Hybrid feature selection of microarray prostate cancer diagnostic system

Nursabillilah Mohd Ali[1], Ainain Nur Hanafi[1], Mohd Safirin Karis[2], Nur Hazahsha Shamsudin[1], Ezreen Farina Shair[1], Nor Hidayati Abdul Aziz[3]
[1]Fakulti Teknologi and Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Fakulti Teknologi and Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[3]Faculty of Engineering and Technology, Multimedia Universiti, Melaka, Malaysia

Article Info

Article history:

Received Jun 30, 2024
Revised Aug 20, 2024
Accepted Aug 31, 2024

Keywords:

Genetic algorithm
Particle swarm optimization
Relief-F
Support vector machine
Whale optimization algorithm

ABSTRACT

DNA microarray prostate cancer diagnosis systems are widely used, and hybrid feature selection methods are applied to select optimal features to address the high dimensionality of the dataset. This work proposes a new hybrid feature selection method, namely the relief-F (RF)-genetic algorithm (GA) with support vector machine (SVM) classification method. The aim is to evaluate the performance of the proposed method in terms of accuracy, computation time, and the number of selected features. The method is implemented using Python in PyCharm and is evaluated on a DNA microarray prostate cancer. The outcome of this work is a performance comparison table for the proposed methods on the dataset. The performance of GA, particle swarm optimization (PSO), and whale optimization algorithm (WOA) is compared in terms of accuracy, computation time, and the number of selected features. Results show that GA has the highest average accuracy (91.17%) compared to PSO (90.52%) and WOA (85.74%). GA outperforms PSO and WOA due to its superior convergence properties and better alignment with complex problems.

Corresponding Author:

Nursabillilah Mohd Ali
Fakulti Teknologi and Kejuruteraan Elektrik, Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
Email: nursabillilah@utem.edu.my

1. INTRODUCTION

Cancer is one of the reasons of mortality around the world, especially prostate cancer [1]. It occurs in the prostate, which is a walnut-shaped gland in men that produces seminal fluid and happens when there is uncontrolled growth of abnormal cells in the prostate gland. Late detection will make the situation worsen because early stages of prostate cancer often present with minimal or no symptoms. Nowadays, doctors detect cancer by conducting X-Ray imaging, laboratory tests (including tests for tumor markers), tumor biopsy, endoscopic examination, and surgery. These methods may lead to some problems, which is the doctors may have missed cancers when reviewing imaging scans and the malfunction of the testing machine may occur [2]. Recently, the advance of machine learning algorithms as well as artificial intelligence (AI) have resulted in diagnosis methods for cancer [3]. Thus, the reduction of dimension of DNA microarray datasets for its use in diagnosis is crucial to detect prostate cancer at an early state [4].

The objective of the study is to propose hybrid feature selection method, which is combination of relief-F (RF) filter with genetic algorithm (GA) as wrapper and support vector machine (SVM) as classifier. Then, evaluate the classification accuracy of the proposed hybrid feature selection method on prostate cancer

DNA microarray dataset [5]. Next, compare the performance of different optimizers in terms of accuracy, computation time, and number of selected features from the proposed hybrid feature selection method.

The filter method is a feature selection technique used in preprocessing phase to reduce the high dimensionality in DNA microarray dataset [6]–[8]. The Relief filter is used extensively to identify relevant features for building accurate diagnosis models [9]. It evaluates the quality of features based on how well their values distinguish between instances that are near to each other. Essentially, it assigns a weight for each feature which indicates the ability of that feature to differentiate between instances that are close to each other in the dataset [10]. The features with the highest weights are considered the most important.

The wrapper method is an optimization technique used to select the optimum features from DNA microarray dataset-based on metaheuristic method [11]–[13]. For example, particle swarm optimization (PSO), GA, and whale optimization algorithm (WOA). GA is applied by the process of natural selection where the fittest individuals are selected for reproduction to produce offspring of the next generation. Firstly, generating an initial population randomly. Individual is characterized by a set of parameters (variables) known as genes. Then, the fitness score of the population is evaluated by the fitness function. Next, individuals with higher fitness scores are selected to be parents and contribute to the next generation [14]. The selected individuals are paired to perform crossover by varying the chromosomes (solution) to create new offspring. Mutation will occur when some of the bits (genes) in the string can flip which introduces random changes to the offspring's genetic information. This helps maintain and improve genetic diversity in the population. The old population will be replaced by the new generation, and this new population will be used in the next iteration of the algorithm. Termination occurs when a satisfactory solution is found, or the algorithm converges to a stable state.

Classification is a type of supervised learning, known as predicting the class of a given data point belongs based on specified classifications label. SVM is applied to solve classification issues [15]. The fundamental idea behind SVM is to find the best boundary (or hyperplane) that can separate data classes with the maximum margin possible. For linearly separable case, SVM can find a straight line (in 2D), plane (in 3D), or hyperplane (in higher dimensions) that separates the classes with the widest possible margin. There are two types of SVM, which is nonlinear and linear SVM [15], [16].

In linear SVM, the data is perfectly linearly separable, while the data in nonlinear SVM is not perfectly linearly separable. For nonlinearly separable cases, SVM uses a technique called the kernel trick. A kernel is a function that maps the data to a higher-dimensional space where a linear hyperplane can effectively perform the separation. To illustrate, SVM is used to find the best line that creates the widest gap between the two groups. Based on existing research [17], the feature selection method used is random-forest based feature selection (RFS) with SVM as classifier. By applying RFS + SVM feature selection method on the DNA microarray prostate cancer dataset, the best prediction result from their work is the average accuracy of 89% and the number of selected features is 11.

