

A hybrid wrapper spider monkey optimization-simulated annealing model for optimal feature selection

Bibhuprasad Sahu¹, Amrutanshu Panigrahi², Bibhu Dash³, Pawan Kumar Sharma³, Abhilash Pati²

¹Department of Artificial Intelligence and Data Science, Vardhaman College of Engineering(Autonomous), Hyderabad, India

²Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be University), Odisha, India

³School of Computer and Information Science, University of the Cumberland, Williamsburg, USA

Article Info

Article history:

Received Aug 23, 2022

Revised Dec 10, 2022

Accepted Feb 13, 2023

Keywords:

Cancer data

Feature selection

ReliefF

Simulating annealing

Spider monkey optimizer

ABSTRACT

In this research, a hybrid wrapper model is proposed to identify the featured gene subset from the gene expression data. To balance the gap between exploration and exploitation, a hybrid model with a popular meta-heuristic algorithm named spider monkey optimizer (SMO) and simulated annealing (SA) is applied. In the proposed model, ReliefF is used as a filter to obtain the relevant gene subset from dataset by removing the noise and outliers prior to feeding the data to the wrapper SMO. To enhance the quality of the solution, simulated annealing is deployed as local search with the SMO in the second phase, which will guide to the detection of the most optimal feature subset. To evaluate the performance of the proposed model, support vector machine (SVM) as a fitness function to recognize the most informative biomarker gene from the cancer datasets along with University of California, Irvine (UCI) datasets. To further evaluate the model, 4 different classifiers (SVM, naïve Bayes (NB), decision tree (DT), and k-nearest neighbors (KNN)) are used. From the experimental results and analysis, it's noteworthy to accept that the ReliefF-SMO-SA-SVM performs relatively better than its state-of-the-art counterparts. For cancer datasets, our model performs better in terms of accuracy with a maximum of 99.45%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Bibhuprasad Sahu

Department of Artificial Intelligence and Data Science, Vardhaman College of Engineering

Kacharam, Shamshabad, 501218 Hyderabad, Telangana, India

Email: prasadnikhil176@gmail.com

1. INTRODUCTION

Cancer is becoming more common and dangerous in today's world. If diagnosed in the early stages of infection, these disorders are treatable and curable. Physicians are helped by the medical diagnostic decision support system to efficiently and correctly identify the cause behind it. The advantages of machine learning algorithms and data mining concepts encourage detecting the disease at an early stage. The data mining concept is inadequate to handle the high dimensional data set due to the curse of dimensions [1]. Feature selection plays a vital role in recognizing irreverent and redundant features. It may reduce the computational cost of the machine learning model by removing them from the original dataset [2]. The most featured genes improve the performance of the classification model, and the selected features almost contain all the features of the dataset. Feature selection problems are typically multi-objective problems, and the main motto is to optimize the number of features selected and minimize the classification error rate. The preceding process is divided into three phases, such as phase 1 involves the generation of feature subsets using various searching methods. In

the second phase, after evaluating each candidate subset and comparing it with the other feature subsets. This iteration continues until the stopping criteria are met, and finally, the best-featured subset is validated with the original dataset.

Gene selection approaches are widely categorized into the filter, wrapper, hybrid, and embedded methods. The filter approaches select the features by using different statistical methods. Information gain (IG), mutual information, chi-square, f-score, and coefficient score are various statistical measures used by the filter approach. In the wrapper approach, the fitness function plays an essential role in identifying the featured gene subset, and the machine learning function performs part of the fitness function. Sequential forward/backward Selection is one of the most popular sampling methods used by various researchers. Basically, in the wrapper, the evaluation of each selected feature subset is focused, so these approaches are time-consuming and computationally costly. Simulated annealing (SA) is one of the single solution-based wrapper approaches used for gene selection. The SA algorithm uses the Monte Carlo algorithm, but its computation requirements are formidable when dealing with high-dimensional datasets.

The machine learning algorithm does not participate in feature selection in the filter technique. After determining ranks for each feature based on statistical measurements, features are deleted or selected. They are frequently univariate and work quickly on high-dimensional datasets. Most of the meta-heuristic optimization algorithms are stuck in local minima or maxima. The authors have used SA to avoid the above condition by controlling the probability function to check whether to accept or reject the new solutions. Indirectly, it reduces the memory space, and it is most accepted as it performs with fewer parameters in comparison with particle swarm optimization (PSO), and Tabu search. The no-free-lunch (NFL) theorem clearly states that one optimizer is not enough to provide a solution to a particular optimization problem [3]. It encourages many researchers to think about different hybrid models to select the optimal genes and improve the accuracy of the classification model. The main goal of this study is to propose a novel model named ReliefF-SMO-SA to identify biomarker genes using ReliefF as a filter embedded with spider monkey optimizer (SMO) as the wrapper [4]. This optimizer is used in various fields and performs well as expected due to enhanced exploration and exploitation [5]. Premature convergence and slow convergence are two limitations of the SMO algorithm. To provide better smoothness between exploration and exploitation while handling high dimensional datasets that may be stuck in local optima. As a result, the authors employed SA as a local search to improve the SMO's performance.

- ReliefF is a pre-processing catalyst for the novel ReliefF-SMO-SA novel model.
- The SMO algorithm is used in collaboration with SA for dealing with microarray datasets only.
- This integration of SA as a local search helps boost the performance of the optimization model and enhance the convergence rate.
- Set of popular bench-marked cancer datasets (UCI and microarray) were used to evaluate the performance of the SMO-SA approach.

The flow of this study is as follows: in section 2, the authors presented the related study. Section 3 and section 4 represent the overall study of the ReliefF, SA, and basic details about the SMO algorithm. The proposed model and its description are presented in section 5. In section 6, the detailed study of the dataset, the experimental description, and the performance comparison of the proposed model with the existing model are presented. Section 6 also includes the results of the statistical performance measure. At last, section 7 presents the conclusion and future scope.

2. STATE OF ART METHOD

Moradi and Gholampour [6] present a hybrid PSO embedded with a local search strategy. The author employed SA as a local search to improve the SMO's performance. The performance of the proposed model is compared with different existing methods. Sahu and Dash [7] proposed a hybrid information gain enhanced jaya optimization based model for optimal feature selection (IG-Jaya). The performance of the hybrid model is better compared with ant bee colony (ABC) or stochastic diffusion search (SDS). Kavitha *et al.* [8] proposed a hybrid model using the Bat algorithm with an extreme learning machine (ELM). The performance of the hybrid model is evaluated using various metrics such as accuracy, precision, F-score, and specificity. Ke *et al.* [9] presented a multi-filter-based feature selection using symmetric uncertainty and relief. According to the findings of the study, the author demonstrates that the proposed model outperforms an individual filter.

Similarly, Ghosh *et al.* [10] proposed a score-based criteria fusion using 2 numbers of filters with k-nearest neighbors (KNN) and support vector machine (SVM) are the two classifiers used to evaluate the performance of the five benchmark microarray data. Yang *et al.* [11] proposed a multifilter-wrapper genetic algorithm (GA) model to identify the most-featured genes from the five benchmark microarray datasets. The author employed a multi-layer perceptron (MLP), a KNN, and an SVM in this study. Yang *et al.* [12] proposed a multifilter genetic algorithm embedded in a wrapper recursive feature elimination (RFE). The proposed model outperforms existing models in the survey, according to the results of this study, which uses nine benchmark cancer datasets. The convergence speed of the model is quite impressive with small datasets.

