Microarray Cancer Gene Selection using Swarm Intelligence: An Ensemble Approach

Isuwa Jeremiah
Department of Computer Science
Federal University of Kashere
Gombe Nigeria
isuwajeremiah@gmail.com
ORCID

Adamu Shamsuddeen
Iya Abubakar Institute of ICT
Ahmadu Bello University
Zaria, Nigeria
shamsu200@yahoo.com

Dima M. Rosemary
Department of Software Engineering,
Faculty of Computing
Federal University Dutsinma
Katsina Nigeria
rocinta976@gmail.com

Sada Wasilah
National Mathematical Center
Kwali, Abuja
wasilahsada@gmail.com

Musa Halima
Iya Abubakar Institute of ICT
Ahmadu Bello University
Zaria, Nigeria
musah@abu.edu.ng

***Abstract*— Genes are essential biomarkers in determining biological conditions, including drug response and identifying diseases like brain, ovarian, and colon cancer. However, many of these genes are not directly associated with the target disease, leading to unnecessary difficulties during data analysis such as computational costs and poor learning performance. To address these issues, various techniques have been developed to select only a subset of relevant genes that aid in interpreting the target disease. Swarm Intelligence (SI) algorithms, such as Salp Swarm Optimization (SSA), have been successful in optimizing complex problems with a large search space, making them suitable for cancer gene datasets. To reduce the high dimensions of these datasets, hybridizing SI algorithms with filter methods for preliminary gene selection is crucial. Furthermore, this approach enhances the classification accuracy of the selected gene subset and minimizes the resources necessary for gene selection. This paper focuses on developing and investigating diverse ensemble microarray cancer gene selection approaches. We combined Chi-Square and Mutual Information filter methods with SSA in multiple singular and dual perspectives, utilizing colon and ovarian cancer datasets. Our findings show that Ensemble_4, which involves combining and re-ranking the top genes selected from each filter method and then passing them to SSA, significantly improves the performance accuracy of cancer gene selection across the datasets.**

Keywords—microarray cancer data, gene selection, salp swarm algorithm, nature-inspired algorithms, filter methods

I. INTRODUCTION

Microarray technology has revolutionized the field of cancer research by enabling the simultaneous analysis of thousands of genes in a single experiment [1], [2]. The ability to assess gene expression patterns on a genome-wide scale has opened up new possibilities for understanding the molecular basis of cancer development and progression [3]. One critical task in microarray analysis is the selection of relevant genes that are associated with cancer, as this helps in identifying potential biomarkers, understanding disease mechanisms, and developing targeted therapies.

Microarray cancer gene selection involves the identification of a subset of genes from the vast pool of gene expression data that are most informative for distinguishing between cancerous and normal tissue samples [4], [5]. The overarching goal of microarray cancer gene selection is to find a minimal set of genes that maximizes the discriminatory power between cancerous and normal samples [6]. The selected genes hold the key to unraveling the underlying molecular mechanisms of cancer, as they represent potential biomarkers [7], [8]. By focusing on these specific genes, researchers can gain insights into the dysregulated cellular pathways and molecular networks driving cancer progression. Moreover, the identified genes can serve as diagnostic tools for early cancer detection, prognostic indicators for patient outcomes, or predictive markers for treatment response [9], [10]. This is typically achieved by applying various feature selection algorithms that assess the relevance and redundancy of genes based on their expression patterns across different samples.

Traditional gene-by-gene approaches to gene selection are time-consuming and often fail to capture the complexity and interactions of genes [3]. Therefore, sophisticated computational methods have been introduced over the years to expedite and enhance the gene selection process. These algorithms take into account statistical measures also called filter methods, such as t-tests, and correlation coefficients, as well as machine learning techniques, such as LASSO regression [11]. Furthermore, Nature-Inspired Algorithms (NIAs) have over again revolutionized the field of microarray gene selection, providing powerful computational tools that mimic the principles of natural processes and organisms [12]. These algorithms draw inspiration from the remarkable efficiency and adaptability observed in biological systems, such as genetic evolution, ant colony behavior, and swarm intelligence [13].

These NIAs have transformed microarray gene selection by providing several benefits. Firstly, they provide efficient and effective search strategies that can handle the high-dimensional and noisy nature of gene expression data [14]. Secondly, they overcome the limitations of traditional gene-by-gene approaches by considering the interactions and synergistic effects among genes [14]. Thirdly, these algorithms are capable of exploring vast search spaces and avoiding premature convergence, increasing the chances of discovering optimal or near-optimal gene subsets [15]. Lastly, NIAs often provide robustness and generalizability, enabling the application of gene selection models across different cancer types and datasets [15]. By leveraging these strengths, researchers are empowered to extract meaningful information from complex gene expression data and unravel the intricate mechanisms underlying cancer.

One of the most rapidly growing NIA in microarray gene selection is the swarm intelligence Salp Swarm Algorithm (SSA) [16]. Inspired by the collective intelligent movement and information-sharing behaviors of salps i.e., marine organisms, SSA offers a unique approach to solving complex optimization problems, including gene subset selection in microarray data analysis [1]. The SSA algorithm utilizes these intelligent behaviors to iteratively update the positions of candidate solutions in the search space i.e., salps, gradually converging towards gene subsets that maximize the classification accuracy or other defined criteria. The algorithm has shown promising results in identifying informative genes associated with cancer and has contributed to the advancement of cancer research [17].

