Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: ijeeemi.org; Vol. 7, No. 1, pp. 154-166, February 2025; e-ISSN: 2656-8624.
https://doi.org/10.35882/ijeeemi.v7i1.49

RESEARCH PAPER

Dimensionality Reduction Using Principal Component Analysis and Feature Selection Using Genetic Algorithm with Support Vector Machine for Microarray Data Classification

Dwi Kartini[1], Rahmat Amin Badali[1], Muliadi[1], Dodon Turianto Nugrahadi[1], Fatma Indriani[1], Setyo Wahyu Saputro[1]
Department of Computer Science, Lambung Mangkurat University, Banjarbaru, South Borneo, Indonesia

ABSTRACT

DNA microarray is used to analyze gene expression on a large scale simultaneously and is a valuable tool for cancer diagnosis. The process of DNA microarray production starts by extracting RNA from the sample, which is converted into cDNA and scanned to generate gene expression data. However, the data acquired through this process is highly dimensional, influencing the performance of predictive models used for cancer detection. Consequently, the complexity of data needs to be minimized by dimensionality reduction. The aim of this research is to assess the effect of applying Principal Component Analysis (PCA) for dimensionality reduction, Genetic Algorithm (GA) for feature selection, and the combination of both in the classification of microarray data with Support Vector Machine (SVM). Datasets used are microarray datasets including breast cancer, ovarian cancer, and leukemia. The methodology of research involves preprocessing, PCA for reducing dimensions, GA feature selection, data splitting, SVM classification, and evaluating performance. According to the results, PCA dimension reduction coupled with GA feature selection and SVM classification gave the best results compared to other classifications. On the breast cancer dataset, the accuracy was highest at 73.33%, recall was 0.74, precision was 0.75, and the F1 score was 0.73. For the ovarian cancer dataset, accuracy was up to 98.68%, recall was 0.98, precision was 0.99, and F1 score was 0.99. For the leukemia dataset, accuracy was up to 95.45%, recall was 0.94, precision was 0.97, and F1 score was 0.95. It can be stated that the use of PCA to reduce features coupled with GA for feature selection in the classification of microarray can simplify the data and improve the SVM classification model accuracy. The finding of this study emphasizes the effectiveness of applying PCA and GA methods for enhancing the classification accuracy of microarray data.

PAPER HISTORY

Received Des. 010, 2024
Revised Feb. 10, 2025
Accepted Feb. 15, 2025
Published Feb. 25, 2025

KEYWORDS

Microarray;
Principal Component Analysis;
Genetic Algorithm;
Classification;

CONTACT:

[1]dwikartini@ulm.ac.id

1. INTRODUCTION

One of the technological advancements in bioinformatics in recent years is DNA Microarray. DNA microarray technology is used to determine gene expression levels in large and varying quantities simultaneously within a single experiment [1]. This technology offers benefits in various biological studies, including cancer detection. In its use for cancer detection, classification methods are employed, allowing medical professionals to diagnose whether a person has cancer. Despite its great utility, DNA microarray has the characteristic of having very high dimensionality with a limited number of samples, which affects the results of gene expression classification [2]. To address this limitation, dimensionality reduction becomes an important step to simplify the data without losing too much information, thus improving the classification model's performance.

Several studies on DNA microarray classification have been conducted. For example, the study [3] compared various classification methods such as SVM, Multi-Layer Perceptron, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). The results showed that SVM achieved the highest accuracy of 99% on the chronic obstructive pulmonary disease (COPD) dataset compared to other classification methods.

The study by [4] used Principal Component Analysis (PCA) dimensionality reduction and Linear Discriminant Analysis (LDA)-based classification for colon cancer detection, achieving a 29.04% increase in accuracy compared to the classification accuracy without PCA, which was 58.06%. This research demonstrates that PCA can improve microarray data classification performance by eliminating noise and reducing the number of features in the dataset. The study [5] used Genetic Algorithm (GA) and Differential Evolution (DE) feature selection methods for microarray data classification. The GA feature selection method achieved an accuracy of 93% on the prostate dataset, higher than the performance of the DE method, which achieved 90%. This shows that using GA feature selection can effectively classify microarray data. The study [6] used machine learning algorithms SVM, KNN, and Naive Bayes for microarray data classification. The results showed that SVM outperformed the KNN and

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

Naive Bayes classification methods, achieving an average accuracy of 88% across the three datasets used. The study [7] used Spider Monkey Optimization to classify microarray data with KNN, Decision Tree, Naive Bayes, Random Forest, and SVM, where SVM achieved the highest accuracy of 100%.

This shows that SVM is the most effective classification method for microarray data, consistently outperforming other algorithms in terms of accuracy. In recent studies on microarray data classification, dimensionality reduction and feature selection are combined to obtain the most significant feature subsets from the data. In the study [8], the Information Gain (IG) feature selection method was combined with the GA method for microarray data classification. The combination of these two methods resulted in higher accuracy (85.71%) for the brain cancer dataset, compared to using the IG method alone, which achieved only 73.81%. The study [9] compared the combination of Independent Component Analysis with PSO feature selection and the combination of Independent Component Analysis with GA feature selection for microarray data classification using Naive Bayes, and the results showed that the ICA+GA combination achieved a higher accuracy of 96.68%, compared to 95.45% for the ICA+PSO combination on the Acute Leukemia dataset. This demonstrates that using a combination of two dimensionality reduction or feature selection techniques can improve accuracy in microarray data classification.

PCA is used to reduce the dimensionality of the dataset by transforming it from correlated dimensions into uncorrelated dimensions [10]. Next, GA will be used as a feature selection method to choose informative features from the results of dimensionality reduction. GA is a feature selection method that applies the concept of natural selection, where only the best individuals survive and are reused in the next selection process. GA is a type of wrapper feature selection method that produces more effective performance compared to filter-based feature selection methods, with a smaller subset of features [11] The classification process of the feature-selected data will be performed using the SVM algorithm. According to [12], SVM uses linear functions in high-dimensional features based on optimization theory to train the classification model. Before building the SVM model, the kernel function needs to be determined, such as polynomial, linear, or radial basis function (RBF), which plays an essential role in the learning process [13]. The study [14] used SVM with the RBF kernel, achieving the highest accuracy of 99.8% compared to other SVM kernels. Based on this, the RBF kernel will be used in this study as the kernel trick in the SVM classification method.

PCA was applied in this research because it is able to reduce the dimension of the dataset without losing variance by converting correlated variables into uncorrelated principal components. PCA is also more tolerant of high-dimensional data and is unsupervised, as opposed to LDA, which needs labeled samples. While ICA was available, it is far more effort and less interpretable. The GA natural selection, however, was used for feature selection because it dynamically optimizes good features based on classification performance and hence is extremely helpful in the scenario of microarray data. SVM with an RBF kernel for classification, GA for feature selection, and PCA for reducing dimensionality are combined in this research. There are four test environments that will be utilized for the comparison of the performance of each method: (1) SVM classification alone, (2) SVM classification plus PCA, (3) SVM classification plus GA, and (4) classification using both PCA and GA together with SVM. Classification performance will be likened by means of the accuracy, precision, recall, and F1-score measures. Comparison will also be conducted with baseline models to determine if PCA+GA enhances microarray classification by SVM.

2. MATERIALS AND METHOD

In this study, the research flow is used to describe the steps taken for each test. Figure 1 shows the research flow used.

A. Data Collection

This study uses three gene expression microarray datasets: Breast Cancer, Ovarian Cancer, and Leukemia [15]. Descriptions of these three datasets can be seen in Table 1.

