JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Fajila, F., & Yusof, Y. (2025). Mutable composite firefly algorithm for microarray-based cancer classification. *Journal of Information and Communication Technology, 24*(1), 102-129. https://doi.org/10.32890/jict 2025.24.1.5

Mutable Composite Firefly Algorithm for Microarray-Based Cancer Classification

[*1]Fathima Fajila & [2]Yuhanis Yusof
[1]Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka
[2]School of Computing, Universiti Utara Malaysia, Malaysia

[*1]fajila@seu.ac.lk
[2]yuhanis@uum.edu.my
*Corresponding author

Received: 4/11/2024 Revised: 15/1/2025 Accepted: 16/1/2025 Published: 31/1/2025

ABSTRACT

Microarray-based cancer biomarker detection is one of the popular trends for cancer classification. Though existing approaches have given competing performance in terms of classification accuracy and reduced feature subsets, the classification of different cancer microarray datasets still requires improvements. Recently, the swarm-based hybrid algorithms have given significant performance in cancer classification. However, the efficiency of a swarm algorithm is dominated by certain factors such as fitness value, convergence, exploration, and exploitation capabilities. Thus, a swarm-based hybrid approach is proposed for cancer classification with a new variant of the Firefly Algorithm (FA) and Correlation-based Feature Selection (CFS) filter. The slow convergence issue in the FA is resolved by non-fixed size solutions termed as mutable size solutions and a composite position update function is designed for the mutable solutions. In addition, the local optima issue is overcome by the population reinitialisation method. The proposed algorithm, named the CFS-Mutable Composite Firefly Algorithm (CFS-MCFA), is evaluated based on two metrics, namely classification accuracy and genes subset size, using a Support Vector Machine (SVM) classifier. Results show that CFS-MCFA-SVM achieved 100% accuracy with only a few biomarkers for all four cancer microarray datasets, indicating the efficiency and the competing performance of the proposed algorithm in biomarker detection for microarray-based cancer classification. Apart from that, the proposed algorithm would also contribute to cancer-related issues upon verifying the relevancy of particular genes via technical analysis from a medical perspective and would be utilised in feature selection applications.

Keywords: Biomarker detection, cancer classification, correlation-based feature selection, firefly algorithm, microarray.

INTRODUCTION

Cancer is one of the deadliest and disease contributing to the high mortality rates all around the world (World Health Organization, 2018). The widely spreading nature of cancer would lead to the worst stage in a short period of time. However, early diagnosis and treatment can reduce the mortality rate to a great extent (World Health Organization, 2018). Nevertheless, early detection and treatment are two more critical tasks. Hence, cancer biomarkers (i.e. informative genes) play a major role in cancer diagnosis, prognosis, early detection, and treatment (Tobore, 2019). Currently, high throughput sequencing such as Deoxyribo Nucleic Acid (DNA) microarrays, is widely used in cancer biomarker detection. In addition to medical perspectives, computerised microarray analysis is a popular and crucial technique for cancer biomarker detection, which facilitates cancer classification. However, since microarrays produce thousands of genes relative to the number of samples (Das et al., 2024; El Akadi et al., 2011; Lai et al., 2016), biomarker detection (i.e. gene selection) is not a trivial task.

Gene selection algorithms such as filters, wrappers, and hybrid methods have been used for biomarker detection in microarray-based cancer classification. Nevertheless, since the existing filter-based (Al-Batah et al., 2019; Mazumder & Veilumuthu, 2019; Wang et al., 2019) and wrapper-based (AlMazrua & AlShamlan, 2022; Panda, 2020; Shekar & Dagnew, 2020) gene selection algorithms suffer with low classification performance and large genes subset size, hybrid methods (Dash, 2021; Elyasigomari et al., 2017) which employ filter and wrapper algorithms outperform both filters and wrappers. A filter-based preprocessing and a wrapper-based subset selection employed in a hybrid approach produce competing performance compared to the single methods (i.e. filter and wrapper) in cancer classification. Besides, swarm-based hybrid gene selection algorithms (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018) are superior to normal hybrid methods (Gao et al., 2017; Mazumder & Veilumuthu, 2018) as the optimal solution is searched instead of exact solution which is less likely to be obtained over a high dimensional feature space such as DNA cancer microarray. Existing studies have employed various swarm algorithms such as Artificial Bee Colony (ABC) algorithm (Alshamlan, 2018), Bat Algorithm (BA) (Al-Betar et al., 2020), Cuckoo Optimization Algorithm (COA) (Elyasigomari et al., 2017), Firefly Algorithm (FA) (Almugren & Alshamlan, 2019b), and Particle Swarm Optimisation (PSO) algorithm (Jain et al., 2018) in cancer classification. Nevertheless, the efficiency of a swarm algorithm depends on certain factors such as fitness value, convergence, exploration, and exploitation capabilities.

FA is a popular swarm-based optimisation algorithm with two special properties: automatic subdivision and multimodality (Yang & He, 2013). FA is widely applied in many recent applications (Dokeroglu et al., 2019). Nevertheless, FA suffers from two major issues: slow convergence and local optimums (Yang, 2014). The iteration of the initial solutions throughout the generation causes increased exploitation, which leads to local optima issues in the conventional FA (Yang, 2010). On the other hand, using a large threshold for the dimension of the fixed-size solutions results in enhanced exploration, which causes slow convergence in the standard FA (Yang, 2010). More precisely, much more exploration with much less exploitation leads to slow convergence, while too much exploitation with too little exploration leads to local optimum (Yang, 2014). Thus, maintaining an adequate balance between exploration and exploitation is crucial for producing better performance. Besides, dynamic solutions (Dashtban & Balafar, 2017; Pashaei & Pashaei, 2019) have been suggested in the literature to achieve fast convergence. Furthermore, existing studies have proposed different kinds of population reinitialisation strategies such as partial reinitialisation (Cheng et al., 2014a; Cheng et al., 2014b; Salgotra et al., 2019), reinitialisation while preserving best individuals (Mostafa Bozorgi & Yazdani, 2019; Sekaj & Oravec, 2009; Sekaj & Perkacz, 2007), and start reinitialisation with a threshold (El-

Abd, 2016; Guo & Tang, 2009) to increase population diversity and exploration capability while resolving local optima issue.

In contrast to existing studies which use initial population throughout the iterations (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018; Emami & Pakzad, 2019; Jabbar, 2019), we propose population reinitialisation which is also known as regeneration formation (Fajila & Yusof, 2022) in order to overcome local optima issue in FA. Further, non-fixed size solutions termed as mutable solutions (i.e. dynamic size solutions) are proposed in this study to resolve the slow convergence issue due to fixed size solutions used in existing works (Almugren & Alshamlan, 2019b; Emami & Pakzad, 2019; Jabbar, 2019). Hence, it is aimed to maintain the exploration and exploitation capabilities to produce better results using the proposed algorithm.

Apart from that, the proposed algorithm also uses a composite scheme for position update in the fireflies. The position update in the standard FA (Yang, 2010) is performed based on Cartesian distance measures and thus applicable only for fixed size solutions. Therefore, a new position update equation is required for the non-fixed size solutions. Furthermore, the position update equation in FA consists of three elements namely the initial position of the less bright firefly, the movement due to the attractiveness, and the movement due to the randomness (Yang, 2010). However, none of these elements focuses on the feature interaction which occurs between the features (i.e. between the elements of the solution) except the distance-based difference.

Thus, a hybrid swarm-based gene selection algorithm is proposed for this study. The proposed algorithm is termed CFS-MCFA-SVM as it integrates a Correlation-based Feature Selection (CFS) filter (Hall, 1999), a Mutable Composite Firefly Algorithm (MCFA), and a Support Vector Machine (SVM) classifier (Vapnik et al., 1997). The next section provides a brief research background, followed by the proposed research methodology, experimental setup with results, and finally, the discussion and conclusion.

RESEARCH BACKGROUND

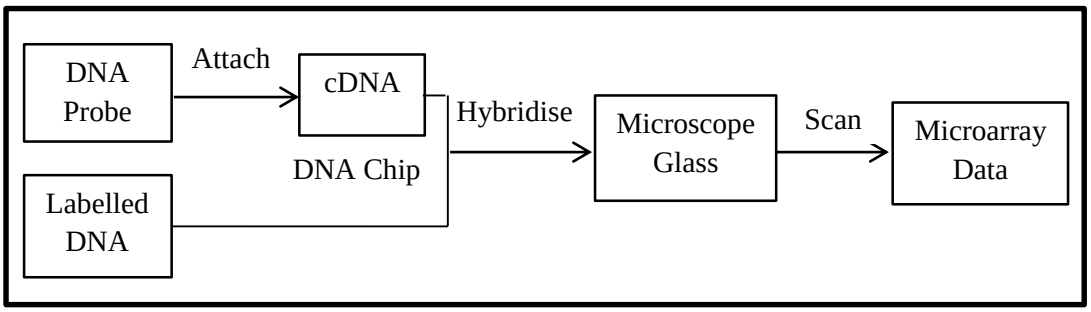
This section describes DNA microarray (Fajriyah, 2021), feature selection, CFS filter (Hall, 1999), SVM (Vapnik et al., 1997), swarm algorithms (Sharma & Kaur, 2021; Shukla et al., 2020), and FA (Yang, 2010). It also discusses related studies relevant to gene selection for each of the aforementioned sub-sections.

DNA Microarray

DNA microarray analysis has recently become a popular field for cancer classification (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018). The huge microarray datasets consist of a large number of genes, which are a collection of relevant, irrelevant, and redundant genes. The typical microarray formation is initialised by letting a DNA probe and a target DNA (i.e., labelled DNA) be hybridised, as shown in Figure 1.