Djellali *et al.* [13] presented an IG-GA model and claimed that the proposed model performed better with nine microarray cancer datasets. In this study, the author used IG as a filter and GA as a wrapper. Alshamlan *et al.* [14] presented a comparative study between two purposeful models named fast correlation based filter (FCBF)-GA and FCBF-PSO. FCBF-PSO performs better as compared to its counterpart. A filter wrapper model called maximum relevance-minimum redundancy (MRMR)-GA for the classification of microarray datasets, similarly, Alzubi *et al.* [15] proposed a multi-layer filter-based hybrid model, MRMR-ABC. In the first layer, three different approaches are used to find the optimal gene subset. Then again another phase filter called mutual information maximization was applied to detect the optimal feature subset with adaptive GA as wrapper. Shukla *et al.* [16] presented a combo feature selection model for the evaluation of the cancer microarray dataset. In this study, the author used enhanced version of PSO (IBPSO). Six cancer datasets and 3 known classifiers were used in this study. According to the author, the proposed model achieves good accuracy without trapping at a local minimum. Han *et al.* [17] created a hybrid filter-wrapper model in which conditional mutual information maximization (CMIM) serves as the filter and SVM-RFE serves as the wrapper. Single nucleotide polymorphisms (SNP) datasets were used in the overall experiment. Shukla *et al.* [18] presented a hybrid model combining teaching learning-based optimization (TLBO) and gravitational search algorithm (GSA). In this study, the author used mRMR as a filter to deal with ten microarray datasets using NB for classification need. A hybrid ReliefF and recursive binary GSA were proposed by Jain *et al.* [19], whereas the SA-enhanced TLBO model was developed with CFS as a filter by Arunkumar and Ramakrishnan [20] for optimal feature selection. A hybrid classification model named CFS-Binary PSO was provided by Jain *et al.* [21]. Similarly, Chinnaswamy and Srinivasan proposed CFS-PSO for the classification of microarray gene expression datasets [22].

A hybrid feature selection (FS) model with minimum redundancy maximum relevance (mRMR) and mouth-fame optimization was used to classify seven cancer microarray datasets. The performance of the hybrid model with a combination of gain ratio (GR) and an improved gene expression programming algorithm (IGEP) as filter and wrapper, respectively [23]. Nine cancer datasets were used in this study. The author has used the combination of IG and binary krill herd algorithms for the classification of nine microarray datasets. It achieves 100 percent accuracy in nine cases [24]. Minimum redundancy maximum relevancy-flower pollination algorithm (MRMR-FPA): a hybrid model was developed to identify the optimal feature subset. Here, MRMR is used as a filter, and FPA acts as a wrapper. The performance of MRMR-FPA is compared with that of the MRMR-GA and other existing counterparts. It seems that MRMR-FPA performs better [25]. A hybrid flower pollination algorithm meta-heuristic model is proposed, in which CFS (correlation feature selection) works as a filter whereas ABC is treated as the wrapper. Six different binary and multi-class cancer datasets are used to evaluate the performance of an SVM classifier [26]. Urbanowicz *et al.* [27] proposed a multifilter-voting concept-based meta-model for the classification of cancer datasets. Alomari *et al.* [28] compared the performance of Spearman's correlation (SC)-MRMR with three distinct filters: ReliefF, joint mutual information (JMI), and MRMR, as well as four well-known classifiers (naïve Bayes (NB), KNN, decision tree (DT), and SVM), and concluded that SC-MRMR outperforms other alternatives for the Lymphoma dataset. Here, they evaluate the performance of the model with six classifiers and six cancer datasets.

From the literature survey, we observed that most of the metaheuristic algorithms suffer premature convergence and slow convergence. SMO is one of the most recent algorithms preferred by various researchers to solve various problems including feature selection. To provide better smoothness between exploration and exploitation while handling high dimensional datasets SMO may be stuck in local optima. So, we employed SA as a local search to boost SMO performance and increase efficiency.

3. RELIEF METHOD

ReliefF is a popular and classification-efficient algorithm used in machine learning filters during feature selection [29]. The ReliefF algorithm performs with small, large, and nominal or continuous feature datasets. It also handles missing data as well as noisy ones. This is a correlation-concerning approach that integrates the features that have a high correlation to each other by ignoring the low-correlated samples. The flow of the ReliefF algorithm and how it carefully handles high-dimensional datasets are discussed below.

- The training sample yields sample y_i , and the P nearest neighbour similar sample of y_i is chosen based on the high correlation value.
- Similarly, non-similar samples (Q) from different classes y_i are identified and denoted as P(c).
- The correlation between the sample and the inter-class as well as the intra-class can be used to figure out how to evaluate and change the weight vector of the feature.
- This process will be repeated until the weights of all features are calculated. The weight value of the features is evaluated using (1).

$$W[D] = W[D_0] - \frac{\sum_{j=1}^k \text{diff}(D, y_i, H)}{mk} + \sum_{D \neq \text{class}(y_i)} \frac{p(D)}{1 - p(\text{class}(y_i))} \cdot \frac{\sum_{j=1}^k \text{diff}(D, y_i, M_j(D))}{mk} \quad (1)$$

Where D_0 and D represent the number of features in the raw and filtered datasets, respectively. $W[D_0]$ is denoted as the weight coefficient (before updation), and $W[D]$ is denoted as the updated weight coefficient; y_i presents the i^{th} sample, and within the intraclass nearest neighbor samples with y_i is denoted in Q. $\text{diff}(D, y_i, Q)$ represents the quantitative difference between y_i and Q on each feature in D. Here, m and k represent the total number of repeats and nearest neighbors, respectively. $p(C)$ is the j^{th} neighbor sample in a different class that contains the target samples C; $p(\text{class}(y_i))$ is the ratio of samples in the same class that contain y_i to the total number of samples; $m_j(D)$ is the jth neighbor sample in a different class that contains the target samples D; and $\text{diff}(D, y_i, M_j(D))$ is the difference between y_i and $M_j(c)$ on each feature in D.

3.1. Improved ReliefF method

The interdependence between the features is the main concept adopted by the Relief algorithm to recognize the most prominent, highly ranked genes from the original dataset samples. The relief method is used as a filter approach to recognize the samples that are most similar to each other. This greatly aids in avoiding redundancy genes, which is a major issue in microarray datasets due to the "curse of dimensions." Not only does it help reduce the processing speed of the model, but it also enhances performance accuracy.

Definition 1: the (2) is used to calculate the distance between sample y_i and other samples within a class, as well as the sample within gene subset D.

$$\text{dis}(D, y_i, Q) = \sum_{i=1}^k \frac{|y_i - \bar{Q}|}{\max(D) - \min(D)} \quad (2)$$

Here Q and \bar{Q} present the distance of the samples of the same class with y_i and the average distance between the neighboring samples with y_i within the same class. The maximal and minimal feature values of gene subset D are described by $\max(D)/\min(D)$.