Table 1. Description of the Datasets Used

Dataset	Number of Data	Number of Features
Breast Cancer	97	24.481
Ovarian Cancer	253	15.154
Leukemia	72	7129

The breast cancer dataset has 97 data and 24,481 features consisting of 2 classes, namely *relapse* with 46 data and *non-relapse* with 51 data. In the ovarian cancer dataset there is a total of 253 data and a total of 15,154 features consisting of 2 classes, namely *normal* with 91 data and *cancer* consisting of 162 data. Meanwhile, the leukemia dataset has a total of 72 data and a total of 7129 features consisting of 2 classes, namely *Acute Lymphoblastic Leukemia* (ALL) with 47 data and *Acute Myeloid Leukemia* (AML) with 25 data.

B. Data Preprocessing

Both data normalization and label encoding were used as preprocessing process. The Binary Encoding method was used to encode labels, transforming categorical labels into a numerical representation between 0 and 1.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

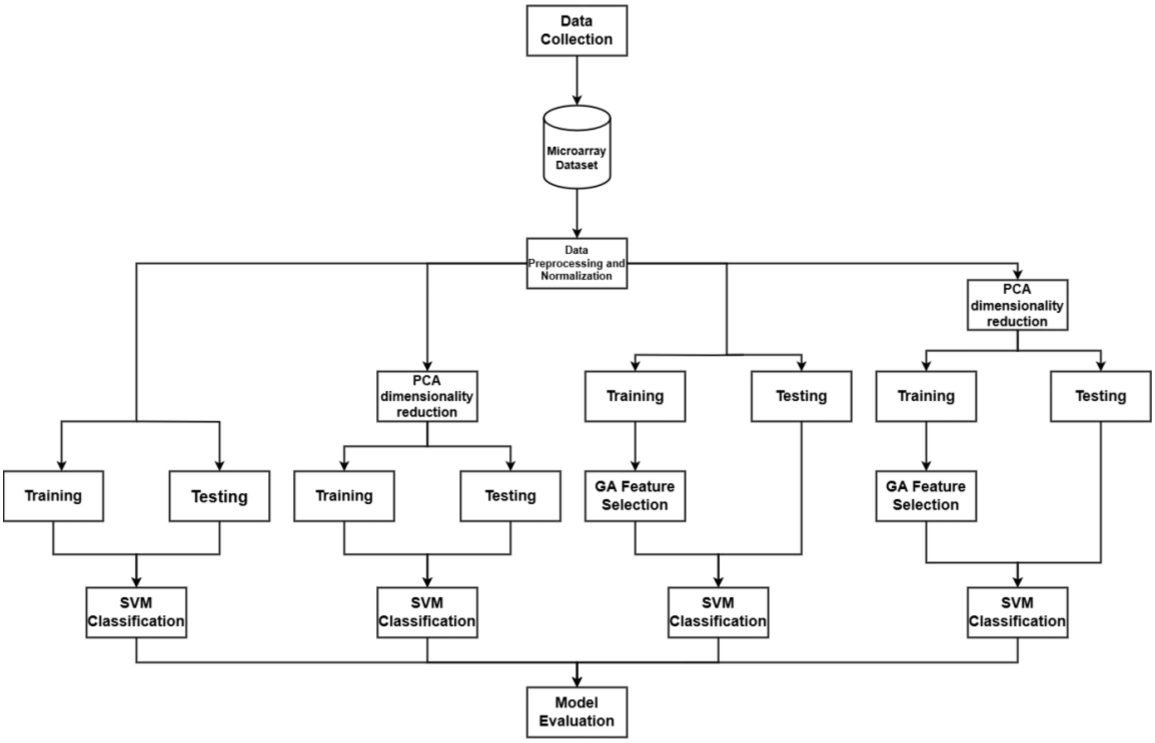

Figure. 1. Research Flow

To ensure that sure all feature values were inside the range [0,1], normalization was executed out using Min-Max Normalization to prevent the dominance of features with larger value ranges over other features in the classification model.

1. Label Encoding

Label encoding is a method used to convert categorical data in the form of text labels into numeric format [16] In this research, the encoding process is carried out using the Binary Encoding method, which converts the data on the "Class" feature which is still a string into binary numbers (0 and 1).

2. Data Normalization

The main purpose of normalization is to scale all features in the dataset, so that no feature with a larger value dominates or exerts excessive influence on the analysis or model that uses the data. In addition, normalization can also help reduce the computation time of the model, as well as ensure that the feature values in the dataset have a balanced range without changing the information contained in it [17].

One of the commonly used data normalization methods for classification is *Min Max Normalization* Min-Max normalization is a method in data processing used to transform the values in a dataset into a specific range, usually between 0 and 1 [18] The following equation is used in *Min Max Normalization*:

$$x' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

C. Dimensionality Reduction

Dimensionality reduction is a reduction in the number of dimensions in a dataset with the consideration that important information in the dataset is maintained after the process is carried out [16]. In this research, the dimensionality reduction method used is Principal Component Analysis (PCA). PCA is paka n a multivariate statistical method that performs a linear transformation of an initial set of features into a smaller, uncorrelated set of features, and is able to represent information from these initial features [19] The steps to perform PCA dimensionality reduction according to [20] are as follows:

1. Suppose X is the input matrix to perform PCA, where X is the normalized data with dimension $.n \times m$

2. Calculate the mean value of each feature in the dataset using equation (2).

$$\bar{X} = \sum_{i=1}^{n} \frac{1}{X_i} \quad (2)$$

3. Calculating the covariance matrix using equation (3)

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \quad (3)$$

4. Next is to calculate the eigenvalue λ_m and eigenvector v_m based on the covariance matrix using equation (4).

$$C_x \cdot v_m = \lambda_m \cdot v_m \quad (4)$$

5. Sort the eigenvalues from largest to smallest. A large eigenvalue shows how much data variance is explained by the eigenvector.

6. The last steps is to transform the data using the eigenvector that has been selected to produce a new data representation with smaller dimensions but still maintaining the largest variance of the original data. The following is the equation:

$$X_{PCA} = x'.v_m \quad (5)$$

To reduce the dimension of PC based on eigenvalues, one of the commonly used parameters is using *proportion of variance* (PPV). PPV calculation can be done using equation 6.

$$PPV = \frac{\lambda_m}{\sum_{i=1}^{p}\lambda_i} \quad (6)$$

D. Feature Selection

Feature selection aims to select the smallest number of feature subsets that include individually relevant features and interactions between features, so as to explain differences between classes with minimal loss of information [21]. In this research, the feature selection method used is *Genetic Algorithm* (GA). GA is a feature selection method that applies the concept of natural selection, where only the best individuals can survive and be reused in the next selection process. According to [8]there are four main steps that are generally used in GA, namely

1. Chromosome encoding: this process is done by randomly generating a series of binary numbers with a length according to the number of features.
2. Population initialization: at the initial stage, the GA randomly generates an initial population of chromosomes representing a subset of potential features where the number of chromosomes formed corresponds to the specified population size.
3. Fitness value evaluation: fitness value states how good the value of an individual or optimal solution is obtained. The fitness value in this study uses the accuracy value of the results of the machine learning method classification process
4. Reproduction: in GA there are 3 genetic operations used, namely:
   Selection: The selection process aims to select the chromosomes that will serve as parents for new offspring.
   Crossover: *The crossover* process is done by randomly swapping some genes from two parents' chromosomes to produce a new chromosome (*offspring*).
   Mutation: Mutation in GA is performed randomly from the *offspring* chromosome by changing the gene value from 0 to 1 or vice versa.

