Multi-Leader Dynamic Composite Salp Swarm Algorithm for Biomarker Selection in Cancer Classification

Mohamed Nisper Fathima Fajila and Yuhanis Yusof

Abstract*—Biomarker selection for microarray-based cancer classification is a recent alternative in cancer diagnosis and prognosis. Though existing approaches have yielded competing performance in terms of classification accuracy and produce smaller gene subsets, a generalized classifier that can operate on various cancer microarray datasets is yet to be reported. Recently, metaheuristic algorithms have demonstrated momentous performance in cancer classification. Nevertheless, the performance of a metaheuristic algorithm is influenced by the fitness value, convergence, exploration, and exploitation capabilities. Thus, this study proposes a new variant of Salp Swarm Algorithm (SSA) which is later integrated with Correlation-based Feature Selection (CFS) filter for gene selection in cancer classification. The slow convergence issue in SSA is solved by introducing dynamic size solutions that are generated using a composite position update function. Further, the local optima issue is overcome by the population reinitialization method. Adding to that, exploitation of best solution search space is enhanced by multi-leaders which are termed as dual leaders. The proposed hybrid algorithm, named the CFS-Multi-Leader Dynamic Composite Salp Swarm Algorithm (CFS-MDCSSA), is evaluated using two metrics, namely classification accuracy and gene subset size, using Support Vector Machine (SVM) classifier. The proposed CFS-MDCSSA-SVM achieved 100% accuracy with only a few biomarkers for all six cancer microarray datasets, reflecting the competitive performance of the proposed algorithm in gene selection for cancer classification.

***Index Terms*—Biomarker Selection, Cancer Classification, Microarray, Salp Swarm Algorithm**

I. Introduction

THE International Agency for Research on Cancer (IARC) has represented the cancer as a growing burden [1]. Cancer has caused for 9.7 million deaths in 2022 [1]. The highly spreading nature of cancer would lead to the worst stage in a short period of time. Nevertheless, early diagnosis and treatment can reduce the mortality rate heavily [2]. However, detecting cancer in its early stage is not a trivial task. Besides, cancer-related informative genes known as cancer biomarkers assist in cancer diagnosis, prognosis, early detection, and treatment [3]. Hence, automated biomarker selection using DNA microarrays has become a popular trend currently. However, since the number of genes resulting from a microarray experiment is very large compared to the number of samples [4], gene selection is a challenging task.

Existing gene selection algorithms such as filters [5], [6] and wrappers [7]-[9] fall off into the two significant issues of gene selection: low classification performance and large gene subset size. Hence, hybrid methods [10], [11] which deploy filter and wrapper algorithms are suggested to have enhanced performance in cancer classification. Beyond that, swarm-based hybrid gene selection algorithms [12]-[15] are better than normal hybrid methods [16], [17]. Swarm-based optimization algorithms focus on searching for the optimal solution instead of the exact solution, which is challenging to be achieved over a large feature space such as a DNA cancer microarray. Swarm intelligence metaheuristics such as Ant Colony Optimization (ACO) algorithm [18], Ant Lion Optimization (ALO) algorithm [19], Artificial Bee Colony (ABC) algorithm [20], Firefly Algorithm (FA) [21], and Salp Swarm Algorithm (SSA) [22] are being exploited in various cancer classification studies. However, the success of a model always depends on the influencing factors such as fitness, convergence, exploration, and exploitation capabilities of the swarm algorithm.

SSA is a recently developed swarm-based optimization algorithm that mimics the swarming nature of salps in a salp chain. Though SSA is characterized by simplicity, it suffers due to slow convergence, premature convergence, and local optimum [23], [24]. However, sufficient amount of diversification (i.e. exploration or global search) together with intensification (i.e. exploitation or local search) would address the issues in the conventional SSA. Existing studies have proposed various population reinitialization strategies such as partial reinitialization [25]-[27], reinitialization while preserving best individuals [28], [29], and start reinitialization with a threshold [30] to enrich the diversity of the population and exploration of search space. This research adapts the population reinitialization method that was suggested in recent work [12], [31] to address the local optima issues in conventional SSA [22].

It is noteworthy that existing studies iterate the initial population throughout the generations [23], [32], [33], which may lead to the stagnation in locally optimal solutions. Further, the fixed size solutions used in existing works [23], [32]-[34] would slow down the convergence. In

Manuscript received March 19, 2025; revised August 14, 2025.

Mohamed Nisper Fathima Fajila is a lecturer (probationary) at the Department of Computer Science, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka (Corresponding author to provide phone: +94768509396; e-mail: fajila@seu.ac.lk).

Yuhanis Yusof is an Associate Professor at the School of Computing, Universiti Utara Malaysia, Malaysia (e-mail: yuhanis@uum.edu.my).

other words, defining a large threshold for the size of the solution not only increases the feature subset size but also allows the algorithm to search exhaustively throughout the feature space, leading to slow converge. Hence, non-fixed size solutions [12], [31], termed dynamic size solutions, are utilized to handle the slow convergence in standard SSA.

On top of all suggested enhancements, in contrast to the conventional SSA [22] that defines a single leader, this study proposes multi-leader concept where two leaders, termed as dual leaders, lead the salp chain. It is noteworthy that multi-leaders are suggested in literature with different techniques such as assigning half of the population as leaders [34] and dynamically changing the number of leaders over the iteration [23], [32] to address the slow convergence issue and ensure the balance between exploration and exploitation. Hence, this study proposes a dual leader concept where two leaders are utilized to guide the follower salps. The leaders are preserved and exploited during the population reinitialization to maintain a balance between exploration and exploitation.

Apart from that, the proposed algorithm uses a composite position update function for the dynamic size salps. The position update in the standard SSA [22] is applicable only for fixed-size solutions thus, a new position update equation is required for dynamic-size solutions. Furthermore, the position update equation in the conventional SSA is related to the initial position of the salp; thus, it is not appropriate for gene selection, which should concern the interactions among the genes.

Therefore, a hybrid swarm-based gene selection algorithm is proposed for this study. The proposed algorithm is termed CFS-MDCSSA-SVM as it integrates a Correlation-based Feature Selection (CFS) filter [35], a Multi-leader Dynamic Composite Salp Swarm Algorithm (MDCSSA), and a Support Vector Machine (SVM) classifier [36]. The contributions of this research are summarized below:

First, the proposed study designs a hybrid gene selection algorithm that combines a CFS filter for data preprocessing and MDCSSA for biomarker selection. Second, the multi-leader (i.e. dual leader) concept is integrated to balance between exploration and exploitation. Third, a composite position update function is adapted for the dynamic size solutions concerning the interactions among the genes. Fourth, the population reinitialization method is modified to address the local optima issues in conventional SSA. Finally, as the fifth contribution, the performance of the algorithm is evaluated on six cancer microarray datasets using SVM and the results are compared with existing work.

The rest of the paper is organized as follows: Section 2 provides a brief background of related studies, whereas Section 3 illustrates the proposed research methodology. The experimental results and discussion are presented in Section 4. Finally, the conclusion is provided in Section 5.