Definition 2: with the exception of the inter-class concept, this definition presents the same concept as the previous one. This means that (3) depicts the distance between sample y_i and M_j (C) in a different class,

with y_i in the aforementioned subset D.

$$dis(D, y_i, M_j(c)) = \sum_{c \neq class(y_i)} \frac{p(C)}{1 - p(class(y_i))} \cdot \sum_{i=1}^k \frac{|y_i - \overline{M_j(C)}|}{\max(D) - \min(D)} \quad (3)$$

Where $p(C)$ represents the target sample ratio and C the total number of samples, and $p(class(y_i))$ the ratio of samples in the subset, including y_i , to the total number of samples. $M_j(c) - \overline{M_j(C)}$ denotes the average distance of sample (non-nearest neighbor) of interclass with y_i . The distance between the samples is evaluated using Euclidean distance. It can be presented in (4).

$$\Delta_A(x, y) = \sqrt{\sum_{k=1}^{|A|} |f(x, a_k) - f(y, a_k)|^2}, \quad (4)$$

The number $|A|$ denotes the cardinality of genes present in A, and $f(x, a_k)$ denotes the effect of sample x on gene a_k . In order to acquire accurate weight values, all individual samples should be compared with the samples of the intraclass as well as the interclass. As the ReliefF algorithm considers the samples randomly even if the training sample remains the same, there is a chance of weight value fluctuation. Using theorems 1 and 2, we can enhance the correctness of the calculation. But when trying to decrease the distance between samples within the class, it directly increases the distance of the sample from other samples. To solve this issue, a new distance coefficient calculation is adopted in definition 3 as presented in (5).

$$Coeff - Distance = \frac{\sqrt{\sum_{i=1}^k (y_i - \bar{x})^2}}{\sum_{i=1}^k y_i}, \quad (5)$$

Where K and \bar{x} represent the number of genes and average numbers of the sample chosen, and $x_1, x_2, \dots, x_i, \dots,$ and x_k represent the distinct gene values.

Definition 3: the following formula is used to update the weight coefficient of genes in the ReliefF algorithm and is defined in (6).

$$W[D] = W[D_0] - \frac{CD_{intra} \sum_{j=1}^k dis(A, y_i, H)}{mk} + \frac{CD_{difference} \sum_{C \neq class(y_i)} \frac{p_i(C)}{1 - p(class(y_i))}}{\sum_{j=1}^k \frac{dis(A, y_i, M_j(C))}{mk}}, \quad (6)$$

D and D_0 represent the number of gene subsets in the filtered and original datasets, respectively. Before updating the weight coefficient, it is denoted as $W[A_0]$. CD_{intra} and $CD_{difference}$ are the distance coefficients of samples within and between classes, respectively. Instead of using standard ReliefF, we have adopted the enhanced version as a filter. The output of the filter is fed as input to the wrapper model called SMO-SA.

4. METHODS

Numerous researchers are inspired by natural behaviours, and as a result, they have devised algorithms that replicate natural activities. These algorithms are known as nature-inspired algorithms (NIAs). In numerous applications, NIAs have been employed in combination with machine learning methods. SMO and SA are two well-known metaheuristic strategies in the field of artificial intelligence. In this section, the fundamentals of SMO and SA algorithms are discussed in addition to the proposed SMO-SA technique.

4.1. Spider monkey optimizer

SMO is one of the recent metaheuristic techniques based on the foraging nature of spider monkeys. The unique foraging nature of the monkeys belongs to a special social structure named fission-fusion. The main concept behind SMO is a single female leader who can take decisions in the social organisation to divide and create groups as needed. Global and local groups or teams comprise the entire organization. Global and local team leaders are referred to as global and local leaders, respectively. As this algorithm is mainly focused on the search for food for the spider monkeys, when food scarcity arises we can consider that there is no progress in the solution. Because this algorithm is an inherited swarm concept optimization model, the swarm population is built using a small group of monkeys. A group of a minimal number of monkeys is fixed for the beginning of the problem space. When a group with fewer monkeys than the minimum number is found, this stage is called "fission," and the leader can lead the fusion at any time to search for food for the team. Different six different phases including the local leader phase (LLP), global leader phase (GLP), local leader learning phase (LLL), global leader learning phase (GLLP), local leader decision phase (LLDP), and global leader decision phase (GLDP) are used in this optimization model to solve a particular problem. The execution steps for SMO are presented in Algorithm 1. The detailed study of SMO's phases is explained below.

Algorithm 1 Spider monkey optimization

Input: initialization of swarm population, local/global leader limit and perturbation rate.

Output: optimal feature subset.

1. Iteration= 0.
 2. While (termination criteria not satisfied)
 3. Identify the local and global leader.
 4. Update the position of local leader and update the position of global leader.
 5. Using global leader learning learn the position.
 6. Using local leader learning learn the position.
 7. Using local leader decision phase update the position.
 8. Using using global leader decision phase decision should taken whether fission or fusion.
 9. if termination condition satisfies stop.
 10. Decide global leader position is the optimal solution one other wise update the local leader position .
-

4.1.1. Initialization

During the initialization phase, SMO generates an equally distributed initial swarm of N spider monkeys (SM), where SPM_i denotes the swarm's i^{th} spider monkey (SPM). The following is how each SM is set up and is derived in (7).

$$SPM_{ij} = SPM_{min_j} + UB(0, 1) \times (SPM_{max_j} - SPM_{min_j}) \quad (7)$$

Where SPM_{min_j} and SPM_{max_j} are the lower and upper bounds of the search space in the j^{th} dimension, respectively, and UB is a uniformly distributed random number in the range (0, 1).

4.1.2. Local leader phase

This is a critical stage in the SMO algorithm process. All spider monkeys have the opportunity to update their skills here. The spider monkey's location has been modified based on the experiences of its local leader and local group members. The fitness value of the individual monkey is evaluated and compared with the present fitness value, if this is less, the fitness value should be updated with the new one; otherwise, ignore it. The position update equation is as mentioned in (8).

$$SPM_{new_{ij}} = SPM_{ij} + UB(0, 1) \times (LL_{kj} - SPM_{ij}) + UDR(-1, 1) \times (SPM_{rj} - SPM_{ij}) \quad (8)$$

Here the SPM_{ij} and LL_{kj} presents the j^{th} dimension of i^{th} monkey spider and position of k^{th} group local leaders position in j^{th} dimension respectively. Whereas SPM_{rj} presents the randomly selected SPM from the k^{th} group in j^{th} dimension with a condition $r \neq i$. UDR(-1,1) denotes the values are uniformly

distributed within the range of $(-1,1)$. The spider monkey, which is about to update its position, is drawn to the local leader while preserving its self-confidence, as shown by (2). The final component helps introduce variations in the search process, which helps retain the algorithm's stochastic nature and avoid premature stagnation.

4.1.3. Global leader phase

In this phase, the fitness function is defined. The fitness function (fit_i) can be defined on the basis of the selection probability $Prob_i$, which is denoted by (9) and (10).

$$fit_i = \begin{cases} \frac{1}{1+f_i}, & \text{if } f_i \geq 0 \\ 1 + abs(f_i), & \text{if } f_i < 0 \end{cases} \quad (9)$$

$$Prob_i = \frac{fitness_i}{\sum_{i=1}^n fitness_i} \quad (or)$$

$$Prob_i = 0.9 \times \frac{fitness_i}{max_{fit}} + 0.1 \quad (10)$$

After gaining the knowledge from the Global Leader the SPM tries to improve its position using (11).

$$SPM_{new_{ij}} = SPM_{ij} + UB(0, 1) \times (GL_j - SPM_{ij}) + UDR(-1, 1) \times (SPM_{rj} - SPM_{ij}) \quad (11)$$

4.1.4. Global leader learning phase

During this phase, the algorithm finds the best possible solution for the entire swarm. The detected SM is widely recognized as the global leader of the swarm. Additionally, the position of the global leader is checked, and if it has not changed, the counter associated with the global leader is reset. If the leader is not set to 0, the global limit count (GLC) is increased by one; otherwise, it remains at zero. The GLC for the global leader is verified and compared to the global leader limit (GLL).

