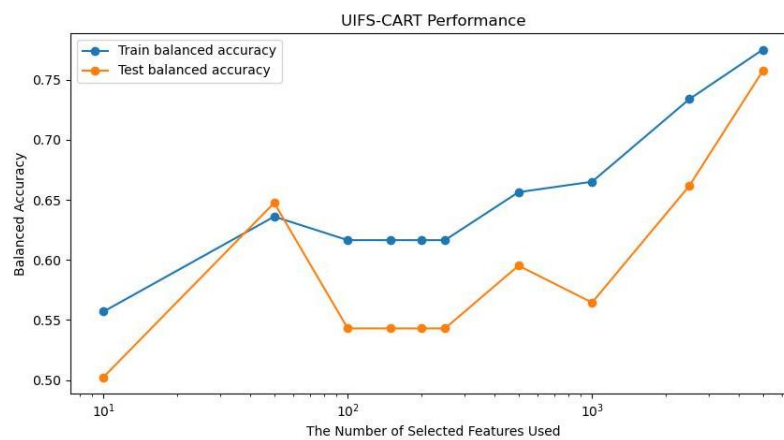


Figure 5. UIFS-CART performance on different numbers of selected features

Figures 6(a) and 6(b) summarize the best BA values and the number of features used in every method. Figure 6(a) shows that SIFS-CART achieved the highest BA compared to CART and UIFS-CART. Interestingly, UIFS-CART does not achieve higher BA than CART although it is a hybrid method. In terms of features used, Figure 6(b) illustrates that CART itself is more complex because it used all features (10935 features) when SIFS-CART and UIFS-CART respectively use 1000 and 5000 features. Even though it only uses 1000 features, Figure 6(a) shows that SIFS-CART is able to achieve higher BA compared to CART and UIFS-CART.

Table 6 shows all runtimes of every method. Because the feature selection stage takes time, the fastest method is CART (single method). However, CART is less effective because it uses all features. SIFS-CART is more effective because it uses the least features but achieves the highest BA compared to the others. It can be concluded that SIFS-CART is more effective to identify ovarian cancer from the OVA_ovary dataset than CART and UIFS-CART.

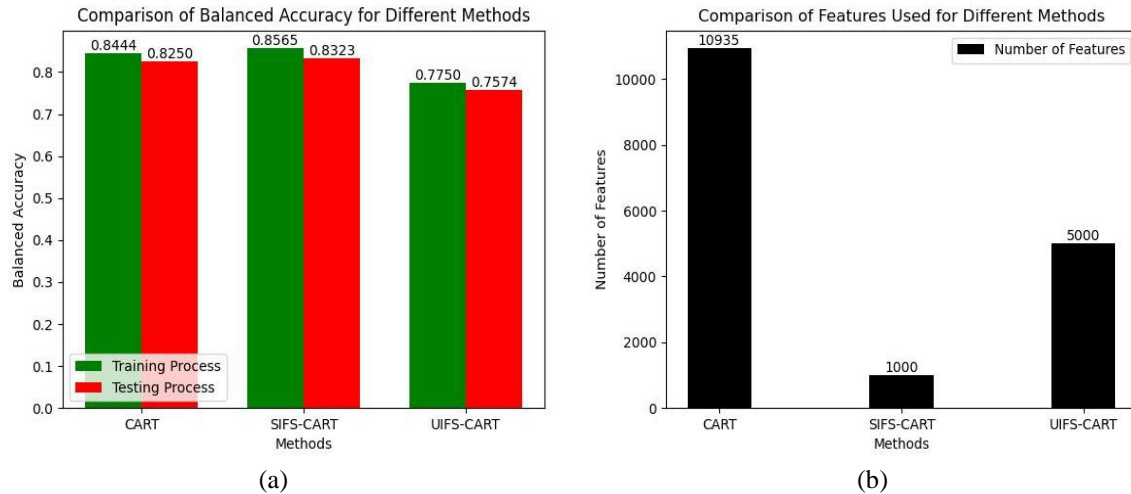


Figure 6. Comparison of all methods performances based on: (a) BA values and (b) the number of selected features used

