BAREKENG: Journal of Mathematics and Its Applications
September 2024 Volume 18 Issue 3 Page 1909–1918
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss3pp1909-1918

IDENTIFYING IMPORTANT GENES IN OVARIAN CANCER FROM HIGH-DIMENSIONAL MICROARRAY DATA USING SIFS-CART METHOD

Ni Kadek Emik Sapitri[1*], Umu Sa'adah[2], Nur Shofianah[3]

[1,2,3]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Brawijaya*
Jl. Veteran, Malang, East Java, 65145, Indonesia

*Corresponding author's e-mail: * **emikpitri@gmail.com***

Article History:

Received: 28th February 2024
Revised: 18th April 2024
Accepted: 11th July 2024
Published: 1st September 2024

Keywords:

CART;
Important Genes;
Machine Learning;
Microarray Data;
Ovarian Cancer;
SIFS.

ABSTRACT

Ovarian cancer can be identified from microarray data using machine learning. Many studies only focus on improving the machine learning classification algorithms to achieve higher performance. The purpose of classification is not only to obtain high performance but also to seek new knowledge from the results. This research focuses on both. By using a hybrid Supervised Infinite Feature Selection (SIFS) method with Classification and Regression Tree (CART) or SIFS-CART, this research aims to predict ovarian cancer and identify potential genes for ovarian cancer cases. The data used is the OVA_ovary dataset. SIFS in the best SIFS-CART model reduced 10935 genes in the initial OVA_ovary dataset to 1000 genes. Then, CART was built with these 1000 genes. Based on the balanced accuracy (BA) metric for imbalanced microarray data, the best SIFS-CART model achieves 85.7% BA in training and 83.2% in testing. The optimal CART in the best SIFS-CART model only needs four genes from 1000 selected genes to build it. Those genes are STAR, WT1, PEG3, and ASPN. Based on studies of several pieces of literature in the medical field, it can be concluded that STAR, WT1, and PEG3 play an important role in ovarian cancer cases. However, the relationship between ASPN and ovarian cancer in more detail has not been studied by medical researchers.

How to cite this article:

N. K. E. Sapitri, U. Sa'adah and N. Shofianah., "IDENTIFYING IMPORTANT GENES IN OVARIAN CANCER FROM HIGH-DIMENSIONAL MICROARRAY DATA USING SIFS-CART METHOD," *BAREKENG: J. Math. & App.*, vol. 18, iss. 3, pp. 1909-1918, September, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Ovarian cancer is the most dangerous cancer among various cancers that attack the female reproductive system [1]. In 2020, Harsono reported that only 20% of ovarian cancer was diagnosed at stage 1 (early) when the disease is limited to the ovaries. 90% of patients in the early stages respond well to existing therapy [2]. Therefore, a method that can detect ovarian cancer quickly and accurately is needed.

Various types of cancer, including ovarian cancer, can be identified from microarray data, which is a high-dimensional data [3]. Machine learning algorithms can identify cancer from microarray data relatively quickly. One way to identify cancer from microarray data is by classifying it.

Several research studies have used machine learning for cancer classification from microarray data [4]–[9]. Those researchers only focus on improving the machine learning algorithms to achieve higher performance based on the evaluation metrics used. The purpose of classification is not only to obtain high performance but also to seek new knowledge from the classification results [10]. This research focuses on both, building model that can achieve higher performance and also seeks for the knowledge from the results.

Research by Rochayani et al. in 2020 [3] identified genes associated with breast cancer using hybrid machine learning algorithms. The algorithm used is a combination of Least Absolute Shrinkage and Selection Operator (LASSO) as feature selection and Classification and Regression Tree (CART) as a classifier called LASSO-CART. In microarray data, the features are genes. The set of genes that LASSO selected is then used for the classification process with CART. Rochayani et al. also show the optimal CART interpretation. Based on several related studies in the medical field about breast cancer, the selected genes in optimal CART interpretation are indeed closely related to breast cancer growth.

In 2021, Roffo et al. [11] proposed an Infinite Feature Selection (IFS) feature selection method. IFS is a graph-based feature selection method. In that research, IFS was tested on 11 datasets, nine high-dimensional data. The results show that IFS performs better than LASSO.

Considering the performance of IFS in [11] and the superiority of CART interpretation in [3], this study uses a hybrid method of IFS and CART. IFS can work in two scenarios, namely unsupervised and supervised. This research uses a supervised scenario, which is now written as SIFS. SIFS was chosen because research [12] in 2022, which is still relatively new, concluded that using SIFS can increase the accuracy of various classifiers in cases of heart disease detection.