Although the NIAs have demonstrated improved performance in microarray gene selection, hybridizing them with filter methods can further enhance their effectiveness and provide even better results [18], [19]. Filter methods are a class of feature selection techniques that offer various statistical and information-theoretic measures to evaluate the relevance and redundancy of individual genes independently of the classification model [20]. By combining the strengths of NIAs such as the SSA and filter methods, researchers can achieve a more comprehensive and robust gene selection process, leading to enhanced performance and more reliable gene subsets associated with cancer [14], [21].

One of the main reasons to hybridize SSA with filter methods is to address the issue of redundancy among selected genes. Although SSA explores the search space efficiently, it may still produce gene subsets that contain redundant genes which can lead to overfitting [22]. By incorporating filter methods into the gene selection process, redundant genes can be identified and eliminated, resulting in a more concise and informative gene subset [23]. Additionally, hybridizing SSA with filter methods can enhance the interpretability of the selected gene subsets by providing a rank or score for each gene, indicating its importance [24], [25]. This ranking enables researchers to focus on the most informative genes in subsequent analyses.

Regardless, it is important to note that the choice of the specific filter method for hybridization with SSA or other NIAs such as Particle Swarm Optimization (PSO) or Genetic Algorithms (GA) depends on the characteristics of the dataset and the objectives of the analysis. Commonly used filter methods include information gain [26], chi-square [27], mutual information [28], and correlation-based approaches. Each of the filter methods has its strengths and limitations, and the hybrid algorithm can be tailored to the specific requirements of the study.

Thus, the exact goal of this study is to develop diverse microarray gene selection approaches from the combination of SSA with two popular filter methods i.e., the Chi-Square and Mutual Information, to achieve an improved classification accuracy as well as a reduced and highly informative gene subset. The combination is performed in multiple faces ranging from singular (one-to-one) to dual (two-to-one) perspectives. To achieve this overall goal, specifically, we:

- Performed an extensive and recent literal study on the hybridizing of SSA with specific filter methods for microarray gene cancer selection.
- Develop four different ensemble cancer gene selection approaches from both singular and dual combinations of the SSA with Chi-Square and Mutual Information filter methods.
- Finally, an extensive experimental study and comparison amongst the developed ensemble approaches as well as other optimization algorithms using the colon and ovarian microarray cancer gene datasets was performed.

The remainder of this paper is organized as follows: Section 2 offers a brief overview of the field's background knowledge as well as discussions of current literal works. The proposed ensemble approaches are described in detail in Section 3. Section 4 includes the experimental results as well as an analysis of the findings. Finally, Section 5 brings the work to a close.

II. BACKGROUND

This section provides an overview as well as discussions of current literal works on key terms used in this study, such as microarray data, SSA, and the gene selection process.

A. Microarray Technology

Microarray technology involves the deposition of thousands of DNA or RNA probes, representing specific genes or sequences, onto a solid surface such as a glass slide or silicon chip [6]. These probes are then hybridized with labeled complementary target molecules derived from the cancer samples [6].

By measuring the intensity of the labeled molecules bound to each probe, researchers can determine the expression level of the corresponding gene in the cancer cells [29]. This process is also termed a microarray experiment.

Figure 1 depicts the three general stages involved in obtaining the gene expression profile from a DNA microarray.

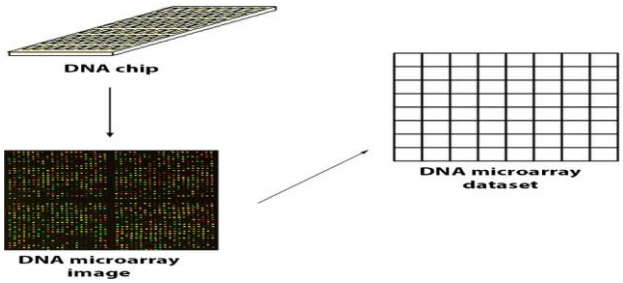

Fig. 1: General stages in gene expression profile generation [29]

The three general processes of obtaining the gene expression profile from a DNA microarray as depicted in Figure 1 involve the design and spotting of DNA probes onto a DNA chip, the scanning of the chip to generate a DNA microarray image, and subsequent image processing and data analysis to derive insights from the resulting DNA microarray dataset [29].

B. Microarray Cancer Data

Microarray cancer data refers to the collection of gene expression profiles obtained from microarray experiments performed on cancer samples [4]. They provide valuable information about the gene expression patterns associated with different types of cancer, tumor subtypes, and stages of disease progression. The data can be used to identify genes that are differentially expressed between cancerous and normal cells, as well as to characterize the molecular mechanisms underlying tumor development and progression [4].

C. Microarray Gene Selection

Microarray gene selection refers to the process of identifying a subset of genes that are most relevant and informative for a particular analysis or research question from a larger set of genes represented on a microarray [30]. The goal is to reduce the dimensionality of the data and focus on a smaller set of genes that have the greatest potential for providing meaningful insights. The two main approaches to microarray gene selection are filter-based methods and NIAs [31].