Table 6. Runtimes of every method

Method	Feature selection runtime (s)	Classifier runtime		Total runtimes (minutes)
		Training runtime (s)	Testing runtime (s)	
CART	-	77.71	0.11	1.30
SIFS-CART	628.39	7.10	0.02	10.59
UIFS-CART	765.57	34.73	0.05	13.34

In this research, UIFS-CART uses 5000 features. Remember that the data after the preprocessing stage has 10935 features columns and 1545 samples rows. The selection of 5000 features by UIFS means that UIFS reduces the data dimensions from 10935 features columns and 1545 rows to 5000 features columns and 1545 rows. This indicates that UIFS has not succeeded in reducing the OVA_ovary data to non-high-dimensional form because there are still more columns than rows. UIFS-CART also has poorer performance than the single model itself. On the other hand, SIFS succeeded in reducing OVA_ovary data to non-high-dimensional. SIFS reduces the dataset dimensions to 1000 features columns and 1545 rows. SIFS-CART superior in terms of BA values and used least features. The runtime of SIFS is also faster than UIFS. In other words, SIFS is more effective as feature selection on OVA_ovary dataset in this research compared to UIFS.

4.2. Discussion

This research proposes two hybrid methods, namely SIFS-CART and UIFS-CART, and also compare the performance of both with CART (without IFS) in cases of ovarian cancer classification from OVA_ovary dataset. This research considers overcoming the high-dimensional and imbalanced problem on the data used, while earlier studies did not address the imbalanced problem. Our results show that SIFS-CART can outperform CART and UIFS-CART in terms of BA values achieved and the number of features used. SIFS also succeeded in reducing OVA_ovary data to non-high-dimensional form. However, further and in-depth studies may be needed to confirm if $\alpha_1 = 0.45$, $\alpha_2 = 0.1$, $\alpha_3 = 0.45$ are best parameters of SIFS and 1000 is the best number of selected features when it is used in the OVA_ovary dataset.

Roffo *et al.* [20] that used IFS did not consider to overcome imbalanced problems on all microarray data tested and not compare their hybrid method performance to single method. Sa'adah *et al.* [17] used OVA_ovary dataset is also not considered to overcome imbalanced problems. For that reason, both used accuracy as an evaluation metric. On the other hand, Abdellatif *et al.* [34] using IFS on non high-dimensional data about heart disease. It makes we can not compare our research results to those related research.

Although we can not compare our research to the others, we can explain that our research indirectly showed that the parameter and the number of selected features is crucial when using IFS. We suspected that the optimal parameter chosen in UIFS makes the results of UIFS-CART not superior to CART and can not even reduce OVA_ovary dataset to non-high-dimensional form. Further research is recommended to use other lists of parameters to test or different optimization techniques to find the alpha parameter that make UIFS successful in reducing the OVA_ovary dataset and produce better BA values. Further research also can focus on testing SIFS-CART on different imbalanced high-dimensional datasets to get more general trends of its performances.

5. CONCLUSION

This research aims to propose two hybrid methods, namely SIFS-CART and UIFS-CART, and also compare the performance of both with CART (without IFS) in cases of ovarian cancer classification from imbalanced high-dimensional microarray data. The data used is the OVA_ovary dataset which has 10937 columns and 1545 rows. The evaluation metric used is BA. Based on testing 10 selected feature sizes, SIFS-CART achieves maximum performance when using 1000 features and UIFS-CART 5000 features. Although the runtime of CART is the fastest because it does not include the feature selection stage, the result shows that SIFS-CART can outperform CART and UIFS-CART in terms of BA values achieved and the number of features used. SIFS also succeeded in reducing OVA_ovary data to non-high-dimensional form, from 10935 features columns and 1545 rows to 1000 features columns and 1545 rows. On the other hand, UIFS has not succeeded in reducing OVA_ovary data to non-high-dimensional form and UIFS-CART can not even outperform CART. Further research is recommended to use other lists of parameters to test or different optimization techniques to find the alpha parameter that make UIFS successful in reducing the OVA_ovary dataset and produce better BA values. Further research also can focus on testing SIFS-CART on different imbalanced high-dimensional datasets to get more general trends of its performances.