2. METHOD

In the methodology there are three steps for implementing the study. The first step involves implementin the filter-bsaed methods in experiment 1. Secondly, experiment 2 focuses on conducting classification. Finally, experiment 3 involves optimization and comparison of performance.

2.1. Experiment 1: perform single RF filter method on prostate cancer dataset

Figure 1 shows RF Filter method. The step-by-step method is briefly dicussed in the following: to analyze a prostate cancer DNA microarray dataset, essential libraries such as Pandas, NumPy, and RF were imported in PyCharm. The dataset was loaded from a comma separated values (CSV) file, and feature variables were extracted by dropping the last column, which contained the labels, and storing them in feature_x. Labels were extracted, converted to numerical values using a Label Encoder, and stored in target_y. The data was then split into training and testing sets with a 0.2 ratio. The RF method was applied to select the 10 most relevant features, and the filtered dataset's dimensions were checked to confirm accuracy. This first step of FS is crucial as it will filter out unrelevant features, update wights and rank features with higher weughts. Features with higher weights are considered more relevant for the classification task.

2.2. Experiment 2: evaluation on RF filter's classification accuracy by using SVM classifier

Figure 2 shows the classifcaiton after performing the RF filter method. The next experiment is dicussion about classification using SVM. With the SVM machine learning model, the libraries in PyCharm were set. The SVM classifier was selected and fitted with the filtered training set. Training and testing accuracy were evaluated using the classifier's prediction function. Steps were repeated with varying numbers of selected features (50, 100, 200, 500, 1000, 2000) to analyze the relationship between the number of features and accuracy. Next, the optimizers were applied and compared against each other using PSO, GA, and WOA.

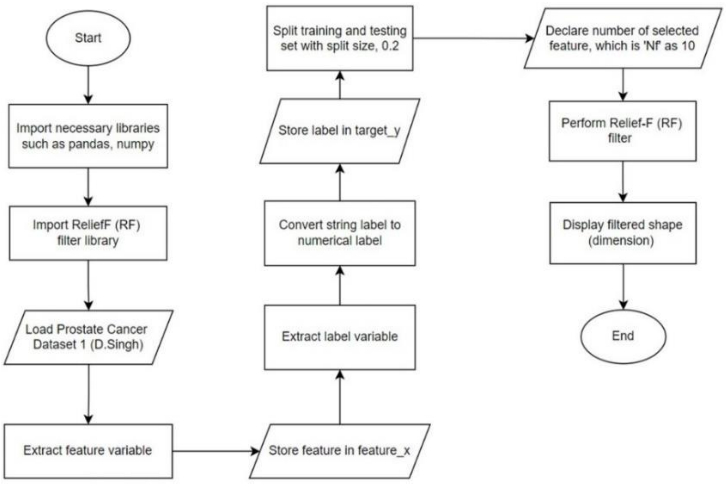
Start
Import necessary libraries such as pandas, numpy
Import ReliefF (RF) filter library
Load Prostate Cancer Dataset 1 (D.Singh)
Extract feature variable
Store feature in feature_x
Extract label variable
Convert string label to numerical label
Store label in target_y
Split training and testing set with split size, 0.2
Declare number of selected feature, which is 'Nf' as 10
Perform Relief-F (RF) filter
Display filtered shape (dimension)
End

Figure 1. Filter method

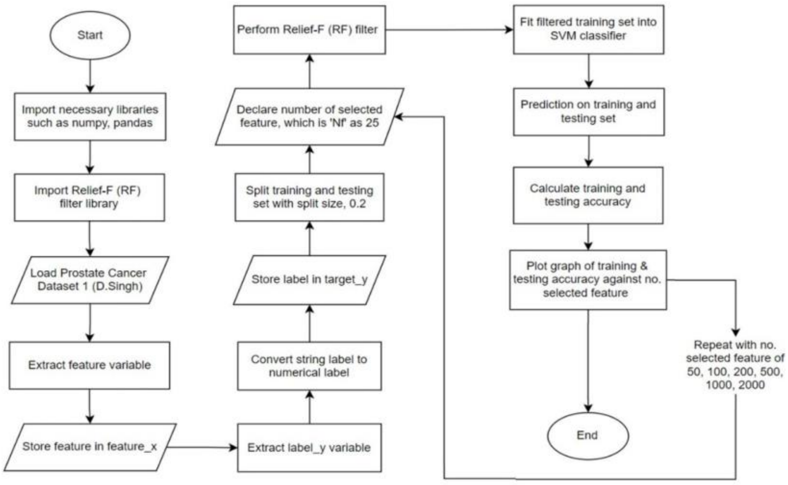
Start
Import necessary libraries such as numpy, pandas
Import Relief-F (RF) filter library
Load Prostate Cancer Dataset 1 (D.Singh)
Extract feature variable
Store feature in feature_x
Extract label_y variable
Convert string label to numerical label
Store label in target_y
Split training and testing set with split size, 0.2
Declare number of selected feature, which is 'Nf' as 25
Perform Relief-F (RF) filter
Fit filtered training set into SVM classifier
Prediction on training and testing set
Calculate training and testing accuracy
Plot graph of training & testing accuracy against no. selected feature
Repeat with no. selected feature of 50, 100, 200, 500, 1000, 2000
End