II. RELATED WORK

This section discusses the basic knowledge of the CFS filter [35], the SVM classifier [36], and the SSA [22]. This section also presents details on feature selection and its application.

A. Correlation-based Feature Selection Filter

Filter-based preprocessing is typically proposed in existing work [14], [15], [37] as a preliminary extraction technique for selecting the relevant genes. The large-dimensional microarrays often comprise a massive collection of relevant, irrelevant, and redundant genes. Hence, removing irrelevant and redundant genes is crucial for gene selection. The two types of filters namely, the univariate filters such as F-score filter [38] and minimum Redundancy Maximum Relevance (mRMR) filter [39] that evaluates the individual features, and the multivariate filters such as CFS filter [35] and Markov blanket filter [40] that evaluate the feature subsets, are utilized in existing work.

The evaluation criteria of the CFS filter are based on the correlations among the genes and the corresponding class, where the preference is provided to the genes with a higher correlation towards the class and a lower correlation within the genes. A gene subset is assigned a score according to (1).

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

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. Apart from that, the Best First Search (BFS) [41] strategy is applied for searching due to its excellent performance over large feature space [42]. CFS filter has been utilized in many classification tasks [43], [44] along with cancer classification [12], [31], [45] giving appreciable output. Hence, the proposed study employs a CFS filter for data preprocessing.

B. Support Vector Machine

SVM is a supervised classification algorithm introduced by Vapnik [36]. Many existing applications have exploited the SVM classifier for different purposes, especially in medical domain, such as for X-ray analysis [46], diabetes prediction [47], Alzheimer's disease classification [48], [49], and also for gene selection [7], [14], [50] to produce competing results. In SVM-based classification, the samples in a dataset are separated by a hyperplane that is drawn with respect to the class. SVM classifier possesses the ability to detect the optimal hyperplane boundary which can separate the different classes with a larger margin [51]. Also, SVM can manage both linear and non-linear separations [52]. In concern to the SVM properties, this study utilizes the SVM classifier to assess the performance of the produced feature set.

C. Salp Swarm Algorithm

Swarm algorithms are an alternative source for solving computationally intensive applications for which obtaining the exact solution is an NP-hard optimization problem. Swarm algorithms provide optimal or near-optimal solutions, thus an appropriate technique for high-

dimensional feature selection. SSA is a recently developed swarm optimization algorithm proposed by Mirjalili et al. [22]. The swarming behaviour of salps in a marine environment, as a salp chain, which helps for foraging and movement, is simulated in SSA [53]. The salp swarm is initially divided into two groups, consisting of a leader and the rest as followers. The leader salp acts as the front salp of the salp chain while the followers represent the other salps in the population. The SSA population is initialized randomly in a similar fashion to many of the other swarm-based algorithms. Further, the positions of each salp are determined in an n dimensional feature space, where n denotes the number of variables in a given problem. Besides, the positions of the leader and the followers are updated according to (2) and (4), respectively [22]

$$x_j^1 = \begin{cases} F_j + c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) c_3 \geq 0.5 \\ F_j - c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) c_3 < 0.5 \end{cases} \quad (2)$$

where x_j^1 is the position of the leader salp in jth dimension, F_j is the food source, c_2 and c_3 are random numbers within the range [0,1], and lb_j and ub_j are the lower and upper bounds, respectively. In addition, the coefficient c_1 to balance between exploration and exploitation is calculated using (3) [22]

$$c_1 = 2e^{-\left(\frac{4t}{t_{\max}}\right)^2}, \quad (3)$$

where t and t_{max} represent the current and maximum iterations, respectively.

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right). \quad (4)$$

where x_j^i is the position of follower i ($i \geq 2$) in dimension j. The basic steps of SSA are given in Algorithm 1.

Various applications have utilized SSA. For instance, SSA has been used for sentiment analysis [54], crude oil price forecasting [55], wind power prediction [56], unrelated parallel machine scheduling [57], digital mammogram classification [58], [59], and for cancer classification [14], [60]. The conventional SSA has been adapted in many ways to produce SSA variants [23], [24], [32], [61] that address its setbacks. For example, the binary SSA [62], multi-leader binary SSA with sub-chains [34], time-varying binary SSA with dynamically changing leaders and followers [23], and multi-objective binary SSA with dynamic time-varying strategy [32] are few versions of SSA proposed in existing studies for feature selection. Similarly, this study also introduces a new variant of SSA, but it is for gene selection in cancer classification.

Algorithm 1: *Salp Swarm Algorithm*

Step 1: Define population size n, maximum number of iteration *maxGeneration*
Step 2: Generate the initial population of salps randomly: x_i, $i=1,2,3,\ldots,n$
Step 3: while ($t <$ *maxGeneration*)
Step 4: Evaluate the fitness of each salp x_i using the fitness function: $f(x)$
Step 5: Determine the best salp and save as food source F.
Step 6: Update c_1 using (3).
Step 7: for $i=1$ to n
Step 8: if (i==1)
Step 9: Update the position of leader using (2).
Step 10: else
Step 11: Update the position of follower using (4).
Step 12: end if
Step 13: end for i
Step 14: Reposition the salps which go out of bounds.
Step 15: end while
Step 16: Return the best solution F.

D. Feature Selection

Feature selection determines a set of relevant features from a huge feature space. Similarly, gene or biomarker selection produces a set of informative biomarkers out of a large gene expression profile, such as a DNA microarray. DNA microarray datasets contain a massive collection of genes: relevant, irrelevant, and redundant genes. Besides, DNA microarrays have emerged into the current cancer research field with amazing benefits such as classification, early detection, fast and non-invasive cancer diagnosis, and prognosis [63]. Thus, even though the process of biomarker selection is not a trivial task, it has become a popular field of research. Typically, there are three types of gene selection methods: filter-based, wrapper-based, and hybrid methods that are later described along with corresponding cancer classification studies.

Typically, filter-based gene selection methods [5], [64], [65] produce low classification accuracy and large subset size, as filters are popular for preprocessing rather than gene selection, thus rarely utilized alone for gene selection. Existing studies have evaluated various filters, such as the mRMR filter [65], Naïve Bayes (NB) [65], CFS filter [5], and mutual information filter [64], for gene selection. However, the produced results show that these filter-based approaches still require improvement. For instance, Ghosh et al. [64] evaluated the performance of ten different filter methods with three classifiers, using ten datasets, where the mutual information filter outperformed the others, yielding 100% accuracy on three out of ten datasets. Similarly, the wrapper-based gene selection algorithms also suffer the same drawbacks: low classification accuracy and a large gene subset [7], [8], [66]. For instance, the wrapper approach suggested by Al-Baity and Al-Mutlaq [66] did not produce 100% accuracy on any datasets and thus, the researchers [66] proposed hybrid method for future analysis. Overall, the single methods, filter-based methods and wrapper-based methods, can be substituted with hybrid methods that aim for significant classification performance.