GA iteratively evolves the chromosomes through a process of fitness evaluation and reproduction in order to obtain a better solution. This process stops when it reaches certain stopping criteria such as the number of iterations or reaches a condition where there is no significant improvement in the solution (convergent). The individual with the highest fitness value is then considered the best solution[22]

E. Data Splitting

Data Splitting data is done by dividing the data into two parts, namely data for the training process and data for the testing process. Training data will be used to train the classification model, while testing data will be used to test the performance of the model. Research [23], using 70% training data division and 30% testing data for SVM classification in Alzheimer's disease, the results obtained an accuracy of 98.16% greater than using k-fold cross validation with an accuracy obtained of 94.65%. Based on this, this research will use a data comparison ratio of 70% for training data and 30% for testing data.

F. Classification

The microarray data classification process is carried out using the Support Vector Machine (SVM) method. The concept of SVM classification is to separate data from various classes by finding the optimal hyperplane that maximizes the margin between classes. In its application, SVM uses a kernel to map non-linear data to a higher dimensional space to make it easier to separate. In this research, the kernel used is the RBF kernel. To build an SVM classification model using the RBF kernel, there are two parameters that must be determined first, namely C and Gamma. In research [24]SVM classification with RBF kernel using parameter C = 1 and gamma = "scale" produces good performance with 77.08% accuracy. Therefore, in this study the parameter values used for SVM classification with RBF kernel are C = 1 and gamma= "scale" which is calculated using equation 7.

$$gamma = \frac{1}{n_{fitur}.var()} \quad (7)$$

The following is the equation used in the SVM classification process [1]:

$$\vec{w}\cdot\vec{x} + b = 0 \quad (8)$$

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

For class +1

$$\vec{w} \cdot \vec{x} + b \ \geq 1 \tag{9}$$

and for class -1

$$\vec{w} \cdot \vec{x} + b \ \leq -1 \tag{10}$$

In order to determine the optimal *hyperplane* for both classes using the equation:

$$min_w \frac{1}{2}\left(\left|\left|\vec{w}\right|\right|\right)^2 \tag{11}$$

with

$$y_i(\vec{x_i}, \vec{w} + b) - 1 \ \geq 0 \tag{12}$$

Where $\vec{x_i}$ is the i-th input data to find the value of w and b parameters, while the value of y_i is the output data. Then the decision function on the RBF kernel is done with the equation:

$$RBF : K\left(\vec{x_i}, \vec{x_j}\right) = \ exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{13}$$

Where xi and xj are vectors from the feature space, and σ is a free parameter that controls how quickly the distance between two points will lose its influence on the kernel value.

G. Model Evaluation

A confusion matrix is used in this study as a metric to evaluate the classification model's quality. An open assessment of the model's correctness is made possible by a confusion matrix, which is a table format that shows the number of test data examples that are properly and wrongly classified [25]. The confusion matrix can be used to get the following metrics [26].

1. Accuracy, which is a metric that measures how accurate the model is in classifying the data correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

2. Precision, a metric that describes how many positive predictions are correct.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

3. Recall is a metric that describes how well the model is able to find positive data.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

4. F1 score, which is the average comparison between precision and recall.

$$F1\ score = 2\ \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

3. RESULTS

A. Data Preprocessing Result

Data preprocessing is done using two stages, namely label encoding and data normalization. Label encoding is done using binary encoding. This label encoding process is necessary because some machine learning methods such as SVM can only work with category labels in the form of numeric data. Table 2 shows the label encoding results for each dataset

Table 2. Label Encoding for Each Dataset

Dataset	Before Label Encoding	After Label Encoding
Breast Cancer	relapse	1
	non-relapse	0
Ovarian Cancer	normal	1
	cancer	0
Leukemia	ALL	0
	AML	1

B. Dimensionality Reduction Result

The dimensionality reduction process is done by selecting the eigen vectors and eigen values generated from the covariance matrix. To determine how many Principal Components (PCs) are retained, proportion of variance (PPV) is used with various threshold values.

The threshold value determines how many PCs are selected during the dimensionality reduction process. According to [27], 70% threshold is the minimum limit point to determine how many PCs should be retained. Meanwhile, research [28] uses a variance proportion of 100% and gets the number of PCs resulting from PCA dimensionality reduction as many as 61 PCs from a dimension of 62 × 2,000.The number of PCs in an analysis is determined by the Proportion of Variance Explained (PPV), with a kumulatif threshold of 70%, 80%, 90%, 95%, and 100%. This threshold is used to determine the minimal PC that is still able to support the largest amount of variation in the dataset, hence reducing the dimensions without compromising important information. In this study, PPV is calculated for the datasets of breast cancer, ovarian cancer, and leukemia in order to determine the number of ideal PCs used in the classification process. The cumulative value of PPV on the leukemia dataset is shown in Table 3. Based on Table 3, It can be seen that to get a PPV of 100%, 71 PCs are required. In addition, PC 1 is known to be the PC with the largest eigenvalue of 41.747810 which represents a variance proportion of 16.077206%. This eigenvalue then continues to decrease until PC 71 of 0.900398, while PCs below 71 are not selected because they do not have enough information.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

Table 3. Cumulative PPV Value on Leukemia Dataset

PC	Eigen Value	Total PPV (%)
1	41.75781	16.077206
2	24.695514	25.585244
3	12.368428	30.347221
4	10.263153	34.298645
5	9.015112	37.769561
6	8.121143	40.896288
7	6.878518	43.544591
…..	….	…..
…..	….	…..
…..	….	…..
…..	….	…..
…..	….	…..
60	1.32768	95.412081
61	1.28187	95.905615
66	1.071141	98.116144
67	1.051194	98.520865
68	1.028532	98.916861
69	0.971105	99.290747
70	0.941766	99.653337
71	0.900398	100

The number of PCs retained for each PPV value can be seen in Table 4. From these results, it is known that as the PPV increases, the more the number of PCs retained on the breast cancer, ovarian cancer, and leukemia datasets if a threshold of 70% is used, the number of PCs retained are 20, 4, and 25, respectively. Meanwhile, if a threshold of 100% is used, the number of PCs retained are 96, 252, and 71, respectively.

Table 4. Number of PC for Each PPV Value

PPV (%)	Number of PC		
	Breast	Ovarian	Leukemia
70	20	4	25
80	33	5	37
90	55	12	51
95	71	24	60
100	96	252	71

C. Feature Selection Result

In the Principal Component Analysis (PCA) dimensionality reduction procedure, the Genetic Algorithm (GA) was utilized for feature selection in this investigation. As indicated in Table 5, the datasets were split into 70% training data and 30% test data prior to the use of GA.To prevent data leaking, which could result in inaccurate model evaluation and excessively optimistic performance estimations, feature selection was limited to the training data. This ensured that information from the test data did not affect the feature selection process.