Figure 2. Classification with SVM

2.3. Experiment 3: optimization on RF filter by using wrapper (GA, PSO, or WOA)

2.3.1. Wrapper based GA

Figure 3 shows the flowchart of GA wrapper-based method. The parameters of the GA wrapper were set, including a population size equal to the number of samples in the dataset (102), a maximum of 50 generations, and a mutation rate of 0.1. The GA was conducted. The SVM classifier was selected and fitted with each new generation [18], [19]. The fitness score of each generation was evaluated using the testing set and the classifier's prediction function. Steps were repeated until the maximum number of generations was reached, and the fitness score (accuracy) of each generation was generated.

2.3.2. Wrapper based PSO

Figure 4 shows the flowchart of PSO wrapper-based method, the step-by-step method is discussed as follows. The necessary library for PSO is set. The parameters for the PSO wrapper are set, including 10 particles and 50 iterations. PSO is then conducted, and the particle positions are fitted into the SVM classifier. The fitness score of each particle's position is evaluated using the testing set in the classifier's prediction function. After every iteration, the position and velocity of each particle are updated. The steps are repeated until the maximum number of iterations is reached. Finally, the fitness score (accuracy) is generated out. The next step involves applying for WOA.

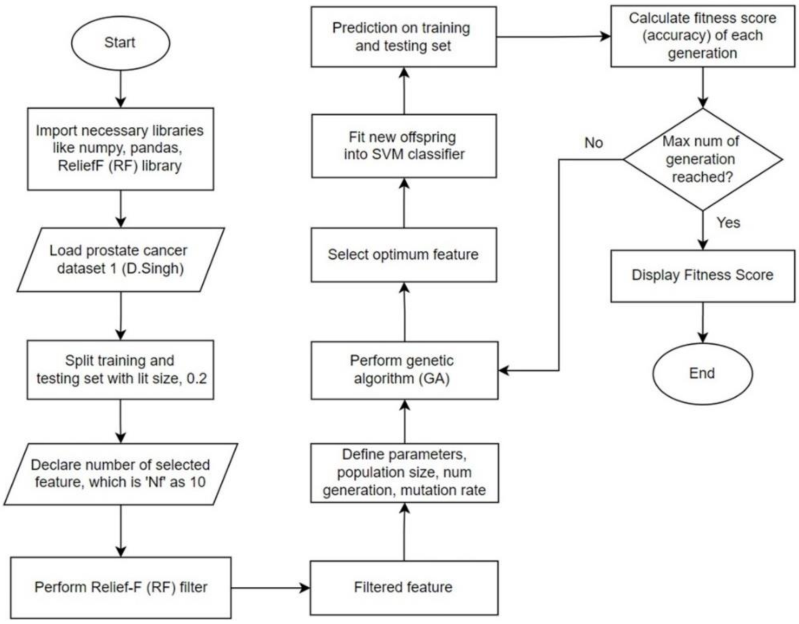
Start
Import necessary libraries like numpy, pandas, ReliefF (RF) library
Load prostate cancer dataset 1 (D.Singh)
Split training and testing set with lit size, 0.2
Declare number of selected feature, which is 'Nf' as 10
Perform Relief-F (RF) filter
Filtered feature
Define parameters, population size, num generation, mutation rate
Perform genetic algorithm (GA)
Select optimum feature
Fit new offspring into SVM classifier
Prediction on training and testing set
Calculate fitness score (accuracy) of each generation
Max num of generation reached?
No
Yes
Display Fitness Score
End

Figure 3. The flowchart of GA

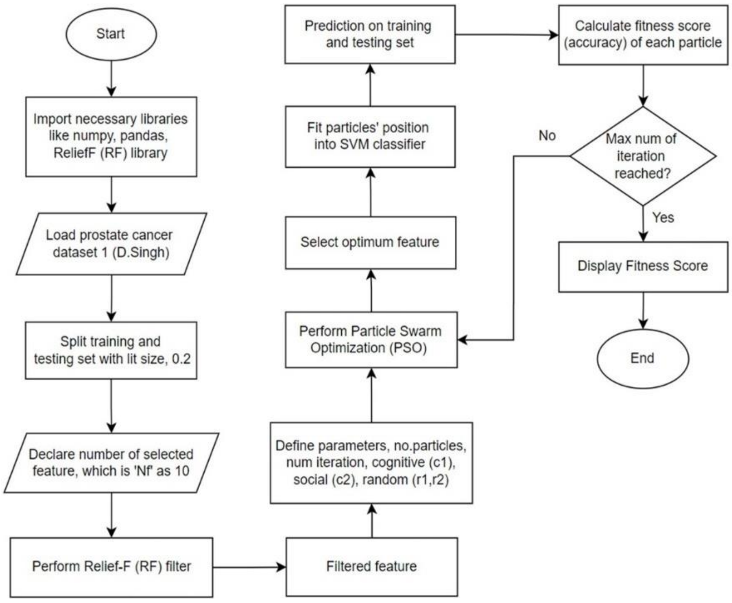
Start
Import necessary libraries like numpy, pandas, ReliefF (RF) library
Load prostate cancer dataset 1 (D.Singh)
Split training and testing set with lit size, 0.2
Declare number of selected feature, which is 'Nf' as 10
Perform Relief-F (RF) filter
Filtered feature
Define parameters, no.particles, num iteration, cognitive (c1), social (c2), random (r1,r2)
Perform Particle Swarm Optimization (PSO)
Select optimum feature
Fit particles' position into SVM classifier
Prediction on training and testing set
Calculate fitness score (accuracy) of each particle
Max num of iteration reached?
No
Yes
Display Fitness Score
End