Swarm-based hybrid algorithms are a good option for optimization and thus for gene selection. Existing swarm-based hybrid approaches [13]-[15], [67] have shown superior performance in gene selection compared to other methods. Specifically, the ABC algorithm [42], BA [37], FA [12], [15], [31], [33], Horse herd Optimisation Algorithm (HOA) [13], and SSA [14], [67] have been utilized in various studies for gene selection. For instance, an improved binary SSA was suggested by Qin et al. [14] in which it produces 100% accuracy on seven cancer datasets out of ten. Further, an improved multilayer binary FA was suggested by Xie et al. [15]. Nevertheless, only one dataset obtained 100% accuracy. Besides, a binary HOA was proposed by Mehrabi et al. [13], which provided 100% accuracy on six datasets out of ten. On the other hand, Panda et al. [67] presented a hybrid approach with SSA; unfortunately, none of the datasets obtained 100% accuracy. Nonetheless, existing results reported in the literature [13]-[15], [33], [37], [67] reflect the potential for growth in classification task.

III. Methodology

The proposed methodology is composed of three phases: data preprocessing, gene selection, and classification, as in Fig. 1. The proposed CFS-MDCSSA-SVM, combines CFS with MDCSSA and then classifies using SVM.

A. Data Preprocessing

Prior to moving the standard microarray datasets into the gene selection phase, the data preprocessing is performed through normalization and filtering. The original value v of each feature is normalized into v', which is in between 0 and 1, obtained by (5), using min-max normalization [68]. Then, the normalized datasets are filtered by the CFS filter [35] to remove the unnecessary features (i.e. irrelevant and redundant features) and reduce the size as given in Table 1 where the gene count is denoted within the parentheses.

B. Gene Selection

The normalized CFS-filtered datasets are further processed through the gene selection phase to produce the biomarkers for classification. The proposed MDCSSA is utilized for biomarker selection which is achieved through four steps: dynamic size salp population initialization, multi-leader identification, composite position update, and population reinitialization as described below.

Dynamic Size Salp Population Initialization

The dynamic size salp population in MDCSSA varies from the standard population generation in SSA as, MDCSSA changes the solution size rather than specifying a threshold for fixing the solution size, similar to the strategy used in standard SSA. The salp population is initialized with dynamic size (i.e. non-fixed size) salps to resolve the slow convergence issue in the standard SSA. Specifically, it is believed that the slow convergence issue is worsened upon fixing the solution size to a large threshold, which may often be defined in a high-dimensional feature space. Hence, the proposed MDCSSA generates the dynamic size salps that would consist of s number of genes where $s=1,2,3,\ldots,D$ in a *D-dimensional* space. A sample dynamic size salp population with n number of salps could be illustrated as in Fig. 2.

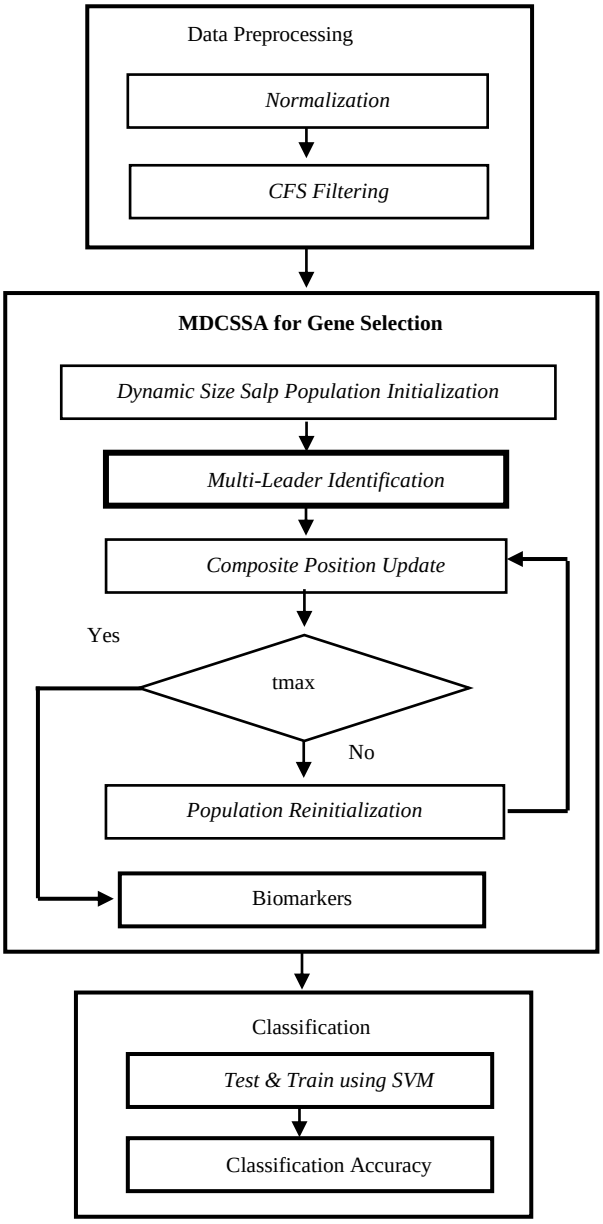

Fig. 1. Flowchart of CFS-MDCSSA-SVM

TABLE I
Normalized CFS-filtered Microarray Datasets

Dataset (No. of genes)	Filtered Dataset (No. of genes)
Colon (2000)	Colon (26)
Leukemia2 (7129)	Leukemia2 (81)
Leukemia3 (7129)	Leukemia3 (104)
MLL (12582)	MLL (149)
Leukemia4 (7129)	Leukemia4 (119)
SRBCT (2308)	SRBCT (111)

$$v' = \frac{v - \min(v)}{\max(v) - \min(v)} \tag{5}$$

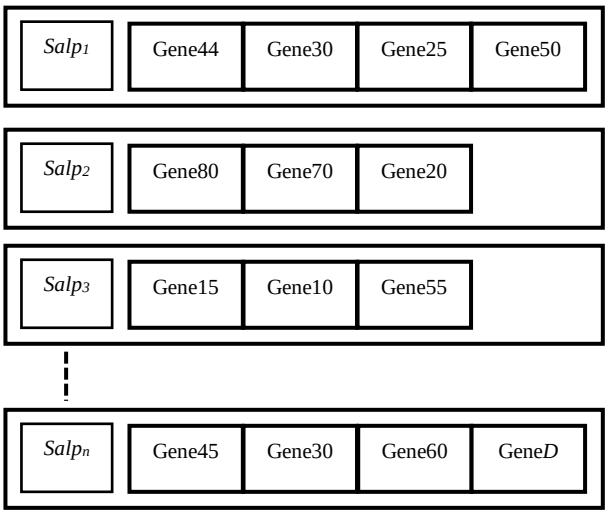

Fig. 2. Dynamic size salp population initialization in MDCSSA

Multi-leader Identification

In contrast to the standard SSA [22], which specifies a single leader, the proposed MDCSSA defines two leaders termed dual leaders, who are followed by the follower salps. The dual leader concept is embedded in the proposed algorithm to balance exploration and exploitation. In other words, the leaders are preserved and exploited during the population reinitialization to balance exploration and exploitation.