Figure 1

Microarray Formation



The DNA probe on the chip is known as complementary DNA (cDNA). The gene expression values of each probe-target hybrid are determined by scanning the gene chip. The scanned results provide quantitative information about each gene corresponding to the cDNA (Fajriyah, 2021; Govindarajan et al., 2012). Then, the microarray dataset of dimension $m \times n$, as shown in Figure 2, is formed as a collection of gene expressions. Generally, the patient samples are denoted row-wise, while the gene information is given column-wise in a microarray data matrix.

Figure 2

Sample Microarray Dataset

	Gene_1	Gene_2	Gene_3	…	Gene_n
Sample_1	0.6	4.4	1.3	...	2.2
Sample_2	1.5	2.6	5.2	...	2.9
Sample_3	0.7	3.7	2.4	...	1.6
.	.	.	.	...	.
Sample_m	0.5	3.4	3.0	...	2.5

Biomarkers are crucial in therapeutics and are utilised in the diagnosis processes of genetic diseases such as cancer (Guo, 2003; Sriram et al., 2011). High throughput technologies such as DNA microarrays facilitate researchers in identifying cancer biomarkers that have a major impact on cancer classification, early detection, and prognosis (Fathi et al., 2014; Kang, 2015; Ramasubramanian, 2020). Further, cancer mortality rate is reduced through early diagnosis and treatment using cancer biomarkers. In addition, fast and non-invasive cancer diagnosis is one of the major advantages of cancer biomarkers (Karley et al., 2011). Nevertheless, the proper analysis of microarray data is the key factor influencing the utilisation of cancer biomarkers in different fields. Hence, feature selection plays a major role in the process of biomarker identification.

Feature Selection

The process of identifying a small number of features out of a large feature space is known as feature selection. Like feature selection, gene selection (or biomarker identification) is a process of identifying biomarker genes from a large gene expression profile. There are three types of gene selection methods: filter, wrapper, and hybrid. Numerous computational applications such as face recognition (Agarwal & Bhanot, 2015), drug dataset analysis (Sayed et al., 2019), and text classification (Marie-Sainte & Alalyani, 2020) have utilised feature selection algorithms. Besides, the existing gene selection algorithms related to cancer classification are discussed, along with their remarks.

In concern to filter-based gene selection approaches, Spearman's correlation and three filters were suggested by Shukla et al. (2019), reporting that Spearman's Correlation (SC) with minimum Redundancy Maximum Relevance (mRMR) and Naïve Bayes (NB) has produced 99.74% accuracy with 20 genes. CFS filter with the best classifier JRip was applied by Al-Batah et al. (2019) on eleven cancer datasets where none produced 100% accuracy. In addition, Al-Batah et al. (2019) produced relatively large gene subsets with more than 25 genes. Besides, Wang et al. (2019) proposed Adaptive Elastic Net with Conditional Mutual Information (AEN-CMI) for Leukemia and Colon cancer classification giving 91.05% and 89.30% accuracy with 26.85 and 25.20 genes, respectively. It is noteworthy that the filter methods (Al-Batah et al., 2019; Shukla et al., 2019; Wang et al., 2019) have resulted in low accuracy (i.e. less than 100%) while producing large gene subsets (i.e. gene subsets with more than 20 genes).

On the other hand, a wrapper-based gene selection algorithm using modified binary FA (BFA) and Random Forest (RF) was suggested by Sawhney et al. (2018) for the classification of three cancer datasets, which produced less than 100% accuracy. Further, Dif and Elberrichi (2019) proposed an enhanced recursive FA with SVM (RFA-SVM), which was evaluated on twelve cancer datasets, giving 100% accuracy on only four datasets with large gene subsets having 9, 13, 18, and 20 genes. The same year, Almugren and Alshamlan (2019a) evaluated FA and SVM on five cancer datasets. Nevertheless, due to the low performance of wrapper-based FA, the authors (Almugren & Alshamlan, 2019a) suggested a hybrid FA-based approach (Almugren & Alshamlan, 2019b) for future work. Besides, a wrapper-based approach using an Elephant Search Algorithm and Deep Neural Network (ESA-DNN) was evaluated on ten cancer datasets by Panda (2020). However, the approach could not produce 100% classification accuracy on any of the ten datasets. In addition, Panda (2020) produced large gene subsets. Al-Baity and Al-Mutlaq (2021) proposed Simulated Annealing (SA) for breast cancer prediction. Three datasets were evaluated using three classifiers – SVM, RF, and Decision Tree (DT), among which SVM outperformed the others. However, none of the datasets produced 100% accuracy. The authors (Al-Baity & Al-Mutlaq, 2021) suggested the hybrid method with an ensemble technique for the future task. Recently, AlMazrua and AlShamlan (2022) utilised the Harris Hawks Optimization (HHO) algorithm together with two classifiers, namely, SVM and K-nearest neighbour (KNN). However, the approach perfectly classified only two datasets out of the six used. Overall, the wrapper-based methods (AlMazrua & AlShamlan, 2022; Almugren & Alshamlan, 2019a; Dif & Elberrichi, 2019; Panda, 2020; Sawhney et al., 2018) have produced low accuracy with large genes subsets similar to filter methods (Al-Batah et al., 2019; Shukla et al., 2019; Wang et al., 2019).

Hybrid algorithms utilise single methods (i.e. filter and wrapper) for gene selection. Hybrid gene selection algorithms (Dash, 2021; Gao et al., 2017; Mazumder & Veilumuthu, 2018) have shown better performance compared to single methods (Al-Batah et al., 2019; Panda, 2020; Sawhney et al., 2018). Furthermore, hybrid methods that utilise swarm algorithms (Al-Betar et al., 2020; Almugren &

Alshamlan, 2019b; Alshamlan, 2018) are superior to normal hybrid methods (Gao et al., 2017; Mazumder & Veilumuthu, 2018) due to their capability in searching the optimal solution instead of the exact solution. Alshamlan (2018) evaluated the CFS filter and ABC using SVM, named Co-ABC, on six cancer datasets and produced 100% accuracy on five datasets with relatively small gene subsets. However, the approach (Alshamlan, 2018) produced low accuracy on the Colon cancer dataset. Besides, a hybrid FA, which uses an F-score filter, was proposed by Almugren and Alshamlan (2019b) for cancer classification. Three datasets out of five were classified with 100% accuracy while giving low accuracy and large gene subsets for Colon and Leukemia datasets.

Apart from that, a hybrid method with robust mRMR filter, modified BA, and SVM (rMRMR-MBA-SVM) was applied by Al-Betar et al. (2020) on ten cancer datasets, giving competing performance, though some datasets produced low accuracy with slightly large gene subsets. Moreover, Qin et al. (2022) presented a hybrid approach named as ISSA that used an improved binary Salp Swarm Algorithm (SSA) and a combination of three filters for cancer classification. The approach (Qin et al., 2022) was evaluated on ten cancer datasets out of which seven produced 100% accuracy. Ali and Saeed (2023) proposed a hybrid approach where three filters were evaluated individually together with GA. The approach (Ali & Saeed, 2023) was evaluated on four cancer datasets out of which only one produced 100% accuracy still with a very large genes subset. Xie et al. (2023) proposed a method with improved mRMR and improved Multilayer Binary FA (MBFA) together with an ensemble classifier. Four microarray datasets were evaluated using the proposed method. However, similar to Ali and Saeed (2023), 100% accuracy was produced on only one dataset.

Besides, Mehrabi et al. (2024) suggested a hybrid framework combining mRMR filter and binary Horse herd Optimisation Algorithm (MRMR-BHOA) for gene selection. Ten datasets were evaluated using the SVM classifier, where 60% of the datasets produced 100% accuracy. Recently, Panda et al. (2025) proposed a machine-learning model that utilises the Boruta form, an improved mRMR filter, and SSA (BIMSSA). Four microarray datasets were used to evaluate the performance of the model using five classifiers. Nevertheless, the BIMSSA model could not produce 100% accuracy on any of the four datasets and as well as all the datasets produced large gene subsets. It is noteworthy that the results obtained in existing studies (Al-Betar et al., 2020; Ali & Saeed, 2023; Almugren & Alshamlan, 2019b; Alshamlan, 2018; Mehrabi et al., 2024; Panda et al., 2025; Qin et al., 2022; Xie et al., 2023) as summarised in Table 1 show that there is still room for improvements in terms of classification accuracy and smaller genes subset through a new hybrid approach. Hence, we propose a hybrid gene selection algorithm based on FA for cancer classification.

Table 1

Summary of Hybrid Gene Selection Methods

Reference - Method	Dataset	Accuracy (%) (No. of Genes)	Remark
Alshamlan (2018) – Co-ABC	Colon (Alon et al., 1999)	96.77(9)	Low accuracy on Colon dataset.
	Leukemia2 (Golub et al., 1999)	100(3)	
	Lung Michigan (Beer et al., 2002)	100(2)	
	SRBCT (Khan et al., 2001)	100(4)	
	Lymphoma (Alizadeh et al., 2000)	100(2)	
	Leukemia3 (Armstrong et al., 2002)	100(6)	

(continued)