Figure 4. PSO wrapper-based flowchart

2.3.3. Wrapper based WOA

Figure 5 shows the flowchart of WOA wrapper-based method. The number of features selected by the RF filter is set as 10. The parameters for WOA are set: 10 whales and 50 iterations. WOA is conducted, and the whale positions are fitted into the SVM classifier. The fitness score of each whale's position is evaluated using the testing set in the classifier's prediction function. The position of each whale is updated after every iteration. Steps are repeated until the maximum number of iterations is reached. Finally, the fitness score (accuracy) using in (1) is produced. The next process is comparing the three optimizers and identified the best results from the three optimizers.

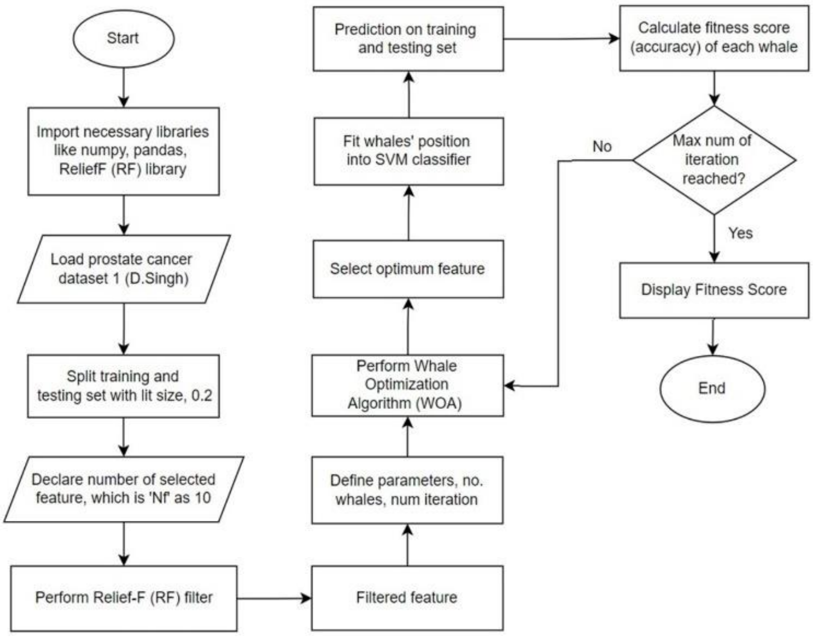
Start
Import necessary libraries like numpy, pandas, ReliefF (RF) library
Load prostate cancer dataset 1 (D.Singh)
Split training and testing set with lit size, 0.2
Declare number of selected feature, which is 'Nf' as 10
Perform Relief-F (RF) filter
Filtered feature
Define parameters, no. whales, num iteration
Perform Whale Optimization Algorithm (WOA)
Select optimum feature
Fit whales' position into SVM classifier
Prediction on training and testing set
Calculate fitness score (accuracy) of each whale
Max num of iteration reached?
No
Yes
Display Fitness Score
End

Figure 5. WOA wrapper-based method flowchart

2.4. Performance comparison between GA, PSO, and WOA

The last step is conducted the comparison between GA, PSO, and WOA optimizer. Figure 6 shows the process. A comparison table of GA, PSO, and WOA is made in terms of accuracy, computation time, and number of selected features. The performance of GA, PSO, and WOA is visualized with graphs showing accuracy, computation time, and number of selected features. The best-performing optimizer is selected for main use based on the metrics such as accuracy, computation time and number of selected features.

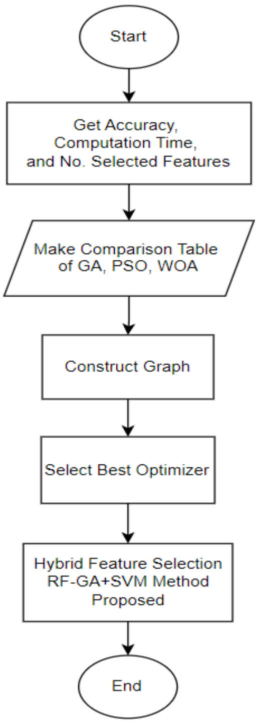
Start
Get Accuracy, Computation Time, and No. Selected Features
Make Comparison Table of GA, PSO, WOA
Construct Graph
Select Best Optimizer
Hybrid Feature Selection RF-GA+SVM Method Proposed
End

Figure 6. Summary work between GA, PSO, and WOA

Finally, the classification accuracy's [20] used for the indicating the performance evaluation and can be computed as (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

in (1) shows the accuracy computation. True positive (TP) represents correctly classified positive samples. True negative (TN) denotes correctly classified negative samples. False positive (FP) refers to negative samples that were misclassified as positive, while false negative (FN) indicates positive samples that were misclassified as negative.