4.1.5. Local leader learning phase

The local leader is kept up to date by a competitive selection process among group members. A counter called local limit count (LLC) associated with the local leader is increased by one if the local leader does not update its position; otherwise, the counter is reset to 0. This process is used to choose the local leader for each of the groups. LLC is increased incrementally until it reaches a certain, predetermined threshold.

4.1.6. Local leader decision phase

Before proceeding with this phase, the positions of the local and global leaders are identified. If the local leader does not reach the local leader limit, then all members need to update their positions by random initialization or by using the knowledge gained by the global leader. It uses a new concept called perturbation rate and is presented in (12).

$$SPM_{new_{ij}} = SPM_{ij} + UB(0, 1) \times (GL_j - SPM_{ij}) + UDR(-1, 1) \times (SPM_{rj} - LL_{kj}) \quad (12)$$

4.1.7. Global leader decision phase

The swarm is split into distinct groups or fused into a single group if the global leader does not reorganise the GLL, which is similar to the local leader decision phase. In this case, GLL is the parameter that checks for premature convergence, and it varies from $N/2$ to $2N$. If GLC is more than GLL, GLC is set to zero and the number of groups are compared to the maximum groups. Global leaders further divide existing groups if their number is less than the pre-defined maximum number of groups, otherwise, they combine to form a single group.

4.2. Simulating annealing

SA is a probabilistic and metaheuristic method that can find a global optimum in a high-dimensional search space. It uses the hill climbing approach. The main focus behind this algorithm is to solve the issues that occur while the metaheuristic algorithms are stuck in the local optimum. From the literature survey, it's noteworthy to state that, most of the researchers used this algorithm for this need. Randomly generates the solution quality given as an initial solution at the start of the technique; each subsequent generation generates a solution that is near to the best according to a preset neighbourhood structure and evaluated by the fitness function. SA performs better whether it is a maximization or minimization problem. While handling maximization, it chooses the new solution as the best solution if the fitness value is better than the current solution. But in the case of minimization problems, the difference between current and new is less than 0. The SA is being used to address the slow convergence and enhance exploitation by searching the high-quality regions discovered by the SMO.

5. PROPOSED METHOD

A hybrid FS model is built on an enhanced version of SMO termed ReliefF-SMO-SA-SVM, which refers to a ReliefF SMO with local search strategy (SA), to cope with the FS problem we suggest in this study in Figure 1. By eliminating irrelevant and redundant features that are significantly correlated, the ReliefF-SMO-SA method aims for great precision with the least number of features possible. In order to fulfill the first objective, the proposed ReliefF-SMO-SA employs a two-phase approach to choose the most appropriate feature subset from a large number of possible options. Initially, the ReliefF is used as a preprocessing step to define the correlation between features and classes. Each attribute is assigned a numerical value (weight) that corresponds to the importance of the feature. The SMO approach is used in addition to the specific local search as a wrapper-based FS methodology. To begin, SMO is initialized using the weighted set of features obtained from the filter stage. As a result, the weights of characteristics are ordered in decreasing order, and the SMO algorithm's initialization is performed in accordance with this ranking order. If a feature's ranking is low, it has a good chance of being chosen. Meanwhile, because the feature ranking is higher, there is a greater chance that it will be excluded from the first population. Following the startup step, the SMO algorithm iteratively looks for the lowest feature subset with the best performance. Even though the ReliefF ranking was the driving force behind the SMO search process, it was insufficient because the ReliefF treats each feature individually and does not take into account the probable relationship between the features, which can lead to redundancy and, as a result, affect classification accuracy. With the goal of overcoming this limitation, we combined the SMO with a local search technique that takes into consideration both feature correlation and weights in order to steer the particle to the ideal feature subset throughout the search process. The candidate particles are next inspected, and the wrapper selection process is repeated as many times as necessary until the termination condition is reached. Finally, the SMO-SA algorithm returns the best collection of features. Algorithm 2 depicts the major steps of ReliefF-SMO-SA.

5.1. Fitness function

Specifically, SMO-SA serves as a wrapper that is critical to the process of gene selection, which is necessary to enhance the classification performance in terms of accuracy. When using SMO-SA, the primary objective is to choose a subset of features in order to obtain greater classification accuracy than when using all of the available features. During the evolutionary process, the fundamental fitness function, as described in (13), is to maximize the algorithm's classification accuracy by using the selected genes.

$$Accuracy = \frac{T_{pos} + T_{neg}}{T_{pos} + T_{neg} + F_{pos} + F_{neg}} \quad (13)$$

Where, T_{pos} , T_{neg} , F_{pos} , and F_{neg} represent as true positive, true negative, false positive, and false negative.

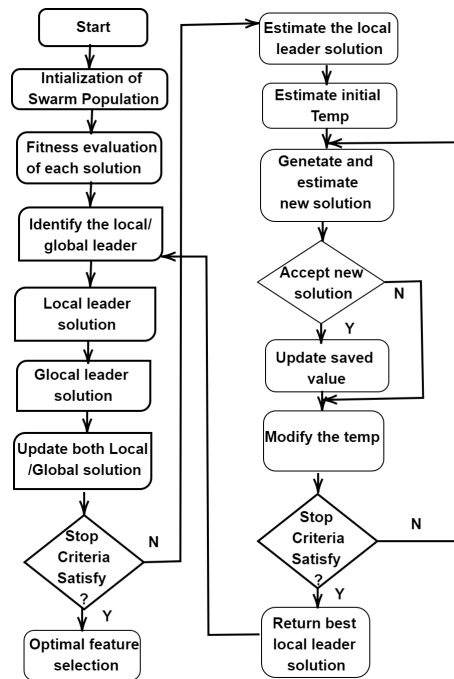


Figure 1. Block diagram of proposed model

Algorithm 2 Pseudo-code of ReliefF-SMO-SA

Input:

1. Load training dataset.
2. Initialize the parameters of SMO and SA.
3. Set Max_itr, Population and Max_feature selection

Stage-1:

1. Let Y the feature subset consists of $Y = (Y_1, Y_2, \dots, Y_n)$.
2. By using ReliefF algorithm, find the weight of individual features.
3. Rank and sort the features on rank basis in descending order.
4. Identify the top features as input for phase-2 (wrapper) SMO.

Stage-2:

1. Evaluate the correlation value for every features from the top features (Selected from phase 1).
 2. Initialize the population of N no's from the top features.
 3. Evaluate the fitness function for each individual feature.
 4. Using greedy solution find the position of global and local leader.
 5. Using LLP algorithm update the position of local leader.
 6. Using GLP algorithm update the position of global leader.
 7. Learning through global/local learning phase.
 8. Learning through global/local learning phase.
 9. Using local leader decision phase update the position of local leader.
 10. Decide fission/fusion using global decision leader position.
 11. If termination criteria satisfied find the optimal feature otherwise use SA for local search.
 12. The output of SA feed to step-4.
-

6. EXPERIMENTAL STUDY

Using two experimental series, the proposed model is assessed for its overall performance. In the first experiment, benchmark datasets of various sizes, as well as reasonably large and small samples, are obtained from the UCI machine learning repository (Table 1) and used. Using a second experimental series consisting of 10 separate high-dimensional microarray datasets, as shown in Table 2, the efficiency of R-SMO-SA in feature selection is investigated.

