International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 12 (2020), pp. 4394-4401
© International Research Publication House. http://www.irphouse.com

The Comparative Analysis of Hybrid Genetic Algorithm Feature Selection Method and Particle Swarm Optimization on the High Dimensional Data

Maryam[1*], Annas Fagiat[2] and Arfian Ardiansyah[3]

[1,2,3] *Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Indonesia*.

Abstract

Cancer data using microarray technology has a data structure consisting of thousands of features. High dimensional data with a small amount of data can cause overfitting. It can be detrimental to the classification process. Therefore, a feature selection method is needed to reduce dimensions so that the classification process becomes easier and more efficient. Feature selection can also improve accuracy. This study used three cancer data taken from the Kent-ridge Bio-medical Data Set Repository. The feature selection method used was the Information Gain filter and the Genetic Algorithm wrapper method. The SVM algorithm was implemented as the evaluator. Meanwhile, the Particle Swarm Optimization wrapper method was utilized as a comparison method in testing performance. The results were compared in terms of accuracy, precision, recall, and f1-score to obtain a reliable feature selection method in reducing high-dimensional data. The performance test results show that the proposed IG-GA hybrid method is superior to the IG-PSO method. The feature selection implementation is able to reduce feature dimensions while increasing performance

Keywords: Microarray data, Information Gain, Genetic Algorithm, Support Vector Machine

I. INTRODUCTION

The World Health Organization (WHO) states that cancer is the second leading cause of death in the world and is responsible for 9.6 million deaths in 2018 [1]. It causes research topics to detect and analyze cancer to become a significant topic in the world of bioinformatics, including the use of DNA microarray technology.

Microarray is a technology that uses simulated analysis of the expression of thousands of genes in a single experiment by monitoring complex biological processes [2]. The microarray analysis process produces data that can be used for the prediction and classification process of genes to be classified into one sub-class in the form of predefined disease types. The main challenges in the microarray data classification process are the high dimensions and complex relationships between various genes. It made the process of extracting information more difficult. The higher the data dimension, the number of sample data required increases exponentially [3]. It does not rule out that of all the dimensions in the data that there are actually unnecessary dimensions during the mining process. The high dimensions and the limited number of samples cause the classification performance at a certain point to decrease. Moreover, the resulting model is too complex and the number of data observations is small. Thus, the risk of overfitting is getting bigger. This phenomenon related to high-dimensional data problems is known as the "curse of dimensionality".[4]. Certain techniques are needed to reduce dimensions to facilitate data processing. One way to reduce dimensions is by feature selection. Feature selection removes irrelevant features and reduces noise. Another advantage is that the amount of time and memory required during the data mining process is also reduced and can increase the accuracy of the classifier [5]. One of the reliable classification algorithms that can be used for prediction is the Support Vector Machine (SVM). SVM performs well for evaluating high dimensional data with a feature selection process [6].

The use of the right dimension reduction method can optimize the performance of data mining classification. The feature selection is divided into three, namely filter, wrapper, and hybrid. The filter method works independently without involving a classification algorithm so it works computationally more efficiently without reducing performance [7]. One of them is the selection of the Information Gain (IG) feature. Meanwhile, the wrapper method works by involving a classification algorithm so that the computation is more complex because the hypothesis from the model is implemented into the training and testing data; it also uses more CPU time and memory to run the program. [7]. The advantage is that wrappers can detect dependencies between features. Popular wrapper methods include Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) [3].

The Hybrid method is a combination of the filter and wrapper method resulting in better performance [8]. Nada's research [8] reviewed the use of microarray data to compare several hybrid methods in feature selection. Abul Hasnat [9] applied CC-MOGA and K-Nearest Neighbor. Hanaa [10] implemented the hybrid IG method with GA. Mahendra [11] used Mutual Information and the Bayes Theory. Pradana [12] utilized ib-PSO and C4.5. Bintang [13] used IG-GA and Gaussian Naïve Bayes. Correlation-based Feature Selection (CFS) and Particle Swarm Optimization (PSO) utilized Naïve Bayes as a classifier [14]. The combination of the GA and SVM methods is tested with six datasets [15]. This study compares the SVM results with k-NN, Decision Tree, and Linear Discriminant Analysis. The results state that the GA-SVM provides better accuracy results than the combination of GA with other classifications.