Reference - Method	Dataset	Accuracy (%) (No. of Genes)	Remark
Almugren and Alshamlan (2019b) - FFF-SVM	Leukemia3 (Armstrong et al., 2002)	97.8(10)	Only three (60%) datasets produced 100% accuracy.
	SRBCT (Khan et al., 2001)	100(8)	
	Lung Michigan (Beer et al., 2002)	100(2)	
	Leukemia2 (Golub et al., 1999)	100(5)	
	Colon (Alon et al., 1999)	94.3(15)	
Al-Betar et al. (2020) - rMRMR-MBA-SVM	Breast (Zhu et al., 2007)	95.4(12.63)	Low accuracy on Breast and Colon datasets.
	MLL (Zhu et al., 2007)	100(8)	
	Colon (Alon et al., 1999)	97.85(12.27)	
	Leukemia2 (Golub et al., 1999)	100(4.07)	
	Leukemia3 (Armstrong et al., 2002)	100(5.33)	Though 80% of the datasets produced 100% accuracy, the gene subset size would be further reduced.
	Leukemia4 (Zhu et al., 2007)	100(6.73)	
	Lymphoma (Alizadeh et al., 2000)	100(8.13)	
	CNS (Pomeroy et al., 2002)	100(11.2)	
	Ovarian (Zhu et al., 2007)	100(3.07)	
	SRBCT (Khan et al., 2001)	100(9.13)	
Qin et al. (2022) – ISSA	Colon (Alon et al., 1999)	97.6(148)	Seven datasets produced 100% accuracy. However, the approach resulted in very large gene subsets.
	CNS (Zhu et al., 2007)	99.2(216)	
	Leukemia2 (Golub et al., 1999)	100(241)	
	Breast (Zhu et al., 2007)	100(325)	
	Lung (Zhu et al., 2007)	99.4(283)	
	Ovarian (Zhu et al., 2007)	100(300)	
	Leukemia3 (Armstrong et al., 2002)	100(182)	
	Leukemia4 (Zhu et al., 2007)	100(336)	
	MLL (Zhu et al., 2007)	100(489)	
	SRBCT (Khan et al., 2001)	100(104)	
Ali and Saeed (2023) - Filter-GA	CNS (Zhu et al., 2007)	93.33(182)	The approach resulted in very large gene subsets.
	Breast (Zhu et al., 2007)	93.81(618)	
	Lung (Zhu et al., 2007)	98.52(327)	
	Brain (Pomeroy et al., 2002)	100(138)	
Xie et al. (2023) – MBFA	Colon (Alon et al., 1999)	90.48(9.4)	Only one dataset produced 100% accuracy.
	Leukemia2 (Golub et al., 1999)	100(4.3)	
	Prostate (Singh et al., 2002)	94.18(6.7)	
	DLBCL (Shipp et al., 2002)	98.75(5.9)	
Mehrabi et al. (2024) – MRMR-BHOA	Colon (Alon et al., 1999)	98.48(7.36)	Only 60% of the datasets produced 100% accuracy.
	Lymphoma (Zhu et al., 2007)	100(2)	
	Leukemia2 (Golub et al., 1999)	100(2.6)	
	DLBCL Stanford (Alizadeh et al., 2000)	100(2.8)	
	Ovarian (Zhu et al., 2007)	100(3)	
	MLL (Zhu et al., 2007)	100(4.1)	
	SRBCT (Khan et al., 2001)	100(5.53)	
	Lung Harvard (Bhattacharjee et al., 2001)	98.66(8.36)	
	Prostate (Singh et al., 2002)	97.72(5.4)	
	Brain Tumor1 (Pomeroy et al., 2002)	96.45(8.81)	
Panda et al. (2025) – BIMSSA	Leukemia3 (Armstrong et al., 2002)	96.7(57)	None of the datasets provided 100% accuracy.
	Lymphoma (Zhu et al., 2007)	96.2(29)	
	MLL (Zhu et al., 2007)	95.1(194)	
	SRBCT (Khan et al., 2001)	97.1(23)	

Correlation-based Feature Selection Filter

Filters are utilised in many studies (Al-Betar et al., 2020; Ali & Saeed, 2023; Qin et al., 2022; Xie et al., 2023) to retain the relevant genes while removing the irrelevant and redundant genes from the large microarray datasets. The univariate filter is a kind of filter that evaluates the features individually. In contrast, multivariate filters evaluate feature subsets. F-score filter (Wright, 1965), mRMR filter (Peng et al., 2005), and mutual information filter (Vergara & Estévez, 2014) are categorised as univariate filters, whereas the CFS filter (Hall, 1999) and Markov blanket filter (Pearl, 2014) are classified under multivariate filters.

The gene subsets are evaluated in terms of correlations among the genes and the corresponding class when a CFS filter is used for filtering. The final genes subset is prioritised based on the correlation concept, where precedence is given to the genes with a higher correlation towards the class and lower correlation within the genes. In addition, the Best First Search (BFS) (Pearl, 1984) technique is applied for searching because of its ability to manage the high dimensional feature space (Alshamlan, 2018). First, the training dataset produces a matrix of correlations: gene-class and gene-gene correlations (Liu et al., 2002). Then, the subset of genes is assessed to provide a score based on Equation 1.

$$Score_s = \frac{N\overline{r_{gc}}}{\sqrt{N + N(N-1)\overline{r_{gg}}}} \quad (1)$$

where $Score_s$ is the score of a gene subset s with N number of genes, $\overline{r_{gc}}$ is the average gene-class correlation, and $\overline{r_{gg}}$ is the average gene-gene correlation.

Many feature selection applications, such as the classification of food grains (Pushpalatha & Karegowda, 2017), the classification of tear film lipid layer (Remeseiro et al., 2015), and retinal image quality assessment (Remeseiro et al., 2017) have utilised CFS filter in addition to cancer classification. Gene selection studies (Alshamlan, 2018; Jain et al., 2018) which have used CFS filter for preprocessing, have produced significant results showing the adequate performance of CFS filter. Besides, the CFS filter is concerned with the gene interactions and their inter-related functionalities (Kohavi & John, 1997; Zhang & Deng, 2007) when producing the correlation matrix. Thus, this study uses a CFS filter for preprocessing.

Support Vector Machine

Machine learning algorithms are typically categorised into two types: supervised learning and unsupervised learning. One of the well-known supervised learning techniques is the classification which classifies the data into different classes. Among the various popular classification algorithms such as artificial neural network (McCulloch & Pitts, 1943), Bayesian belief networks (Pearl, 2014), decision trees (Quinlan, 1986), and KNN method (Aha et al., 1991); SVM is the point of interest in this research.

Vapnik invented the SVM (Vapnik et al., 1997) which is a supervised classification algorithm used in many fields such as in regression (Cherkassky & Ma, 2004), fuel price prediction (Mustaffa et al., 2013; Zhao et al., 2006), pattern recognition (Weston & Watkins, 1999), forecasting (Kazem et al., 2013) and also in gene selection (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018; Qin et al., 2022) to produce significant results. In the process of SVM-based classification, the samples in a

dataset are separated by a hyperplane drawn with respect to the class. Besides, SVM can manage linear and non-linear separations (Abdulqader et al., 2020). Thus, this study applies the SVM classifier to evaluate the performance of the proposed algorithm.

Swarm Algorithms

Finding the exact solution for an NP-hard optimisation problem is a computationally expensive task (Basir et al., 2019; Hoang, 2008; Yab et al., 2024). Thus, swarm algorithms that provide optimal or near-optimal solutions are widely used for high-dimensional feature analysis. Swarm algorithms are also called meta-heuristic optimisation algorithms, nature-inspired optimisation algorithms or bio-inspired optimisation algorithms (Yang, 2013) that imitate the natural behaviour of some organisms (Fong et al., 2018). For instance, ABC algorithm (Karaboga, 2005), Ant Colony Optimization (ACO) algorithm (Dorigo et al., 2006), COA (Yang & Deb, 2009), FA (Yang, 2010), GWO algorithm (Mirjalili et al., 2014), SSA (Mirjalili et al., 2017), and Whale Optimisation Algorithm (WOA) (Mirjalili & Lewis, 2016) are few swarm algorithms that are designed to mimic the behaviours of relevant organisms in the nature. Nevertheless, selecting the best swarm algorithm from the varieties of swarm-based algorithms is influenced by factors such as the fitness function, convergence property, exploitation, and exploration capabilities.

Firefly Algorithm

FA simulates the light-emitting behaviour of fireflies with three idealised rules (Yang, 2010).

- All fireflies are of the same sex (i.e. unisex) and attracted towards each other.
- Attractiveness of a firefly is proportional to the brightness of the firefly. Hence, the fireflies with low brightness will be attracted toward the fireflies with high brightness. The brightness is inversely proportional to the distance.
- A fitness or an objective function calculates the brightness.

The brightness or the light intensity of a firefly at the position x is determined by the fitness function $f(x)$. The attraction is given by β that changes with the distance between the fireflies. The steps in the FA is illustrated in Algorithm 1. Equation 2 shows how the position of a less bright firefly x_i is updated when it moves towards a brighter firefly x_j. The parameters β_0, γ, and α denote the initial attractiveness, light absorption coefficient, and randomisation parameter, respectively while ε_i indicates a vector of random numbers drawn from either a Gaussian or uniform distribution (Yang, 2010).

$$x_i(t+1) = x_i(t) + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j(t) - x_i(t)\right) + \alpha\varepsilon_i \tag{2}$$

where, $x_i(t+1)$ is a new position and r_{ij} is the distance between firefly i and j. The distance r_{ij} is calculated using Cartesian distance (Yang, 2010) as in Equation 3.

$$Cartesian\_distance(x_j, x_i) = \sqrt[2]{(x_j - x_i)^2} \tag{3}$$

Algorithm 1: *Firefly Algorithm.*

Step 1: Define the population size n, β_0, γ, α_0, and maximum number of iteration: tmax
Step 2: Generate the initial firefly population randomly: $x_i,\ i = 1, 2, 3, \ldots, n$
Step 3: Evaluate each firefly using the fitness function: $f(x)$
Step 4: while (t < tmax)
Step 5: for $i = 1$ to n
Step 6: for $j = 1$ to n
Step 7: if $(f(x_i) < f(x_j))$
Step 8: Move firefly i towards j
Step 9: Calculate the distance r using Equation (3)
Step 10: Calculate the new position x_i using Equation (2)
Step 11: Update the fitness of firefly i
Step 12: end if
Step 13: Evaluate new solutions and update light intensity
Step 14: end for j
Step 15: end for i
Step 16: Rank the fireflies and find the best firefly
Step 17: end while