REFERENCES

- [1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.
- [2] A. Arnaout *et al.*, "Baseline staging imaging for distant metastasis in women with stages I, II, and III breast cancer," *Current Oncology*, vol. 27, no. 2, pp. e123–e145, 2020, doi: 10.3747/co.27.6147.
- [3] C. Goebel, C. L. Loudon, R. McKenna Jr., O. Onugha, A. Wachtel, and T. Long, "Diagnosis of non-small cell lung cancer for early stage asymptomatic patients," *Cancer Genomics & Proteomics*, vol. 244, pp. 229–244, 2019, doi: 10.21873/cgp.20128.
- [4] T. Takikawa *et al.*, "Clinical features and prognostic impact of asymptomatic pancreatic cancer," *Scientific Reports*, vol. 12, pp. 1–11, 2022, doi: 10.1038/s41598-022-08083-6.
- [5] F. Kusuma, M. Riani, and F. Witjaksono, "Association between risk of malnutrition and surgical outcome in ovarian cancer patients," *eJournal Kedokteran Indonesia*, vol. 9, no. 3, pp. 203–207, 2021, doi: 10.23886/ejki.9.71.203-7.
- [6] C. Slatnik and E. Duff, "Ovarian cancer: Ensuring early diagnosis," *The Nurse Practitioner*, vol. 40, no. 9, pp. 47–54, 2015, doi: 10.1097/01.NPR.0000450742.00077.a2.
- [7] J. Huang *et al.*, "Worldwide Burden, risk factors, and temporal trends of ovarian cancer: a global study," *Cancers*, vol. 14, 2022, doi: 10.3390/cancers14092230.
- [8] D. Žilović, R. Čiurlienė, R. Sabaliauskaitė, and S. Jarmalaitė, "Future screening prospects for ovarian cancer," *Cancers*, vol. 13, no. 15, pp. 1–17, 2021, doi: 10.3390/cancers13153840.
- [9] M. Y. Rochayani, U. Sa'adah, and A. B. Astuti, "Two-stage gene selection and classification for a high-dimensional microarray data," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 9–18, 2020, doi: 10.15575/join.v5i1.569.
- [10] 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.
- [11] M. Ramachandro and R. Bhramaramba, "Classification of gene expression data set using support vectors machine with RBF kernel," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 2907–2913, 2019, doi: 10.35940/ijrteB2463.078219.
- [12] C. Liu and H. S. Wong, "Structured penalized logistic regression for gene selection in gene expression data analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 1, pp. 312–321, 2019, doi: 10.1109/TCBB.2017.2767589.
- [13] G. Kunapuli, *Ensemble methods for machine learning*, 6th ed. New York: Manning Publications, 2022.
- [14] Y. He, J. Zhou, Y. Lin, and T. Zhu, "A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data," *Computational Biology and Chemistry*, vol. 80, pp. 121–127, 2019, doi: 10.1016/j.compbiolchem.2019.03.017.
- [15] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, "Evolutionary machine learning: a survey," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–35, 2021, doi: 10.1145/3467477.
- [16] Y. Wang, X. Li, and R. Ruiz, "Weighted general group lasso for gene selection in cancer classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2860–2873, 2019, doi: 10.1109/TCYB.2018.2829811.
- [17] U. Sa'adah, M. Y. Rochayani, and A. B. Astuti, "Knowledge discovery from gene expression dataset using bagging lasso decision tree," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 1151–1159, 2020, doi: 10.11591/ijeecs.v21i2.pp1151-1159.
- [18] Z. Yahya, R. Alhamzawi, and H. T. M. Ali, "Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression," *Computers in Biology and Medicine*, vol. 97, pp. 145–152, 2018, doi: 10.1016/j.compbiomed.2018.04.018.
- [19] Z. Y. Algarnal and M. H. Lee, "Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification," *Expert Systems With Applications*, vol. 42, pp. 9326–9332, 2015, doi: 10.1016/j.eswa.2015.08.016.
- [20] G. Roffo, S. Melzi, U. Castellani, A. Vinciarelli, and M. Cristani, "Infinite feature selection: a graph-based feature filtering approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4396–4410, 2021, doi: 10.1109/TPAMI.2020.3002843.
- [21] N. A. Al-thanoon, O. S. Qasim, and Z. Y. Algarnal, "Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification," *Computers in Biology and Medicine*, vol. 103, pp. 262–268, 2018, doi: 10.1016/j.compbiomed.2018.10.034.
- [22] 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: 10.20895/infotel.v12i3.485.
- [23] A. M. Alharthi, M. H. Lee, and Z. Y. Algarnal, "Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty," *Informatics in Medicine Unlocked*, vol. 24, 2021, doi: 10.1016/j.imu.2021.100622.
- [24] 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, doi: 10.1186/s12864-019-6413-7.