Hopefully, this research can be a reference for using the SIFS-CART method for ovarian cancer classification cases. In addition, it is hoped that the interpretation of optimal CART in this study can give insights into important and potential genes in ovarian cancer.

2. RESEARCH METHODS

The data used in this research is the OVA_ovary dataset, an open-source microarray data available on the OpenML website. The OVA_ovary dataset file is in the ".arff" or Attribute-Relation File Format. The file consists of metadata and a dataset. The available dataset consists of 10937 columns and 1545 observations (rows). These columns include 1 ID_REF column, 10935 gene columns, and 1 Tissue column (target column) according to **Table 1**. The "Other" value in the target column indicates no ovarian cancer tissue sample, and "Ovary" indicates an ovarian cancer tissue sample. The OVA ovary dataset is imbalanced because the 'Ovary' class only contains 198 samples while the 'Other' class has 1347 samples.

Table 1. 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

Data source: *https://www.openml.org/search?type=data&status=active&id=1166*

The research steps generally consist of four stages: data preprocessing, gene selection, classification, and evaluation and interpretation. This research uses Python as the programming language in the data processing process. The software used is JupyterLab version 3.6.3. The hardware used is a laptop with 13th-generation Intel Core i7 processor specifications and dual-channel 8GB RAM (or 16GB RAM). The research steps can be seen in **Figure 1**.

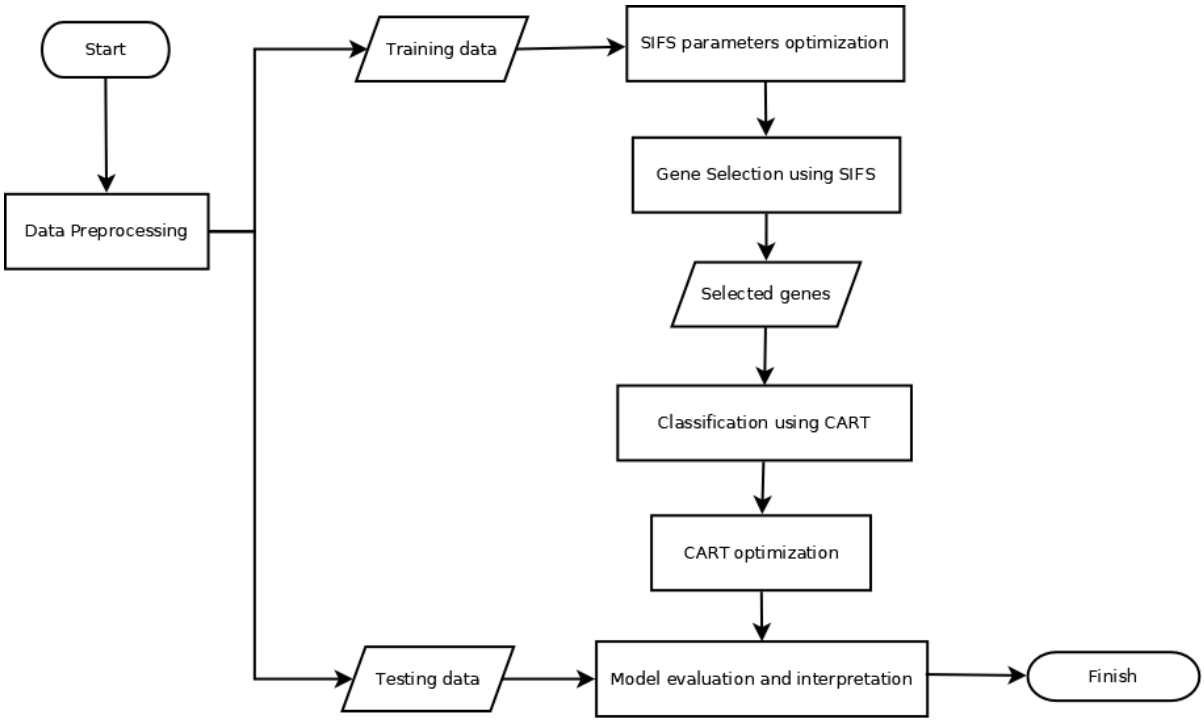


Figure 1. The research steps

2.1 Data Preprocessing

OVA_ovary dataset does not have missing values. Data preprocessing conducted in this research included data extraction, deleting irrelevant columns (ID_REF column), adjusting data types, data scaling with min-max normalization [13], and data splitting to training data and testing data using 80%:20% proportion [14]. The data preprocessing step produces two groups of data, namely training data and testing data.