Table 5. Data Splitting for Each Dataset

Dataset	Total Data	
	Training Data (70%)	Testing Data (30%)
Breast Cancer	67	30
Ovarian Cancer	177	76
Leukemia	50	22

GA was set up to have a population of 20 people and 20 iterations per generation.Rank selection was used for the individual selection process, and the crossover mechanism was used with 0.6, 0.8, and 0.9 probability as well as 0.02, 0.05, 0.1, and 0.5 for mutation.After a predetermined number of iterations, convergence criteria were established based on the fitness value not showing any discernible improvement. References to earlier research and pilot experiments served as the basis for choosing these settings in order to optimize feature selection while preserving crucial dataset information. The GA feature selection process begins with random population initialization using binary coding. The population size was set at 20 individuals, and the chromosome length was adjusted to the number of dataset features. Each chromosome had its fitness value calculated based on the SVM classification accuracy value. Selection is performed using the rank selection method to select the parent chromosome, which then undergoes crossover with a predetermined probability in Table 6

Table 6. GA Parameters

Dataset	Values	Source
Number of Population	20	[29], [8], [30], [31], [32], [33]
Generation	20	
Crossover Probability	0.6, 0.8, and 0.9	
Mutation Probability	0.02, 0.05, 0.1 and 0.5	

This Figure 2 shows how the Genetic Algorithm's (GA) fitness value changed over 20 generations using the breast cancer dataset. Early on, the solution has not improved much, as evidenced by the fitness value being steady at about 0.75 until the 13th generation. The 14th generation showed some slight gains, suggesting that the crossover and selection processes were starting to yield better feature combinations. Mutation also helped to keep

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

the population diverse so that it wouldn't become stuck in locally inefficient solutions.Generations 18 to 20 see a notable jump in fitness, rising to 0.82, suggesting that selection, crossover, and mutation have all worked together to find an alternate solution. The best condition is reached when the fitness value rises noticeably and eventually stabilizes, signifying that the ideal solution has been discovered. In this procedure, the GA progressively finds the best answer from generation to generation.The individual with the highest fitness value is then considered the best solution[22].

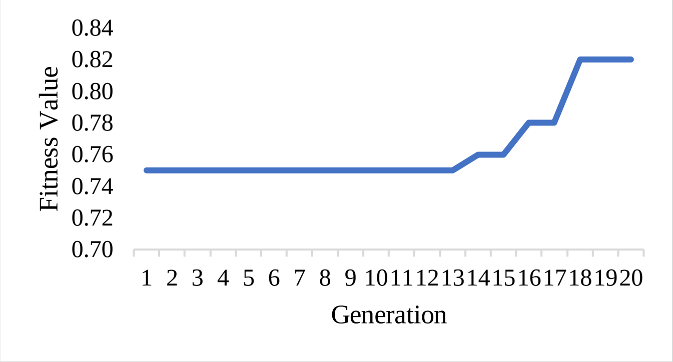

Figure. 2. Graph of Fitness Value on Breast Cancer Dataset

This Figure 3 displays a development of the GA's fitness value during 20 generations for the Ovarian Cancer dataset. So until about the eighth generation, the fitness value initially hovers around 0.945 with minor oscillations, suggesting that the solution has not made much progress. Between the ninth and eleventh generations, there was a noticeable increase, suggesting that the crossover and mutation process was successful in producing a superior feature combination, which raised the fitness value considerably. The fitness value stays constant until the 20th generation, after peaking at about 0.965 in the 11th generation. This suggests that the ideal solution has been discovered and that no more progress is made by following iterations.

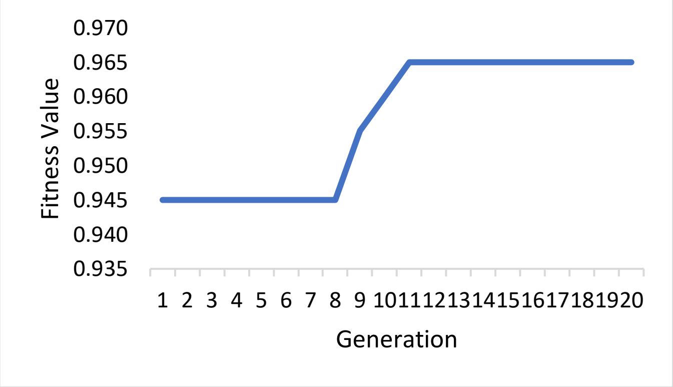

Figure. 3. Graph of Fitness Value on Ovarian Cancer Dataset

This Figure 4 shows the change in fitness value in the GA on the leukemia dataset over 20 generations. In the early stages, the fitness value stabilizes around 0.86 until the 5th generation, indicating that the solution has not improved significantly. A sharp spike occurred in the 6th generation, where the fitness value increased significantly to reach 0.88. This increase indicates that the selection, crossover and mutation processes have successfully optimized a better combination of features, thereby drastically improving the performance of the model in a short period of time.After this spike, the fitness value remains stable until the 20th generation, indicating that the optimal solution has been found early, and subsequent iterations no longer provide further improvement in performance.

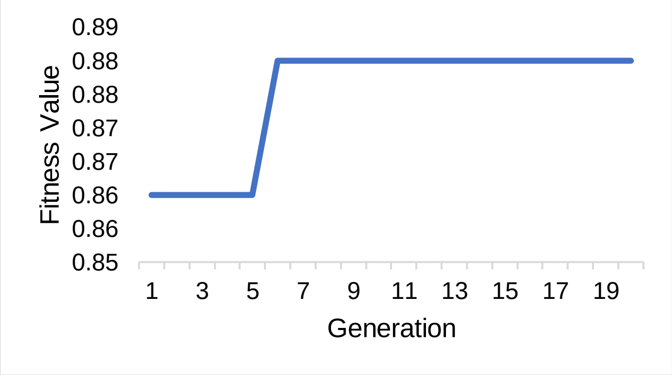

Figure. 4. Graph of Fitness Value on Leukemia Dataset

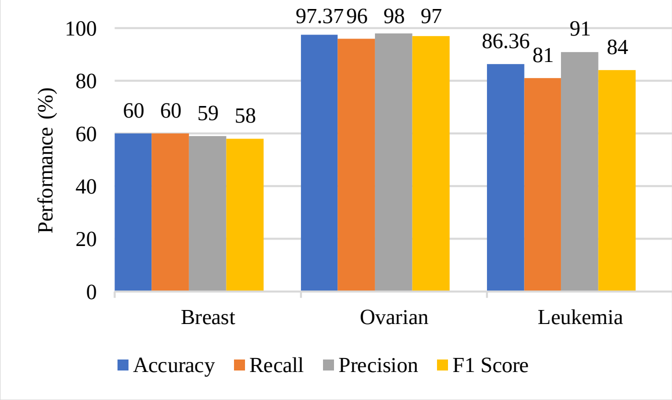

Figure 5. SVM Classification Evaluation Graph

D. SVM Classification

The model is built using all data and features from the dataset. The data is divided into 2 parts, namely training data used to train the model with a 70% ratio, and testing data to test the performance of the model with a 30% ratio. The kernel used is the RBF kernel with a parameter value of C = 1 and a gamma value calculated using equation 7. Based on the accuracy and evaluation results of SVM classification displayed in Figure 5, on the breast cancer dataset the resulting performance is quite low, namely 60% accuracy, recall 60%, precision 59%, and F1 score 58%. On the ovarian cancer dataset, the performance is very good, namely 97.37% accuracy, recall 96%, precision 98%, and F1 score 97% and on the leukemia

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

dataset the accuracy value is 86.36%. recall 81%, precision 91%, and F1 score 81%.

E. SVM Classification with PCA

In order to minimize the dataset's dimensions, the PCA approach is used to build the model first. How much PC is kept is determined by the PPV parameter. It is visible from analyzing Table 7 PPV parameter that the SVM classification parameters—in particular, PC and Gamma—differ depending on the dataset. The findings show that, for all datasets, PC values continue to improve as PPV does, but Gamma values continue to fall. Table 7 shows that the SVM classification parameters utilized are still the same as those from the previous model used.