1) Filter methods

Filter-based methods use statistical or computational filters to rank genes based on certain criteria and select the top-ranked genes for further analysis [32]. Filter-based gene selection techniques are further classified as either univariate or multivariate [33]. With the univariate method, individual feature evaluations are performed to estimate the strength of the relationship between the feature and the underlying target disease. Chi-Square, Mutual information (MI), and information gain (IG) are some of the common univariate filter methods. On the contrary, multivariate methods are capable of simultaneously considering multiple features [28]. They take into account the interrelationships among these features, ranking them as groups rather than individually as seen in univariate analysis. Two of its most popular techniques are the minimum redundancy maximum relevance (mRmR) [34], and the fast correlation-based filter (FCBF) [35].

- Chi-Square: The chi-square statistic, as described by Alrefai & Ibrahim, [3], quantifies the relationship or independence between two categorical variables. Specifically, it assesses the deviation from the expected distribution under the assumption that the feature under consideration is independent of the class label. Ghosh et al., [37]presented a mathematical expression of the chi-square test, wherein the expected range is divided into intervals. The chi-square (X^2) value of feature f is calculated as follows:

$$X^2 f = \sum_{j=1}^{r}\sum_{s=1}^{c}\frac{(n_{js} - \mu_{js})^2}{\mu_{js}} \qquad (1)$$

where r denotes the count of dissimilar values present in the feature, c represents the number of dissimilar values within a class, n_{js} represents the frequency of j^{th} element with s^{th} class and $\mu_{js} = \frac{n_{*s}n_{j*}}{n}$, n_{j*} frequency of j^{th} element, n_{*s} is the total number of elements present in the s^{th} class. To summarize, larger chi-square values indicate a greater degree of significance or importance.

The Chi-Square filter method has several strengths that make it a popular choice for feature selection in various data-driven domains such as microarray analysis. Some of its key strengths include its ease of interpretability, nonparametric nature, Robust to outliers, simplicity, and efficiency [27].

- Mutual Information (MI): MI is a metric that quantifies the relationship between two random variables, namely X (the feature) and Z (the associated class label), as explained by Ayham et al., [38]. It is derived from Shannon entropy, which aims to estimate the level of uncertainty within the distribution of events related to feature X. When a particular event occurs more frequently, the entropy is lower, indicating reduced uncertainty. Conversely, when all events have equal chances of occurring, the entropy is at its highest level, signifying the absence of certainty regarding any specific outcome. Zhou et al., [28] defined MI mathematically as:

$$I(x;y)\ \sum_{i=1}^{n}\sum_{j=1}^{n} p\big(x(i),y(j)\big).\log\left(\frac{p(x(i),y(j))}{p(x(i)).p(y(j))}\right) \qquad (2)$$

In situations where the MI is zero, x and y are statistically independent, i.e., $p\big(x(i),y(j)\big) = p\big(x(i)\big).p\big(y(j)\big)$. Equations 3 and 4 demonstrate the linear relationship between MI and the entropies of the variables. In Figure 2, a Venn diagram is utilized to visually depict the relationships between these variables [28]

$$I(x;y) = \begin{cases} H(x) - H(x|y) \\ H(y) - H(y|x) \\ H(x) + H(y) - H(x,y). \end{cases} \quad (3)$$

Let's consider z as a discrete random variable. To measure its interaction with the other two variables, x and y, we can

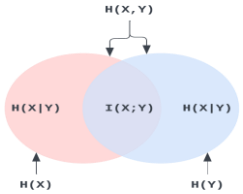

Fig. 2. Venn diagram depicting the relationship between MI and the entropies of the variables [28]

employ conditional MI. This measure is defined as follows:

$$I(x;y|z) = \sum_{i=1}^{n} p(z(i)) I(x;y|z = z(i)) \quad (4)$$

where $I(x;y|z = z(i))$ is the MI between x and y when $z = z(i)$. The Conditional MI allows the information between two variables to be measured in the context of a third, but the information among the three variables cannot be measured.

The Mutual Information (MI) filter method also possesses several strengths that contribute to its effectiveness in feature selection. Some of these key strengths include its ability to capture nonlinear relationships, no assumptions of linearity or distribution, and independence on label labels, among others [28].

2) Nature-inspired algorithms (NIAs)

Nature-inspired algorithms, also known as metaheuristic algorithms, draw inspiration from natural phenomena to solve complex optimization problems. These algorithms explore large solution spaces to identify optimal or near-optimal subsets of genes. Examples of long-standing and widely used NIAs for gene selection include Genetic Algorithms (GA)[39] that simulate the process of natural selection, including selection, crossover, and mutation, to evolve a population of gene subsets. Particle Swarm Optimization (PSO)[40] mimics the behavior of a swarm of particles moving in a multidimensional search space where each particle represents a potential solution, and their movements are guided by the best-performing particles in the swarm. Ant Colony Optimization (ACO)[13] simulates the foraging behavior of ants to identify promising gene subsets. Each Ant represents a potential solution, and pheromone trails guide their search. Brezočnik et al., [41] in their work extensively classify NIAs based on specific characteristics, providing a more comprehensive understanding of these algorithms.