2.2 Gene Selection

Infinite Feature Selection, abbreviated as Inf-FS or IFS, is a graph-based feature selection method proposed by Roffo et al. A complete explanation of IFS can be seen in [11]. IFS can work in two scenarios, namely unsupervised or supervised. In this research, we use a supervised scenario (SIFS). The SIFS steps are summarized below [11]:

1. The first stage of SIFS is building undirected fully-connected weighted graph $G = (V, E)$. On graph G, the set of vertices $V = \{\vec{v}_1, \vec{v}_2, \vec{v}_3, \dots, \vec{v}_n\}$ represents a set of feature distributions $F = \{f_1, f_2, f_3, \dots, f_n\}$ and the set of edges E modeling the relationship between pairs of nodes (in this case the relationship between distributions).
2. The second stage is making weighted adjacency matrix. Let A be matrix with elements $A(i, j)$ where $1 \leq i, j \leq n$. Matrix A is the weighted adjacency matrix of the graph G defined by a function φ_S as in **Equation (1)** where $0 \leq \alpha_k \leq 1, \sum_k a_k = 1$.

$$\varphi_S(\vec{v}_i, \vec{v}_j) = A(i,j) = (\alpha_1 h_i + \alpha_2 m_i + \alpha_3 \sigma_i)(\alpha_1 h_j + \alpha_2 m_j + \alpha_3 \sigma_j). \quad (1)$$

The function φ_S has positive real value that defines the weight in each edge. The φ_S function is formed from three factors, including Fisher criterion (h_i), normalized mutual information (m_i), and normalized standard deviation (σ_i).

3. Matrix A is then used to calculate partial scores matrix ($\check{C}$) from the feature set using **Equation (2)**. This formula utilizes the concepts of regularization and eigenvalues.

$$\check{C} = (I - rA)^{-1} - I, \quad (2)$$

where I stands for identity matrix and r is real-valued regularization factor. This research use $r = \frac{0.9}{\rho(A)}$, where $\rho(A)$ is the largest eigenvalue (spectral radius) of A.

4. The partial score from step 3 is used to calculate the final score vector ($\check{c}$) using **Equation (3)**.

$$\check{c} = \check{C}\boldsymbol{e}, \quad (3)$$

where $\boldsymbol{e}$ is a 1D vector of ones. The purpose of SIFS is to provide a score of importance for each feature as a function of the importance of its neighboring features. The most discriminating and relevant features will get a higher rank.

In this research, we use SIFS as a gene selection method on the OVA_ovary dataset. All concepts of SIFS used are the same as the original SIFS method in **[11]**, except for counting mutual information in step 2. Mutual information in this research was obtained using the Nearest Neighbor Method in **[15]**, applying the 'mutual_info_classif' package in Python.

SIFS requires 3 parameters, namely α_1, α_2, and α_3. Those parameters optimized using k-fold cross validation **[16]** with $k = 5$. On the other hand, SIFS does not select genes directly but gives them a score of importance based on the weights obtained. Thus, the number of the selected features (or selected genes) needs to be specified manually. This research tested six selected gene sizes, namely 10, 50, 100, 500, 1000, and 5000 highest-ranked genes.

2.3 Classification

Decision trees are considered as data mining method. It can be applied in classification and regression tasks **[17]**. There are various decision tree algorithms, such as Iterative Dichotomizer 3 (ID3), C4.5, and CART. This research uses CART as a classifier after the gene selection step because the literature **[18]** explained that CART is superior to ID3 and C4.5 since it can handle outlier data.

CART uses Gini impurity as splitting criterion to create branches of the tree. Let D be a node, $K = 1,2, \ldots, k$ represents the number of classes, and p_K is the proportion of class K observations in node D. The Gini impurity of D denoted $Gini(D)$ is expressed by **Equation (4)**.

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2. \quad (4)$$

The branching process continues until all the leaves are formed, resulting in the maximum tree. The form of the maximum tree is often too complex. The maximum tree needs to be pruned to prevent overfitting. Overfitting occurs when a model performs very well in training but poorly in testing **[19]**. The final CART obtained is called the optimal CART.

This research created CART with the 'DecisionTreeClassifier' package from the 'sklearn.tree' library by setting the parameters criterion = 'gini' and splitter = 'best'. The pruning parameter, which in Python is called ccp_alpha, is searched with 'cost_complexity_pruning_path' in the 'DecisionTreeClassifier'. It prunes the tree using a minimum cost-complexity pruning technique **[20]**. The optimal CART is selected based on the evaluation metric used.

2.4 Evaluation and Interpretation

The classifier determines the class of each data sample. By the end of the classification process, each sample is categorized into one of four cases **[21]**. The four cases are summarized in the confusion matrix displayed in **Table 2**. 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. The values of TP, TN, FP, and FN used to calculate evaluation metrics.