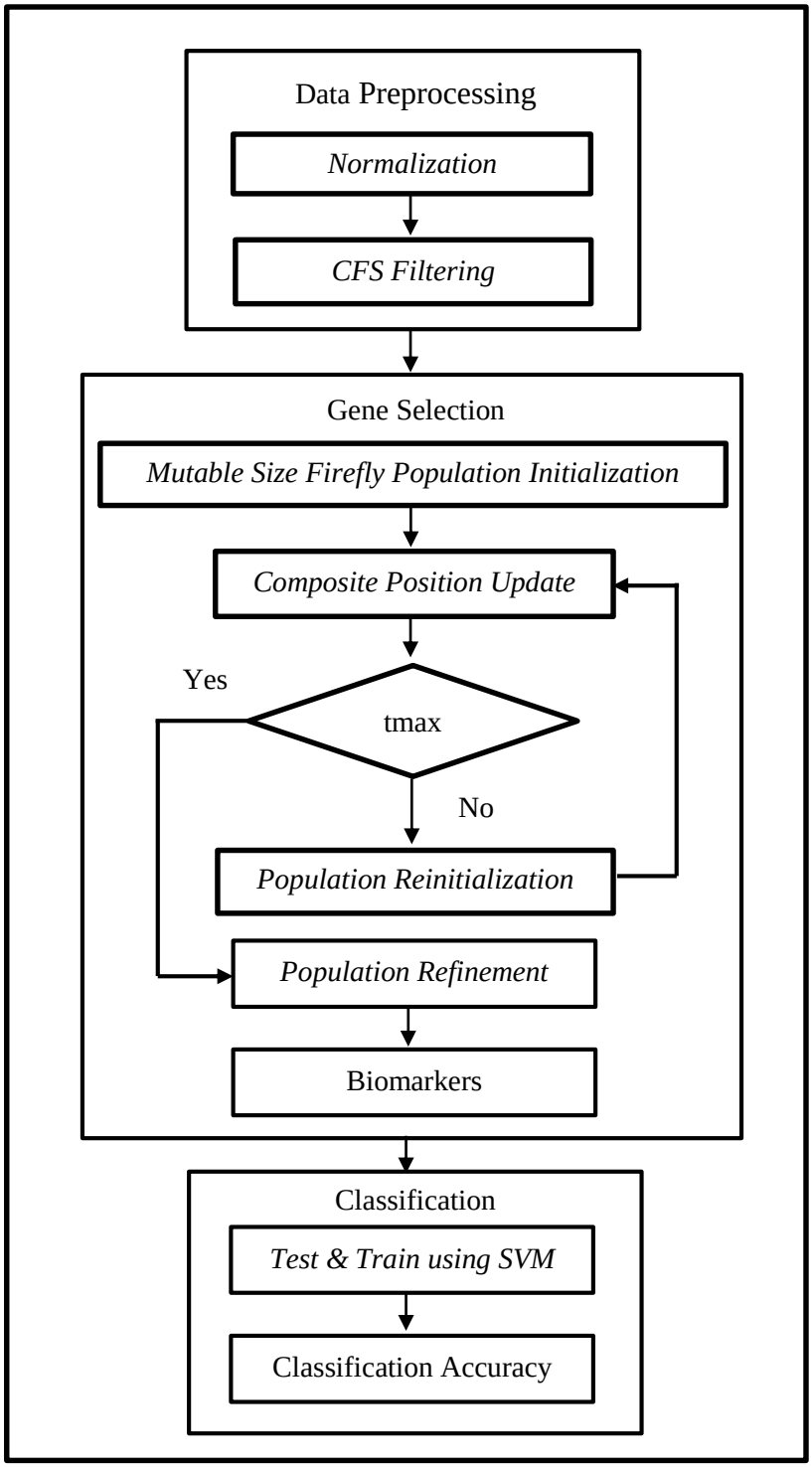
Numerous applications have utilised FA for different purposes such as churn prediction (Ahmed & Maheswari, 2017), document clustering (Mohammed et al., 2014), heart disease prediction (Deepika & Balaji, 2022), protein detection (Zhang et al., 2016), stock market price forecasting (Kazem et al., 2013), speech recognition (Nuha & Abido, 2016), and text classification (Sainte & Alalyani, 2020). However, the standard FA has focused on optimisation (Yang, 2010); hence, it needs to be modified according to the gene interactions when it is applied to gene selection tasks (Fajila & Yusof, 2022). Furthermore, suitable techniques should be used to address the issues in the standard FA: slow convergence and local optimum.

METHODOLOGY

The proposed hybrid approach, CFS-MCFA-SVM, hybridises a filter approach and a wrapper approach. More precisely, CFS filter-based preprocessing and MCFA-based subset selection are utilised to find the biomarkers from the microarray datasets, and then, the cancer samples are classified using the SVM classifier. CFS-MCFA-SVM uses mutable solutions and a population reinitialisation strategy to resolve slow convergence issues and avoid trapping into local optimums. Further, the proposed algorithm also uses a composite position update function for the mutable solutions. As illustrated in Figure 3, there are three steps: data preprocessing, gene selection, and classification. The detailed descriptions of each step are given below.

Figure 3

Flow of CFS-MCFA-SVM



Data Preprocessing

The data preprocessing consists of two steps: normalisation and filtering. At the initial step, the popular min-max normalisation (Al Shalabi et al., 2006) technique was applied to the standard microarray datasets in order to normalise the values of each feature in the range of 0 and 1 using Equation 4 in

which the original value v is normalised into v'. Then, in the second step, the normalised microarray datasets were further processed using the CFS filter (Hall, 1999). CFS filtering reduces the large cancer microarray datasets by eliminating the irrelevant and redundant features, and consequently, the reduced CFS-filtered cancer microarray datasets are produced. After that, the normalised CFS-filtered datasets will undergo further processing through the gene selection step. Besides, the number of genes present in each cancer microarray dataset after filtering is tabulated in Table 2. The values inside the parentheses in Table 2 denote the number of genes in the particular dataset.

$$v' = \frac{v - min(v)}{max(v) - min(v)} \quad (4)$$

Table 2

Normalised Filtered Microarray Datasets

Dataset (No. of genes)	Filtered Dataset (No. of genes)
Colon (2000)	Colon (26)
Leukemia2 (7129)	Leukemia2 (81)
Leukemia3 (7129)	Leukemia3 (104)
SRBCT (2308)	SRBCT (111)

Gene Selection

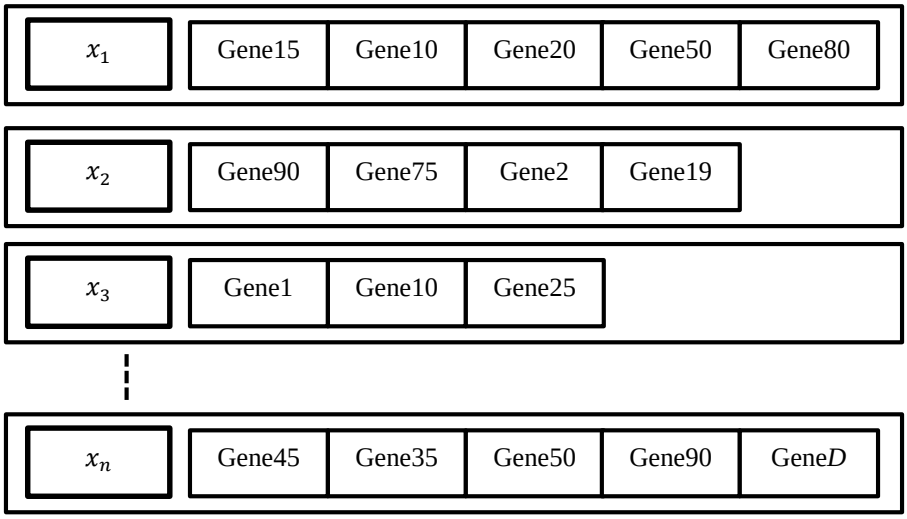
Selecting the biomarkers from the large microarray datasets is not a trivial task. Thus, feature selection using swarm optimisation algorithms should be handled effectively to produce global optimal solutions. Accordingly, the proposed MCFA has been designed with four steps: mutable size firefly population initialisation, composite position update, population reinitialisation, and population refinement. Each of the steps are discussed as given below.

Mutable Size Firefly Population Initialization

The population generation in the proposed MCFA is differentiated from the population generation in the standard FA. More precisely, MCFA uses mutable size solutions (i.e. non-fixed size solutions or dynamic size solutions) to initialise the population. In contrast, the standard FA generates the population with fixed-size solutions. The purpose of applying the property of mutable solutions is to overcome the slow convergence issue in the standard FA due to fixed-size solutions. Besides, a mutable solution would consist of s number of features where $s = 1,2,3, \dots, D$ in a D dimensional feature space. A firefly population with n number of non-fixed size fireflies would be depicted as in Figure 4.

Figure 4

Population Generation with Mutable Solutions in MCFA



Composite Position Update

The proposed composite position update strategy comprises two functions, namely, integrative position update and discriminative position update, as represented in Equations 5 and 6, respectively. According to the integrative function, the discriminant genes in the brighter firefly x_j are used to create a new firefly, which is used to enhance the brightness of the less bright firefly x_i. In contrast, in discriminative function, the discriminant genes in x_j create a new firefly itself to enhance the x_i.

$$x_i(t+1) = x_i(t) + x_j(t) - \left(x_i(t) \cap x_j(t)\right) \quad (5)$$

$$x_i(t+1) = x_j(t) \cap \left(x_i(t)\right)^c \quad (6)$$

where $x_i(t+1)$ is the new position of firefly i in iteration $(t+1)$ while $x_i(t)$ and $x_j(t)$ represent the positions of fireflies i and j in iteration (t), respectively in both Equation 5 and Equation 6. However, it is important to mention that the position of x_i would be updated if and only if the new firefly increases the brightness of xi according to the relevant position update function. Besides, in case, none of the position update functions will lead to a higher fitness of x_i, then the less bright firefly xi will remain as before. Nevertheless, since the population is reinitialised in each iteration while preserving the best-fit firefly of the current generation (as represented in the next section), the less bright fireflies are automatically updated randomly. Moreover, in contrast to the integrative position update, the discriminative position update results in a smaller gene subset, which is the aim of gene selection.