Table 1. UCI datasets

No.	Dataset	Features	Instances
D_1	Breastcancer	9	699
D_2	BreastEW	30	569
D_3	CongressEW	16	435
D_4	Exactly	13	1000
D_5	Exactly2	13	1000
D_6	HeartEW	13	270
D_7	IonosphereEW	34	351
D_8	Lymphography	18	148
D_9	M-of-n	13	1000
D_{10}	PenglungEW	325	73
D_{11}	SonarEW	60	208
D_{12}	SpectEW	22	267
D_{13}	Tic-tac-toe	9	958
D_{14}	Vote	16	300
D_{15}	WineEW	13	178
D_{16}	Zoo	16	101

Table 2. Microarray datasets

Sl.No.	Dataset	Features	Instances
DS_1	l1_tumors	12533	699/11
DS_2	Brain_tumors1	5920	90/5
DS_3	Brain_tumors2	10367	50/4
DS_4	Lung_Cancer	12600	203/5
DS_5	Colon	2000	62/2
DS_6	Leukemia1	7029	72/2
DS_7	Leukemia2	7029	72/2
DS_8	SRBCT	2308	83/4
DS_9	DLBCL	5469	77/11
DS_{10}	Prostate_Tumor	10509	102/2

6.1. Performance evaluation criteria and experimental parameter setting

The performance of the proposed model is evaluated by focusing on three basic criteria: the number of selected features, the best fitness value, and the accuracy achieved. Here, we have used two parameters named mean (average of the fitness value obtained by the FS approach among M runs) and mean feature selection number (number of the selected features (avg) after M runs) to evaluate the performance of the proposed model. Table 3 presents a performance comparison between R-SMO and R-SMO-SA in terms of the average number of features selected, best fitness achieved, and accuracy. Similarly, Table 4 presents the average number of features selected, the best fitness achieved, and accuracy in terms of standard deviation, whereas Table 5 denotes the performance of repetitive SMO (R-SMO) and R-SMO-SA with different UCI datasets by calculating the P-value after 20 iterations. Finally, Table 6 presents a comparative study of the proposed model with different existing models such as TLBO-SA, IG-modified binary krill herd (MBKH), random ant colony optimization (R-ACO), binary shuffled frog leaping algorithm (BSFLA)-PSO, and ant lion optimizer (ALO). The proposed model's performance is compared to five different existing models. The parameters considered throughout the experimental analysis are shown in Table 7. As all metaheuristic algorithms use stochastic-based algorithms, we have used the same approximate parameter for unbiased comparison between each other.

Table 3. Comparative study of average no of features selected, best fitness value with accuracy on R-SMO and R-SMO-SA

Dataset	Selected features		Best fitness		Accuracy	
	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA
<i>D</i> ₁	6.020	6.618	0.035	0.018	0.872	0.891
<i>D</i> ₂	13.580	12.587	0.097	0.086	0.784	0.814
<i>D</i> ₃	5.300	5.587	0.089	0.068	0.987	0.991
<i>D</i> ₄	6.280	6.445	0.354	0.025	0.953	0.974
<i>D</i> ₅	1.550	1.975	0.018	0.011	0.967	0.988
<i>D</i> ₆	5.980	5.929	0.355	0.297	0.960	0.926
<i>D</i> ₇	11.672	11.857	0.078	0.068	0.914	0.957
<i>D</i> ₈	18.360	19.998	0.018	0.011	0.953	0.993
<i>D</i> ₉	7.700	7.890	0.008	0.072	0.987	0.996
<i>D</i> ₁₀	121.340	118.592	0.097	0.082	0.909	0.931
<i>D</i> ₁₁	27.550	28.885	0.078	0.062	0.958	0.979
<i>D</i> ₁₂	8.650	8.657	0.018	0.009	0.959	0.961
<i>D</i> ₁₃	6.750	6.889	0.078	0.068	0.914	0.935
<i>D</i> ₁₄	21.650	21.693	0.038	0.025	0.953	0.961
<i>D</i> ₁₅	6.840	6.899	0.042	0.037	0.962	0.968
<i>D</i> ₁₆	8.850	8.891	0.098	0.086	0.919	0.935

Table 4. Comparative study of average no of features selected, best fitness value with accuracy on R-SMO and R-SMO-SA in standard deviation

Dataset	Selected features		Best fitness		Accuracy	
	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA
<i>D</i> ₁	0.000000	0.000000	0.006389	0.005974	0.014274	0.028456
<i>D</i> ₂	2.518876	2.626894	0.005165	0.004859	0.003624	0.004875
<i>D</i> ₃	1.716790	1.889965	0.015687	0.005964	0.001785	0.002785
<i>D</i> ₄	1.234300	1.293365	0.002897	0.001987	0.004859	0.004985
<i>D</i> ₅	2.459675	2.475368	0.013910	0.012298	0.003632	0.003264
<i>D</i> ₆	1.576138	4.986217	0.017450	0.016895	0.039875	0.046314
<i>D</i> ₇	4.006245	4.986217	0.081205	0.080263	0.073124	0.083214
<i>D</i> ₈	2.518876	2.498617	0.005121	0.006972	0.004631	0.005785
<i>D</i> ₉	7.727429	7.92387	0.000938	0.000897	0.001785	0.002963
<i>D</i> ₁₀	0.698683	0.832146	0.063070	0.068597	0.004821	0.005746
<i>D</i> ₁₁	1.142481	1.156982	0.011450	0.011357	0.005987	0.005687
<i>D</i> ₁₂	2.680800	2.787913	0.015586	0.014968	0.002831	0.003971
<i>D</i> ₁₃	0.910460	0.997159	0.004565	0.003265	0.004827	0.004966
<i>D</i> ₁₄	1.234300	1.178249	0.005190	0.005118	0.028255	0.008141
<i>D</i> ₁₅	1.142481	1.69784	0.000000	0.000000	0.042238	0.048125
<i>D</i> ₁₆	0.998683	0.863172	0.004160	0.004111	0.096478	0.095851

Table 5. Comparative study of features selected, best fitness value with accuracy on R-SMO and R-SMO-SA in (p_value > 0.05 after 20 iteration)

Dataset	Selected features		Best fitness		Accuracy	
	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA	R-SMO	R-SMO-SA
<i>D</i> ₁	0.000000	0.000000	0.005231	0.005974	0.185232	0.278457
<i>D</i> ₂	0.005329	0.005698	0.004852	0.004859	0.003896	0.432556
<i>D</i> ₃	0.004965	0.005982	0.052312	0.005964	0.185425	0.254223
<i>D</i> ₄	0.006398	0.007158	0.004852	0.001987	0.856521	0.964875
<i>D</i> ₅	0.065241	0.078521	0.055952	0.012298	0.089663	0.095477
<i>D</i> ₆	0.051284	0.059631	0.004851	0.056895	0.057889	0.046314
<i>D</i> ₇	0.006547	0.006385	0.005328	0.080263	0.077852	0.081215
<i>D</i> ₈	0.005785	0.006315	0.005946	0.006315	0.004631	0.005785
<i>D</i> ₉	0.005974	0.005267	0.059381	0.068521	0.017856	0.029634
<i>D</i> ₁₀	0.052147	0.051462	0.069041	0.068597	0.048213	0.057468
<i>D</i> ₁₁	0.051478	0.056982	0.059862	0.066741	0.05787	0.056452
<i>D</i> ₁₂	0.058749	0.587913	0.063152	0.068547	0.002953	0.003389
<i>D</i> ₁₃	0.005698	0.897159	0.004983	0.005967	0.004281	0.004525
<i>D</i> ₁₄	0.002841	0.178249	0.053645	0.062542	0.514785	0.558225
<i>D</i> ₁₅	0.001825	0.001691	0.000000	0.000000	0.051551	0.048582
<i>D</i> ₁₆	0.005361	0.863172	0.001365	0.003975	0.092589	0.091545