Table 7. the SVM classification parameters PCA

Dataset	PPV (%)	PC	Gamma
Breast Cancer	70	20	0.002281
	80	33	0.002053
	90	55	0.001858
	95	71	0.001771
	100	97	0.001693
Ovarian Cancer	70	4	0.002562
	80	5	0.002447
	90	12	0.002182
	95	24	0.002067
	100	252	0.001968
Leukemia	70	25	0.005186
	80	37	0.004563
	90	51	0.004142
	95	60	0.003940
	100	71	0.003768

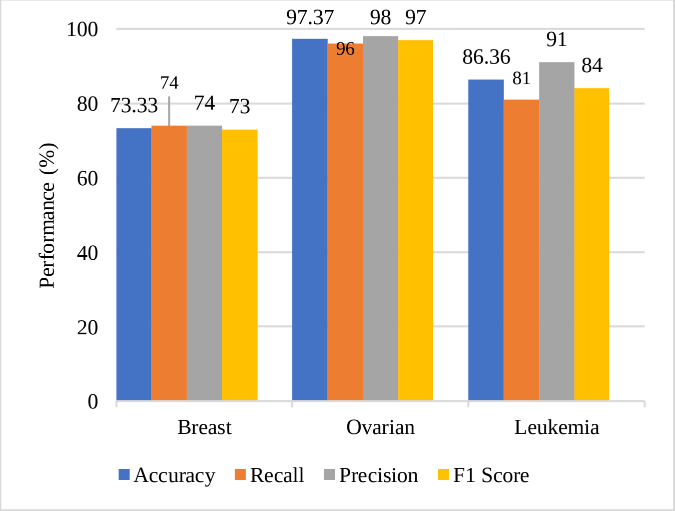

Figure 6. Classification Evaluation Graph of PCA + SVM

Based on the PCA + SVM classification results shown in Table 7, it can be seen that increasing PPV along with the number of PCs generated does not always indicate better performance. The performance increases gradually from 70% - 90% PPV, but decreases afterwards because there are more PCs with small eigenvalues that can interfere with PCs with informative data in the classification process. The best performance obtained from this test is then evaluated and the results can be seen in Figure 6.

In the breast cancer dataset, the best performance was obtained using 90% PPV which resulted in 55 PCs, 73.33% accuracy, precision 74%, recall 74% and F1-score 73%. In the ovarian cancer dataset, the best performance was achieved by producing 24 PCs, 97.37% accuracy, precision 98%, recall 96%, and F1-score 97% obtained when using PPV 95%. While in the leukemia dataset, the best performance was obtained by using 70% PPV which resulted in 25 PCs, 86.36% accuracy, precision 91%, recall 81%, and F1-score 84%.

F. SVM Classification with GA

This test was conducted using GA feature selection first to select the best features from the dataset. Parameter combinations of crossover probabilities with values of 0.6, 0.8 and 0.9 and mutation probabilities with values of 0.02, 0.05 0.1, and 0.5 are used to obtain more chromosome variations and determine their effect on classification performance. Table 8 shows that the SVM classification setup is still the same as it was for the previous model being used.

Table 8. Parameter Optimization GA for SVM Classification

Dataset	Pc	Pm	Gamma
Breast Cancer	0.6	0.02	0.000928
	0.6	0.05	0.000926
	0.6	0.1	0.000944
	0.6	0.5	0.000930
	0.8	0.1	0.000925
	0.9	0.1	0.000938
Ovarian Cancer	0.6	0.02	0.002953
	0.6	0.05	0.002975
	0.6	0.1	0.002977
	0.6	0.5	0.002994
	0.8	0.1	0.003002
	0.9	0.1	0.002923
Leukemia	0.6	0.02	0.004843
	0.6	0.05	0.004749
	0.6	0.1	0.004724
	0.6	0.5	0.004822
	0.8	0.1	0.004690
	0.9	0.1	0.004719

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia**.**

Based on the evaluation results of GA+SVM classification evaluation results shown in Figure 7, on breast cancer dataset, the best performance is obtained with 63.33% accuracy, precision 63%, recall 62% and F1-score 62%. While on the ovarian cancer dataset, the best performance was obtained with 97.37% accuracy, precision 98%, recall 96%, and F1-score 97%. And on the leukemia dataset, the best performance was obtained with an accuracy of 90.91%, precision 94%, recall 88%, and F1-score 90%.

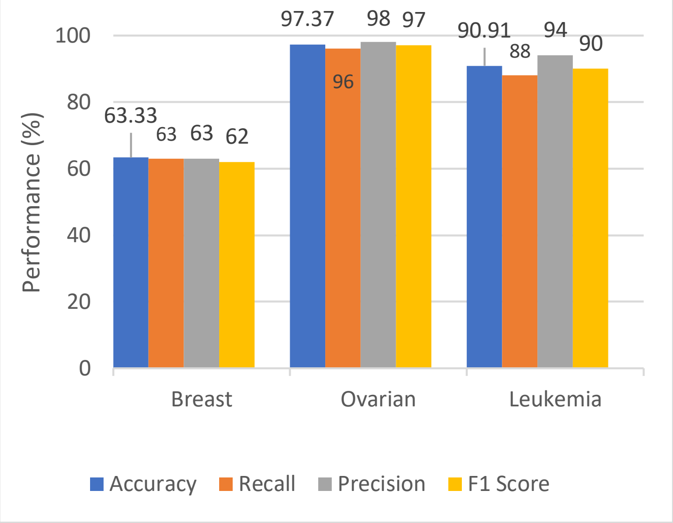

Figure 7. GA + SVM Classification Evaluation Graph

G. SVM Classification with PCA and GA

The last test is carried out using PCA for dimensionality reduction first, the dataset that has been reduced in dimension will then be selected for its best features through GA feature selection and finally the dataset with the selected feature subset is used as input in SVM classification. The PPV parameter value used for PCA is taken based on the best performance in the previous test, namely 90% PPV for breast cancer dataset, 95% PPV for ovarian cancer dataset, and 70% PPV for leukemia dataset. The GA and classification parameters used are the same as the previous test. The classification accuracy results of PCA, GA, and SVM can be seen in Table 9.

Table 9. Classification Accuracy of PCA + GA + SVM

Dataset	Pc	Pm	C	Gamma	Accuracy (%)
Breast Cancer	0.6	0.02	1	0.005549	63.33
	0.6	0.05		0.002555	73.33
	0.6	0.1		0.005132	66.67
	0.6	0.5		0.004574	70.00
	0.8	0.1		0.005881	66.67
	0.9	0.1		0.004542	60.00
Ovarian Cancer	0.6	0.02		0.003411	94.74
	0.6	0.05		0.006222	97.37
	0.6	0.1		0.003324	96.05
	0.6	0.5		0.006228	96.05
	0.8	0.1		0.006767	98.68
	0.9	0.1		0.007404	96.05
Leukemia	0.6	0.02		0.007507	90.91
	0.6	0.05		0.010325	90.91
	0.6	0.1		0.012214	90.91
	0.6	0.5		0.009615	90.91
	0.8	0.1		0.012167	95.45
	0.9	0.1		0.010094	90.91

Based on the PCA+GA+SVM classification evaluation results shown in Figure 8, on the breast cancer dataset, the best performance was obtained with an accuracy of 73.33%, precision 75%, recall 74%, and F1-score 73%. On the ovarian cancer dataset, the best performance was obtained using a combination of crossover probability 0.8 and mutation probability 0.1, namely accuracy of 98.68%, precision 99%, recall 98%, and F1-score 99%. While in the leukemia dataset, the best performance was obtained by 95.45% accuracy, precision 97%, recall 94%, and F1-score 95%.