Composite Position Update

A composite position update function is suggested for the dynamic size salps in the proposed MDCSSA because the position update in the standard SSA [22] is not applicable for non-fixed size solutions, and at the same time, it concerns the initial positions of the salps, hence, inapplicable for gene selection. Therefore, the proposed algorithm employs a composite position update with two functions: an integrative and a discriminative position update, considering the gene interconnections. Especially, the integrative position update function represented in (6) is designed for leaders. In contrast, the discriminative position update is employed on both the leaders (as in (7)) and the follower salps (as represented in (8)).

Position Update in Leaders

The integrative function combines the genes in both leaders (i.e. $x_{leader\_1}(t)$ and $x_{leader\_2}(t)$) to create a new leader salp (i.e. $x_{lnew\_1}(t+1)$) whereas the discriminative function (refer (7)) extracts the discriminant genes in the second leader (i.e. $x_{leader\_2}(t)$) to create a new leader salp (i.e. $x_{lnew\_2}(t+1)$).

$$x_{\ln ew\_1}(t+1) = x_{leader\_1}(t) + x_{leader\_2}(t) - \left(x_{leader\_1}(t) \cap x_{leader\_2}(t)\right) \quad (6)$$

$$x_{\ln ew\_2}(t+1) = x_{leader\_2}(t) \cap \left(x_{leader\_1}(t)\right)^c \quad (7)$$

where $x_{lnew\_1}(t+1)$ and $x_{lnew\_2}(t+1)$ are the new positions of the leaders according to integrative and discriminative functions, respectively in iteration $(t+1)$ while $x_{leader\_1}(t)$ and $x_{leader\_2}(t)$ are the positions of the first and the second leaders in iteration (t), respectively in (6) and (7).

Further, during the position update, priority is given to the first leader than the second leader. More precisely, if the fitness of the new leader salp (i.e. $x_{lnew\_1}(t+1)$), produced from the integrative position update is higher than that of the discriminative position update (i.e. $x_{lnew\_2}(t+1)$), then, the particular new leader salp is utilized for the position update in the first leader. In contrast, the second-fit new leader salp is used for the position update in the second leader. Nevertheless, it is noteworthy that the current positions of leaders would be updated if and only if the new positions enhance the fitness of the leaders; otherwise, the current positions will remain the same as before. Besides, the discriminant genes in the second leader is utilized in the discriminative function since the first leader salp is more likely to be selected for the next population than the second leader salp, thus, would be exploited more than the second leader.

Position Update in Followers

The discriminative function, as represented in (8), extracts the discriminant genes in the follower salp compared to the leaders. The discriminative function reduces the gene subset size more, compared to the integrative function, and is thus employed for position update in the followers. Together with classification accuracy, the gene subset size is also important for performance evaluation.

$$x_{fnew}(t+1) = x_{follower}(t) \cap \left(x_{leaders}(t)\right)^c \quad (8)$$

where $x_{fnew}(t+1)$ is the new position of the follower salp in iteration $(t+1)$ while $x_{follower}(t)$ and $x_{leaders}(t)$ are the positions of the follower salp and the leader salps in iteration (t), respectively in (8).

Population Reinitialization

Generally, a swarm-based optimization algorithm, including the standard SSA, undergoes several iterations on the same population, regardless of the fact that this would reduce the population diversity, result in a local optimum. To overcome this issue, the proposed MDCSSA employs the population reinitialization method which initializes the population for the next iteration while conserving the leaders. The proposed algorithm intends to improve the population diversity through reinitializing the followers. At the same time, the proposed MDCSSA also intensifies the leaders to maintain a balance between exploration and exploitation.

C. Classification

The proposed MDCSSA uses the SVM classifier to train and test (70:30) the samples. The algorithm's performance is assessed by its classification accuracy, as represented in (9). The steps in the proposed MDCSSA are also presented in Algorithm 2.

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

where, *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

Algorithm 2:
Multi-leader Dynamic Composite Salp Swarm Algorithm

Step 1: Define parameters - population size *n*, dimension *D*, and maximum iteration: *maxGeneration*
Step 2: Randomly generate salp population with dynamic size solutions: x_i, i=1,2,3,…,n;
Step 3: Evaluate the fitness of each salp using fitness function: *f(x)*
Step 4: Determine the best two salps and save as the dual leaders.
Step 5: while (*t* < *maxGeneration*)
Step 6: for *i=1* to *n*
Step 7: if *(i==1 || i==2)*
Step 8: Calculate the new position_1 x_{lnew_1} using (6)
Step 9: Calculate the new position_2 x_{lnew_2} using (7)
Step 10: Calculate the fitness of x_{lnew_1} using (9)
Step 11: Calculate the fitness of x_{lnew_2} using (9)
Step 12: if fitness of x_{lnew_1} is greater than the fitness of x_{lnew_2}
Step 13: if fitness of x_{lnew_1} is greater than the fitness of Leader_1
Step 14: Update the position of Leader_1 using (6)
Step 15: Update the fitness of Leader_1 using (9)
Step 16: if fitness of x_{lnew_2} is greater than the fitness of Leader_2
Step 17: Update the position of Leader_2 using (7)
Step 18: Update the fitness of Leader_2 using (9)
Step 19: end if
Step 20: end if
Step 21: else
Step 22: if fitness of x_{lnew_2} is greater than the fitness of Leader_1
Step 23: Update the position of Leader_1 using (7)
Step 24: Update the fitness of Leader_1 using (9)
Step 25: if fitness of x_{lnew_1} is greater than the fitness of Leader_2
Step 26: Update the position of Leader_2 using (6)
Step 27: Update the fitness of Leader_2 using (9)
Step 28: end if
Step 29: end if
Step 30: end if
Step 31: else
Step 32: Calculate the new position x_{fnew} of the follower using (8)
Step 33: Update the position of follower $x_{follower}$ using (8)
Step 34: Update the fitness of follower using (9)
Step 35: end if
Step 36: end for *i*
Step 37: Determine the new leaders and return Leader_1
Step 38: Reinitialize the salp population
Step 39: end while

IV. Experimental Results and Discussion

This section presents the experimental results produced by the proposed algorithm and subsequently a comparative discussion on the performance of the proposed CFS-MDCSSA-SVM. The execution was conducted using WEKA and MATLAB software on a PC with an Intel Core i3 processor, 4.00 GB RAM, and a Windows 10 operating system.

A. Experimental Results

A brief description of the microarray datasets used in this study, parameter setting applied to the proposed algorithm, and the results produced on the different cancer datasets are described in this section.

Dataset Description

Six publicly available cancer microarray datasets were applied in this study. Among the six datasets, two datasets: Colon [69] and Leukemia2 [70] are of binary class while the rest: Leukemia3 [71], MLL [72], Leukemia4 [72], and Small Round Blue Cell Tumor (SRBCT) [73] are of multiclass datasets. The number of classes, number of genes, and number of samples along with a description of the datasets are given in Table 2.