- [25] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: An overview,” *arXiv*, pp. 1–17, 2020.
- [26] N. K. E. Sapitri, U. Sa’adah, and N. Shofianah, “Knowledge Discovery from Confusion Matrix of Pruned CART in Imbalanced Microarray Data Ovarian Cancer Classification,” *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 227–236, 2024, doi: 10.15294/sji.v11i1.50077.
- [27] S. Singh, “Comparative study Id3, Cart and C4.5 decision tree algorithm: a survey,” *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, pp. 97–103, 2014.
- [28] T. Wong, “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation,” *Pattern Recognition*, vol. 48, pp. 2839–2846, 2015, doi: 10.1016/j.patcog.2015.03.009.
- [29] K. B. Ravi and J. Serra, “Cost-complexity pruning of random forests,” *arXiv- Statistics*, pp. 1–10, Mar. 2017.
- [30] G. Stiglic and P. Kokol, “Stability of ranked gene lists in large microarray analysis studies,” *Journal of Biomedicine and Biotechnology*, vol. 2010, pp. 1–9, 2010, doi: 10.1155/2010/616358.
- [31] 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: 10.1109/ICSICT.2018.8565736.
- [32] M. Y. Rochayani, U. Sa’adah, and A. B. Astuti, “Simulation study of imbalanced classification on high-dimensional gene expression data,” *Scientific Journal of Informatics*, vol. 10, no. 1, pp. 45–54, 2023, doi: 10.15294/sji.v10i1.40589.
- [33] 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 Mining*, vol. 14, pp. 1–22, 2021, doi: 10.1186/s13040-021-00244-z.
- [34] A. Abdellatif, H. Abdellatif, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, “Improving the heart disease detection and patients’ survival using supervised infinite feature selection and improved weighted random forest,” *IEEE Access*, vol. 10, pp. 67363–67372, 2022, doi: 10.1109/ACCESS.2022.3185129.

BIOGRAPHIES OF AUTHORS






Ni Kadek Emik Sapitri    received her bachelor’s degree in mathematics from Udayana University in 2020 and master’s degree in mathematics from Brawijaya University in 2024. This research is part of her final project as a postgraduate student. Her research interest is about machine learning applications in medical fields, especially bioinformatics. She is also interested in graph data science. She can be contacted at email: emikpitri@gmail.com.



Umu Sa’adah    received the bachelor, master, and doctoral degree in mathematics from Universitas Gadjah Mada, Indonesia in 1993, 2002, and 2015, respectively. She is currently an Associate Professor at the Department of Mathematics in the Faculty of Mathematics and Natural Sciences, Brawijaya University, Indonesia. Her research interests are in artificial neural networks, bootstrap, data science, data mining, machine learning, statistics, and risk theory. She can be contacted at email: u.saadah@ub.ac.id.



Nur Shofianah    received the doctoral degree in Graduate School of Natural Science and Technology, Kanazawa University, Japan. She received bachelor and master’s degrees from Sepuluh Nopember Institute of Technology, Indonesia. Currently she is a lecturer in the Faculty of Mathematics and Natural Sciences, Brawijaya University, Indonesia, since 2010. Her research interests are in dynamical analysis and optimal control of biomathematical models, numerical methods for PDE, predictive modeling, and machine learning. She can be contacted at email: nur_shofianah@ub.ac.id.