Table 2. Confusion matrix

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

Some research that classifies cancer from microarray data ignores the problem of imbalanced data. Some studies, such as [14], [22], and [23]. Those studies used accuracy as a measuring tool. In fact, accuracy can produce overly optimistic results on imbalanced data [19]. In other words, accuracy is sensitive to imbalanced data.

There is another evaluation metric called balanced accuracy (BA). BA is a tool for evaluating classification results insensitive to imbalanced class distribution [24]. Sapitri et al. [25] have shown that BA can work more fairly than accurately in imbalanced data classification cases. Since OVA_ovary is considered imbalanced data, this research uses BA as an evaluation metric. BA is defined in Equation (5) [26].

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{5}$$

The optimal CART is chosen based on the BA value in every pruning iteration of each selected gene size (10, 50, 100, 500, 1000, and 5000). The optimal CART criteria in this research are the same as in [25] which chooses CART, which produces the second lowest difference in BA values at the training and testing stage. Those CARTs met the three criteria: produce a reasonably high BA value at the testing stage, have a low difference in BA scores at the training and testing stages, and the CART interpretation is not very simple (it does not consist of only the root node, to prevent underfitting). From the optimal CARTs that used different selected gene sizes, this research selected one CART that has the highest BA value as the best SIFS-CART model.

The final stage is to make a CART interpretation. CART interpretations are made from optimal CART of all selected gene sizes (10, 50, 100, 500, 1000, and 5000), so there are 6 CARTs. Furthermore, the genes selected to build the best SIFS-CART model were further dissected by literature studies of some researchers in medical fields related to ovarian cancer.

3. RESULTS AND DISCUSSION

After going through the data preprocessing stage, the OVA_ovary dataset has a value in [0, 1]. The outputs in this stage are training data and testing data. There are 1236 rows in training data and 309 rows in testing data. The training data is then used to determine the optimal SIFS parameters using 5-fold cross validation.

The 5-fold cross-validation technique divides the training data into five folds. The five folds were processed in five iterations. In each iteration, one-fold acts as testing data, and the rest is used as training data. Using 5-fold cross-validation, we obtained the optimal SIFS parameter values: $\alpha_1 = 0.45, \alpha_2 = 0.1$, and $\alpha_3 = 0.45$. Those parameters determine the weights and rank of genes in the OVA_ovary dataset. Runtime of SIFS as gene selection is 628.39 seconds or around 10 minutes.

SIFS, which uses optimal parameters, reduces data dimension. The number of columns in the original data is 10935 gene columns plus one target column (Tissue). The gene columns were reduced based on the selected gene sizes tested. This research tested six selected gene sizes, namely 10, 50, 100, 500, 1000, and 5000. The changing of data dimensions is listed Table 3. The number of columns in reduced data equals the number of selected genes used plus one target column.

Table 3. The Dimensions of data in each SIFS-CART model

The number of selected genes used	Training data dimension		Testing data dimension	
	Columns	Rows	Columns	Rows
10	11	1236	11	309

The number of selected genes used	Training data dimension		Testing data dimension	
	Columns	Rows	Columns	Rows
50	51	1236	51	309
100	101	1236	101	309
500	501	1236	501	309
1000	1001	1236	1001	309
5000	5001	1236	5001	309

The new training and testing data (reduced data) is used in the classification process with CART. After optimizing CART with minimum cost-complexity pruning and using optimal criterion in [25], we got ccp_alpha results and computing time as in Table 4.

Table 4 shows that the number of genes used is directly proportional to the computational time required for the classifier CART. The more genes used, the longer the computing time needed for CART training and testing. Furthermore, the optimal ccp_alpha size required for each gene size varies. This is because differences in the genes used will form different maximum CART. The BA values obtained are summarized in Figure 2.

Table 4. Cost-Complexity parameter and runtimes of CART in SIFS-CART

The number of selected genes	The optimal value of ccp_alpha	Runtime (s)	
		Training	Testing
10	0.004644897	0.062249	0.013062
50	0.006490471	0.262553	0.010346
100	0.00730178	0.443611	0.010645
500	0.013897171	3.101056	0.012987
1000	0.005779011	7.096078	0.019628
5000	0.034140344	35.18662	0.049774