3. DISCUSSION

This section discussed the result and discussion of the experiment. There are five experiments have been conducted namely performing filter method using Relief method, classification, wrapper method, observed the performance and the number of selected features.

3.1. Experiment 1: perform single RF filter method on prostate cancer dataset

In experiment 1, the DNA microarray dataset used is the DNA microarray prostate cancer dataset. Based on Table 1, the dimension of the prostate cancer dataset before filtered is (102, 12600), which means that the dataset has 102 rows and 12600 columns represents prostate genes [5]. This can be proven by loading the csv file of prostate cancer dataset named 'prostate_cancer_DNA_microarray.csv' in PyCharm and using pandas library to read the csv file.This means that the dataset is separated into features and label variable, where the label variable is at the last column of the dataset, and the features variable is the columns except the last column.

3.2. Calculation of the training set and testing set ratio of 80/20

The RF filter started by initializing the weights of all features in prostate cancer DNA microarray dataset to zero. Then, a sample instance is selected from the dataset randomly. This instance is referred as the target instance. After that, RF filter will find the nearest hits and nearest misses based on the target instance. Nearest hit is the closest instance to the target instance that has the same class label, while nearest miss is the closest instance to the target instance but with a different class label. Then, each feature will update its weight based on how well it distinguishes the target instance from the nearest hit and the nearest miss.

Based on the RF update rule above, the weight is decreased if the difference between the target instance and its nearest hit are larger (implying this feature is less important for class similarity). In contrast, the weight is increased if the difference between the target instance and its nearest miss are larger (implying this feature is important for distinguishing between different classes). After iterating through the dataset, the top 10 features with the highest weights are considered the most important and will be used for machine training model. In other words, this is done by selecting the first 10 of the features based on their entropy measure ranking in descending order. The entropy measure represents randomness in a feature, 0 means the most, and 1 means least, thus the higher the entropy measures, the more useful the feature is. The calculations can be easily made to prove the ratio of the training set to the testing set will be 80/20.

Testing set samples = (split size) (no.samples) = (0.2) (102) = 20.4 = 21

Training set samples = (no.samples) – (no.testing samples) = 102 – 21 = 81

3.3. Experiment 2: evaluation on RF Filter's classification accuracy by using SVM classifier

In Experiment 2, the evaluation of RF filter's classification accuracy is made by using SVM classifier. After RF filters the features in prostate cancer DNA microarray dataset, SVM will calculate the training and testing accuracy. The number of the training feature is ranges from 25, 50, 100, 200, 500.

Based on Figure 7, when the number of selected features increase, the training and testing accuracy will decrease. It is because the RF filter does not filter out the less useful features in dataset and this will decrease the efficiency of the results. Thus, higher the number of selected features, higher the possibility of less useful features not filtered out, and may lower the accuracy. To overcome this issue, the hybrid feature selection method is proposed to make optimization to further select the optimum features by using a wrapper method like GA, PSO, and WOA, which is more accurate due to its efficient optimization algorithm.

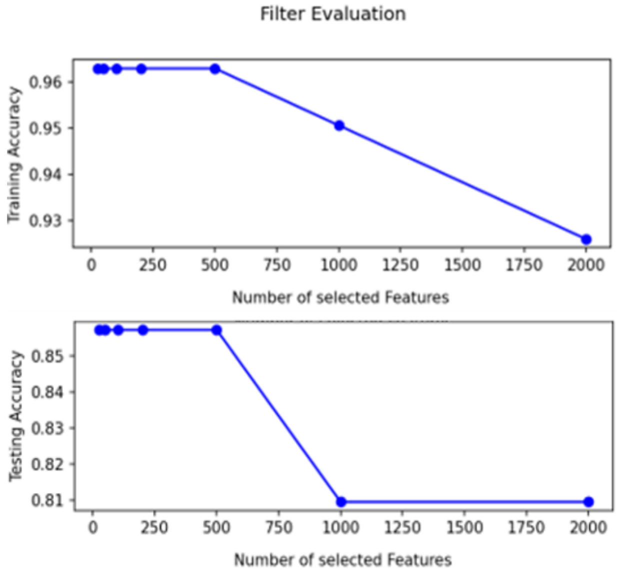

Figure 7. Training and testing accuracy against no. selected feature

3.4. Experiment 3: optimization on RF filter by using different wrapper (GA, PSO, or WOA)

Recent years have seen extensive exploration of various technique and method for analyzing and classifying prostate cancer. The comparative analysis in Table 1 suggests a superior method to previous studies on the same dataset. The finding indicate that the Relife-GA+SVM hybrid approach performs promisingly.

Based on Table 1, the performance between GA, PSO, WOA is compared to each other in terms of accuracy, computation time, and number of selected features. PSO has the fastest computation time [21] (4.66s) if compared to GA (17.63s) and WOA (4.74s) as in Table 1. PSO and WOA have almost the same speed, which is about four seconds, when handling the optimization of prostate cancer DNA microarray dataset. While GA needs the longest computation time, which is 17.63s, to handle the optimization of dataset.