TABLE II
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
MLL	3	12582	72	ALL: 24, MLL: 20, and AML: 28
Leukemia4	4	7129	72	BM: 21, PB: 4, B-cell: 38, and T-cell: 9
SRBCT	4	2308	83	EWS: 29, BL: 11, NB: 18, and RMS: 25

Note - ALL: Acute Lymphoblastic Leukemia, AML: Acute Myeloid Leukemia, BL: Burkitt's Lymphoma, BM: Bone Marrow, EWS: Ewing's Sarcoma, MLL: Mixed Lineage Leukemia, NB: Neuroblastoma, PB: Peripheral Blood, and RMS: Rhabdomyosarcoma

Parameter Settings

The parameters of an algorithm should be practically employable and as well as optimal. Hence, the proposed CFS-MDCSSA-SVM uses the parameter settings as given in Table 3. The salp population consists of 80 salps with dynamic size where the maximum size would be equal to the feature dimension. Further, there are 100 iterations in a given execution, and the algorithm was executed for 30 independent runs to validate the proposed algorithm's robustness and performance.

TABLE III
Parameter Settings for CFS-MDCSSA-SVM

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

Results

The performance of CFS-MDCSSA-SVM was evaluated for gene selection on six high-dimensional cancer microarray datasets. This study considered the classification accuracy and the gene subset size as evaluation metrics. The best, average, and worst results produced by the proposed algorithm are tabulated in Table 4. Further, the informative genes generated by CFS-MDCSSA-SVM are presented in Table 5. The values mentioned in the parentheses in Table 4-6 denote the number of genes in each dataset.

The proposed CFS-MDCSSA-SVM has provided significant results as shown in Tables 4 and 5. Explicitly, all six cancer datasets were classified ideally with 100% accuracy with very small gene subsets. The Leukemia2

dataset was classified with one gene while Leukemia4 and SRBCT were classified with four and five genes, respectively. Further, gene subsets produced for Colon, Leukemia3, and MLL were equal in size of three. Gene selection aims to produce a few biomarkers that would exactly classify the cancer samples, just as achieved in the proposed algorithm. Hence, it is comprehensible that CFS-MDCSSA-SVM is robust and efficient for biomarker selection. Furthermore, the comparative discussion presented in the next section would clarify the significance and the contribution of the results produced by the proposed algorithm for gene selection.

TABLE IV
CLASSIFICATION PERFORMANCE OF CFS-MDCSSA-SVM

Dataset	Run	Accuracy (%)		
		Best	Average	Worst
Colon (2000)	5	**100(3)**	**99(8)**	94.74(4)
	10	100(3)	97(7)	89.47(8)
	15	100(3)	97(7)	89.47(8)
	20	100(3)	97(7)	89.47(8)
	25	100(3)	96(7)	84.21(6)
	30	100(3)	96(8)	84.21(6)
Leukemia2 (7129)	5	100(2)	**100(3)**	100(4)
	10	100(2)	100(3)	100(4)
	15	100(2)	100(3)	100(4)
	20	100(2)	100(3)	100(4)
	25	**100(1)**	100(3)	100(4)
	30	100(1)	100(3)	100(5)
Leukemia3 (7129)	5	100(5)	**100(7)**	100(10)
	10	100(5)	100(8)	100(12)
	15	100(5)	100(8)	100(12)
	20	100(5)	100(8)	100(12)
	25	**100(3)**	100(8)	100(12)
	30	100(3)	100(8)	100(12)
MLL (12582)	5	100(5)	**100(5)**	100(6)
	10	100(4)	100(5)	100(9)
	15	**100(3)**	100(5)	100(9)
	20	100(3)	100(5)	100(9)
	25	100(3)	100(6)	100(9)
	30	100(3)	100(6)	100(9)
Leukemia4 (7129)	5	100(16)	100(20)	100(22)
	10	100(14)	100(19)	100(22)
	15	100(14)	100(19)	100(22)
	20	**100(4)**	**100(18)**	100(22)
	25	100(4)	100(18)	100(22)
	30	100(4)	100(18)	100(22)
SRBCT (2308)	5	100(6)	100(9)	100(10)
	10	100(6)	**100(8)**	100(10)
	15	**100(5)**	100(8)	100(10)
	20	100(5)	100(8)	100(11)
	25	100(5)	100(9)	100(11)
	30	100(5)	100(9)	100(11)

TABLE V
BIOMARKERS GENERATED BY CFS-MDCSSA-SVM

Dataset	Genes
Colon (3)	A377, A1042, A1423 OR A682, A765, A1560
Leukemia2 (1)	attribute3252
Leukemia3 (3)	D88270_at, X60992_at, Z49194_at
MLL (3)	x38097_at, x40191_s_at, x480_at
Leukemia4 (4)	M13792_at, M89957_at, X61587_at, U90546_at
SRBCT (5)	gene251, gene742, gene774, gene1327, gene1924

B. Discussion

Surprisingly, the proposed CFS-MDCSSA-SVM has classified all six datasets perfectly with 100% accuracy using small gene subsets, as presented in Tables 4 and 5. Table 6 compares the results produced in this study with those of recent related studies. Further, a detailed discussion is also provided in this section. Besides, the comparative evaluation is based on the classification accuracy and the gene subset size as is popular in existing work.

Regarding the Colon cancer classification, the proposed algorithm has produced 100% accuracy with a gene subset consisting of three genes while the same accuracy has been reported by Fajila and Yusof [31] and Fajila and Yusof [12] using five and one gene, respectively. Other related studies [13]-[15], [17], [33], [37], [42], [45], [74]-[76] compared in Table 6 have provided low classification accuracy which is below 100%. Hence, the perfect result for the Colon cancer dataset was the one presented in Fajila and Yusof [12] compared with other studies.

In concern to the Leukemia2 dataset, all the algorithms in Table 6, except Mazumder and Veilumuthu [17], have offered 100% accuracy. However, the ideal output: 100% accuracy with a single gene, has been yielded in the proposed algorithm and as well as by Fajila and Yusof [12], [31] while the other existing works have produced gene subsets with more than one gene.

Similarly, all the algorithms except Mazumder and Veilumuthu [17], Almugren and Alshamlan [33], and Panda et al. [67] have provided 100% accuracy on Leukemia3 dataset. However, the best result, as highlighted in Table 6, has been reported by Fajila and Yusof [12], [31]. Meanwhile, the proposed algorithm provided 100% accuracy with a gene subset consisting of three genes, which is slightly bigger with only one more gene compared to the best result. Besides, the gene subset sizes reported in other existing studies [14], [37], [42], [45], [74]-[76] are bigger than that of the proposed approach.

Regarding the classification of MLL, all the algorithms except Panda et al. [67] have provided 100% accuracy. Nevertheless, the best result: 100% accuracy with three genes, is achieved by the proposed algorithm. In contrast, the gene subset size in other related studies is bigger than that of CFS-MDCSSA-SVM.

Likewise, all the algorithms except Jain et al. [45] have provided 100% accuracy on Leukemia4 dataset, but the proposed algorithm produces a smaller gene subset. Specifically, the proposed algorithm has produced 100% accuracy with only four genes highlighting the best result for Leukemia4 classification compared to the existing works [14], [17], [37], [74].