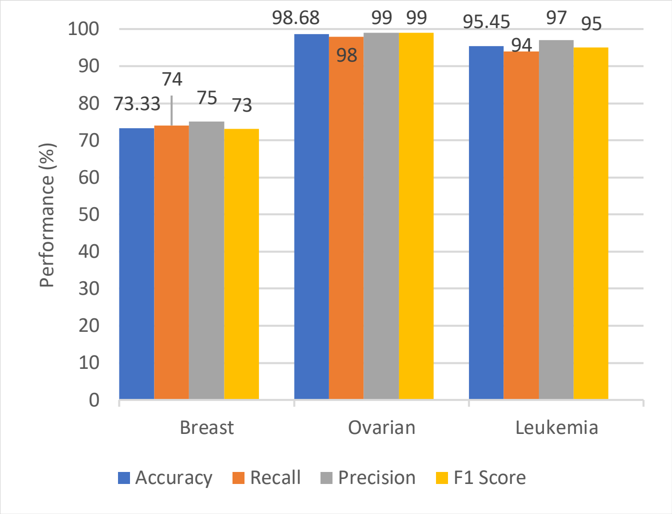

Figure 8. PCA + GA + SVM Classification Evaluation Graph

4. DISCUSSION

Based on the research results that have been described, there are four testing models carried out, namely SVM classification, SVM classification with PCA, SVM classification with GA, and SVM classification with a combination of PCA and GA. After all testing models are evaluated for performance, the evaluation results will be compared to determine whether there is an increase or decrease in the performance of each model. In addition, the number of features obtained in each test is also

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

compared. A comparison of the number of features generated from each test model is shown in Table 11.

Table 11. Comparasion of the Number of Features in Each Model

Dataset	Model	Number of Features
Breast Cancer	SVM	24481
	PCA+SVM	55
	GA+SVM	12127
	PCA+GA+SVM	28
Ovarian Cancer	SVM	15154
	PCA+SVM	24
	GA+SVM	7465
	PCA+GA+SVM	11
Leukemia	SVM	7129
	PCA+SVM	25
	GA+SVM	3541
	PCA+GA+SVM	9

SVM classification testing uses all features from the dataset, while in PCA + SVM classification testing using PPV, the number of features on the breast cancer dataset is reduced to 55 features, on the ovarian dataset the number of features is reduced to 24 features, while on the leukemia dataset the number of features is reduced to 25 features. In the GA + SVM test, the number of features was successfully reduced by approximately 50% using a combination of crossover probability and mutation probability. While in the PCA+GA+SVM test, the dataset that has been reduced by PCA is then selected by GA, on the breast cancer dataset the number of features is reduced to only 28 features, on the ovarian cancer dataset the number of features is reduced to only 11 features while on the leukemia dataset the number of features is reduced to only 9 features.

Based on the performance comparison results in Figure 9, it can be seen that the SVM model without optimization has the lowest performance, especially on the breast cancer dataset, with only 60% accuracy, 60% recall, 59% precision, and 58% F1-score. However, SVM performed better on the ovarian and leukemia datasets, with accuracies of 97.37% and 86.36%, respectively. This shows that although SVM can perform quite well on datasets with small features such as the ovarium dataset, this model still has limitations when applied to datasets with a high number of features such as breast cancer.

Applying PCA for dimensionality reduction, there was a significant improvement on the breast cancer dataset, with an increase in accuracy of 13.33%, recall of 14%, precision of 15%, and F1-score of 15%. As for the ovarian cancer and leukemia datasets, the performance remained stable compared to the previous test, even though the number of features used was much less. This shows that PCA is effective in reducing data complexity without impacting SVM classification accuracy.

The GA+SVM model results showed that even though the number of features was reduced to 50% of the initial dataset, the performance was still lower than PCA+SVM. This is due to the presence of irrelevant features, which impacts the SVM prediction accuracy. On the breast cancer dataset, the performance of the GA+SVM model decreased compared to PCA+SVM, with a 10% decrease in accuracy, 12% recall, 11% precision, and 11% F1-score. Nevertheless, this model is still better than SVM without optimization. Meanwhile, on the ovarian cancer dataset, the performance remained stable with no significant change, while on the leukemia dataset, there was an increase of 4.55% in accuracy recall 7%, precision 3%, and F1-score 6%. Although GA successfully reduced the number of features, the results show that without combining it with PCA, the performance is still suboptimal. GA+SVM was more effective in certain datasets, such as leukemia, but less competitive in other datasets than the model combining PCA and GA. This indicates that although GA is able to extract more relevant features, initial dimensionality reduction such as PCA is still required to improve classification efficiency.

The final evaluation showed that the PCA+GA+SVM model provided the best performance across all datasets, with more optimized feature selection and improved SVM accuracy. The combination of PCA and GA resulted in a dataset with fewer features, but retained important information from the original dataset. While the final test evaluation results are PCA + GA + SVM classification, this model is able to reduce features more optimized and improved SVM classification accuracy performance. This is because the dataset formed from the combination of these two methods has far fewer features, even so this dataset retains as much information and characteristics from the original dataset as possible, besides that the features in this dataset are the most optimal features subset after going through the selection process. On the breast dataset in terms of performance, there was an increase in accuracy to 73.33%, an increase in recall to 74%, an increase in precision to 75%, an increase in F1 score to 73%. While in the ovary dataset in terms of performance there was an increase in accuracy to 98.68%, an increase in recall to 98%, an increase in precision to 99%, and an increase in F1 score to 99%. On the leukemia dataset in terms of performance, there was an increase in accuracy to 95.45%, an increase in recall to 94%, an increase in precisions to 97% and an increase in F1 score to 95%.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.