Figure 8 shows that GA-SVM has the highest average accuracy (91.17%) if compared to PSO(90.52%) and WOA (85.74%). GA outperforms PSO and WOA due to its superior convergence properties and better alignment with complex problem landscapes. Its ability to balance exploration and exploitation effectively helps avoid local optima, making it particularly suitable for challenging optimization domains. principal component analysis (PCA) and kernel-based principal component analysis (KPCA) were utilized by Xiaoxue *et al.* [22]

to simplify and identify prostate cancer. In their findings, KPCA outperformed PCA with SVM models based on cumulative contribution. Despite KPCA being more efficient, PCA was more effective in dimensionality reduction, although it had a high running time of 18.31 seconds and the accuracy of 84.31% was achieved. Highlighting their significant contribution. Notable limitations include dataset dimension, potentioal challenges in clinical implemntation and the problem in the simplification of the feature reduction process. However, the number of selected features was not reported in their findings. Experimental results indicate that the proposed Relief-GA+SVM hybrid approach achieved the highest accuracy of 91.17% surpassing previous methods such as those by Xiaoxue *et al.* [22].

For GA, it works through mechanisms, such as crossover, and mutation. Each of these steps involves significant computation time. Crossover involves pairing selected individuals (solution) and exchanging portions of their genetic information to create new offspring. The individual selected is based on their fitness scores and this usually requires sorting or ranking operations, which are computationally costly. After crossover, mutation alters the genes of the new generation randomly, and it requires additional computation for each gene mutation.

Thus, GA tends to explore the solution space more thoroughly due to its diverse operations, so that slower the convergence. For PSO and WOA, they can converge faster to a solution because of their simpler update rules, which is directly integrate the best-known positions (solution) without the disruptive steps like mutation in GA. Although the PSO and WOA computation speed is fast, lack of crossover and mutation steps might influence the accuracy of feature selection.

Table 1. Performance comparison between GA, PSO, and WOA

Study	Method	Accuracy (%)	Compution time (s)	No.selected features
Proposed	GA	91.17	17.63	4
	PSO	90.52	4.66	4
	WOA	85.74	4.74	4
[22]	KPCA-SVM	84.31%	18.31	----

Note*: -----indicate not avaliable

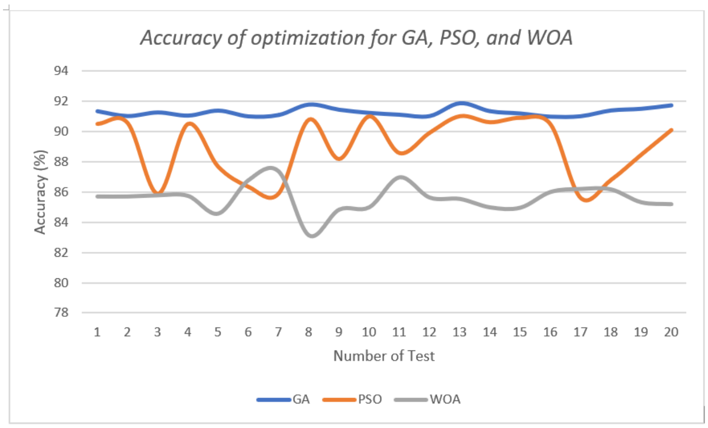

Figure 8. Accuracy of optimization for GA, PSO, and WOA

By using crossover, a chromosome in the new generation can inherit the best genetic characteristics from the parents' generation. Mutation is a process when some of the bits (genes) in the binary string can flip which introduce random changes to the offspring's genetic information. This mutation is important because it helps improving the next generation and leading to diversity and more innovative solutions that might not be discovered by other methods [23].

In DNA microarray dataset, a set of features with high fitness will be selected as parents and perform crossover to yield new generation, which is the set of features with greater fitness. When the number of generations increases, the accuracy of the selected features increases, because only the most useful features are retained in the new generation and the less useful features already eliminated. Moreover, a mutation occurs on the selected features in each generation. This is crucial to ensure the diversity and accuracy of solutions produced, which is the feature selected.

3.5. Number of selected features

Table 2 displays the number of genes selected by GA, PSO, and WOA. As anticipated, GAs, as metaheuristic approaches, tend to outperform other algorithms constrained by issues like local minima. From the employed optimizer, the number of selected genes is 4. Form the affymetric features, can be used to identify the gene and symbol for prostate cancer [24]. These selected features can be further analysed as potential biomarkers for patient targeted therapy [25]–[27].

Table 2. Selected gene name for GA, PSO, and WOA

GA		PSO		WOA	
Gene affymetrix	Gene name	Gene affymetrix	Gene name	Gene affymetrix	Gene name
32598_at	neural EGFL like 2(NELL2)	32598_at	neural EGFL like 2(NELL2)	37394_at	complement C7(C7)
38406_f_at	prostaglandin D2 synthase (PTGDS)	37366 _at	PDZ and LIM domain 5(PDLIM5)	38406_f_at	prostaglandin D2 synthase (PTGDS)
38634_at	retinol binding protein 1(RBP1)	38634_at	retinol binding protein 1(RBP1)	39054_at	glutathione S-transferase mu 1(GSTM1)
39054_at	glutathione S-transferase mu 1(GSTM1)	39054_at	glutathione S-transferase mu 1(GSTM1)	40282_s_at	complement factor D(CFD)

4. CONCLUSION

Prostate cancer is a leading cause of death among men, highlighting the importance of early identification. Robust optimizers and classifiers can significantly improve prostate classification. Hybrid method that uses metaheuristic as feature selection generally outperform other methods. A major challenge is combining the most accurate feature selection method with the classifier model. To address this, a hybrid Relief-GA and SVM has been proposed for accurate prostate cancer classification. Furthermore, the model identifies four selected features, which can be further analyzed for biomarker discovery. The GA was determined to be the most effective optimizer in hybrid feature selection, as it achieved better accuracy than both PSO and WOA. Future work will assess the applicability of the proposed model across diverse disease high dimensional datasets for comprehensive validation. Additionally, efforts will focus on enhancing model performance by exploring other metaheuristic as feature selection methods and considering hyperparameter tuning in machine learning models to tackle strength and capacity challenges on larger datasets.