Furthermore, all the algorithms except Panda et al. [67] have produced 100% accuracy on SRBCT dataset. Nevertheless, the best result: 100% accuracy with four genes, is presented in Fajila and Yusof [12] and Alshamlan [42]. Besides, the proposed algorithm provided 100% accuracy with a gene subset consisting of five genes, which is larger than just a single gene compared to the best result.

TABLE VI
PERFORMANCE COMPARISON BETWEEN CFS-MDCSSA-SVM AND RELATED WORKS

Algorithms	Colon	Leukemia2	Leukemia3	MLL	Leukemia4	SRBCT
CFS-MDCSSA-SVM	100(3)	**100(1)**	100(3)	**100(3)**	**100(4)**	100(5)
Fajila and Yusof [12]	**100(1)**	**100(1)**	**100(2)**	-	-	**100(4)**
Panda et al. [67]	-	-	96.7(57)	95.1(194)	-	97.1(23)
Mehrabi et al. [13]	98.48(7.36)	100(2.6)	-	100(4.1)	-	100(5.53)
Xie et al. [15]	90.48(9.4)	100(4.3)	-	-	-	-
Fajila and Yusof [31]	100(5)	**100(1)**	**100(2)**	-	-	100(7)
Qin et al. [14]	97.6(148)	100(241)	100(182)	100(489)	100(336)	100(104)
Fajila and Yusof [74]	95.23(4)	-	100(4)	100(4)	100(5)	100(8)
Al-Betar et al. [37]	97.85(12.27)	100(4.07)	100(5.33)	100(8)	100(6.73)	100(9.13)
Almugren and Alshamlan [33]	94.3(15)	100(5)	97.8(10)	-	-	100(8)
Alshamlan [42]	96.77(9)	100(3)	100(6)	-	-	**100(4)**
Mazumder and Veilumuthu [17]	-	98.61(3)	98.61(3)	100(6)	100(7)	100(6)
Jain et al. [45]	94.89(4.2)	100(4.3)	100(6)	100(30.8)	97.63(20.7)	100(34.1)
Alshamlan et al. [75]	96.77(15)	100(14)	100(20)	-	-	100(10)
Alshamlan et al. [76]	98.38(10)	100(4)	100(8)	-	-	100(6)

Accordingly, the results highlight the proposed algorithm's competency compared to the existing related studies. The overall achievement of CFS-MDCSSA-SVM is better than the existing competitor studies on all the datasets despite the fact that Fajila and Yusof [12] also have shown identical achievement. Moreover, the significant steps in the proposed algorithm would have facilitated the production of the optimal solutions and as well as enhanced the performance through resolving the issues in the standard SSA.

V. CONCLUSION

A hybrid gene selection algorithm namely CFS-MDCSSA-SVM is presented in this study for high dimensional gene selection in cancer classification. The proposed algorithm applies a CFS filter-based preprocessing and a new variant of SSA: MDCSSA for gene selection. The SVM classifier is used for evaluation. The algorithm's convergence is enhanced through the dynamic size solutions while the population reinitialization method avoids the local optima issue. Further, the balance between the exploration and exploitation is maintained using the multi-leader concept and the composite position update functions. Especially, the best solutions (i.e. dual leaders) are exploited by the composite functions, and also preserved during the reinitialization, to be exploited further in the next generation. This strategy enhances the exploitation property together with the exploration capabilities come from the reinitialization; thus, maintaining the balance between exploration and exploitation.

The contribution of the significant steps suggested in the proposed algorithm is strengthened via the superiority of the experimental results, which show that 100% accuracy is given just by a few biomarker genes for all the six cancer datasets. In addition, comparing the results with that of the existing related algorithms emphasizes the substantial performance of the proposed algorithm. Furthermore, as a theoretical contribution, a hybrid swarm-based algorithm CFS-MDCSSA was developed in this study. At the same time, the findings would be beneficial practically in real-world clinical applications such as cancer classification, prognosis, diagnosis, and therapy. Even though there are many contributions from this study, there are few limitations as well. Firstly, the proposed algorithm was evaluated on microarray data, thus, it needs to be tested on other data in future, to validate the effectiveness and robustness of the proposed algorithm in feature selection. Further, the algorithm was evaluated using the percentage splitting technique; hence, other techniques such as cross-validation are required to be applied to ensure the generalizability of the approach. Besides, it is also intended to explore and embed the properties of other swarm-based algorithms to enhance the potentials of the proposed algorithm.

REFERENCES

[1] World Health Organization. (2024, February 1). *Global cancer burden growing, amidst mounting need for services*. Available: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

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

[3] T. O. Tobore, "On the need for the development of a cancer early detection, diagnostic, prognosis, and treatment response system," *Future Sci. OA.*, vol. 6, no. 2, pp. FSO439, 2019. https://doi.org/10.2144/fsoa-2019-0028

[4] C. M. Lai, W. C. Yeh, and C. Y. Chang, "Gene selection using information gain and improved simplified swarm optimization," *Neurocomputing*, vol. 218, pp. 331-338, 2016. https://doi.org/10.1016/j.neucom.2016.08.089

[5] M. Al-Batah, B. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers," *Int. J. Online Biomed. Eng.*, vol. 15, no. 8, pp. 62-73, 2019. https://doi.org/10.3991/ijoe.v15i08.10617

[6] D. H. Mazumder and R. Veilumuthu, "An enhanced feature selection filter for classification of microarray cancer data," *ETRI J.*, vol. 41, no. 3, pp. 358-370, 2019. https://doi.org/10.4218/etrij.2018-0522

[7] H. AlMazrua and H. AlShamlan, "A new algorithm for cancer biomarker gene detection using Harris Hawks optimization," *Sensors*, vol. 22, no. 19, pp. 7273, 2022.

[8] M. Alzaqebah, K. Briki, N. Alrefai, S. Brini, S. Jawarneh, M. K. Alsmadi, R. M. A. Mohammad, I. ALmarashdeh, F. A. Alghamdi, N. Aldhafferi, and A. Alqahtani, "Memory based cuckoo search algorithm for feature selection of gene expression dataset," *Inform. Med. Unlocked.*, vol. 24, pp. 100572, 2021. https://doi.org/10.1016/j.imu.2021.100572

[9] B. H. Shekar and G. Dagnew, "L1-regulated feature selection and classification of microarray cancer data using deep learning," in *Proc. of 3rd Int. Conf. on Computer Vision and Image Processing*, Springer, 2020, pp. 227-242. https://doi.org/10.1007/978-981-32-9291-8_19

[10] R. Dash, "An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 2, pp. 195-207, 2021. https://doi.org/10.1016/j.jksuci.2018.02.013

[11] V. Elyasigomari, D. A. Lee, H. R. C. Screen, and M. H. Shaheed, "Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification," *J. Biomed. Inform.*, vol. 67, pp. 11–20, 2017. https://doi.org/10.1016/j.jbi.2017.01.016

[12] F. Fajila and Y. Yusof, "Mutable Composite Firefly Algorithm for Microarray-Based Cancer Classification," *J. Inf. Commun. Technol.*, vol. 24, no. 1, pp. 105-130, 2025. https://doi.org/10.32890/jict2025.24.1.5