Table 6. Comparison between R-SMO-SA and other algorithms in terms of best fitness

Dataset	Binary particle swarm optimization (BPSO)	Best fitness				ALO	R-SMO-SA
		Firefly algorithm (FA)	BAT	Grey wolf optimization (GWO)			
D_1	0.038	0.046	0.216	0.189	0.008	0.028	
D_2	0.044	0.035	0.176	0.048	0.046	0.047	
D_3	0.034	0.041	0.038	0.059	0.017	0.078	
D_4	0.037	0.496	0.126	0.019	0.085	0.045	
D_5	0.243	0.026	0.084	0.022	0.063	0.032	
D_6	0.135	0.276	0.012	0.165	0.087	0.044	
D_7	0.113	0.312	0.182	0.026	0.012	0.084	
D_8	0.145	0.412	0.045	0.067	0.063	0.078	
D_9	0.005	0.089	0.039	0.189	0.075	0.096	
D_{10}	0.165	0.041	0.166	0.068	0.045	0.074	
D_{11}	0.093	0.123	0.045	0.013	0.312	0.068	
D_{12}	0.134	0.016	0.058	0.046	0.096	0.097	
D_{13}	0.201	0.037	0.056	0.035	0.141	0.049	
D_{14}	0.032	0.192	0.041	0.058	0.257	0.084	
D_{15}	0.200	0.179	0.040	0.016	0.109	0.085	
D_{16}	0.045	0.045	0.046	0.041	0.246	0.091	

Table 7. Parameter description of algorithms

Algorithm	Parameters details
TLBO-SA	Pop-size: 20, No of generation: 100, No of run: 10, performance: accuracy.
IG-MBKH	Pop-size: 20, Iteration: 100, Top M: 80, N_{max} : 4, V_f : 0.02, D_{max} : 0.005
R-ACO	No of ants(r): 100, number of iterations: 80, Q: 100
BSFLA-PSO	Pop-size: 25, the number of memplexes:5, intra-updates of memplexes: 8, number of improvisations: 100
ALO	σ, λ , Loudness:0.9, r_0 :0.5, F_{max} :1, F_{min} :0.5, C_{max} : 35000, C_{min} : 0.01, σ_{min} :0.01, σ_{max} : 100, sd_{min} : 100
R-SMO-SA	Max no of groups: 10, GLL: (50,100), LLL: D*100, perturbation rate (pr): (0.1-0.8)

6.2. Result study

Table 3 to Table 6 presents the performance of the UCI datasets. Table 3 focus the performance of the proposed model comparison with R-SMO in term of average no of features selected, best fitness and accuracy. From Table 3 it is noteworthy to state that the proposed methodology R-SMO-SA perform better in case of all datasets except breastEW dataset in term of selecting the optimal no of features R-SMO-SA performs well good as expected in case of best fitness and accuracy in case of all datasets. In Table 4, we have presented the same attributes with standard deviation. In case of no of feature selected the standard deviation of Lymphography, PenglungEW, Vote, Zoo is not good as compare to the R-SMO-SA. Similarly while evaluating the best fitness for Lymphography and PenglungEW the proposed model unable perform better comparing to the other datasets. Except Exactly2, SonarEW, Zoo the standard deviation of accuracy is not good in case of R-SMO with R-SMO-SA. From the Table 3 and it is clear that the proposed model is R-SMO-SA approach is better than R-SMO approach on half of the cases. To compare the overall results produced from binary dragonfly algorithm (BDA) and BDA-SA, the average and standard deviation were employed as metrics. To see if the discrepancies in the results are significant. Whether the non-parametric results are statistically significant or not Wilcoxon significance threshold of 0.05 was used in the test. This test is suitable for comparing algorithms with stochastic behaviour. On majority of the data sets, the p-values for accuracy and fitness reveal that R-SMO-SA achieved considerably superior outcomes than R-SMO, as shown in Table 5. In term of accuracy out of 16 datasets except exactly dataset, its shows that p value is statistically significant in R-SMO-SA but in R-SMO three results is not statistically significant. There are 2 no of results and 5 no of are not statistically significant in term of selected features and best fitness with R-SMO. Where as 4 number of result are not significant in case of R-SMO-SA in term of no of features and best fitness respectively. In order to avoid falling into the trap of local optima, SMO's capacity to explore highly relevant regions in the feature space is employed, followed by SA's ability to intensify surrounding regions until the optimum solution achieved by the SMO algorithm is reached.

TLBO-SA, IG-MBKH, R-ACO, BSFLA-PSO, ALO and R-SMO-SA were used to compare the performance of SMO-SA. From the Table 6 it is clear that R-SMO-SA had the lowest averages of best fitness on 15 numbers of dataset out of 18 numbers taken into consideration. In addition, in terms of the smallest number of selected attributes, Table 8 shows that R-SMO-SA surpassed all existing models in terms of accuracy rates.

At the end of the study we have tested the proposed model with different existing machine learning models in the studied in the literature survey. It is clear that our proposed model performed better. In Comparison to IG-MBKH model for Leukemia1, Leukemia2, small round blue cell tumors (SRBCT) and diffuse large B-cell lymphoma (DLBCL), prostate tumor for TLBO-SA model, our model performed almost close to it. The graphical presentation of the Table 8 is presented in Figure 2.

Table 8. Comparison between R-SMO-SA and other algorithms in terms of accuracy (%)

Dataset	TLBO-SA [30]	IG-MBKH [31]	R-ACO [32]	BSFLA-PSO [33]	ALO [34]	R-SMO-SA-SVM
<i>DS₁</i>	95.31	-	95.06	-	94.79	96.79
<i>DS₂</i>	96.98	-	-	-	87.41	98.11
<i>DS₃</i>	-	-	-	-	88.17	95.91
<i>DS₄</i>	99.87	-	99.50	94.91	89.57	99.21
<i>DS₅</i>	99.01	96.47	94.00	-	-	96.98
<i>DS₆</i>	95.31	100	95.80	95.78	87.39	98.45
<i>DS₇</i>	99.54	100	-	-	91.44	96.78
<i>DS₈</i>	99.91	100	-	-	-	94.31
<i>DS₉</i>	99.52	-	-	-	-	99.45
<i>DS₁₀</i>	99.13	-	89.20	96.76	89.44	99.11