ACKNOWLEDGEMENTS

This work is supported by Universiti Teknikal Malaysia Melaka and Multimedia University, Malaysia. Special thanks to Mr. Loh Yi Cong for his valuable assistance with the simulation work during his final year project for Bachelor of Mechatronics Engineering. We also thank to the refered reviewer for their valuable comments and suggestions, which have significantly improved our work.

REFERENCES

[1] M. Sekhoacha, K. Riet, P. Motloung, L. Gumenku, A. Adegoke, and S. Mashele, "Prostate cancer review: genetics, diagnosis, treatment options, and alternative approaches," *Molecules*, vol. 27, no. 17, p. 5730, Sep. 2022, doi: 10.3390/molecules27175730.

[2] Y. J. Tan, K. S. Sim, and F. F. Ting, "Breast cancer detection using convolutional neural networks for mammogram imaging system," in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*, IEEE, Nov. 2017, pp. 1–5. doi: 10.1109/ICORAS.2017.8308076.

[3] P. Esmaeilzadeh, "Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 170, Dec. 2020, doi: 10.1186/s12911-020-01191-1.

[4] A. Gumaei, R. Sammouda, M. Al-Rakhami, H. AlSalman, and A. El-Zaart, "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression," *Health Informatics Journal*, vol. 27, no. 1, p. 146045822198940, Jan. 2021, doi: 10.1177/1460458221989402.

[5] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, Mar. 2002, doi: 10.1016/S1535-6108(02)00030-2.

[6] P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," *Applied Soft Computing*, vol. 43, pp. 117–130, Jun. 2016, doi: 10.1016/j.asoc.2016.01.044.

[7] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, Mar. 2017, doi: 10.1016/j.ygeno.2017.01.004.

[8] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification," *Computational Biology and Chemistry*, vol. 56, pp. 49–60, 2015, doi: 10.1016/j.compbiolchem.2015.03.001.

[9] A. K. Shukla, S. K. Pippal, S. Gupta, B. R. Reddy, and D. Tripathi, "Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6637–6648, May 2020, doi: 10.3233/JIFS-179743.

[10] A. Coletta *et al.*, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9(4), pp. 1106–1119, 2012.

[11] N. A. Zamri, N. A. Ab. Aziz, T. Bhuvaneswari, N. H. A. Aziz, and A. K. Ghazali, "Feature selection of microarray data using simulated kalman filter with mutation," *Processes*, vol. 11, no. 8, p. 2409, Aug. 2023, doi: 10.3390/pr11082409.

[12] N. Mohd Ali, R. Besar, and N. A. Ab. Aziz, "Hybrid feature selection of breast cancer gene expression microarray data based on metaheuristic methods: a comprehensive review," *Symmetry*, vol. 14, no. 10, p. 1955, Sep. 2022, doi: 10.3390/sym14101955.

[13] N. Mohd Ali, N. A. Ab Aziz, and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, p. 712, Nov. 2020, doi: 10.11591/ijeecs.v20.i2.pp712-719.

[14] J. McCall, "Genetic algorithms for modelling and optimisation," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 205–222, Dec. 2005, doi: 10.1016/j.cam.2004.07.034.

[15] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data," *International Journal of Neural Systems*, vol. 15, no. 6, pp. 475–484, Dec. 2005, doi: 10.1142/S0129065705000396.

[16] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, Jul. 2021, doi: 10.3390/technologies9030052.

[17] E. Glaab, J. Bacardit, J. M. Garibaldi, and N. Krasnogor, "Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data," *PLoS ONE*, vol. 7, no. 7, p. e39932, Jul. 2012, doi: 10.1371/journal.pone.0039932.

[18] Irawansyah, Adiwijaya, and W. Astuti, "Comparative analysis of support vector machine (SVM) and random forest (RF) classification for cancer detection using microarray," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2021, pp. 650–656. doi: 10.1109/ICoICT52021.2021.9527458.

[19] E. Alba, J. Garcia-Nieto, L. Jourdan, and E.-G. Talbi, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," in *2007 IEEE Congress on Evolutionary Computation*, IEEE, Sep. 2007, pp. 284–290. doi: 10.1109/CEC.2007.4424483.

[20] P. Mohapatra, S. Chakravarty, and P. K. Dash, "Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system," *Swarm and Evolutionary Computation*, vol. 28, pp. 144–160, Jun. 2016, doi: 10.1016/j.swevo.2016.02.002.

[21] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, "Particle swarm optimization: a comprehensive survey," *IEEE Access*, vol. 10, pp. 10031–10061, 2022, doi: 10.1109/ACCESS.2022.3142859.