Population Reinitialisation

Typically, the execution of the algorithm up to the maximum number of iterations takes place over the initially generated population which would decrease the opportunity for exploration, which in turn would lead to local optimum in the standard FA. Thus, MCFA utilises the property of population reinitialisation, which regenerates the population for each iteration with the aim of increasing the exploration capabilities while preserving the best-fit firefly of the current iteration with an eye towards

enhancing the exploitation capability. Besides, the identification of the best firefly is also based on the size of the firefly, in addition to its fitness. In other words, the firefly with the highest fitness and the lowest number of genes is considered the best firefly in the population.

Population Refinement

The best fireflies found in each iteration are collected at the end of the iteration to create a new population of best fireflies, which is refined further using the proposed MCFA to find the global best firefly. Population refinement allows the population of best fireflies (which are expected to be more informative) to be exploited well to produce the optimal solution.

Classification

The third step of the proposed algorithm is classification, which trains and tests the different cancer samples using SVM and the splitting (70:30) technique. The outcome is then analysed based on classification accuracy. Algorithm 2 provides the steps of the proposed MCFA.

Algorithm 2: *Mutable Composite Firefly Algorithm.*

Step 1: Define parameters - population size n, dimension D, and maximum iteration: tmax
Step 2: Randomly generate firefly population with mutable size solutions: $x_i,\ i = 1, 2, \dots, n$
Step 3: Evaluate the fitness of each firefly: $f(x)$
Step 4: while (t < tmax)
Step 5: for $i = 1$ to n
Step 6: for $j = 1$ to n
Step 7: if $(f(x_i) < f(x_j))$
Step 8: Calculate the new position_1 x_{i_1} using Equation 5
Step 9: Calculate the new position_2 x_{i_2} using Equation 6
Step 10: Calculate the fitness of x_{i_1} using Equation 7
Step 11: Calculate the fitness of x_{i_2} using Equation 7
Step 12: if fitness of x_{i_1} is greater than the fitness of old x_{i_2}
Step 13: if fitness of x_{i_1} is greater than the fitness of old x_i
Step 14: Move firefly i towards j
Step 15: Update the position of firefly i using Equation 5
Step 16: Update the fitness of firefly i using Equation 7
Step 17: end if
Step 18: else
Step 19: if fitness of x_{i_2} is greater than the fitness of old x_i
Step 20: Move firefly i towards j
Step 21: Update the position of firefly i using Equation 6
Step 22: Update the fitness of firefly i using Equation 7
Step 23: end if
Step 24: end if
Step 25: end if
Step 26: Evaluate new solutions and update light intensity
Step 27: end for j
Step 28: end for i
Step 29: Rank the fireflies and find the best firefly
Step 30: Reinitialise the firefly population
Step 31: Refine the best fireflies and find the global best firefly
Step 32: end while

EVALUATION AND RESULTS

The performance of CFS-MCFA-SVM was analysed on four cancer microarray datasets, two of which are of binary and the rest are multiclass datasets. The classification accuracy (as provided in Equation 7) was produced using SVM classifier (Vapnik et al., 1997). Further, the proposed algorithm was implemented using WEKA and MATLAB software on a PC with an Intel Core i3 processor, 4.00 GB RAM, and a Windows 10 operating system.

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

where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. The higher classification accuracy would reflect the algorithm's efficiency.

Dataset Description

Four secondary cancer microarray datasets, Colon (Alon et al., 1999), Leukemia2 (Golub et al., 1999), Leukemia3 (Armstrong et al., 2001), and Small Round Blue Cell Tumor (SRBCT) (Khan et al., 2001), were used to evaluate the performance of CFS-MCFA-SVM. Table 3 provides details of the datasets.

Table 3

Details of Cancer Microarray Datasets

Dataset	No. of Classes	No. of Genes	No. of Samples	Description
Colon	2	2000	62	Tumor: 40 and Healthy: 22
Leukemia2	2	7129	72	ALL: 47 and AML: 25
Leukemia3	3	7129	72	B-cell: 38, T-cell: 9, and AML: 25
SRBCT	4	2308	83	EWS: 29, BL: 11, NB: 18, and RMS: 25

Note: ALL: Acute Lymphoblastic Leukemia, AML: Acute Myeloid Leukemia, EWS: Ewing's Sarcoma, BL: Burkitt's Lymphoma, NB: Neuroblastoma, and RMS: Rhabdomyosarcoma

The binary class dataset (i.e. two-class dataset), namely the Colon cancer dataset (Alon et al., 1999), contains 2000 genes and 62 samples. The binary dataset Leukemia2 (Golub et al., 1999) and multiclass (i.e. three class) dataset Leukemia3 (Armstrong et al., 2002) have the same number of genes (7129 genes) and an equal number of samples (72 samples). Further, the SRBCT (Khan et al., 2001) is a four-class dataset with 2308 genes and 83 samples. All four cancer datasets have many gene expressions which insist on the need for gene selection. Besides, these high-dimensional cancer microarray datasets have been popularly evaluated in existing studies (Al-Betar et al., 2020; AlMazrua & AlShamlan, 2022; Almugren & Alshamlan, 2019b; Alshamlan, 2018; AlShamlan & AlMazrua, 2024; Jain et al., 2018; Mazumder & Veilumuthu, 2018). Therefore, the proposed hybrid algorithm is evaluated on these datasets.

Parameter Settings

The parameter settings of CFS-MCFA-SVM as tabulated in Table 4 are based on preliminary studies and existing works (Al-Betar et al., 2020; Alshamlan, 2018). In Table 4, the parameter, namely

population size, which is equal to 80 fireflies, denotes the size of the firefly population. Besides, the feature space dimension is set to be equal to the number of genes in the dataset. In addition, the algorithm is iterated up to 100 times over 30 independent executions to produce better results.

Table 4

Parameter Settings applied in this Study

Parameter	Value
Population size	80
Dimension	Number of genes
Number of iterations	100
Number of runs	30

Results

This section evaluates the performance of the proposed algorithm in terms of the produced classification results on the four datasets. Table 5 presents the best, average, and worst classification accuracy results obtained on the four cancer datasets using the CFS-MCFA-SVM algorithm proposed parameter settings, while the biomarkers selected using CFS-MCFA-SVM are provided in Table 6. The gene count presented in each dataset is provided inside the parentheses in Table 5-7.

Table 5

Classification Performance of CFS-MCFA-SVM

Dataset	Run	Accuracy (%)		
		Best	Average	Worst
Colon (2000)	5	**100(1)**	**100(4)**	100(9)
	10	100(1)	98.95(5)	94.74(3)
	15	100(1)	99.3(5)	94.74(3)
	20	100(1)	98.95(5)	94.74(3)
	25	100(1)	98.74(5)	94.74(3)
	30	100(1)	98.6(5)	89.47(5)
Leukemia2 (7129)	5	**100(1)**	**100(1)**	100(2)
	10	100(1)	100(2)	100(3)
	15	100(1)	100(2)	100(3)
	20	100(1)	100(2)	100(3)
	25	100(1)	100(2)	100(3)
	30	100(1)	100(2)	100(3)
Leukemia3 (7129)	5	**100(2)**	100(6)	100(9)
	10	100(2)	100(6)	100(9)
	15	100(2)	100(6)	100(9)
	20	100(2)	100(6)	100(9)
	25	100(2)	**100(5)**	100(9)
	30	100(2)	100(5)	100(9)

(continued)

Dataset	Run	Accuracy (%)		
		Best	Average	Worst
SRBCT (2308)	5	**100(4)**	**100(7)**	100(11)
	10	100(4)	100(7)	100(11)
	15	100(4)	100(8)	100(11)
	20	100(4)	100(8)	100(11)
	25	100(4)	100(8)	100(11)
	30	100(4)	100(8)	100(11)

Table 6

Informative Genes obtained using CFS-MCFA-SVM

Dataset	Genes
Colon (1)	A249
Leukemia2 (1)	attribute3252 OR attribute2020 OR attribute1685 OR attribute1779
Leukemia3 (2)	X95735_at, U50327_s_at
SRBCT (4)	gene509, gene545, gene796, gene1662

According to Table 5 and Table 6, the results produced by the proposed CFS-MCFA-SVM reflect the remarkable contribution of the proposed algorithm for biomarker selection by producing 100% accuracy on all four datasets with only a few genes. Giving 100% accuracy for all the datasets shows the robustness of the proposed algorithm in classifying the cancer samples correctly, while the number of genes (which is less than 5) in the subset indicates the level of efficiency of the proposed algorithm in identifying a small set of informative biomarkers. It is worthwhile mentioning that the potential capability of minimising the number of genes in the final subset is nontrivial and essential for decreasing the computational complexity related to cancer diagnosis and treatment.

DISCUSSION

Remarkably, the proposed CFS-MCFA-SVM algorithm has produced 100% accuracy on all four datasets with only a few biomarkers (refer to Tables 5 and 6). The results reveal the algorithm's efficiency in perfecting the classification of different cancer microarray samples. Besides, the perfect classification justifies that the selected biomarkers are informative, which would be beneficial in cancer diagnosis, treatment, and therapeutics.