The combination of the GA method and the filter method is tested using three classifiers, namely Multi-Layer Perceptron (MLP), SVM, and K-Nearest Neighbor (k-NN) [16]. The results obtained by GA and MLP achieve the highest accuracy. The GA method is also used in the microarray data dimension reduction process and compared with the Wavelet Harr extraction method; it produces superior performance [17].

This study implemented a dimensional reduction process in high dimensional data, which aims to reduce the computational load. The method used was a hybrid method using feature selection with the filter method, namely Information Gain (IG), and the wrapper method, namely Genetic Algorithm (GA). As a comparison method, the Particle Swam Optimization (PSO) wrapper method was used. The performance of the IG-GA and IG-PSO hybrid method combination was analyzed to obtain the best method. Meanwhile, Support Vector Machines (SVM) was used as the evaluator algorithm.

II. METHOD

II.I. Dataset Research Tools and Materials

The dataset used in this study was 3 microarray datasets of cancer. The source of the dataset came from the Kent_ridge Bio-medical Data Set Repository. Details of each dataset can be seen in Table 1.

Table 1. Details of the dataset

No.	Data set	Data		Feature	Class
1	Colon Tumor	40 Negative	20 Positive	2000	2
2	Lung Cancer	31 Mesothelioma	150 ADCA	12533	2
3	Ovarian	91 Normal	162 Cancer	15154	2

II.II. Research Flow

The main objective of this study is to analyze the increase in classification performance by applying the dimensional reduction method. Figure 1 provides an overview of the research course. The research process consisted of several steps, namely data preparation, dimensional reduction, classification processes, and evaluation of test results.

II.III. Data Preparation

Data preparation needs to be done so that the dataset used can be classified more easily. The steps taken included feature normalization. Scaling feature is a method used to standardize the range of data for each feature. The method used for the normalization process was the Min-Max method. Min-Max is a normalization method by performing linear transformations of the original data. In general, the formula for this method is shown in Equation (1).

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

The range of values commonly used for the feature normalization process is [0; 1] and [-1; 1]. However, this study utilized a range of values [0; 1] because it is faster in terms of computation time.

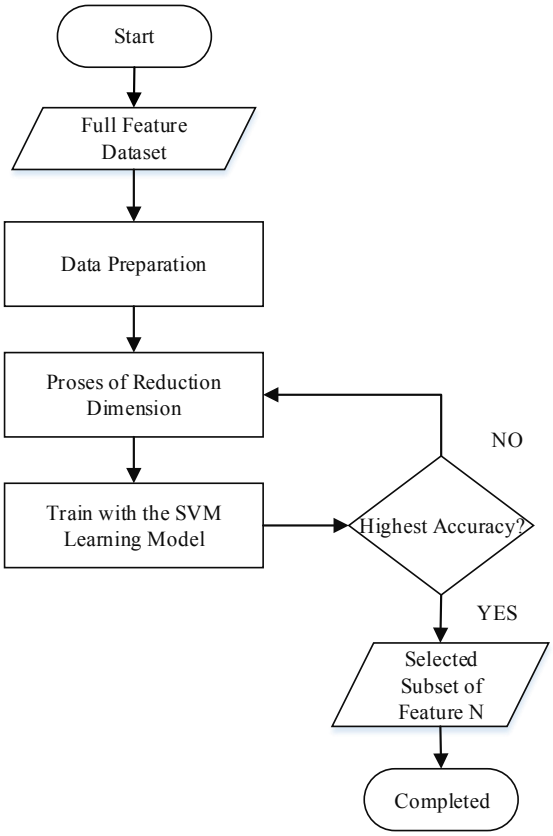

Fig. 1. Research Process Flow chart

II.IV. Dimensional Reduction

II.IV.I. Information Gain (IG)

Information Gain feature selection is a filter method that works independently without a classifier so that it works computationally more efficiently [7]. The way it works is by calculating the weight value of each feature. The formula for calculating the weight on Information Gain is as follows.