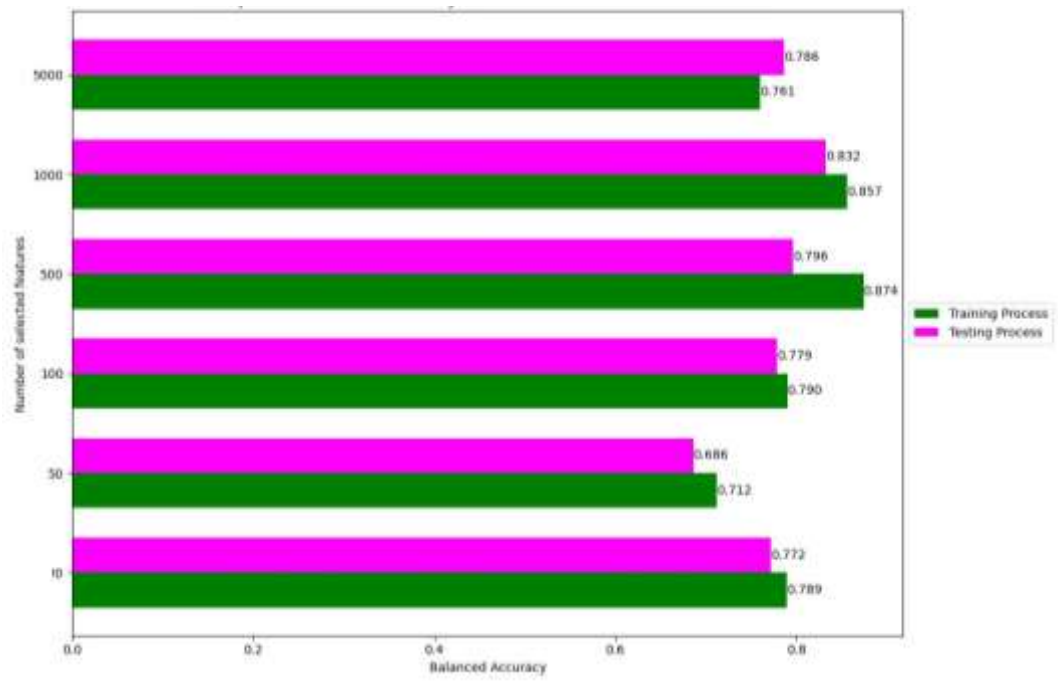


Figure 2. Comparison of BA from different selected genes

In Figure 2, BA values in the training and testing process are volatile. However, CART, which uses 5000 selected gene sizes, apparently only produces more excellent BA than 1000 selected gene sizes. It may indicate that in the group of 5000 genes, many genes are less relevant to ovarian cancer. Considering that BA is not sensitive to imbalanced data, information can be obtained that the SIFS algorithm in calculating the

weights and ranking of gene features by SIFS still has the opportunity to investigate. This is because, in **Figure 2**, the model with 50 genes has a much lower BA than 10 and 100 genes.