In concern to Colon cancer dataset, the proposed algorithm has classified the samples with 100% accuracy with a single gene. It is noteworthy that producing 100% accuracy with one gene is the best performance that could ever be achieved. Besides, none of the existing results (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018; Alshamlan et al., 2015a; Alshamlan et al., 2015b; Fajila & Yusof, 2021; Jain et al., 2018; Mehrabi et al., 2024; Qin et al., 2022; Xie et al., 2023) that are compared in Table 7 has yielded 100% classification accuracy for Colon cancer dataset except Fajila and Yusof (2022) which selected a bigger gene subset with five genes. Thus, the results obtained for the Colon cancer dataset using the proposed algorithm specify the significance of CFS-MCFA-SVM compared to the existing studies.

Similar to the Colon cancer classification, the proposed approach has produced 100% accuracy with a single gene for the Leukemia2 dataset. Identically, Fajila and Yusof (2022) also have achieved the same results. Meanwhile, Alshamlan et al. (2015a), Alshamlan et al. (2015b), Alshamlan (2018), Jain et al. (2018), Almugren and Alshamlan (2019b), Al-Betar et al. (2020), Qin et al. (2022), Xie et al. (2023), and Mehrabi et al. (2024) also have produced 100% accuracy but with 14, 4, 3, 4.3, 5, 4.07, 241, 4.3, and 2.6 genes, respectively. Nevertheless, the proposed algorithm and our previous work reported in Fajila and Yusof (2022) have achieved the optimum results (i.e. 100% accuracy with one gene) while the number of genes resulted by existing studies is greater than CFS-MCFA-SVM though the accuracy is the same. Apart from that, one existing work (Mazumder & Veilumuthu, 2018) still lack of capabilities to produce a perfect classification.

Further, CFS-MCFA-SVM has classified the Leukemia3 dataset with 100% accuracy using only two genes. In a similar fashion, our existing work published in Fajila and Yusof (2022) also has produced the same results - 100% accuracy using only two genes. Besides, the same accuracy has been obtained in existing studies (Al-Betar et al., 2020; Alshamlan, 2018; Alshamlan et al., 2015a; Alshamlan et al., 2015b; Fajila & Yusof, 2021; Jain et al., 2018; Qin et al., 2022) but with larger genes subsets. Moreover, the performance of the proposed approach is greater than that presented by Mazumder and Veilumuthu (2018), Almugren and Alshamlan (2019b), and Panda et al. (2025). Eventually, the proposed algorithm and our previous work reported in Fajila and Yusof (2022) are recognised with the best performance for Leukemia3 classification compared to existing works.

Interestingly, all the algorithms compared in Table 7 have provided 100% accuracy on SRBCT except Panda et al. (2025) which produced 97.1% accuracy using 23 genes. However, in concern to the number of selected genes, only the proposed CFS-MCFA-SVM and Alshamlan (2018) have shown the best performance (i.e. 100% accuracy with four genes) compared to related studies (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan et al., 2015a; Alshamlan et al., 2015b; Fajila & Yusof, 2021; Fajila & Yusof, 2022; Jain et al., 2018; Mazumder & Veilumuthu, 2018; Mehrabi et al., 2024; Qin et al., 2022). As a consequence, the overall classification performance of CFS-MCFA-SVM (as given in Table 7) compared to the existing methods is superior on all four datasets. The performance evaluation criteria are based on classification accuracy and the number of biomarker genes.

Table 7

Classification Performance Comparison between CFS-MCFA-SVM and Related Works

Algorithms	Colon	Leukemia2	Leukemia3	SRBCT
CFS-MCFA-SVM	**100(1)**	**100(1)**	**100(2)**	**100(4)**
Panda et al. (2025)	-	-	96.7(57)	97.1(23)
Mehrabi et al. (2024)	98.48(7.36)	100(2.6)	-	100(5.53)
Xie et al. (2023)	90.48(9.4)	100(4.3)	-	-
Fajila and Yusof (2022)	100(5)	**100(1)**	**100(2)**	100(7)
Qin et al. (2022)	97.6(148)	100(241)	100(182)	100(104)
Fajila and Yusof (2021)	95.23(4)	-	100(4)	100(8)
Al-Betar et al. (2020)	97.85(12.27)	100(4.07)	100(5.33)	100(9.13)
Almugren and Alshamlan (2019b)	94.3(15)	100(5)	97.8(10)	100(8)
Alshamlan (2018)	96.77(9)	100(3)	100(6)	**100(4)**
Mazumder and Veilumuthu (2018)	-	98.61(3)	98.61(3)	100(6)
Jain et al. (2018)	94.89(4.2)	100(4.3)	100(6)	100(34.1)
Alshamlan et al. (2015a)	96.77(15)	100(14)	100(20)	100(10)
Alshamlan et al. (2015b)	98.38(10)	100(4)	100(8)	100(6)

Notably, the proposed algorithm has demonstrated outstanding contribution in biomarker detection for cancer classification since the performance of the proposed algorithm was excellent on all four datasets, as emphasised in Table 7. The contribution of data preprocessing with CFS filter has been reflected through the ultimate performance of CFS-MCFA-SVM, which would have benefited from the preprocessing step. Selecting the relevant set of features from a large pool of features is crucial for determining the informative features. Besides, the non-fixed size population has resolved the slow convergence problem in the standard FA by providing better exploration and exploitation. More specifically, CFS-MCFA-SVM does not limit the size of the solution in contrast to the fixed-size solutions in existing studies (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018) thus, the solution generation would not penalise from too much exploration with too little exploitation so, the algorithm will converge appropriately. Further, a composite position update function is designed for the mutable fireflies. The composite strategy is used to facilitate the exploration and exploitation capabilities. Furthermore, in contrast to existing studies that iterate the initial population throughout the generations (Al-Betar et al., 2020; Almugren & Alshamlan, 2019b; Alshamlan, 2018; Emami & Pakzad, 2019; Jabbar, 2019), the proposed algorithm reinitialises the population hence, the population diversity is enhanced to overcome the local optima issue in standard FA. At the same time, the exploitation capability is also increased via the process of preserving the best solution during population reinitialisation. In addition, firefly population refinement also increases the exploitation and exploration capabilities of the proposed CFS-MCFA-SVM. Consequently, the optimal solution or a small subset of biomarkers that can accurately classify the cancer samples with the highest accuracy is produced using the proposed algorithm.

Regarding the genes' correlations, it is recognised that a strongly relevant gene is crucial for an informative genes subset, and further, the elimination of a strongly relevant gene will affect the classification results (Zhang & Deng, 2007). At the same time, a weakly relevant gene can enhance the classification results in some scenarios (Kohavi & John, 1997). Besides, even an informative genes subset can hardly ever contain any irrelevant genes (Kohavi & John, 1997). Accordingly, the optimality

indicates neither, all the genes in the subsets are relevant or all the genes will increase the classification results. Thus, it is observable that the composite position update functions have significantly contributed for biomarker selection.

CONCLUSION

The unsettled cancer-related issues in terms of treatment, early detection, and growing rate of mortality have stimulated the utilisation of high throughput technologies for cancer classification. However, the unfilled gaps in the related works have motivated the invention of this hybrid algorithm CFS-MCFA-SVM. Accordingly, CFS-MCFA-SVM put forward a new non-fixed size solution generation method, namely mutable solution generation for the firefly population, which was aimed in order to settle down the slow convergence issue in the standard FA, which would cause, upon defining a large threshold for the dimension of solution. In addition, CFS-MCFA-SVM proposes a composite position update function for the mutable solutions in order to facilitate the exploration and exploitation. Moreover, a population reinitialisation method that initialises the population of the next generation while retaining the current best solution is suggested by CFS-MCFA-SVM in order to avoid the local optimums through diversifying the population while improving the exploration and exploitation capabilities via utilising the best solution. Besides, a population refinement method that exploits the best fireflies of each iteration is also recommended in CFS-MCFA-SVM concerning the identification of the global best firefly. Correspondingly, the balance between exploration and exploitation is maintained in the proposed CFS-MCFA-SVM in order to overcome the major issues in the standard FA.

The properties of the algorithm are verified through the produced results as all four datasets have shown the best performance (i.e. 100% accuracy with only a few biomarkers) compared to the existing related works. More specifically, 100% accuracy with a single gene (which is considered as the optimal solution) was produced on the Colon and Leukemia2 datasets, while the Leukemia3 and SRBCT datasets were classified with 100% accuracy using only 2 and 4 informative genes, respectively. However, though the proposed CFS-MCFA-SVM operationalised well in gene selection for cancer classification, this research is limited to cancer DNA microarray data classification using FA. Thus, as for the future work, it would be beneficial if other classification problems apart from cancer classification is addressed to assess the performance of the proposed algorithm. Further, it is aimed to extend the properties of the proposed CFS-MCFA-SVM to be evaluated on various microarray datasets to validate the robustness of the algorithm in gene selection. Besides, the biomarkers produced by CFS-MCFA-SVM would contribute to cancer-related issues upon verifying the relevancy of particular genes via technical analysis from a medical perspective.

ACKNOWLEDGMENT

This research was supported by the Ministry of Higher Education (MoHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2022/ICT02/UUM/02/1).

REFERENCES

Agarwal, V., & Bhanot, S. (2015, August). Firefly inspired feature selection for face recognition. In *2015 Eighth International Conference on Contemporary Computing (IC3)* (pp. 257-262). IEEE. https://doi.org/10.1109/IC3.2015.7346689

Aha, D. W., Kibler, D., & Albert, M. K. (1991, January). Instance-based learning algorithms. *Machine Learning, 6*, 37-66. https://doi.org/10.1007/BF00153759

Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*, *18*(3), 215-220.