For all NIAs, fitness functions evaluate the performance of each gene subset, and over multiple generations, the algorithm converges towards a high-performing subset. Figure 3 illustrates the process of NIAs, while Figure 4 showcases additional NIAs that were not covered in this study.

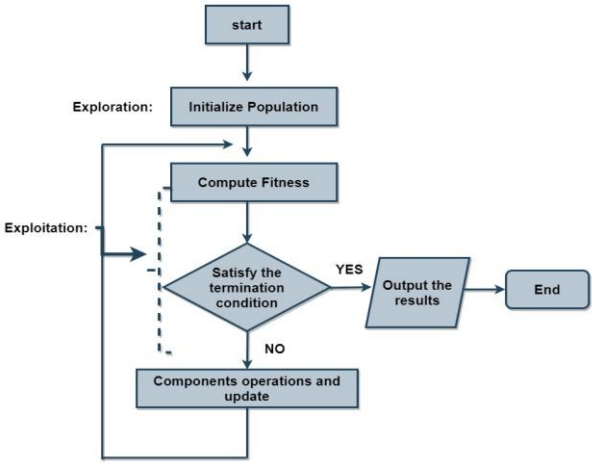

Fig. 3: Procedure for nature-inspired algorithms [42]

NIAs have the advantage of exploring a wide range of possible solutions and can handle complex relationships between genes. However, they may require more computational

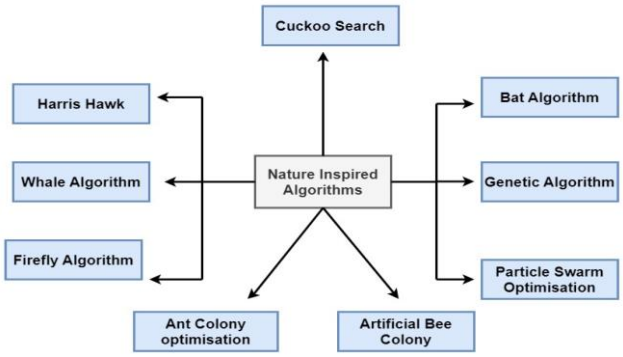

Fig. 4: Other popular nature-inspired algorithms [42]

resources and parameter tuning compared to filter-based methods [42].

- Salp swarm algorithm (SSA): Salps, which are marine organisms with a cylindrical body shape, are known to inhabit the ocean in groups or chains while intentionally floating. Existing literature confirms that this collective movement of Salp swarms enables them to achieve better locomotion and locate food more effectively [43]. The concept of SSA, introduced by Mirjalili et al., [43], has been utilized to address various optimization problems. In SSA, all salps form a chain-like structure to navigate the search space and locate the target or food source [1]. To mathematically represent the chain formation of salps, the swarm is divided into two parts: the leader and the

followers. The leading Salp always assumes the front position and guides the other followers within the chain.

Let's consider a collection of n Salps denoted as $Y = \{Y_1, Y_2, \ldots, Y_i, \ldots . Y_n\}$, where each Salp is represented by a d-dimensional vector $(Y_i = y_1, y_2, \ldots, y_d)$. The target vector or food source is denoted as F_s. The position of the leader Salp is updated according to (5).

$$Y_1 \begin{cases} F_s + \alpha 1\big((Y_{max} - Y_{min})\alpha 2 + Y_{\min}\big)\alpha 3 \geq 0 \\ F_s - \alpha 1\big((Y_{max} - Y_{min})\alpha 2 + Y_{min}\big)\alpha 3 < 0 \end{cases} \quad (5)$$

In this context, we have random values, $\alpha 1, \alpha 2$, and $\alpha 3$, and Y_1 represents the location of the leading Salp. Y_{max}, and Y_{min} denote the upper and lower boundaries for each Salp, respectively. In SSA, the balance between exploration and exploitation is regulated by $\alpha 1$, and its value is updated in each cycle using (6).

$$\alpha 1 = 2e^{-(\frac{4 * c_{iter}}{Max_{iter}})^2} \quad (6)$$

Here, we have the variables Max_{iter} and c_{iter}, which represent the total number of iterations and the current iteration, respectively. The positions of the follower Salps, excluding Y_1, are enhanced using Newton's law of motion, as described in (7).

$$Y_j(i) = \frac{1}{2} a t^2 + v_o t \quad (7)$$

In this scenario, we have j ranging from $2\ to\ n$, and $Y_j(i)$ represents the i^{th} dimension of the j^{th} Salp. v_o, t, and a correspond to the initial velocity, time, and acceleration, respectively, which are calculated using (8).

$$a = \frac{v_{end}}{v_o}\ where\ v = \frac{y - y_o}{t} \quad (8)$$

In the context of optimization problems, the term "time" represents the iteration count, and the initial velocity is set to 0. Consequently, the positions of the followers are updated using a modified equation provided in (9).

$$Y_j(i) = \frac{1}{2}\ (Y_j(i) + Y_{J-1}(i) \quad (9)$$

The step-by-step procedure of the SSA is described in Algorithm 1. In SSA, the initial population is randomly generated, and the Salp that best fits the problem is considered as the food source. The remaining Salps then move towards this food source. The position of the food source (F_s) is updated in each iteration.