$$Inf(S) = \sum k\ i{=}1 - P(Ci, S) * log2(P(Ci, S)) \quad (2)$$

$Info\ (S)$ is the formula for finding the value of Entropy while $P\ (Ci, S)$ is the class probability Ci on the set S.

$$Info(S) = \sum v\ i{=}1 - ||SSi|| * Info(Si) \quad (3)$$

Si is the number of cases in the partition to i and Ai is the value of the attribute or feature A.

$$Gai(A) = Info(S) - InfoA(S) \quad (4)$$

$Gain\ (A)$ is the information gain boost to the feature

II.IV.II. Genetic Algorithm (GA)

A genetic algorithm is an optimization algorithm based on the mechanism of natural selection and gene selection [18]. The general framework of a Genetic Algorithm can be described as follows [19]:

1) Early population initiation

The initial population is randomly initiated to obtain the initial solution. This population is determined from a number of chromosomes to present the best solution.

2) The formation of a new generation

There are three operators in the formation of the new generation, namely operator selection, crossover, and mutation. This process is repeated until a sufficient number of chromosomes is obtained to form a new generation.

3) Solution evaluation

Each population is evaluated using the results of the fitness value on the chromosomes and will stop if the criteria are met. This process will be repeated as long as the criteria have not been met by repeating step 2.

The selection process for the Genetic Algorithm feature used the following criteria:

Table 2. Criteria for elimination

Elimination Criteria	Score
Population size	100
Maximum generation	10
Encoding scheme	Binary encoding
Fitness function	SVM evaluation
Crossover	Single Point Crossover
Crossover rate (probability)	0,8
Mutation	Fit Bit Mutation
Mutation rate (probability)	0,1
Mutation Mechanism	Roulette Wheel
Selection of Survivors	Generational Replacement

II.IV.II. Particle Swarm Optimization (PSO)

The PSO method is a global heuristic optimization method; it is a population-based iterative algorithm [20]. The population consisted of many particles where the initial initiation was completed using a random population to find a solution. Each particle represented a candidate solution and moves towards the optimal position by changing its position according to the dynamic moving velocity space according to the historical behavior of the problem. The particles searched for a better search area during the process [21]. The formula for finding the displacement and velocity of a particle is as follows:

$$V_i(t) = V_i(t-1) + c_1 r_1 [(X_{pbest} - X_i(t))] + c_2 r_2 [(X_{gbest} - X_i(t))] \quad (5)$$

$$X_i(t) = X_i(t-1) + V_i(t) \quad (6)$$

With,

Vi(t) = particle velocity i by iteration t

Xi(t) = the position of particle i by iteration t

c1 and c2 = learning individual (cognitive) and social (group) skills

r1 and r2 = random numbers with intervals of 0 and 1

Xpbesti = best position particle i

Xgbest = global best particle position

II.V. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification by utilizing a hypothetical model using linear functions in feature space that has high dimensions. The learning algorithm is based on the optimization theory by using the statistical learning theory, namely learning bias [22].

The purpose of the learning process in SVM is to obtain a hypothesis with the result that the best hyperplane by minimizing empirical risk (average error) and having good generalizations. Figure 2 describes a hyperplane that separated positive and negative samples based on the maximum margin value using a simple linear form. The margin was calculated using the closest hyperplane distance between positive and negative samples.

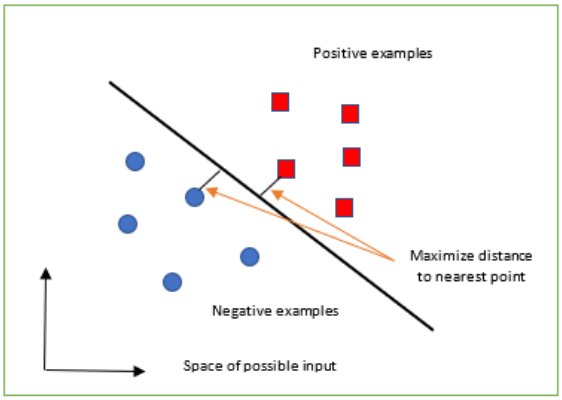

Fig. 2. Linear Support Vector Machine

II.VI. Classification

II.VI.I. Data Allocation

The data set was divided into 2 parts, namely validation data, and learning data. The two data were selected using the stratified random technique. Validation data must be different from training data to obtain good performance in the optimization stage, and training data must be different from testing data to obtain a reliable estimate of the error rate.