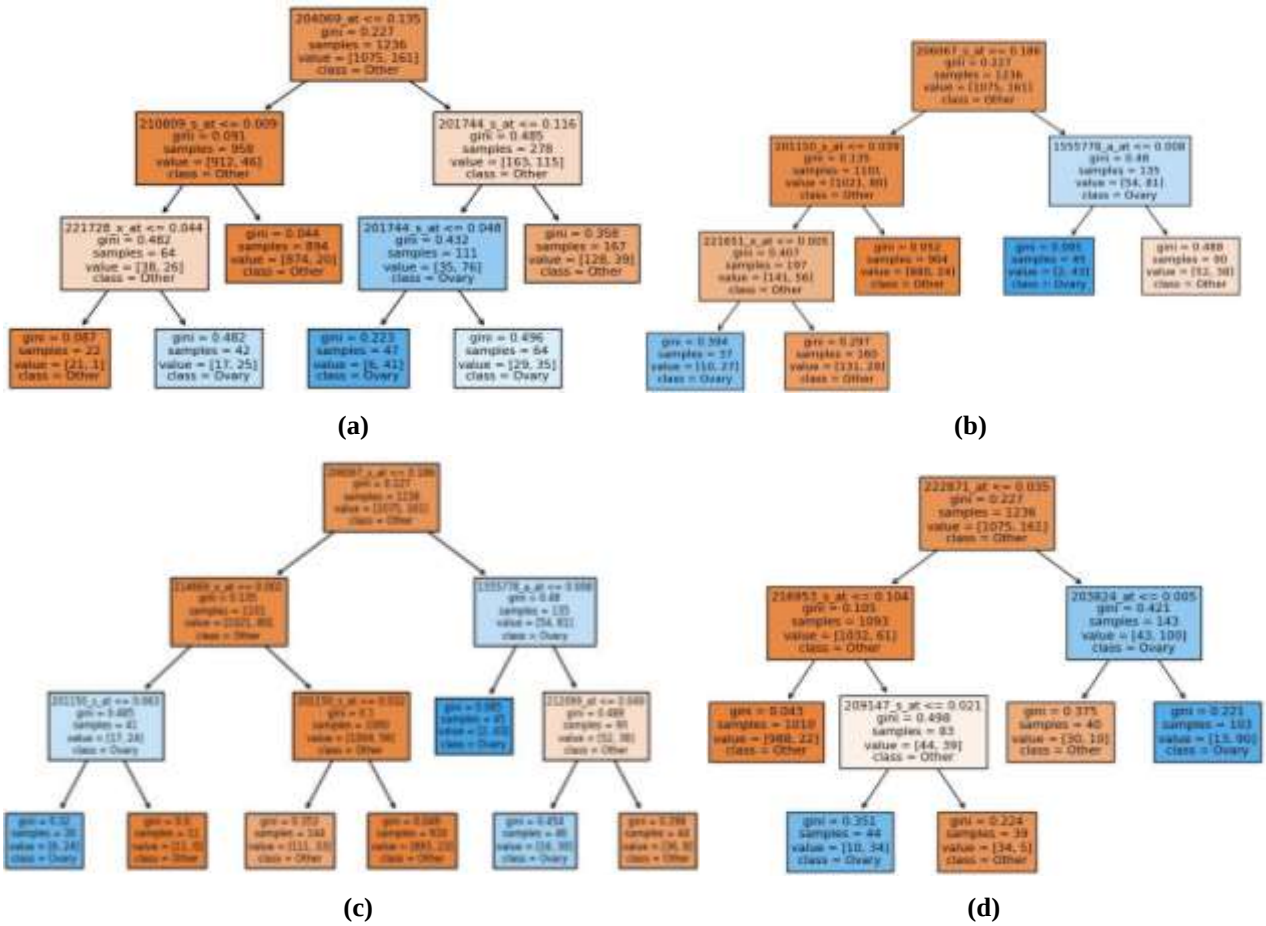
Based on BA values, CART, which uses 1000 selected genes, was chosen as the best model because it achieves the highest BA results in the testing process. The BA in the training process is also relatively high compared to the others. The best SIFS-CART model for ovarian cancer classification from OVA_ovary used 1000 selected genes and achieved 85.7% BA in training and 83.2% in testing. The classification results are listed in **Table 5**.

Table 5. The classification results of best SIFS-CART model

Stage	TP	TN	FP	FN
Training	116	1067	8	45
Testing	25	269	3	12

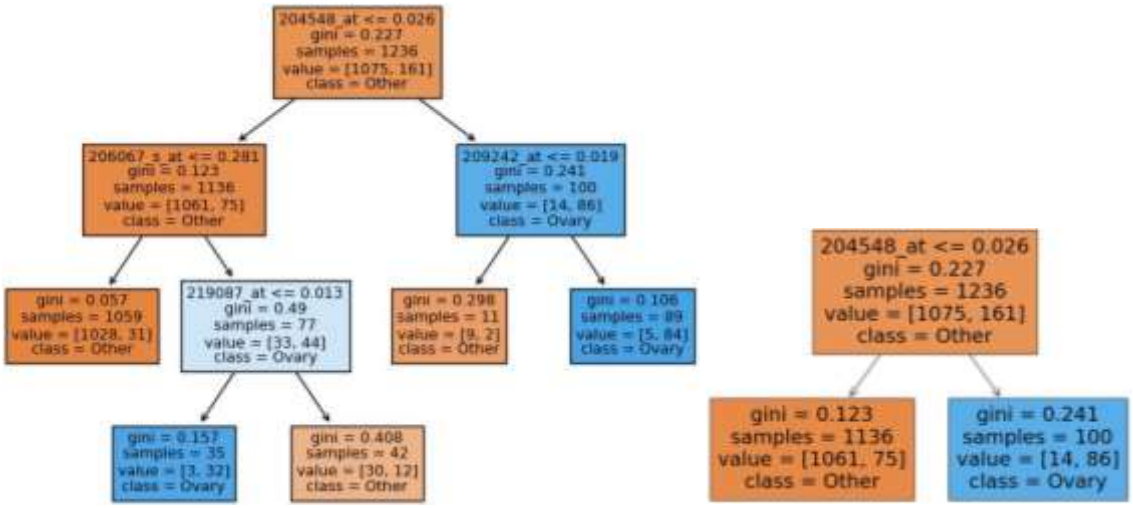
It obtained the best SIFS-CART model containing 1000 genes. In other words, SIFS selected 1000 out of 10935 genes for 1545 samples. It indicates that after going through the gene selection stage, the dataset (including the target column) has dimensions of 1001 columns and 1545 rows. The number of columns now is lower than the number of rows. It means that the OVA_ovary dataset has been successfully reduced to no longer have a high-dimensional form.