D. Related Literary Works

Hybridization of the NIAs with filter methods for microarray cancer gene selection is an area of research that aims to combine the advantages of both approaches to improve the effectiveness and efficiency of gene selection in cancer-related studies. Several related works have explored this hybridization approach, offering valuable insights and advancements in the field. By hybridizing SSA with filter methods, researchers aim to leverage the exploration capabilities of SSA and the gene ranking abilities of filter methods to improve the accuracy and efficiency of microarray cancer gene microarray cancer microarray selection. The hybridization approach involves combining the SSA with one or more filter methods, as well as integrating one or more other NIAs with SSA and filter methods. This approach combines the strengths of SSA, filter methods, and additional NIAs to enhance the gene selection process for microarray analysis.

Algorithms 1: Salp Swarm Algorithm [17]

Input: Generate the initial population (Y_i), where $i = 1,2,\ldots,N$ randomly
Output: The solution vector F_s, fittest solution

```
While stopping criteria do
    Compute Salp's fitness and find the current best
    Calculate α1 using Eq. (6)
    For each Salp (Xj) In the chain do
        if (j == 1) then
            Modify the position of leader Salp using Eq. (5)
        else
            Modify locations of followers Salps according to Eq. (9)
        End if
    End for
End while
Return the fittest solution Fs
```

For example, in their study, Prabhakar & Lee, [26] introduced a transformation-based Tri-level feature selection approach that utilizes wavelets for prostate cancer classification. The researchers employed four distinct filter methods, namely Relief-F, Fisher's Score, Information Gain, and SNR, to conduct an initial gene selection process. Subsequently, the SSA was applied in the final phase of gene selection. This hybrid approach combines the advantages of filter methods and SSA to enhance the accuracy and effectiveness of prostate cancer classification. Sharma & Rani, [44] employed a gene selection method based on Fisher-Score-SSA-SHO and evaluated it using seven diverse microarray cancer gene datasets. They harnessed the strengths of both the Spotted Hyena Optimizer (SHO) and the SSA within a multi-objective framework. To address the high dimensionality of the microarray cancer datasets, they first utilized the Fisher-Score filter method for preliminary feature selection. Subsequently, the hybridization of SSA and SHO was employed for further gene selection and optimization. Shekhawat et al., [1] employed Principal Component Analysis (PCA) and fast Independent Component Analysis (fastICA) as dimensionality reduction methods, along with the SSA. The authors utilized PCA and fastICA to transform the microarray cancer data from a higher to a lower dimension. This reduced-dimensional data was then fed into SSA for identifying the most relevant features. Qin et al., [17] presented a gene selection framework that incorporates the Minimum Redundancy Maximum Relevance (mRmR) filter method as the initial stage of feature selection. Following the mRmR-based selection, the chosen subset of genes is then fed into an enhanced version of the SSA to dynamically refine and select the final set of the most

informative features. This two-stage approach combines the strengths of the mRmR filter method and the improved SSA to optimize gene selection and enhance the identification of relevant and informative features.

In addition to SSA, several other NIAs have been explored in combination with filter methods. For instance, Alrefai & Ibrahim, [36] investigated the use of Particle Swarm Optimization (PSO) in conjunction with filter methods such as Chi-Square, Information Gain, and Relief. Similarly, Azadifar & Ahmadi, [45] employed PSO with Fisher-Score, while Ayham et al., [38] utilized PSO with Mutual Information for gene selection. In addition to PSO, other NIAs, including the Jaya Algorithm (JA), Ant Colony Optimization (ACO), and Moth Flame Optimization (MFO), have also been combined with filter methods in related works. Baliarsingh et al., and Chaudhuri & Sahu, [46], [47] explored the combination of JA with filter methods. Ke et al., Ma et al., and Meenachi & Ramakrishnan, [13], [48] utilized ACO in combination with filter methods. Furthermore, Dabba, Tari, & Meftali, ; Dabba, Tari, Meftali, et al., and Nadimi-Shahraki et al., [12], [50], [51] investigated the integration of MFO with filter methods for microarray cancer gene selection.

III. PROPOSED ENSEMBLE GENE SELECTION APPROACH

In this section, we present the proposed ensemble approaches for microarray cancer gene selection. These methods as mentioned in earlier sections are composed of both singular (One-to-one) and dual (two-to-one) perspectives.

A. Singular Perspective (One-to-One)

The ensemble approach from a singular perspective involves combining the SSA with a specific filter method i.e., Chi-Square or Mutual Information. This leads to two ensembled approaches, namely Ensemble_1 (Chi-Square × SSA) and Ensemble_2 (Mutual Information × SSA). Figure 5 illustrates the general structure of the one-to-one perspective for these ensembled approaches. The Chi-Square filter method assesses the dependence between features and the target variable based on statistical significance. By incorporating SSA, the feature selection process can be further enhanced by considering the collective behavior of the Salp swarm. The SSA can provide additional insights into the relevance of features based on their interactions and dependencies within the feature subset.