[13] N. Mehrabi, S. P. Haeri Boroujeni, and E. Pashaei, "An efficient high-dimensional gene selection approach based on the Binary Horse Herd Optimization Algorithm for biologicaldata classification," *Iran J. Comput. Sci*., pp. 1-31, 2024. https://doi.org/10.1007/s42044-024-00174-z

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

[15] W. Xie, L. Wang, K. Yu, T. Shi, and W. Li, "Improved multi-layer binary firefly algorithm for optimizing feature selection and classification of microarray data," *Biomed. Signal Process. Control*., vol. 79, pp. 104080, 2023. https://doi.org/10.1016/j.bspc.2022.104080

[16] L. Gao, M. Ye, X. Lu, and D. Huang, "Hybrid method based on information gain and support vector machine for gene selection in cancer classification," *Genom. Proteom. Bioinform*., vol. 15, no. 6, pp. 389-395, 2017. https://doi.org/10.1016/j.gpb.2017.08.002

[17] D. H. Mazumder and R. Veilumuthu, "Cancer Classification with a Novel Hybrid Feature Selection Technique," *Int. J. Simulat. Syst. Sci. Tech*., vol. 19, no. 2, 2018. https://doi.org/10.5013/IJSSST.a.19.02.07

[18] M. Dorigo, M., Birattari, and T. Stutzle, "Ant colony optimization," *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28-39, 2006.

[19] S. Mirjalili, "The ant lion optimizer," *Adv. Eng. Softw*., vol. 83, pp. 80-98, 2015. https://doi.org/10.1016/j.advengsoft.2015.01.010

[20] D. Karaboga, "An Idea Based on Honey Bee Swarm for Numerical Optimization," vol. 200, Technical report-tr06, Erciyes University, 2005, pp. 1-10.

[21] X. S. Yang, *Nature-inspired Metaheuristic Algorithms*. Luniver press, 2010.

[22] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Adv. Eng. Softw*., vol. 114, pp. 163-191, 2017. https://doi.org/10.1016/j.advengsoft.2017.07.002

[23] H. Faris, A. A. Heidari, A. Z. Ala'M, M. Mafarja, I. Aljarah, M. Eshtay, and S. Mirjalili, "Time-varying hierarchical chains of salps with random weight networks for feature selection," *Expert Syst. Appl*., vol. 140, pp. 112898, 2020. https://doi.org/10.1016/j.eswa.2019.112898

[24] A. E. Hegazy, M. A. Makhlouf, and G. S. El-Tawel, "Feature selection using chaotic salp swarm algorithm for data classification," *Arab. J. Sci. Eng*., vol. 44, no. 4, pp. 3801-3816, 2019. https://doi.org/10.1007/s13369-018-3680-6

[25] S. Cheng, Y. Shi, Q. Qin, T. O. Ting, and R. Bai, "Maintaining population diversity in brain storm optimization algorithm," in *2014 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2014, pp. 3230-3237. https://doi.org/10.1109/CEC.2014.6900255

[26] S. Cheng, Y. Shi, Q. Qin, Q. Zhang, and R. Bai, "Population diversity maintenance in brain storm optimization algorithm," *J. Artif. Intell. Soft Comput. Res*., vol. 4, no. 2, pp. 83-97, 2014. https://doi.org/10.1515/jaiscr-2015-0001

[27] R. Salgotra, U. Singh, and S. Saha, "On some improved versions of whale optimization algorithm," *Arab. J. Sci. Eng*., vol. 44, no. 11, pp. 9653-9691, 2019. https://doi.org/10.1007/s13369-019-04016-0

[28] S. Mostafa Bozorgi and S. Yazdani, "IWOA: An improved whale optimization algorithm for optimization problems," *J. Comput. Des. Eng*., vol. 6, no. 3, pp. 243-259, 2019. https://doi.org/10.1016/j.jcde.2019.02.002

[29] I. Sekaj and M. Oravec, "Selected population characteristics of fine-grained parallel genetic algorithms with re-initialization," in *Proc. of the First ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 2009 pp. 945-948. https://doi.org/10.1145/1543834.1543980

[30] M. El-Abd, "Brain storm optimization algorithm with re-initialized ideas and adaptive step size," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2016, pp. 2682-2686. https://doi.org/10.1109/CEC.2016.7744125

[31] M. N. F. Fajila and Y. Yusof, "Hybrid gene selection with mutable firefly algorithm for feature selection in cancer classification," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 3, pp. 24-35, 2022. https://doi.org/10.22266/ijies2022.0630.03

[32] I. Aljarah, M. Habib, H. Faris, N. Al-Madi, A. A. Heidari, M. Mafarja, M. A. Elaziz, and S. Mirjalili, "A dynamic locality multi-objective salp swarm algorithm for feature selection," *Comput. Ind. Eng*., vol. 147, pp. 106628, 2020. https://doi.org/10.1016/j.cie.2020.106628

[33] N. Almugren and H. M. Alshamlan, "New bio-marker gene discovery algorithms for cancer gene expression profile," *IEEEAccess*, vol. 7, pp. 136907-136913, 2019. https://doi.org/10.1109/ACCESS.2019.2942413

[34] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris, Y. Zhang, and S. Mirjalili, "Asynchronous accelerating multi-leader salp chains for feature selection," *Appl. Soft Comput.*, vol. 71, pp. 964-979, 2018. https://doi.org/10.1016/j.asoc.2018.07.040

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

[36] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Advances in Neural Information Processing Systems 9*, MIT Press, 1997, pp. 281-287.

[37] M. A. Al-Betar, O. A. Alomari, and S. M. A. Abu-Romman, "TRIZ-inspired bat algorithm for gene selection in cancer classification," *Genomics*, vol. 112, no. 1, pp. 114-126, 2020. https://doi.org/10.1016/j.ygeno.2019.09.015

[38] S. Wright, "The interpretation of population structure by F-statistics with special regard to systems of mating," *Evolution*, vol. 19, no. 3, pp. 395-420, 1965. https://doi.org/10.2307/2406450

[39] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 27, no. 8, pp. 1226-1238, 2005. https://doi.org/10.1109/TPAMI.2005.159

[40] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014.

[41] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. United States, 1984.

[42] H. M. Alshamlan, "Co-ABC: Correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile," *Saudi J. Biol. Sci.*, vol. 25, no. 5, pp. 895-903, 2018. https://doi.org/10.1016/j.sjbs.2017.12.012

[43] K. R. Pushpalatha and A. G. Karegowda, "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)*, IEEE, 2017, pp. 1-6. https://doi.org/10.1109/ICECIT.2017.8453403

[44] B. Remeseiro, V. Bolón-Canedo, A. Alonso-Betanzos, and M. G. Penedo, "Learning features on tear film lipid layer classification," in *Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2015, pp. 195-200.