This study used evaluation data to evaluate the stability of the performance of the resulting model during the SVM learning process. Figure 3 shows a data allocation diagram.

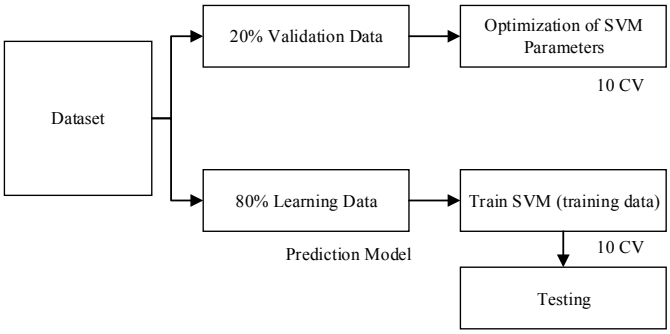

Fig. 3. Data Allocation Diagram

II.VI.II. Evaluation of Support Vector Machine (SVM)

SVM is a classification algorithm that requires adjustments to its kernel functions. This study used the RBF kernel function.

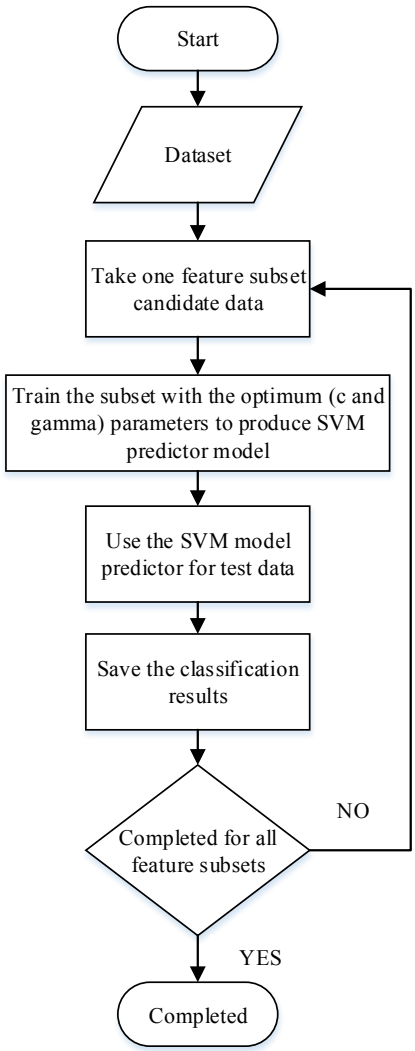

Fig. 4. SVM Flowchart as an Algorithm Evaluator

The RBF kernel requires settings in the c and γ. To find parameters, validation data was used to determine the best parameter values. Optimization of SVM parameters was performed using the grid search method and 10-fold cross-validation using a subset of data taken by stratified random sampling. The grid search method searched for a combination of values from C and γ. Then, each combination was trained with SVM to estimate learning precision. Parameter range C and γ that is recommended is $\log_2 C \in \{-5,-3,...,15\}$ and $\log_2 \gamma \in \{-15,-13....,3\}$, with an exponential increase in value [23]. The evaluation process with the SVM algorithm is shown in Figure 4. The dataset used was a learning dataset. Then, one candidate was taken as the selected feature subset. The subset was trained using the parameters C and γ, and the classification results were stored. All feature subsets were tested and compared, the highest accuracy result is the best feature subset.

II.VI. Evaluation

The type of evaluation used in this study was accuracy, specificity, and sensitivity by determining indicators:

True Positive (TP) is the number of positive instances and the data is actually positive. False Negative (FN) is the number of negative instances and the data is actually positive. False Positive (FP) counts positive instances and the data is actually negative. True Negative (TN) is the number of negative instances and actually negative data.