Similarly, combining the SSA with the Mutual Information filter method offers advantages in terms of enhanced feature relevance assessment, detection of nonlinear relationships, efficient exploration and exploitation, and handling of high-dimensional data among others. This combination can improve the overall effectiveness and efficiency of feature selection, particularly in scenarios where nonlinear associations and high-dimensional data are involved.

As depicted in Figure 5, the process consists of two stages. The initial stage focuses on employing the specific filter method to obtain a smaller set of genes, which serves as a preliminary step before proceeding to the SSA. Additionally, it is important to highlight that only the top 100 performing genes are selected

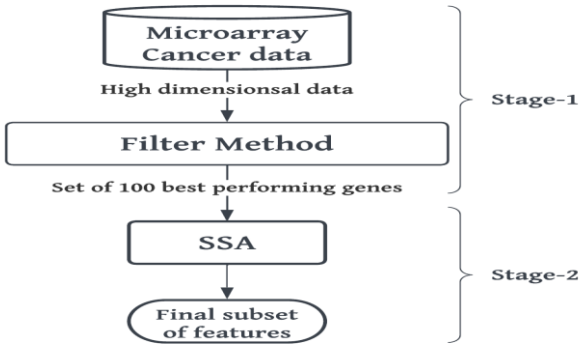

Fig. 5: General Structure of the One-to-One Perspective

and forwarded to the SSA for the final stage of gene selection. The decision to choose solely the top 100 genes aligns with the findings of Almugren & Alshamlan, [52], [53]. Their work involved conducting experiments using different gene subsets of 100, 200, 300, and 500 genes. Remarkably, the 100-gene subsets consistently demonstrated superior performance, leading to the rationale behind this selection approach.

B. Dual Perspective (Two-to-One)

The ensemble approach from a dual perspective involves combining both filter methods, i.e., Chi-Square and Mutual Information, simultaneously with the SSA. However, this combination is carried out using two distinct strategies.

- *First Strategy (((ChiSquare - MutualInfo) × SSA), Ensemble_3)*

The strategy referred to as Ensemble_3 involves a sequential process. Initially, the Chi-Square filter method is applied as the first stage to filter out less relevant genes. The selected subset of genes is then further subjected to the Mutual Information filter method for a second round of selection. Finally, the gene subset obtained from Mutual Information is passed on to the SSA for the final selection. In this approach, the features are passed sequentially from Chi-Square to Mutual Information,

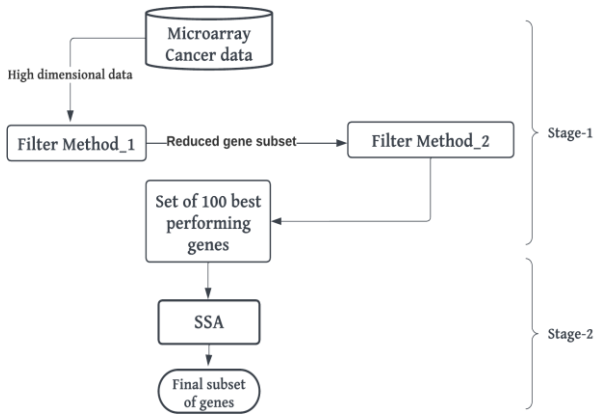

Fig. 6: General Structure of the two-to-One Perspective of Ensemble_3

indicated by the hyphen sign ($'-'$). The structure of Ensemble_3 is depicted in Figure 6.

- *Second Strategy (((ChiSquare + MutualInfo) × SSA), Ensemble_4)*

The strategy known as Ensemble_4, alternatively referred to as the 'Combination Approach', incorporates the simultaneous selection of top genes through the utilization of both Chi-Square and Mutual Information methods. By combining these two techniques, a set of relevant genes is formed. It is important to note that the addition sign '+' between Chi-Square and Mutual Information signifies that the features from both methods are merged concurrently, rather than being passed sequentially as in Ensemble_3. The structure of Ensemble_4 is illustrated in Figure 7.

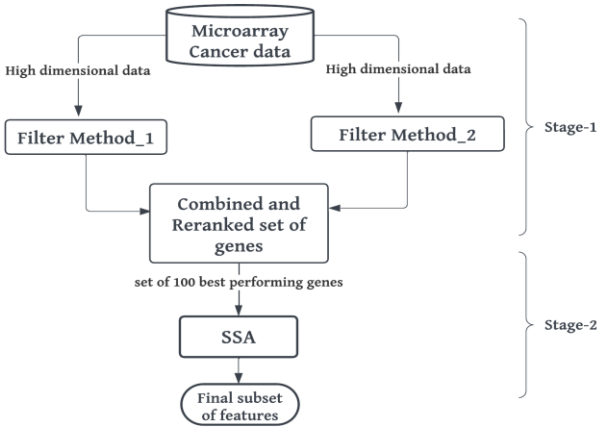

Figure 7: General Structure of the two-to-One Perspective of Ensemble_4

It is crucial to emphasize that in this strategy, each filter method is limited to selecting 200 genes with outstanding performance. However, since the filter methods inherently rank the genes based on their scores, only a unique set of 100 top-performing genes was chosen during the combination stage. This stage involved merging the gene sets from each filter method and removing any potential duplicates. Ultimately, only the top 100 genes were retained. These selected genes were subsequently passed to the next phase of selection, known as the SSA, to undergo the final stage of gene selection.