Al-Baity, H. H., & Al-Mutlaq, N. (2021). A new optimized wrapper gene selection method for breast cancer prediction. *Computers, Materials & Continua*, *67*(3). https://doi.org/10.32604/cmc.2021.015291

Al-Batah, M., Zaqaibeh, B., Alomari, S. A., & Alzboon, M. S. (2019). Gene microarray cancer classification using correlation based feature selection algorithm and rules classifiers. *International Journal of Online & Biomedical Engineering* (*iJOE), 15*(8), 62-73. https://doi.org/10.3991/ijoe.v15i08.10617

Al-Betar, M. A., Alomari, O. A., & Abu-Romman, S. M. (2020). A TRIZ-inspired bat algorithm for gene selection in cancer classification. *Genomics*, *112*(1), 114-126. https://doi.org/10.1016/j.ygeno.2019.09.015

Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. *Processes*, *11*(2), 562. https://doi.org/10.3390/pr11020562

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., ... Staudt, L. M. (2000, February 3). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503-511. https://doi.org/10.1038/35000501

AlMazrua, H., & AlShamlan, H. (2022). A new algorithm for cancer biomarker gene detection using Harris Hawks optimisation. *Sensors*, *22*(19), 7273.

Almugren, N., & Alshamlan, H. (2019a, July). FF-SVM: New FireFly-based gene selection algorithm for microarray cancer classification. *In 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-6). https://doi.org/10.1109/CIBCB.2019.8791236

Almugren, N., & Alshamlan, H. M. (2019b). New bio-marker gene discovery algorithms for cancer gene expression profile. *IEEEAccess*, *7*, 136907-136913. https://doi.org/10.1109/ACCESS.2019.2942413

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., & Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS. 96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735-739.

AlShamlan, H., & AlMazrua, H. (2024). Enhancing cancer classification through a hybrid bio-inspired evolutionary algorithm for biomarker gene selection. *Computers, Materials & Continua*, *79*(1).

Alshamlan, H., Badr, G., & Alohali, Y. (2015a). mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *BioMed Research International*, *2015*, http://dx.doi.org/10.1155/2015/604910

Alshamlan, H. M. (2018). Co-ABC: Correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile. *Saudi Journal of Biological Sciences, 25*(5), 895-903. https://doi.org/10.1016/j.sjbs.2017.12.012

Alshamlan, H. M., Badr, G. H., & Alohali, Y. A. (2015b). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. *Computational Biology and Chemistry, 56*, 49-60. https://doi.org/10.1016/j.compbiolchem.2015.03.001

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., & Korsmeyer, S. J. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet, 30*(1), 41–47. https://doi.org/10.1038/ng765

Basir, M. A., Yusof, Y., & Hussin, M. S. (2019). Optimisation of attribute selection model using bio-inspired algorithms. *Journal of Information and Communication Technology*, *18*(1), 35-55. https://doi.org/10.32890/jict2019.18.1.3

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D., Orringer, M. B., & Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med, 8*(8), 816-824. https://doi.org/10.1038/nm733

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., & Meyerson, M. (2001, November 20). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, *98*(24), 13790-13795. https://doi.org/10.1073/pnas.191502998

Cheng, S., Shi, Y., Qin, Q., Ting, T. O., & Bai, R. (2014a, July 6-11). Maintaining population diversity in brain storm optimisation algorithm. In *2014 IEEE Congress on Evolutionary Computation (CEC)* (pp. 3230-3237). IEEE. https://doi.org/10.1109/CEC.2014.6900255

Cheng, S., Shi, Y., Qin, Q., Zhang, Q., & Bai, R. (2014b). Population diversity maintenance in brain storm optimisation algorithm. *Journal of Artificial Intelligence and Soft Computing Research, 4*(2), 83-97. https://doi.org/10.1515/jaiscr-2015-0001

Das, A., Neelima, N., Deepa, K., & Özer, T. (2024). Gene selection based cancer classification with adaptive optimisation using deep learning architecture. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3392633

Dash, R. (2021). An adaptive harmony search approach for gene selection and classification of high dimensional medical data. *Journal of King Saud University – Computer and Information Sciences, 33*(2), 195-207. https://doi.org/10.1016/j.jksuci.2018.02.013

Dashtban, M., & Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics*, *109*(2), 91-107. https://doi.org/10.1016/j.ygeno.2017.01.004

Deepika, D., & Balaji, N. (2022). Effective heart disease prediction with Grey-wolf with Firefly algorithm-differential evolution (GF-DE) for feature selection and weighted ANN classification. *Computer Methods in Biomechanics and Biomedical Engineering*, *25*(12), 1409-1427.

Dif, N., & Elberrichi, Z. (2019). An enhanced recursive firefly algorithm for informative gene selection. *International Journal of Swarm Intelligence Research (IJSIR), 10*(2), 21–33. https://doi.org/10.4018/IJSIR.2019040102

Dokeroglu, T., Sevinc, E., Kucukyilmaz, T., & Cosar, A. (2019). A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering, 137*, 106040. https://doi.org/10.1016/j.cie.2019.106040

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimisation. *IEEE Computational Intelligence Magazine*, *1*(4), 28-39.

El-Abd, M. (2016, July 24-29). Brain storm optimisation algorithm with re-initialised ideas and adaptive step size. In *2016 IEEE Congress on Evolutionary Computation (CEC)* (pp. 2682-2686). IEEE. https://doi.org/10.1109/CEC.2016.7744125

El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2011). A two-stage gene selection scheme utilising MRMR filter and GA wrapper. *Knowledge and Information Systems, 26*(3), 487-500. https://doi.org/10.1007/s10115-010-0288-x

Elyasigomari, V., Lee, D. A., Screen, H. R. C., & Shaheed, M. H. (2017). Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimisation algorithm and harmony search for cancer classification. *Journal of Biomedical Informatics*, *67*, 11–20, https://doi.org/10.1016/j.jbi.2017.01.016

Emami, N., & Pakzad, A. (2019). A new knowledge-based system for diagnosis of breast cancer by a combination of the affinity propagation and firefly algorithms. *Journal of AI and Data Mining*, *7*(1), 59-68. https://doi.org/10.22044/JADM.2018.6489.1763

Fajila, F., & Yusof, Y. (2021). Incremental search for informative gene selection in cancer classification. *Annals of Emerging Technologies in Computing (AETiC), 5*(2), 15-21, https://doi.org/10.33166/AETiC.2021.02.002

Fajila, M. N. F., & Yusof, Y. (2022). Hybrid gene selection with mutable firefly algorithm for feature selection in cancer classification. *International Journal of Intelligent Engineering and Systems (IJIES), 15*(3), 24-35. https://doi.org/10.22266/ijies2022.0630.03

Fajriyah, R. (2021). Paper review: An overview on microarray technologies. *Bulletin of Applied Mathematics and Mathematics Education*, *1*(1), 21-30. https://doi.org/10.12928/bamme.v1i1.3854

Fathi, E., Mesbah-Namin, S. A., & Farahzadi, R. (2014). Biomarkers in medicine: an overview. *Journal of Advances in Medicine and Medical Research*, *4*(8), 1701-1718. https://doi.org/10.9734/BJMMR/2014/6917

Fong, S., Biuk-Aghai, R. P., & Millham, R. C. (2018, February). Swarm search methods in Weka for data mining. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 122-127). https://doi.org/10.1145/3195106.3195167

Gao, L., Ye, M., Lu, X., & Huang, D. (2017). Hybrid method based on information gain and support vector machine for gene selection in cancer classification. *Genomics, Proteomics & Bioinformatics*, *15*(6), 389-395. https://doi.org/10.1016/j.gpb.2017.08.002

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, L., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy and Bioallied Sciences*, *4*(Suppl 2), S310-S312. https://doi.org/10.4103/0975-7406.100283

Guo, J., & Tang, S. J. (2009, August 26-27). An improved particle swarm optimisation with re-initialisation mechanism. In *2009 International Conference on Intelligent Human-Machine Systems and Cybernetics* (Vol. 1, pp. 437-441). IEEE. https://doi.org/10.1109/IHMSC.2009.117

Guo, Q. M. (2003). DNA microarray and cancer. *Current opinion in oncology*, *15*(1), 36-43.

Hall, M. A. (1999). Correlation-based feature selection for machine learning [Doctoral dissertation, University of Waikato]. University of Waikato. https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Hoang, D. T. (2008). *Metaheuristics for NP-hard combinatorial optimisation problems* [Doctoral dissertation, National University of Singapore].

Jabbar, S. F. (2019). A classification model on tumor cancer disease based mutual information and firefly algorithm. *Periodicals of Engineering and Natural Sciences*, *7*(3), 1152-1162. https://doi.org/10.21533/pen.v7i3.656

Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimisation for gene selection and cancer classification. *Applied Soft Computing*, *62*, 203-215. https://doi.org/10.1016/j.asoc.2017.09.038

Kang, J. (2015). Clinical implications of microarray in cancer medicine. *International Journal of Cancer Research*, *11*(4), 150-158. https://doi.org/10.3923/ijcr.2015.150.158

Karaboga, D. (2005). An idea based on honey bee swarm for numerical optimization. *Technical Report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department*, *200*, 1-10.