Evaluation of classification accuracy is the probability of instances being correctly classified in the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

Precision evaluation is a type of measure of how many instances labeled as a positive class are correctly defined against all positive predictions.

$$\text{Precision} = \frac{TN}{TP+FP} \quad (8)$$

Recall evaluation is a measurement related to how many positive class instance values are classified correctly against all positive data.

$$Recall = \frac{TN}{TP+FN} \quad (9)$$

F1-Score evaluation is a measurement to determine the weighted average comparison value of precision and recall.

$$F1-score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (10)$$

Calculation of the value of accuracy, precision, recall, and F1-Score was done using 10-fold cross-validation.

III. RESULTS AND DISCUSSION

This chapter analyzed the use of the Hybrid feature selection method, namely Information Gain (IG) with Genetic Algorithm (GA), and was compared with the Particle Swarm Optimization (PSO) method.

III.I. Classification with Grid search SVM

The first method used to classify data using the SVM algorithm was to find optimal parameters using the grid search principle. This study applied the Radial Basis Function (RBF) kernel function by determining the value of two parameters, namely parameter C and γ. The optimal parameter value is the parameter that produces the highest accuracy. The data used was the validation data. Table 3 shows the optimal parameter values of the RBF kernel function using the SVM grid search principle on Colon Tumor, Lung Cancer, and Ovarian data. Furthermore, these parameters were used for the classification process in learning data.

Table 3. Search Results from Optimal Parameters with the SVM Grid Search Principle

Data set	Parameter Optimal	
	C	γ
Colon Tumor	2^{13}	2^{-15}
Lung Cancer	2^{12}	2^{-15}
Ovarian	2^{11}	2^{-15}

III.II. Test result

The following are the test results of the Hybrid method (IG-GA) and (IG-PSO) and the SVM classification algorithm with predetermined parameters.

Table 4. Feature Selection Using the Information Gain Method

Data set	Number of Features	After IG	Accuracy	Precision	Recall	F1-Score
Colon Tumor	2000	59	80.24%	72.72%	71.67%	72.72%
Lung Cancer	12533	70	98.89%	98.75%	100%	99.40%
Ovarian	15154	101	98.88%	98.92%	100%	99.25%

Table 5. Feature Selection Using Information Gain and Particle Swarm Optimization Methods

Data set	After IG	After PSO	Accuracy	Precision	Recall	F1-Score
Colon Tumor	59	30	90.95%	93.33%	83.33%	86.67%
Lung Cancer	70	35	99.47%	99.37%	100%	99.67%
Ovarian	101	51	99.50%	99.45%	100%	99.67%

Table 6. Feature Selection Using Information Gain Method and Genetic Algorithm

Data set	After IG	After GA	Accuracy	Precision	Recall	F1-Score
Colon Tumor	59	29	93.81%	95.00%	88.33%	91.00%
Lung Cancer	70	54	99.50%	99.45%	100%	99.67%
Ovarian	101	95	100%	100%	100%	100%

The test results in Table 4 show that the microarray data using IG feature selection obtained the feature subset with the highest weight. Furthermore, the feature subset used for PSO is shown in Table 5 and GA feature selection is shown in Table 6. The number of features that were previously numbered in the thousands could be reduced using hybrid feature selection and produced a feature subset with less than 100 features and was able to increase performance. The performance results in terms of accuracy, precision, and recall show that the combination of IG-GA feature selection was comprehensively superior to the IG-PSO. Hence, it can be stated that the IG-GA hybrid method using SVM as the evaluator is a reliable method in dimensional reduction.

III.III. Effect of Hybrid Method on Accuracy

The study was conducted on three microarray datasets and tested using the feature selection method with predetermined parameters. The results of the accuracy comparison can be seen in Figure 5 for Colon Tumor data, Figure 6 for Lung Cancer data, and Figure 7 for ovarian data. The use of the Hybrid IG-GA and IG-PSO methods could improve performance compared to the IG method. Meanwhile, the IG-GA method had a comprehensive superior level of accuracy compared to the IG-PSO when classified with SVM.