C. Datasets Description

In this study, we assess the effectiveness of the developed ensembled gene selection techniques using two microarray datasets of different sizes. The performance of the developed techniques is compared based on criteria such as classification accuracy, standard deviation, and the number of selected genes. The chosen datasets are the Colon and Ovarian cancer datasets, which have been widely utilized in the literature and serve as a common benchmark for evaluating competing algorithms. The datasets can be accessed publicly on the website https://figshare.com/. Table 1 provides a summary of the dataset characteristics, including an additional notation (x, y) for the class to indicate the distribution of instances in each class. Here, $'x'$ denotes the number of instances in the positive class (tumor or malignant), and $'y'$ represents the number of instances in the negative class (normal or benign).

TABLE 1. OVERVIEW OF THE MICROARRAY CANCER DATASETS USED

Datasets	Instances	Features	Class
Colon Cancer	62	2000	2 (40,22)
Ovarian Cancer	253	15154	2 (162,91)

It is worth noting that the datasets underwent minimal preprocessing, which involved assigning numeric labels to the class and normalizing the data before training the learning algorithm. This preprocessing step was implemented to prevent any bias towards particular genes, enhance data interpretability, and optimize the performance of learning algorithms, especially those utilizing distance metrics such as KNN [50].

D. Learning Algorithm

In this study, the KNN (K-Nearest Neighbors) machine learning algorithm is employed to continuously assess the effectiveness of selected gene subsets obtained from the developed optimization algorithms. Its fundamental principle involves classifying sample points based on the most frequently represented class among their nearest neighbors. In cases where there are multiple such classes, the minimum average distance is utilized [54]. The performance of the KNN algorithm is influenced by the choice of the *K* parameter, which determines the number of neighbors considered during classification [24].

IV. RESULT PRESENTATION AND ANALYSIS

A. Experimental Setup

For all experiments conducted, the Jupyter Notebook platform with Python version 3.10 was utilized. The experiments were carried out on a computer featuring an Intel(R) Core (TM) i7-6600U processor operating at a frequency of 2.60GHz and equipped with 8GB of RAM.

Concerning the developed algorithms, the random initialization technique was employed in the SSA to generate the initial population. This approach was chosen to mitigate potential bias that can arise from using a predetermined initialization method, ensuring a more unbiased and exploratory search process [55]. Additionally, the SSA was adopted in a single-objective formulation and a binary representation. The objective was to simultaneously optimize classification accuracy and the number of selected genes. To achieve this, a Transfer Function (TF) was employed to convert continuous values into binary forms that indicate the selection status of genes. Particularly, the sigmoid function [56], a member of the S-shape family, was chosen as the threshold function to effectively differentiate between selected and unselected genes. The sigmoid function, known for its advantages such as continuity, differentiability, interpretability in terms of probability or likelihood, and robustness to noise and outliers [57], proved to be a suitable choice for this optimization task.

In feature selection, the main aim is to strike a balance between maximizing classification accuracy and minimizing the number of selected features. To quantitatively evaluate the quality of candidate solutions, a fitness function is employed. This mathematical representation assesses the suitability of solutions in optimization problems, measuring how well a solution aligns with the objectives of the task. By assigning a fitness score to each candidate solution, the optimization algorithm is guided toward finding improved solutions. In this study, the performance of the SSA in feature selection was evaluated using the fitness function described in (10).

$$fitness\ function = \alpha\Delta_R(D) + \beta\frac{|Y|}{|T|} \quad (10)$$

where $\Delta_R(D)$ denotes the classifier's error rate. $|Y|$, size of the selected gene subset, and $|T|$, the total number of genes in the dataset. The $'\alpha' \in [0,1]$ and represents the weight of the classifier's error rate. $'\beta' = (1-\alpha)$ represents the significance of the gene reduction.

B. Evaluation Metrics and Hyperparameter Settings

The performance evaluation of the proposed gene selection methods considers the following performance metrics:

- Mean classification accuracy: This metric calculates the average fitness function results (classification accuracy) by running the algorithm $'P'$ number of times.
- F1 Score: This metric calculates a model's performance by considering both precision and recall, enabling a balanced assessment of its predictive capabilities.
- Mean number of selected genes: This metric computes the average number of selected genes after running the algorithm $'P'$ times.

In addition to these evaluation measures, the Standard Deviation (SD) of the competing methods will be presented, along with the results of the T-test [58]. These additional analyses aim to showcase the stability of the competing methods and assess any statistical differences among them.

The performance of NIAs and learning algorithms such as the KNN is influenced by hyperparameters. These hyperparameters are configuration settings that are adjusted before running the algorithm on a specific problem. Selecting appropriate values for these hyperparameters can have a significant impact on the quality of the algorithm's solutions. For instance, in PSO, in addition to parameters like swarm size and the number of generations, there are settings such as acceleration coefficients that guide the particles toward their personal or global best positions. These settings play a crucial role in ensuring effective exploration of the search space and convergence toward optimal solutions.