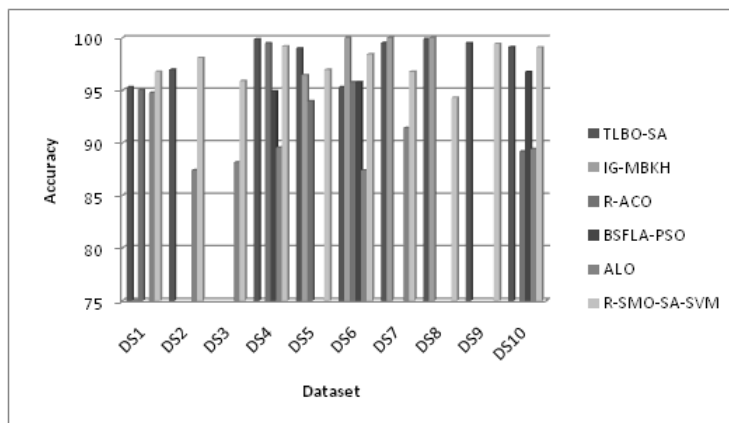


Figure 2. Comparison between R-SMO-SA and other algorithms in terms of accuracy

To compare the performance of the proposed method R-SMO-SA we have compared with different classifiers like NB, linear discriminant analysis (LDA), KNN, DT and SVM. The performance of the different classifiers with microarray data set is presented in Table 9 and graphical presentation is provided in. The highest accuracy achieved by the classifier and lowest one accuracy is presented in bold and italic in the Table 9. The performance of the random forest classifier with the proposed model is impressive in 11 tumors dataset with an accuracy of 98.31%, where as NB, KNN, SVM achieves the accuracy of almost equal to 95% to 98%. But with LDA the achieved accuracy is very poor with 91.77%. Except KNN classifier other classifiers achieved with an accuracy ranging from 95%-98% where as KNN performs with an accuracy of 99.50% in case of Brain tumors1 dataset. For Brain tumors2 dataset also KNN performs better as compare other classifier counterparts. But for rest datasets SVM performs quite impressive with respect to NB, LDA, KNN, DT With a accuracy of 99%. Finally, we can conclude that out of 10 number of datasets, SVM performs better in 7 datasets except 11 tumors, Brain tumors1, Brain tumors2 dataset. Figure 3 represents the graphical presentation of the comparative study of the proposed model with various classifiers.

Table 9. Comparison between R-SMO-SA with different classifiers in terms of accuracy (%)

Dataset	R-SMO-SA-NB	R-SMO-SA-LDA	R-SMO-SA-KNN	R-SMO-SA-DT	R-SMO-SA-SVM
DS_1	95.65	91.77	95.06	98.31	96.79
DS_2	96.75	95.50	99.50	94.28	98.11
DS_3	94.84	91.38	98.61	97.51	95.91
DS_4	96.12	98.50	96.47	95.41	99.21
DS_5	93.94	96.00	96.43	95.97	96.98
DS_6	97.51	95.27	97.31	92.10	98.45
DS_7	96.42	94.90	86.29	91.47	96.78
DS_8	91.33	89.31	89.71	94.23	94.31
DS_9	92.17	98.64	89.34	95.97	99.45
DS_{10}	91.67	97.74	97.63	92.45	99.11

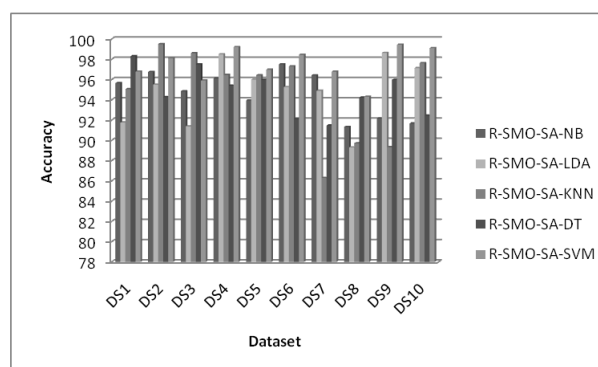


Figure 3. Performance of the proposed method R-SMO-SA with different classifiers

6.3. Statistical study

Generalized Friedman testing, which is one of the most extensively used non-parametric analytical approaches, is employed for ranking the algorithm performance in order to determine its relative merits. Its purpose is to identify any statistically significant discrepancies between the outputs of several algorithms. It is premised on the null hypothesis, which states that there is no difference in the way algorithms are presented in different situations. The algorithm with the best performance gets the lowest rank, while the algorithm with the poorest performance receives the highest rank, as a result of this ranking system.

7. CONCLUSION AND FUTURE WORK

In spite of the fact that existing wrapper techniques are capable of identifying informative genes from high-dimensional datasets, they have a number of flaws, including a lack of exploitation capability and a tendency to become stuck in local optima. To address the flaws in existing wrapper approaches, we have developed R-SMO-SA, a hybrid feature selection method. The major goal was to improve the spider monkey optimization algorithm's performance, particularly in terms of classification accuracy. The best solution identified so far by the R-SMO algorithm was used as an initial solution by the SA algorithm to conduct a local search to find a solution that was better than SMO's. The performance of SMO-SA was compared to that of the native SMO algorithm as well as TLBO-SA, IG-MBKH, R-ACO, BSFLA-PSO, and ALO, among other algorithms. The R-SMO-SA-SVM algorithm outperformed with the other algorithms in tests. It would be worthwhile to test the proposed hybrid technique on real world high dimensional datasets in future and measure its efficiency.

REFERENCES




- [1] B. Sahu, S. Mohanty, and S. Rout, "A hybrid approach for breast cancer classification and diagnosis," *ICST Transactions on Scalable Information Systems*, vol. 6, no. 20, p. e2, Jul. 2019, doi: 10.4108/eai.19-12-2018.156086.
- [2] B. Sahu, "A combo feature selection method (filter + wrapper) for microarray gene classification," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 16, pp. 389-401, 2018.
- [3] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67-82, Apr. 1997, doi: 10.1109/4235.585893.