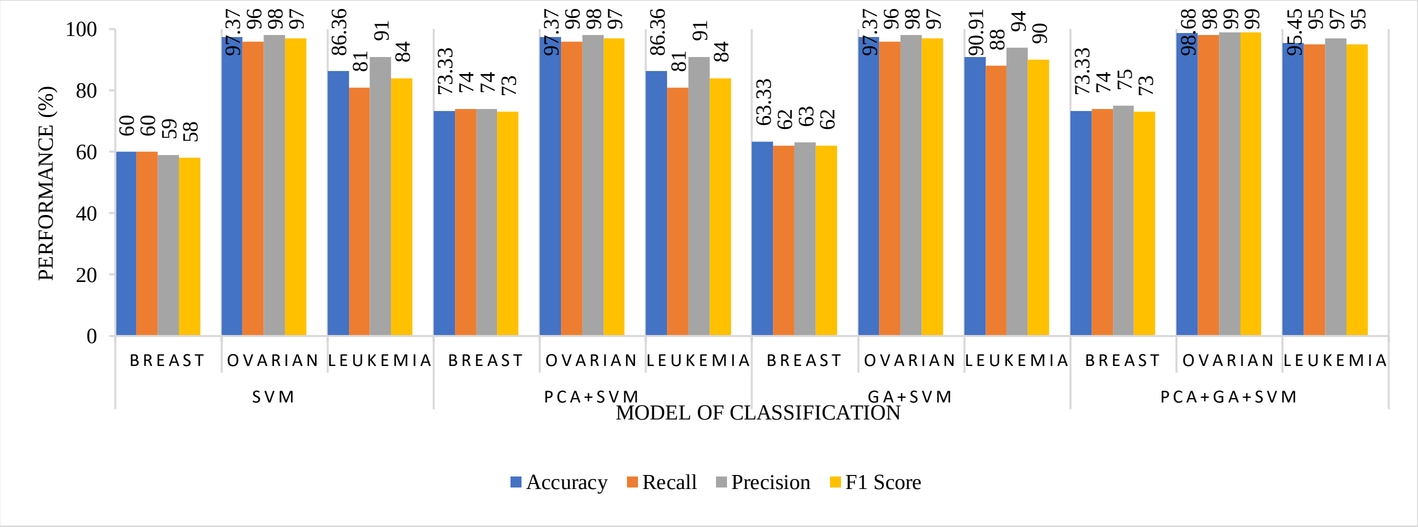

Figure 9. Graph In comparison with Each The model's Best Performance

The resulting performance also exceeds the previous research for the same data set. Combination of PCA and GA can increase the accuracy value by 10.17% compared with the study [34] which combines Features selection methods CFS and SVM for breast cancer data collection. While in ovarian cancer dataset, the proposed method can improve accuracy value of 0.26% compared to the study [29] which uses MI and GA hybrid feature selection with 10 features. Research [35] using elephants Deep Search Optimization (ESO) learning obtained accuracy of 92.11% for leukemia dataset, while in this study proposed method can increase its accuracy by 3.34%.

5. CONCLUSION

This study uses PCA dimensional reduction and GA feature selection to reduce complexity microarray data so as to improve the classification of SVM performance. There are four tests used: SVM classification, SVM classification with PCA dimensions reduction, SVM classification with GA features selection, and classification of SVM with a combination reduction of PCA dimensions and GA features selection. Based on the test results and evaluation that has been done on each classification, the method used is able to overcome weaknesses found in microarray data and improve SVM classification performance. The the best classification in this study is PCA + GA + SVM classification. The use of the PCA method can reduce data dimension to be 97 × 55 for breast cancer data set, 253 × 24 dimensions for ovarian cancer dataset and 72 × 25 dimensions to leukemia dataset. Furthermore, the application of GA feature selection to get the best subset of features can reduce the features to 28 features on the breast cancer data set, 11 features on the ovarian cancer data set, and 9 features on the leukemia dataset. While in terms of performance on breast cancer datasets, the highest accuracy was obtained at 73.33%, recall 74%, precision 75%, F1 score 73%. While in the ovarian cancer datasets with the highest accuracy are 98.68%, drawdown 98%, precision 99%, F1 score 99% and at the highest accuracy leukemia dataset of 95.45%, 94% recall, 97% precision and score F1 95%.

According to the findings in this study, several future research directions can be explored to improve SVM classification performance on high-dimensional microarray data. First, exploration of various GA parameter configurations, such as population size and number of generations, can be done with optimization techniques such as grid search, random search, or Bayesian optimization to obtain more optimal results. In addition, the use of additional feature selection techniques, such as mutual information, recursive feature elimination (RFE), or LASSO regression, can be combined with GA and PCA to improve classification accuracy. Deep learning-based approaches, such as autoencoders or deep feature selection methods, can also be an alternative in reducing the dimensionality of complex microarray data. In addition, validation tests on larger and more diverse datasets, including data from RNA-Seq or single-cell sequencing, can improve the generalizability of the model across different cancer types. Further research can also compare the performance of SVM with other machine learning models, such as random forests, gradient boosting (XGBoost, LightGBM), or deep neural networks, and explore ensemble learning methods to improve classification accuracy. Finally, the development of AutoML (Automated Machine Learning) pipelines can assist in automating the selection of the best feature selection methods, dimensionality reduction, and classification models based on dataset characteristics, thus improving efficiency and scalability in bioinformatics analysis. By exploring this research direction, SVM classification on microarray data can be further optimized, providing more accurate and reliable results in cancer gene expression analysis.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia**.**

REFERENCES

[1] V. Mutiara Purnama, W. Astuti, and Adiwijaya, 'Analisis Perbandingan Klasifikasi Microarray menggunakan Naïve Bayes dan Support Vector Machine (SVM) untuk Deteksi Kanker dengan Feature Extraction PCA', *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 9974–9986, 2021.

[2] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, 'Classification of breast cancer using microarray gene expression data: A survey', *J Biomed Inform*, vol. 117, May 2021, doi: 10.1016/j.jbi.2021.103764.

[3] J. Kim, Y. Yoon, H. J. Park, and Y. H. Kim, 'Comparative Study of Classification Algorithms for Various DNA Microarray Data', *Genes (Basel)*, vol. 13, no. 3, p. 18, Mar. 2022, doi: 10.3390/genes13030494.

[4] W. Astuti and A. Adiwijaya, 'Principal Component Analysis Sebagai Ekstraksi Fitur Data Microarray Untuk Deteksi Kanker Berbasis Linear Discriminant Analysis', *Jurnal Media Informatika Budidarma*, vol. 3, no. 2, p. 72, Apr. 2019, doi: 10.30865/mib.v3i2.1161.

[5] P. K. Ram and P. Kuila, 'Feature selection from microarray data : Genetic algorithm based approach', *Journal of Information and Optimization Sciences*, vol. 40, no. 8, pp. 1599–1610, Nov. 2019, doi: 10.1080/02522667.2019.1703260.

[6] A. Razzaque and D. A. Badholia, 'PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification', *Measurement: Sensors*, vol. 31, Feb. 2024, doi: 10.1016/j.measen.2023.100945.

[7] R. Ranjani Rani and D. Ramyachitra, 'Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using SVM', in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 108–116. doi: 10.1016/j.procs.2018.10.358.

[8] W. Ali and F. Saeed, 'Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data', *Processes*, vol. 11, no. 2, p. 22, Feb. 2023, doi: 10.3390/pr11020562.

[9] R. A. Musheer, C. K. Verma, and N. Srivastava, 'Novel machine learning approach for classification of high-dimensional microarray data', *Soft comput*, vol. 23, no. 24, pp. 13409–13421, Dec. 2019, doi: 10.1007/s00500-019-03879-7.

[10] R. Pujianto and A. A. Rahmawati, 'Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification And Regression Trees', *e-Proceeding of Engineering*, vol. Vol.6, No.1, pp. 2368–2379, Apr. 2019.

[11] 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.

[12] I. M. Parapat, M. T. Furqon, and Sutrisno, 'Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3163–3169, 2018.

[13] D. Fitria, T. H. Saragih, Muliadi, D. Kartini, and F. Indriani, 'Classification of Appendicitis in Children Using SVM with KNN Imputation and SMOTE Approach to Improve Prediction Quality', *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 3, pp. 302–311, Jul. 2024, doi: 10.35882/jeeemi.v6i3.470.

[14] S. Sucharita, B. Sahu, and T. Swarnkar, 'An Empirical Analysis of PCA-SVM Model for Cancer Microarray Data Classification', in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 495–504. doi: 10.1007/978-981-16-0695-3_47.

[15] Z. Zhu, Y. S. Ong, and M. Dash, 'Markov blanket-embedded genetic algorithm for gene selection', *Pattern Recognit*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007, doi: 10.1016/j.patcog.2007.02.007.

[16] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, 'Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi', *Technologia : Jurnal Ilmiah*, vol. 15, no. 1, p. 93, Jan. 2024, doi: 10.31602/tji.v15i1.13457.

[17] F. Adams, R. Agsar, D. Anggoro, M. B. Satria, A. W. Oktavia, and N. Chamidah, 'Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine', *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, pp. 260–268, Sep. 2021.