[22] X. Xiaoxue, L. Fu, S. Weiwei, Li Wenwen, and Z. Yu, "Research of PCA and KPCA in the characteristics simplicity of the gene data," in *Proceedings of 2013 2nd International Conference on Measurement, Information and Control*, IEEE, Aug. 2013, pp. 669–672. doi: 10.1109/MIC.2013.6758051.

[23] M. T. Arslan, D. Arslan, and B. Haznedar, "Training anfis system with genetic algorithm for diagnosis of prostate cancer," *NWSA Academic Journals*, vol. 13, no. 4, pp. 301–309, Oct. 2018, doi: 10.12739/NWSA.2018.13.4.2A0159.

[24] M. T. Vietri *et al.*, "Hereditary prostate cancer: genes related, target therapy and prevention," *International Journal of Molecular Sciences*, vol. 22, no. 7, p. 3753, Apr. 2021, doi: 10.3390/ijms22073753.

[25] J. Isuwa *et al.*, "Optimizing microarray cancer gene selection using swarm intelligence: recent developments and an exploratory study," *Egyptian Informatics Journal*, vol. 24, no. 4, p. 100416, Dec. 2023, doi: 10.1016/j.eij.2023.100416.

[26] C. Bendtsen and S. Petrovski, "How data and AI are helping unlock the secrets of disease," 2019, [Online]. Available: https://www.astrazeneca.com/what-science-can-do/labtalk-blog/uncategorized/how-data-and-ai-are-helping-unlock-the-secrets-of-disease.html

[27] J. Nahar, T. Imam, K. S. Tickle, A. B. M. Shawkat Ali, and Y.-P. P. Chen, "Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12371–12377, Nov. 2012, doi: 10.1016/j.eswa.2012.04.045.

BIOGRAPHIES OF AUTHORS

Nursabillilah Mohd Ali was born in Melaka, Malaysia, in 1985. She received the B.Eng. (Hons.) and M.Sc. degrees from Universiti Teknikal Malaysia Melaka and International Islamic University, Malaysia in 2009 and 2014, respectively, all in mechatronic engineering. She was awarded a PhD in engineering from Multimedia University in September 2024. She has been an academic staff since 2009, where now she is a Senior Lecturer of Universiti Teknikal Malaysia Melaka. She is a Chartered Engineer of the Engineering Council, UK and a Graduate Engineer of the Board of Engineers Malaysia. Her research interests include bioinformatics systems, DNA gene expression, optimization algorithm and machine learning. She can be contacted at email: nursabillilah@utem.edu.my.

Ainain Nur Hanafi holds a Ph.D in Electrical Engineering from University of Newcastle, Australia, specialising in fault tolerant control. She currently serving as a senior lecturer and head of programme (Mechatronics) at the Department of Engineering of Faculty of Technology and Electrical Engineering (FTKE) at Universiti Teknikal Malaysia Melaka (UTeM). Her research interests include control systems, automation and energy management system. She can be contacted at email: ainain@utem.edu.my.

Mohd Safirin Karis was born in Melaka, Malaysia, in 1986. He received the B.Eng. (Hons.) and M.Sc. degrees from Universiti Teknologi Malaysia, Malaysia in 2009 and 2012, respectively, all in control and robotic engineering. He has been an academic staff since 2009, where now he is a Senior Lecturer of Universiti Teknikal Malaysia Melaka. He is a Professional Engineer of Board of Engineers Malaysia. His research interests include rehabilitation systems, control ystem, optimization algorithm and machine learning. He can be contacted at email: safirin@utem.edu.my.

Nur Hazahsha Shamsudin received a B.Eng. degree in Electrical Engineering from Universiti Teknologi Mara, Malaysia, in 2008, M.Eng. from Universiti Malaya in 2012, and Ph.D. degree from Universiti Putra Malaysia in 2022. Currently, she is a senior lecturer at the Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka. She is a Certified Energy Manager Under ASEAN Energy Management Scheme. Her research interests include gas sensors, solar energy, nanomaterial and energy efficiency. She can be contacted at email: nurhazahsha@utem.edu.my.

Ezreen Farina Shair is a senior lecturer at the Department of Electrical Engineering, Faculty of Electrical Technology & Engineering, Universiti Teknikal Malaysia Melaka (UTeM). She received her B.Eng. (Electrical, Control & Instrumentation) (2009) and M. Eng. (Electrical-Mechatronics & Automatic Control) (2011) from Universiti Teknologi Malaysia (UTM) and Ph.D. in Electronics Engineering (2019) from Universiti Putra Malaysia (UPM). Her research interests include bio-signal processing, machine learning, deep learning, artificial intelligence, and the Internet of Things. Dr. Ezreen is currently an executive committee member of the IEEE Engineering in Medicine & Biology Society (IEEE-EMBS) Malaysia Chapter. She can be contacted at email: ezreen@utem.edu.my.

Nor Hidayati Abdul Aziz graduated from Multimedia University in 2002 in Electronics Engineering majoring in Computer and completed her Master's Engineering at Universiti Teknologi Malaysia in 2005. She started her career at Telekom Malaysia Berhad as an engineer after graduating. She then moved on to Multimedia University as a lecturer after completing her master's degree. She was awarded her PhD from Universiti Malaysia Pahang in Computational Intelligence. She is currently the chairperson for the Centre for Engineering Computational Intelligence at the Faculty of Engineering and Technology, Melaka campus. She can be contacted at email: hidayati.aziz@mmu.edu.my.