- [4] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine learning supervised algorithms of gene selection: a review," *Machine Learning*, vol. 62, no. 3, pp. 233–244, 2020.
- [5] J. C. Bansal, H. Sharma, S. S. Jadon, and M. Clerc, "Spider monkey optimization algorithm for numerical optimization," *Mematic Computing*, vol. 6, no. 1, pp. 31–47, Mar. 2014, doi: 10.1007/s12293-013-0128-0.
- [6] P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," *Applied Soft Computing*, vol. 43, pp. 117–130, Jun. 2016, doi: 10.1016/j.asoc.2016.01.044.
- [7] B. Sahu and S. Dash, "Optimal feature selection from high-dimensional microarray dataset employing hybrid IG-Jaya model," *Current Materials Science*, vol. 16, Jan. 2023, doi: 10.2174/2666145416666230124143912..
- [8] K. R. Kavitha, A. Prakasan, and P. J. Dhrishya, "Score-based feature selection of gene expression data for cancer classification," in *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Mar. 2020, pp. 261–266, doi: 10.1109/ICCMC48092.2020.ICCMC-00049.
- [9] W. Ke, C. Wu, Y. Wu, and N. N. Xiong, "A new filter feature selection based on criteria fusion for gene microarray data," *IEEE Access*, vol. 6, pp. 61065–61076, 2018, doi: 10.1109/ACCESS.2018.2873634.
- [10] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Medical Biological Engineering Computing*, vol. 57, no. 1, pp. 159–176, Jan. 2019, doi: 10.1007/s11517-018-1874-4.
- [11] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, Dec. 2010, doi: 10.2174/157489310794072508.
- [12] C. H. Yang, L. Y. Chuang, and C. H. Yang, "IG-GA: a hybrid filter/wrapper method for feature selection of microarray data," *Journal of Medical and Biological Engineering*, vol. 30, no. 1, pp. 23–28, 2010.
- [13] H. Djellali, S. Guessoum, N. Ghoulmi-Zine, and S. Layachi, "Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection," in *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)*, Oct. 2017, pp. 1–6, doi: 10.1109/ICEE-B.2017.8192090.
- [14] H. Alshamlan, G. Badr, and Y. Alohalı, "mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed Research International*, vol. 2015, pp. 1–15, 2015, doi: 10.1155/2015/604910.
- [15] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018, doi: 10.1109/ACCESS.2017.2778268.
- [16] A. K. Shukla, P. Singh, and M. Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," *Swarm and Evolutionary Computation*, vol. 54, p. 100661, May 2020, doi: 10.1016/j.swevo.2020.100661.
- [17] X. H. Han, D. A. Li, and L. Wang, "A hybrid cancer classification model based recursive binary gravitational search algorithm in microarray data," *Procedia Computer Science*, vol. 154, pp. 274–282, 2019, doi: 10.1016/j.procs.2019.06.041.
- [18] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid wrapper TLBO and SA with SVM approach for gene expression data," *Information Sciences*, vol. 503, pp. 238–254, Nov. 2019, doi: 10.1016/j.ins.2019.06.063.
- [19] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Applied Soft Computing*, vol. 62, pp. 203–215, Jan. 2018, doi: 10.1016/j.asoc.2017.09.038.
- [20] C. Arunkumar and S. Ramakrishnan, "Prediction of cancer using customised fuzzy rough machine learning approaches," *Healthcare Technology Letters*, vol. 6, no. 1, pp. 13–18, Feb. 2019, doi: 10.1049/htl.2018.5055.
- [21] A. Dabba, A. Tari, and S. Meftali, "Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2731–2750, Feb. 2021, doi: 10.1007/s12652-020-02434-9.
- [22] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, "Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, no. 3, pp. 288–301, Sep. 2020, doi: 10.1007/s12539-020-00372-w.
- [23] O. A. Alomari, A. T. Khader, M. A. Al-Betar, and Z. A. A. Alyasseri, "A hybrid filter-wrapper gene selection method for cancer classification," in *2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, Jul. 2018, pp. 113–118, doi: 10.1109/ICBAPS.2018.8527392.
- [24] H. M. Alshamlan, "Co-ABC: correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile," *Saudi Journal of Biological Sciences*, vol. 25, no. 5, pp. 895–903, Jul. 2018, doi: 10.1016/j.sjbs.2017.12.012.
- [25] B. Seijo-Pardo, V. Bolón-Canedo, I. Porto-Díaz, and A. Alonso-Betanzos, "Ensemble feature selection for rankings of features," in *Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science()*, Cham: Springer, 2015, pp. 29–42, doi: 10.1007/978-3-319-19222-2_3.
- [26] 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. Advances in Intelligent Systems and Computing*, Singapore: Springer, 2019, pp. 783–792, doi: 10.1007/978-981-10-8055-5_69.
- [27] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, Sep. 2018, doi: 10.1016/j.jbi.2018.07.014.
- [28] O. A. Alomari *et al.*, "Gene selection for microarray data classification based on gray wolf optimizer enhanced with TRIZ-inspired operators," *Knowledge-Based Systems*, vol. 223, p. 107034, Jul. 2021, doi: 10.1016/j.knsys.2021.107034.
- [29] L. Sun, X. Kong, J. Xu, Z. Xue, R. Zhai, and S. Zhang, "A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification," *Scientific Reports*, vol. 9, no. 1, p. 8978, Jun. 2019, doi: 10.1038/s41598-019-45223-x.
- [30] B. Sahu, J. C. Badajena, A. Panigrahi, C. Rout, and S. Sethi, "An intelligence-based health biomarker identification system using microarray analysis," in *Applied Intelligent Decision Making in Machine Learning*, 1st ed., CRC Press, 2020, doi: 10.1201/9781003049548-7.
- [31] H. M. Zawbaa, E. Emary, and B. Parv, "Feature selection based on antlion optimization algorithm," in *2015 Third World Conference on Complex Systems (WCCS)*, Nov. 2015, pp. 1–7, doi: 10.1109/ICoCS.2015.7483317.
- [32] S. Mishra, A. Dash, P. Ranjan, and A. K. Jena, "Enhancing heart disorders prediction with attribute optimization," in *Advances in Electronics, Communication and Computing. ETAERE 2020. Lecture Notes in Electrical Engineering*, Singapore: Springer, 2021, pp. 139–145, doi: 10.1007/978-981-15-8752-8_14.




- [33] C. N. Aher and A. K. Jena, "Rider-chicken optimization dependent recurrent neural network for cancer detection and classification using gene expression data," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 2, pp. 174–191, Mar. 2021, doi: 10.1080/21681163.2020.1830436.
- [34] C. N. Aher and A. K. Jena, "Soft computing based approaches for classifying diseases using medical diagnosis dataset," in *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Mar. 2020, pp. 77–81, doi: 10.1109/ESCI48226.2020.9167518.

BIOGRAPHIES OF AUTHORS






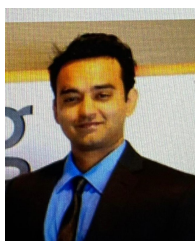
Bibhuprasad Sahu    received his M.Tech degree in Computer Science and Engineering from National Institute of Science and Technology, and his B.Tech in Information Technology from SSIET, Chennai, and pursuing his Ph.D. in the Department of CS&IT, Maharaja Sriram Chandra Bhanja Deo University (MSCB University), formerly North Orissa University (NOU). He is a full-time Assistant professor at the Artificial Intelligence and Data Science graduate program, Vardhaman College of Engineering, Hyderabad. His research lines are the application of evolutionary algorithms for disease diagnosis. He is a senior member of IEEE and member of the ISTE, CSI, IE and IEANG. He is also a reviewer for reputed journals like Springer Nature, Hindawi, and IEEE Access. He published more than 45 research papers in various reputed journals and conferences (Scopus /ESCI). He can be contacted at email: prasadnikhil176@gmail.com.






Amrutanshu Panigrahi    has obtained an M.Tech in information technology from the College of Engineering and Technology, Govt. of Odisha, and B.Tech from BPUT Odisha. He is pursuing his Ph.D. in the Department of Computer Science and Engineering at Siksha 'O' Anusandhan University, Bhubaneswar. He is a member of the ISTE, IEEE, CSI and IEANG. He is also acting as a reviewer for a reputed journal like WILEY. He published more than 25 research papers in various reputed journals and conferences (Scopus /ESCI). He can be contacted at email: amrutansup89@gmail.com.






Bibhu Dash    is a lead solutions architect-data analytics in a Fortune 100 financial organization in Madison, WI. He completed his Ph.D. in information technology from the University of the Cumberlands, KY, USA. Bibhu has also completed his master of engineering in electronics and communication Eng. and MBA from Illinois State University, Normal, IL. Bibhu's research interests include AI, NLP, IoT, cloud computing, big data, and blockchain technologies. He can be contacted at email: bdash6007@ucumberlands.edu.



Pawan Kumar Sharma    is a staff product manager for Walmart in San Bruno, California. He is currently on his Ph.D. in information technology at the University of the Cumberlands, Kentucky. Pawan kumar completed his master of science in management information systems from the University of Nebraska at Omaha in 2015. He also holds another master of science in information systems security from the University of the Cumberlands, Kentucky, and graduated in 2020. His research interests are cyber security, artificial intelligence, retail analytics, and cloud computing. He can be contacted at email: psharma8877@ucumberlands.edu.



Abhilash Pati    is currently working as an assistant professor in the Department of Computer Science and Engineering at GITA Autonomous College, Bhubaneswar, Odisha, and a Ph.D. research scholar in the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Odisha, India. He completed his B.Tech. and M.Tech. in Computer Science and Engineering with Biju Patnaik University of Technology, Odisha, India in 2009 and 2012 respectively. His research interests include the internet of things, fog computing, machine learning, and deep learning, and he has more than 20 papers to his account. He can be contacted at email: er.abhilash.pati@gmail.com.