[45] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Appl. Soft Comput*., vol. 62, pp. 203-215, 2018. https://doi.org/10.1016/j.asoc.2017.09.038

[46] S. Nagashree and B. S. Mahanand, "Pneumonia chest X-ray classification using support vector machine," in *Proc. of Int. Conf. on Data Science and Applications: ICDSA 2022*, Springer, 2023, pp. 417-425. https://doi.org/10.1007/978-981-19-6634-7_29

[47] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Appl. Intell*., vol. 52, no. 3, pp. 2411-2422, 2022. https://doi.org/10.1007/s10489-021-02533-w

[48] H. Alshamlan, S. Omar, R. Aljurayyad, and R. Alabduljabbar, "Identifying effective feature selection methods for Alzheimer's disease biomarker gene detection using machine learning," *Diagn*., vol. 13, no. 10, pp. 1771, 2023. https://doi.org/10.3390/diagnostics13101771

[49] H. A. A. Mohammed, A. A. K. Jizany, I. M. Mahmood, and Q. K. Kadhim, "Predicting Alzheimer's Disease Using a Modified Grey Wolf Optimizer and Support Vector Machine," *Ing. Syst. Inf*., vol. 29, no. 2, pp. 669, 2024. https://doi.org/10.18280/isi.290228

[50] H. AlShamlan and H. AlMazrua, "Enhancing Cancer Classification through a Hybrid Bio-inspired Evolutionary Algorithm for Biomarker Gene Selection," *Comput. Mater. Contin.*, vol. 79, no. 1, 2024.

[51] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare

applications: a review," *Information*, vol. 15, no. 4, pp. 235, 2024. https://doi.org/10.3390/info15040235

[52] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," *Mach. Learn.*, vol. 62, no. 03, pp. 233-244, 2020.

[53] R. A. Ibrahim, A. A. Ewees, D. Oliva, M. Abd Elaziz, and S. Lu, "Improved salp swarm algorithm based on particle swarm optimization for feature selection," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 8, pp. 3155-3169, 2019. https://doi.org/10.1007/s12652-018-1031-9

[54] A. Alzaqebah, B. Smadi, and B. H. Hammo, "Arabic sentiment analysis based on salp swarm algorithm with s-shaped transfer functions," in *2020 11th Int. Conf. on Information and Communication Systems (ICICS)*, IEEE, 2020, pp. 179-184. https://doi.org/10.1109/ICICS49469.2020.239507

[55] M. Abd Elaziz, A. A. Ewees, and Z. Alameer, "Improving adaptive neuro-fuzzy inference system based on a modified salp swarm algorithm using genetic algorithm to forecast crude oil price," *Nat. Resour. Res.*, vol. 29, pp. 2671-2686, 2020.

[56] J. S. Pan, J. Shan, S. G. Zheng, S. C. Chu, and C. K. Chang, "Wind power prediction based on neural network with optimization of adaptive multi-group salp swarm algorithm," *Clust. Comput.*, pp. 1-16, 2021.

[57] A. A. Ewees, M. A. Al-qaness, and M. Abd Elaziz, "Enhanced salp swarm algorithm based on firefly algorithm for unrelated parallel machine scheduling with setup times," *Appl. Math. Model.*, vol. 94, pp. 285-305, 2021. https://doi.org/10.1016/j.apm.2021.01.017

[58] F. Mohanty, S. Rup, B. Dash, B. Majhi, and M. N. S. Swamy, "An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine," *Appl. Soft Comput.*, vol. 91, pp. 106266, 2020. https://doi.org/10.1016/j.asoc.2020.106266

[59] S. Thawkar, "A hybrid model using teaching–learning-based optimization and Salp swarm algorithm for feature selection and classification in digital mammography," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 8793-8808, 2021. https://doi.org/10.1007/s12652-020-02662-z

[60] I. Y. Shallangwa, A. A. Ahmad, and J. Isuwa, "Swarm intelligent optimization algorithms for precision gene selection in microarray-based cancer classification," *Sci. World J.*, vol. 19, no. 3, pp. 842-854, 2024.

[61] M. Tubishat, S. Ja'afar, M. Alswaitti, S. Mirjalili, N. Idris, M. A. Ismail, and M. S. Omar, "Dynamic salp swarm algorithm for feature selection," *Expert Syst. Appl.*, vol. 164, pp. 113873, 2021. https://doi.org/10.1016/j.eswa.2020.113873

[62] H. Faris, M. M. Mafarja, A. A. Heidari, I. Aljarah, A. Z. Ala'M, S. Mirjalili, and H. Fujita, "An efficient binary salp swarm algorithm with crossover scheme for feature selection problems," *Knowl.-Based Syst.*, vol. 154, pp. 43-67, 2018. https://doi.org/10.1016/j.knosys.2018.05.009

[63] S. Ramasubramanian, "Role of microarray in cancer biology," *J. Complement. Med. Res.*, vol. 11, no. 3, pp. 262-268, 2020. https://doi.org/10.5455/jcmr.2020.11.03.33

[64] K. K. Ghosh, S. Begum, A. Sardar, S. Adhikary, M. Ghosh, M., Kumar, and R. Sarkar, "Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data," *Expert Syst. Appl.*, vol. 169, pp. 114485, 2021. https://doi.org/10.1016/j.eswa.2020.114485

[65] A. K. Shukla, P. Singh, and M. Vardhan, "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*, Springer, 2019, pp. 783-792. https://doi.org/10.1007/978-981- 10-8055-5_69

[66] H. H. Al-Baity and N. Al-Mutlaq, "A new optimized wrapper gene selection method for breast cancer prediction," *Comput. Mater. Contin.*, vol. 67, no. 3, 2021. https://doi.org/10.32604/cmc.2021.015291

[67] P. Panda, S. K. Bisoy, A. Panigrahi, A. Pati, B. Sahu, Z. Guo, H. Liu, and P. Jain, "BIMSSA: Enhancing Cancer Prediction with Salp Swarm Optimization and Ensemble Machine Learning Approaches," *Front. Genet.*, vol. 15, pp. 1491602, 2025. https://doi.org/10.3389/fgene.2024.1491602

[68] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "Data mining: A preprocessing engine," *J. Comput. Sci.*, vol. 2, no. 9, pp. 735-739, 2006.

[69] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci.*, vol. 96, no. 12, pp. 6745–6750, 1999. https://doi.org/10.1073/pnas.96.12.6745

[70] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, L. Coller, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999. https://doi.org/10.1126/science.286.5439.531

[71] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nat. Genet.*, vol. 30, no. 1, pp. 41–47, 2002. https://doi.org/10.1038/ng765

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

[73] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, vol. 7, no. 6, pp. 673–679, 2001. https://doi.org/10.1038/89044

[74] F. Fajila and Y. Yusof, "Incremental search for informative gene selection in cancer classification," *Ann. Emerg. Technol. Comput.*, vol. 5, no. 2, pp. 15-21, 2021. https://doi.org/10.33166/AETiC.2021.02.002

[75] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed Res. Int.*, vol. 2015, 2015. http://dx.doi.org/10.1155/2015/604910

[76] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification," *Comput. Biol. Chem.*, vol. 56, pp. 49-60, 2015. https://doi.org/10.1016/j.compbiolchem.2015.03.001