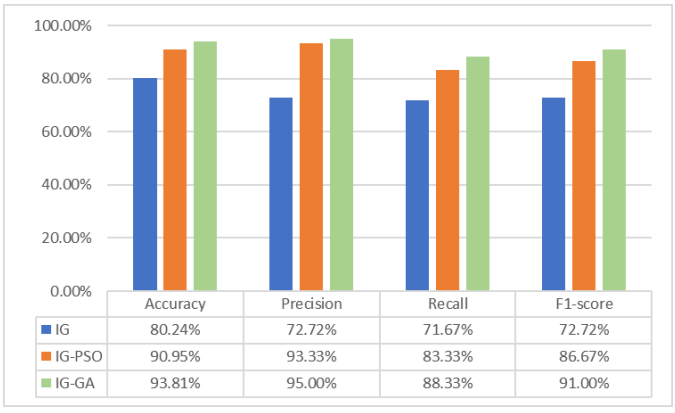

Fig. 5. Comparison of the mean value of accuracy, precision, recall, and F1-score on Colon Tumor data

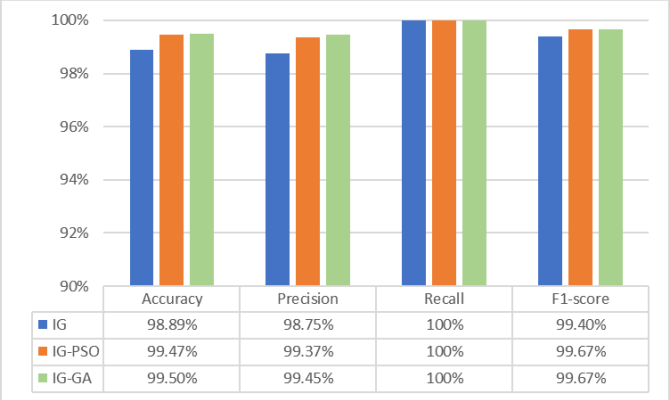

Fig. 6. Comparison of the mean value of accuracy, precision, recall, and f1-score on Lung Cancer data

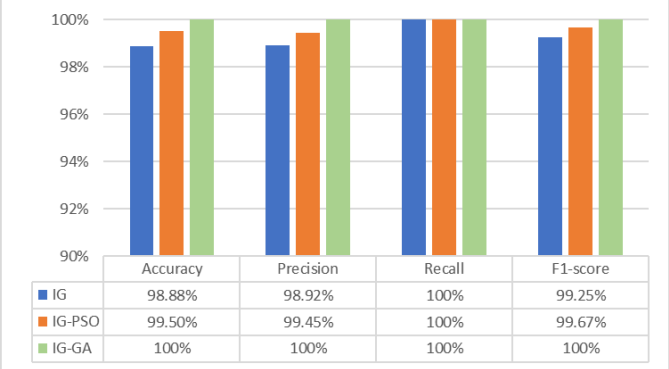

Fig. 7. Comparison of the mean value of accuracy, precision, recall, and f1-score on the Ovarian data

III.IV. Comparison with other methods

Researchers also made comparisons of accuracy results with other methods used in previous studies. The purpose of this comparison is to obtain which method has the maximum performance according to the microarray data used.