Karley, D., Gupta, D., & Tiwari, A. (2011). Biomarkers: The future of medical science to detect cancer. *J Mol Biomark Diagn*, *2*(5), 1-7. https://doi.org/10.4172/2155-9929.1000118

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied soft computing*, *13*(2), 947-958. http://dx.doi.org/10.1016/j.asoc.2012.09.024

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* *7*(6), 673–679. https://doi.org/10.1038/89044

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

Lai, C. M., Yeh, W. C., & Chang, C. Y. (2016). Gene selection using information gain and improved simplified swarm optimisation. *Neurocomputing*, *218*, 331-338. https://doi.org/10.1016/j.neucom.2016.08.089

Liu, H., Li, J., & Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics*, *13*, 51-60. https://doi.org/10.11234/gi1990.13.51

Marie-Sainte, S. L., & Alalyani, N. (2020). Firefly algorithm based feature selection for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, *32*(3), 320-328. https://doi.org/10.1016/j.jksuci.2018.06.004

Mazumder, D. H., & Veilumuthu, R. (2018). Cancer classification with a novel hybrid feature selection technique. *International Journal of Simulation: Systems, Science & Technology*, *19*(2). https://doi.org/10.5013/IJSSST.a.19.02.07

Mazumder, D. H., & Veilumuthu, R. (2019). An enhanced feature selection filter for classification of microarray cancer data. *ETRI Journal*, *41*(3), 358-370. https://doi.org/10.4218/etrij.2018-0522

McCulloch, W. S., & Pitts, W. (1943, December). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF0247825

Mehrabi, N., Haeri Boroujeni, S. P., & Pashaei, E. (2024). An efficient high-dimensional gene selection approach based on the Binary Horse Herd Optimisation Algorithm for biologicaldata classification. *Iran Journal of Computer Science*, 1-31. https://doi.org/10.1007/s42044-024-00174-z

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., & Mirjalili, S. M. (2017). Salp swarm algorithm: A bio-inspired optimiser for engineering design problems. *Advances in Engineering Software*, *114*, 163-191. https://doi.org/10.1016/j.advengsoft.2017.07.002

Mirjalili, S., & Lewis, A. (2016). The whale optimisation algorithm. *Advances in Engineering Software*, *95*, 51-67.

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimiser. *Advances in Engineering Software*, *69*, 46-61.

Mohammed, A. J., Yusof, Y., & Husni, H. (2014). Weight-based firefly algorithm for document clustering. In *Proc. of the First International Conference on Advanced Data and Information Engineering* (pp. 259-266).

Mostafa Bozorgi, S., & Yazdani, S. (2019). IWOA: An improved whale optimisation algorithm for optimisation problems. *Journal of Computational Design and Engineering*, *6*(3), 243-259. https://doi.org/10.1016/j.jcde.2019.02.002

Mustaffa, Z., Yusof, Y., & Kamaruddin, S. S. (2013). Enhanced Abc-Lssvm for energy fuel price prediction. *Journal of Information and Communication Technology*, *12*, 73-101.

Nuha, H. H., & Abido, M. (2016). Firefly algorithm for log-likelihood optimisation problem on speech recognition. In *Proc. of 2016 4th International Conference on Information and Communication Technology* (pp. 1-6).

Panda, M. (2020). Elephant search optimisation combined with deep neural network for microarray data analysis. *Journal of King Saud University-Computer and Information Sciences*, *32*(8), 940-948. https://doi.org/10.1016/j.jksuci.2017.12.002

Panda, P., Bisoy, S. K., Panigrahi, A., Pati, A., Sahu, B., Guo, Z., ... & Jain, P. (2025). BIMSSA: Enhancing cancer prediction with salp swarm optimization and ensemble machine learning approaches. *Frontiers in Genetics*, *15*, 1491602. https://doi.org/10.3389/fgene.2024.1491602

Pashaei, E., & Pashaei, E. (2019, November 28-30). Gene selection using intelligent dynamic genetic algorithm and random forest. In *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 470-474). IEEE. https://doi.org/10.23919/ELECO47770.2019.8990557

Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. United States.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., … Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*(6870), 436-442. https://doi.org/10.1038/415436a

Pushpalatha, K. R., & Karegowda, A. G. (2017, December). CFS based feature subset selection for enhancing classification of similar looking food grains-a filter approach. In *2017 2nd International Conference on Emerging Computation and Information Technologies (ICECIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICECIT.2017.8453403

Qin, X., Zhang, S., Yin, D., Chen, D., & Dong, X. (2022). Two-stage feature selection for classification of gene expression data based on an improved Salp Swarm Algorithm. *Math. Biosci. Eng*, *19*, 13747-13781. https://doi.org/10.3934/mbe.2022641

Quinlan, J. R. (1986). Induction of decision trees. *Mach Learn*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Ramasubramanian, S. (2020). Role of microarray in cancer biology. *Journal of Complementary Medicine Research*, *11*(3), 262-268. https://doi.org/10.5455/jcmr.2020.11.03.33

Remeseiro, B., Bolón-Canedo, V., Alonso-Betanzos, A., & Penedo, M. G. (2015). Learning features on tear film lipid layer classification. In *ESANN*.

Remeseiro, B., Mendonça, A. M., & Campilho, A. (2017, May). Objective quality assessment of retinal images based on texture features. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 4520-4527). IEEE. https://doi.org/10.1109/IJCNN.2017.7966429

Sainte, S. L. M., & Alalyani, N. (2020). Firefly algorithm based feature selection for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, *32*(3), 320-328.

Salgotra, R., Singh, U., & Saha, S. (2019). On some improved versions of whale optimisation algorithm. *Arabian Journal for Science and Engineering*, *44*(11), 9653-9691. https://doi.org/10.1007/s13369-019-04016-0

Sawhney, R., Mathur, P., & Shankar, R. (2018, May). A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In International Conference on Computational Science and Its Applications (pp. 438-449). Springer, Cham. https://doi.org/10.1007/978-3-319-95162-1_30

Sayed, G. I., Tharwat, A., & Hassanien, A. E. (2019). Chaotic dragonfly algorithm: An improved metaheuristic algorithm for feature selection. *Applied Intelligence*, *49*(1), 188-205. https://doi.org/10.1007/s10489-018-1261-8

Sekaj, I., & Oravec, M. (2009, June). Selected population characteristics of fine-grained parallel genetic algorithms with re-initialisation. In *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation* (pp. 945-948). https://doi.org/10.1145/1543834.1543980

Sekaj, I., & Perkacz, J. (2007, September 25-28). Some aspects of parallel genetic algorithms with population re-initialisation. In *2007 IEEE Congress on Evolutionary Computation* (pp. 1333-1338). IEEE. https://doi.org/10.1109/CEC.2007.4424625

Sharma, M., & Kaur, P. (2021). A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. *Archives of Computational Methods in Engineering*, *28*, 1103-1127.

Shekar, B. H., & Dagnew, G. (2020). L1-regulated feature selection and classification of microarray cancer data using deep learning. In *Proceedings of 3rd International Conference on Computer Vision and Image Processing* (pp. 227-242). Springer, Singapore. https://doi.org/10.1007/978-981-32-9291-8_19

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., & Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, *8*(1), 68-74. https://doi.org/10.1038/nm0102-68

Shukla, A. K., Singh, P., & Vardhan, M. (2019). DNA gene expression analysis on diffuse large b-cell lymphoma (dlbcl) based on filter selection method with supervised classification method. In Computational Intelligence in Data Mining (pp. 783-792). Springer, Singapore. https://doi.org/10.1007/978-981- 10-8055-5_69

Shukla, A. K., Tripathi, D., Reddy, B. R., & Chandramohan, D. (2020). A study on metaheuristics approaches for gene selection in microarray data: Algorithms, applications and open challenges. *Evolutionary intelligence*, *13*, 309-329.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M.., Kantoff, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*(2), 203-209. https://doi.org/10.1016/S1535-6108(02)00030-2

Sriram, K. B., Larsen, J. E., Yang, I. A., Bowman, R. V., & Fong, K. M. (2011). Genomic medicine in non-small cell lung cancer: Paving the path to personalised care. *Respirology*, *16*(2), 257-263. https://doi.org/10.1111/j.1440-1843.2010.01892.x

Tobore, T. O. (2019). On the need for the development of a cancer early detection, diagnostic, prognosis, and treatment response system. *Future science OA*, *6*(2), FSO439. https://doi.org/10.2144/fsoa-2019-0028

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. In *Advances in neural information processing systems 9* (pp. 281-287). MIT Press.

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Comput & Applic*, *24*(1), 175-186. https://doi.org/10.1007/s00521-013-1368-0

Wang, Y., Yang, X.-G., & Lu, Y. (2019). Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. *Applied Mathematical Modelling*, *71*, 286–297. https://doi.org/10.1016/j.apm.2019.01.044

World Health Organization. (2018, September 12). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer

Wright, S. (1965, September). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, *19*(3), 395-420. https://doi.org/10.2307/2406450

Xie, W., Wang, L., Yu, K., Shi, T., & Li, W. (2023). Improved multi-layer binary firefly algorithm for optimising feature selection and classification of microarray data. *Biomedical Signal Processing and Control*, *79*, 104080. https://doi.org/10.1016/j.bspc.2022.104080

Yab, L. Y., Wahid, N., & Hamid, R. A. (2024). Impact of balanced exploration and exploitation on high-dimensional feature selection with hierarchical whale optimisation algorithm. *Journal of Information and Communication Technology*, *23*(4), 593-626. https://doi.org/10.32890/jict2024.23.4.2

Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver Press.

Yang, X.-S. (2013). Metaheuristic optimisation: Nature-inspired algorithms and applications. In Artificial Intelligence, Evolutionary Computing and Metaheuristics (pp. 405-420). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29694-9_16

Yang, X. S. (2014). Swarm intelligence based algorithms: A critical analysis. *Evolutionary intelligence*, *7*(1), 17-28.

Yang, X.-S., & Deb, S. (2009, December). Cuckoo Search via Lévy flights. In *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)* (pp. 210-214). IEEE. https://doi.org/10.1109/nabic.2009.5393690

Yang, X. S., & He, X. (2013). Firefly algorithm: Recent advances and applications. *International Journal of Swarm Intelligence*, *1*(1), 36-50.

Zhang, J., Gao, B., Chai, H., Ma, Z., & Yang, G. (2016). Identification of DNA-binding proteins using multi-features fusion and binary firefly optimisation algorithm. *BMC Bioinformatics*, *17*(323), 1-12.

Zhang, J. G., & Deng, H. W. (2007). Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics*, *8*(1), 1-9. https://doi.org/10.1186/1471-2105-8-370

Zhu, Z., Ong, Y. S., & Dash, M. (2007, November). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, *40*(11), 3236-3248. https://doi.org/10.1016/j.patcog.2007.02.007