The final step is interpreting. The visualization of the optimal CARTs is shown in **Figure 4**. The first line at each non-leaf node (node with branches) in **Figure 4** contains the splitting criteria for creating tree branches. The first word refers to the name of the gene column. The word "gini" refers to the Gini impurity value at that node. "Samples" indicates the number of samples at that node. The first value in "value" states the number of 'Other' class samples at that node, and the second value is the number of 'Ovary' class samples at that node. Furthermore, "class" indicates the dominant class at that node. The more dominant the 'Other' class, the more orange the node color. On the other hand, the more dominant the 'Ovary' class, the bluer the node color.



(a) **(b)**

(c) **(d)**



(e) **(f)**

Figure 4. Optimal CART in SIFS-CART interpretations that using:
(a) 10 genes, (b) 50 genes, (c) 100 genes, (d) 500 genes, (e) 1000 genes, and (f) 5000 genes

The selected optimal ccp_alpha values successfully make the optimal CARTs in **Figure 4** look simple. Only a maximum of 6 genes are needed to create each tree. The best model, which uses 1000 genes, only requires four genes. Those genes are listed in **Table 6**. The information in **Table 6** obtained by inputting the gene column name (Probe ID) from CARTs to an annotation genes website named BioGPS which was initiated by Wu et al. on 2009. After that, we searched some literature in Google and Google Scholar using keywords built from the probe ID, gene symbol, gene name, and word 'cancer'.

Table 6. Important genes in ovarian cancer based on the results of best SIFS-CART model

Probe ID	Gene Symbol	Gene Name
204548_at	STAR	Steroidogenic acute regulatory protein
206067_s_at	WT1	Wilms tumor 1
209242_at	PEG3	Paternally expressed gene 3
219087_at	ASPN	Asporin

***Data source**: BioGPS (http://biogps.org/#goto=welcome)*

Based on the literature study, we got some information about gene interpretations. In **[27]**, it was explained that STAR predominantly regulates steroid synthesis. Steroid hormones are a large group of regulatory molecules produced in steroidogenic cells in various organs, including the ovaries. Estrogen is a steroid hormone. It is known that estrogen plays an important role in cancer cases, for example, endometrial, breast, and ovarian cancer.

A study **[28]** concluded that high WT1 expression levels correlated with aggressive clinical features in ovarian cancer. These results are supported by research **[29]** which found that WT1 expression levels can help doctors predict how aggressive a certain type of ovarian cancer, called high-grade serous ovarian carcinoma, will be.

PEG3 dysfunction commonly occurs in various types of cancer **[30]**. Decreased PEG3 expression levels have been detected in 18 cancer types **[31]**. In addition, PEG3 belongs to a group of genes that frequently change their activity due to epigenetic changes in the context of ovarian cancer **[32]**.

ASPN is one of the genes that may represent a new prognostic biomarker for gastric cancer **[33]**. Increased ASPN expression is associated with immune infiltration in endometriosis. ASPN can be used as a diagnostic biomarker and a potential immunotherapy target in endometriosis **[34]**. On the other hand, the association of ASPN with ovarian cancer specifically still has not been studied. This research found that ASPN is one of the candidate genes in the optimal CART. Thus, research in the medical field related to ovarian cancer can consider ASPN for further research.

Based on the literature study above, it can be concluded that STAR, WT1, and PEG3 do play a role in ovarian cancer cases. The results of this study support this information because STAR, WT1, and PEG3 are the top three selected genes used in the best optimal CART interpretation, namely SIFS-CART, which uses

1000 chosen genes. The results of this research also support previous research by Sa'adah et al. [14] which also used the OVA_ovary dataset. Research [14] using LASSO-CART also concluded that STAR and WT1 are ovarian cancer biomarkers because they are the top two genes on their optimal CART. However, since Sa'adah et al. [14] use different evaluation metrics (accuracy) and different data scaling methods (z-score normalization) and do not mention the TP, TN, FP, and FN results, the result on the best SIFS-CART model in this research cannot be compared.

4. CONCLUSIONS

Based on the findings gathered from the research, it can be inferred that:

a) The results of gene selection using SIFS succeeded in reducing the OVA_ovary dataset from 10935 gene columns to 1000 gene columns.

b) The best SIFS-CART model for ovarian cancer classification from OVA_ovary used 1000 genes. It achieves 85.7% BA in training and 83.2% BA in testing.

c) Optimal CART in the best SIFS-CART model only needs four genes to build it. Those genes are STAR, WT1, PEG3, and ASPN. Based on this literature study, STAR, WT1, and PEG3 play a role in ovarian cancer cases. However, the relationship between ASPN and ovarian cancer has not been studied by medical researchers.