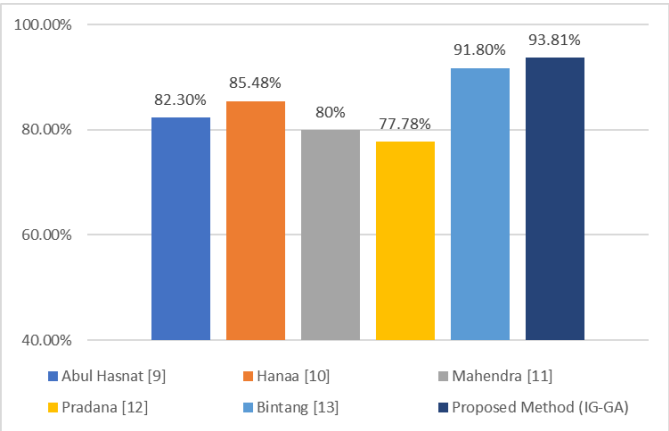

Fig. 8. Comparison of Accuracy on Colon Tumor Data

The comparison of the methods in Figure 8 using Colon Tumor data shows that the proposed IG-GA method with the SVM classifier had superior accuracy, compared to 5 studies, namely Abul Hasnat [9] who used CC-MOGA with the k-Nearest Neighbor classifier, Hanaa [10] who utilized the IG hybrid method with Genetic Algorithm, Mahendra [11] who applied Mutual Information with the classifier Bayes Theorem, Pradana [12] who implemented iBPSO with the C4.5 Decision Tree classifier, and Bintang [13] who used IG-GA with the Naïve Bayes classifier.

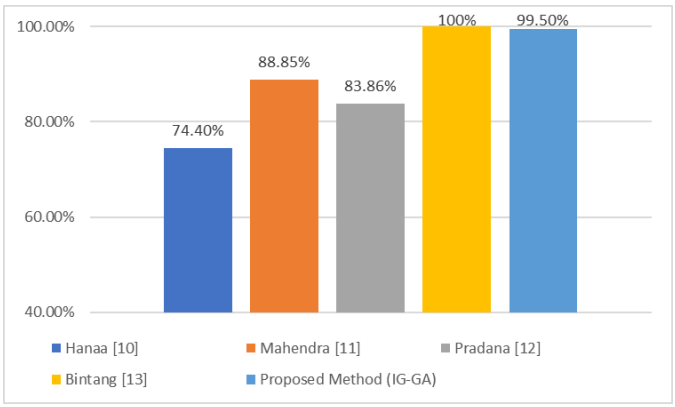

Fig. 9. Ccomparison of accuracy on Lung Cancer Data

The comparison of the methods in Figure 9 using Lung Cancer data shows that the proposed IG-GA method with the SVM classifier had lower accuracy than Bintang's research [13] with 100% accuracy. However, it was superior to three other studies, namely Hanaa [10] who obtained an accuracy of 74.40%, Mahendra [11] with an accuracy of 88.85%, and Pradana [12] who got an accuracy of 83.86%.

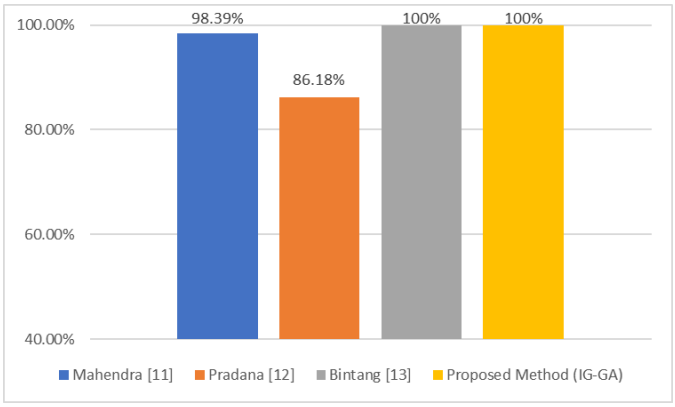

Fig. 10. Comparison of accuracy on Ovarian Data

In comparison to the accuracy of the Ovarian data, the proposed method obtained an accuracy of 100%, having the same results as the study of Bintang [13]. Also, it was superior to two other studies, namely Mahendra [11] with an accuracy of 98.39% and Pradana [12] who obtained an accuracy of 86.18%.

IV. CONCLUSION

Based on the research that has been done, the Hybrid method (IG-GA) using the SVM classifier works well for microarray data. There are several factors that influenced the test results. First, the Information Gain feature selection method was used to obtain the weight of each feature so that the feature subset with the highest weight could be obtained. Second, from the results of IG feature selection, feature selection was conducted using the wrapper method using GA and the SVM classifier. The combination of the two methods was able to provide superior performance. The results of the study using three microarray data show the acquisition of good performance in terms of accuracy, precision, recall, and F1-score. The performance test was obtained from the comparison of IG-GA feature selection and the IG-GA hybrid method. To get a significant test result, a comparison method was used, namely IG-PSO. The test results show that the IG-GA hybrid method had significantly superior performance. The performance on data for Colon tumor was 93.81% accuracy, 95% precision, 88.33% recall, and 91 %% F1-score. Meanwhile, the data for Lung Cancer were 99.5% accuracy, 99.45% precision, 100% recall, and 99.67% F1-score. Also, the ovarian data was 100% accuracy, 100% precision, 100% recall, and 100% F1-score. Comparison with previous research also shows the proposed method had a higher level of accuracy. This research can be developed by maximizing the feature selection method used. There are several parameters in both the IG and GA methods that can be tested to improve performance as a form of comparison.