[18] B. G. Chepino, R. R. Yacoub, A. Aula, M. Saleh, and B. W. Sanjaya, 'Effect Of Minmax Normalization On Orb Data For Improved ANN Accuracy', *Journal of Electrical Engineering, Energy, and Information Technology (J3EIT)*, vol. 11, no. 2, pp. 29–35, Aug. 2023, doi: 10.26418/j3eit.v11i2.68689.

[19] M. Wangge, 'Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA', *Jurnal Cendekia: Jurnal Pendidikan Matematika*, vol. 05, no. 02, pp. 974–988, 2021.

[20] A. Naji Hussain, S. A. Abboud, B. A. baki Jumaa, and M. N. Abdullah, 'Impact of feature reduction techniques on classification accuracy of machine learning techniques in leg rehabilitation', *Measurement: Sensors*, vol. 25, pp. 1–9, Feb. 2023, doi: 10.1016/j.measen.2022.100544.

[21] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, 'A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction', 2022, *Frontiers Media SA*. doi: 10.3389/fbinf.2022.927312.

[22] N. Behera, 'Analysis of microarray gene expression data using information theory and stochastic algorithm', in *Handbook of Statistics*, vol. 43, Elsevier B.V., 2020, ch. 8, pp. 349–378. doi: 10.1016/bs.host.2020.02.002.

[23] H. Alshamlan, A. Alwassel, A. Banafa, and L. Alsaleem, 'Improving Alzheimer's Disease Prediction with Different Machine Learning Approaches and Feature Selection Techniques', *Diagnostics*, vol. 14, no. 19, pp. 1–7, Oct. 2024, doi: 10.3390/diagnostics14192237.

[24] N. Y. Nhu, T. Van Ly, and D. V. Truong Son, 'Churn prediction in telecommunication industry using kernel Support Vector Machines', *PLoS One*, vol. 17, no. 5 May, p. 18, May 2022, doi: 10.1371/journal.pone.0267935.

[25] R. Nurhidayat and K. E. Dewi, 'Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek', vol. 12, no. 1, 2023, [Online]. Available: https://www.kaggle.com/datasets/hafidahmusthaanah/skincare-review?select=00.+Review.csv.

[26] P. , Romadloni, B. , Adhi Kusuma, and W. Maulana Baihaqi, 'Komparasi Metode Pembelajaran Mesin Untuk Implementasi Pengambilan Keputusan Dalam Menentukan Promosi Jabatan Karyawan', *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. Vol. 6 No. 2, pp. 622–628, Sep. 2022, doi: https://doi.org/10.36040/jati.v6i2.5238.

[27] I. T. Jollife and J. Cadima, 'Principal component analysis: A review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 16, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[28] W. Astuti and Adiwijaya, 'Support vector machine and principal component analysis for microarray data classification', in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Apr. 2018, pp. 1–7. doi: 10.1088/1742-6596/971/1/012003.

[29] M. Jansi Rani and D. Devaraj, 'Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification', *J Med Syst*, vol. 43, no. 8, p. 11, Aug. 2019, doi: 10.1007/s10916-019-1372-8.

[30] O. A. Alomari, A. T. Khader, M. A. Al-Betar, and Z. A. Alkareem Alyasseri, 'A hybrid filter-wrapper gene selection method for cancer classification', in *2nd International Conference on BioSignal Analysis, Processing and Systems, ICBAPS 2018*, Institute of Electrical and Electronics Engineers Inc., Nov. 2018, pp. 113–118. doi: 10.1109/ICBAPS.2018.8527392.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia**.**

[31] H. Salem, G. Attiya, and N. El-Fishawy, 'Classification of human cancer diseases by gene expression profiles', *Applied Soft Computing Journal*, vol. 50, pp. 124–134, Jan. 2017, doi: 10.1016/j.asoc.2016.11.026.

[32] E. Alba, J. García-Nieto, L. Jourdan, and E.-G. Talbi, 'Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms', *2007 IEEE Congress on Evolutionary Computation*, pp. 284–290, 2007, doi: 10.1109/CEC.2007.4424483.

[33] C. Gunavathi and K. Premalatha, 'Performance analysis of genetic algorithm with KNN and SVM for Feature Selection in Tumor Classification', *International Journal of Computer and Information Engineering*, vol. 8, pp. 1490–1497, Jan. 2014, doi: 10.5281/zenodo.1096103.

[34] M. A. N. Hakim, Adiwijaya, and W. Astuti, 'Comparative analysis of ReliefF-SVM and CFS-SVM for microarray data classification', *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3393–3402, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3393-3402.

[35] M. Panda, 'Elephant search optimization combined with deep neural network for microarray data analysis', *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 8, pp. 940–948, Oct. 2020, doi: 10.1016/j.jksuci.2017.12.002.

AUTHOR BIOGRAPHY

Dwi Kartini received her bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia University "YPTK" Padang, Indonesia. Her research interests include the applications of artificial intelligence and data mining. She is an assistant professor in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia. She can be contacted at email: dwikartini@ulm.ac.id.

Rahmat Amin Badali is originally from Banjarbaru, South Borneo. After graduating from high school in 2020, he continued his education in the Computer Science at Lambung Mangkurat University. His current research interests are in the field of data science. In additional, his final project focuses on research to predict cancer using microarray data. He can be contacted at email: aminbadali2001@gmail.com.

Muliadi is a lecturer in the Department of Computer Science at Lambung Mangkurat University, where he specializes in Artificial Intelligence, Decision Support Systems, and Data Science. His academic journey began with a bachelor's degree in Informatics Engineering from STMIK Akakom in 2004, followed by the attainment of a master's degree in Computer Science from Gadjah Mada University in 2009. With expertise in Data Science, he also brings valuable skills in Start-up Business Development, Digital Entrepreneurship, and Data Management Staff. He can be contacted at email: muliadi@ulm.ac.id.

Dodon Turianto Nugrahadi is a dedicated academic contributing to the Department of Computer Science at Lambung Mangkurat University. His scholarly interests converge on the dynamic fields of Data Science and Computer Networking. Having established a strong foundation with a bachelor's degree in Informatics Engineering from UK Petra, Surabaya in 2004, He furthered his academic pursuits by obtaining a master's degree in Information Engineering from Gadjah Mada University, Yogyakarta in 2009. His current research endeavors delve into the complexities of networks, data science, the Internet of Things (IoT), and network Quality of Service (QoS), demonstrating a commitment to advancing knowledge in these critical areas. He can be contacted at email: dodonturianto@ulm.ac.id.

Fatma Indriani is a lecturer in the Department of Computer Science at Lambung Mangkurat University. Her research interest is focused on data science. Before becoming a lecturer, she completed her undergraduate program in the Department of Informatics at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia, in 2012. Her latest education is a doctorate in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The research fields she focuses on are data science and bioinformatics. She can be contacted at email: f.indriani@ulm.ac.id.

Setyo Wahyu Saputro is a lecturer in the Computer Science Department, Faculty of Mathematics and Natural Science, Lambung Mangkurat University in Banjarbaru. He received a bachelor's degree also in Computer Science from Lambung Mangkurat University and received his master's degree in Informatics from STMIK Amikom University. His research interests include software engineering and artificial intelligence applications. He can be contacted at email: setyo.saputro@ulm.ac.id.

Corresponding author: Dwi Kartini, dwikartini@ulm.ac.id, Department of Computer Science, Lambung Mangkurat University, Banjarbaru, Jalan Jenderal Ahmad Yani KM 36, Banjarbaru, Kalimantan Selatan 70714, Indonesia.