Future research can analyze or fix the SIFS algorithm based on balanced accuracy to ensure it produces a more relatable trend. Another choice is testing another optimization algorithm to optimize SIFS parameters. Thus, research in the medical field related to ovarian cancer can consider ASPN for further research.

REFERENCES

[1] C. Slatnik and E. Duff, "Ovarian cancer: Ensuring early diagnosis," *Nurse Pract.*, vol. 40, no. 9, pp. 47–54, 2015.

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

[3] 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.

[4] F. Han, D. Tang, Y. Sun, Z. Cheng, J. Jiang, and Q. Li, "A hybrid gene selection method based on gene scoring strategy and improved particle swarm optimization," *BMC Bioinformatics*, vol. 20, no. 8, pp. 1–13, 2019.

[5] 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.

[6] A. Lacalamita, E. Piccinno, V. Scalavino, R. Bellotti, G. Giannelli, and G. Serino, "A Gene-Based Machine Learning Classifier Associated to the Colorectal Adenoma-Carcinoma Sequence.," *Biomedicines*, vol. 9, no. 12, 2021.

[7] X. Qin, S. Zhang, D. Yin, D. Chen, and X. Dong, "Two-stage feature selection for classification of gene expression data based on an improved Salp Swarm Algorithm," *Math. Biosci. Eng.*, vol. 19, no. 12, pp. 13747–13781, 2022.

[8] 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.

[9] C. Lai and H. Huang, "A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique," *Appl. Soft Comput. J.*, vol. 100, p. 106994, 2021.

[10] 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.

[11] 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.

[12] A. Abdellatif, H. Abdellatef, 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.

[13] 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.

[14] 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, 2020.

[15] B. C. Ross, "Mutual Information between Discrete and Continuous Data Sets," *PLoS One*, vol. 9, no. 2, pp. 1–5, 2014.

[16] T. Wong, "Performance evaluation of classi fi cation algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, pp. 2839–2846, 2015.

[17] Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015.

[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] G. Kunapuli, *Ensemble Methods for Machine Learning*, 6th ed. New York: Manning Publications, 2022.
[20] B. R. Kiran and J. Serra, "Cost-complexity pruning of random forests," *Artif. Intell. Med.*, pp. 222–232, 2017.
[21] 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.
[22] 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.
[23] 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.
[24] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," *arXiv*, pp. 1–17, 2020.
[25] 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," *Sci. J. Informatics*, vol. 11, no. 1, pp. 227–236, 2024.
[26] 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.
[27] P. R. Manna, C. L. Stetson, A. T. Slominski, and K. Pruitt, "Role of the steroidogenic acute regulatory protein in health and disease," *Endocrine*, vol. 51, pp. 7–21, 2016.
[28] Z. Liu *et al*., "High levels of wilms' tumor 1 (WT1) expression were associated with aggressive clinical features in ovarian cancer," *Anticancer Res.*, vol. 34, pp. 2331–2340, 2014.
[29] E. T. Taube *et al*., "Wilms tumor protein 1 (WT1) - Not only a diagnostic but also a prognostic marker in high-grade serous ovarian carcinoma," *Gynecol. Oncol.*, vol. 140, pp. 494–502, 2016.
[30] M. Zhang and J. Zhang, "PEG3 mutation is associated with elevated tumor mutation burden and poor prognosis in breast cancer," *Biosci. Rep.*, vol. 40, pp. 1–9, 2020.
[31] M. Li, Q. Sun, and X. Wang, "Transcriptional landscape of human cancers," *Oncotarget*, vol. 8, no. 21, pp. 34534–34551, 2017.
[32] A. Singh, S. Gupta, and M. Sachan, "Epigenetic biomarkers in the management of ovarian cancer: Current prospectives," *Front. Cell Dev. Biol.*, vol. 7, pp. 1–35, 2019.
[33] K. Jiang, H. Liu, D. Xie, and Q. Xiao, "Differentially expressed genes ASPN, COL1A1, FN1, VCAN and MUC5AC are potential prognostic biomarkers for gastric cancer," *Oncol. Lett.*, vol. 17, pp. 3191–3202, 2019.
[34] L. Wang and J. Sun, "ASPN Is a Potential Biomarker and Associated with Immune Infiltration in Endometriosis," *Genes (Basel).*, vol. 13, p. 1352, 2022.