REFERENCES

[1] WHO, "https://www.who.int/news-room/fact-sheets/detail/cancer."

[2] Siang TC, Soon TW, Kasim S, Mohamad MS, Howe CW, Deris S, Zakaria Z, Shah ZA, Ibrahim Z. A review of cancer classification software for gene expression data. International Journal of Bio-Science and Bio-Technology. 2015;7(4):89-108.

[3] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. InProceedings of the 20th international conference on machine learning (ICML-03) 2003 (pp. 856-863).

[4] Bishop CM. Neural Network for Pattern Recognition. Oxford University Press. 1995.

[5] Tan PN, Steinbach M, Kumar V. Introduction to data mining, Pearson education. Inc., New Delhi. 2006.

[6] Du LM, Xu Y, Zhu H. Feature selection for multi-class imbalanced data sets based on genetic algorithm. Annals of Data Science. 2015 Sep 1;2(3):293-300.

[7] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics. 2015;2015.

[8] Almugren N, Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access. 2019 Jun 14;7:78533-48.

[9] Hasnat A, Molla AU. Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient. In2016 International Conference on Emerging Technological Trends (ICETT) 2016 Oct 21 (pp. 1-6). IEEE.

[10] Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing. 2017 Jan 1;50:124-34.

[11] Purbolaksono MD, Widiastuti KC, Mubarok MS, Ma'ruf FA. Implementation of mutual information and bayes theorem for classification microarray data. InJournal of Physics: Conference Series 2018 Mar 1 (Vol. 971, No. 1, p. 012011). IOP Publishing.

[12] Pradana AC, Aditsania A. Implementing binary particle swarm optimization and C4. 5 decision tree for cancer detection based on microarray data classification. InJournal of Physics: Conference Series 2019 Mar (Vol. 1192, No. 1, p. 012014). IOP Publishing.

[13] Peryoga B, Adiwijaya A, Astuti W. Deteksi Kanker Berdasarkan Data Microarray Menggunakan Metode Naïve Bayes dan Hybrid Feature Selection. JURNAL MEDIA INFORMATIKA BUDIDARMA. 2020 Jul 20;4(3):486-94.

[14] Jain I, Jain VK, Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing. 2018 Jan 1;62:203-15.

[15] Fernando M, Halim K, Sanjaya G. Optimization Features Using GA-SVM Approach. Int. J. Sci. Res. 2015;4(9):193–197.

[16] Ghosh M, Adhikary S, Ghosh KK, Sardar A, Begum S, Sarkar R. Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & biological engineering & computing. 2019 Jan 1;57(1):159-76.

[17] Sarmilah M, Adiwijaya A, Atiqi A. Analisis Seleksi Fitur Genetic Algorithm Dan Ekstraksi Fitur Wavelet Pada Klasifikasi Microarray Data Menggunakan Naïve Bayes. eProceedings of Engineering. 2018 Apr 1;5(1).

[18] Desiana A. Konsep Kecerdasan Buatan. Yogyakarta: Andi Offset. 2006.

[19] Setiawan H, Thiang HF. Aplikasi Algoritma Genetika Untuk Merancang Fungsi Keanggotaan Pada Kendali Logika Fuzzy. 2001

[20] Bai Q. Analysis of particle swarm optimization algorithm. Computer and information science. 2010 Feb 1;3(1):180.

[21] Abraham A, Grosan C, Ramos V, editors. Swarm intelligence in data mining. Springer; 2007 Jan 12.

[22] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000 Mar 23.

[23] Staelin C. Parameter selection for support vector machines. Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1. 2003 Nov 19;1.