Table 2 contains an exhaustive list of the hyperparameters used in this study, along with their values. It is important to highlight that the SSA does not possess any specific hyperparameters apart from the general swarm size and the maximum number of iterations. These hyperparameters, along with the fitness function, K value in KNN, and the number of runs $'P'$, were determined based on intuitive considerations.

TABLE II. PARAMETER SETTINGS AND VALUES USED

	Hyperparameters	Parameter values
General	Swarm Size	20
	Number of generation/iterations	20
	Train-Test Split	70-30
	$'\alpha'$	0.99
	$'\beta'$	0.01
	Dimension (D)	Gene count in datasets
	K value in KNN	5
	Number of runs (P)	10
SSA	None	None

C. Results and Discussions

In this study, the average classification accuracy, average F1 score, and the average number of selected genes were used as performance evaluation criteria. The SD of each result is provided to demonstrate the variability across the ten independent runs. Results are presented using the notation $x \pm y$, where $'x'$ represents the average classification accuracy $(in\ \%)$, and $' \pm y'$ represents the SD value. In the evaluation, boldface and underlined results indicate the best performance from each experiment. While feature selection is typically considered a multiobjective optimization problem that simultaneously considers both accuracy and the number of selected genes, this study prioritizes accuracy. This decision is motivated by the criticality of accurate predictive models in life-dependent applications such as healthcare. In applications like cancer diagnosis, accurate predictions play a crucial role in early detection and intervention, ultimately leading to improved patient outcomes. Therefore, in identifying a superior gene selection method in this study, higher accuracy takes precedence over the number of selected genes. Even if a method has significantly fewer selected genes, it will only be considered superior if it achieves better accuracy.

- *Results and Observations from the Colon Cancer Dataset*

Table 3 presents the results obtained from the analysis of the Colon cancer dataset, showcasing the number of selected genes, classification accuracy, F1-score, and additional metrics including SD, precision, and recall. The results highlight the effectiveness of the Ensemble_4 approach, also known as the 'combination approach,' which combines Chi-Square and Mutual Information methods for simultaneous gene selection. This approach demonstrates superior performance across all measures, achieving high accuracy of 94.20%, a low SD value of ±0.0539, and selecting only 16 genes. To capture the trade-off between false positives and false negatives and provide a comprehensive metric for comparison, the F1-score is utilized. Once again, the Ensemble_4 approach outperforms other competing methods, attaining an impressive F1-score of 93.60%.

Overall, the Ensemble_4 approach proves to be highly effective in the gene selection process, delivering accurate results with minimal gene selection. It consistently demonstrates superior performance across various evaluation measures,

highlighting its efficacy in the context of the Colon cancer dataset.

TABLE III. RESULTS OBTAINED FROM THE COLON CANCER DATASET

Dataset	Algorithms	No. Genes	Accuracy & SD	Precision	Recall	F1-Score
Colon Cancer	Ensemble_1 ChiSquare×SSA	21	90.90 ± 0.0384	92.90	89.00	89.90±0.0420
	Ensemble_2 MutualInfo×SSA	25	82.50 ±0.0474	87.00	77.70	78.90±0.0755
	Ensemble_3 ((ChiSquare - MutualInfo) +SSA)	21	84.50±0.0369	89.00	80.10	81.90±0.0504
	Ensemble_4 ((ChiSquare + MutualInfo) +SSA)	**16**	**94.20±0.0539**	**94.40**	**93.80**	**93.60±0.0578**

Additionally, to determine whether the observed differences between the performance of the competing gene selection methods are likely due to actual effects or merely the result of random chance, the T-test statistical test is employed. By conducting a t-test, we can evaluate whether the differences observed in performance metrics, such as accuracy or other evaluation measures, are statistically significant. The t-test compares the means of two groups and calculates the probability that the observed differences between them occurred by chance. The significance of the t-test results provides insights into the reliability and validity of the observed differences. If the p-value obtained from the t-test is below a predetermined threshold (commonly 0.05), it suggests that the differences between the groups are statistically significant. This implies that the observed variations in performance metrics are likely due to the influence of different factors or conditions being compared.

In the context of the Colon cancer dataset, Table 4 presents the p-values obtained from Ensemble_4 and Ensemble_1 for classification accuracy and the F1-score, respectively. The inclusion of Ensemble_1 in the comparison is motivated by its competitive performance in certain evaluation measures. Using a predetermined critical value of 0.05, the recorded p-values indicate that there exists a statistically significant difference between the mean results of the two methods for both measures, as both yield p-values < 0.05.

TABLE IV. T-TEST RESULTS ON THE ACCURACY AND F1 SCORE USING THE TOP TWO METHODS FOR THE COLON CANCER DATASET

Measures	Competing algorithms	P_values
Accuracy	Ensemble_1 Vs. Ensemble_4	0.0002466
F1-Score	Ensemble_1 Vs. Ensemble_4	0.0001426

- *Results and Observations from the Ovarian Cancer Dataset*

Table 5 presents the results obtained from analyzing the Ovarian cancer dataset, including the number of selected genes, classification accuracy, F1-score, and additional metrics such as SD, precision, and recall. Once again, the results underscore the effectiveness of the Ensemble_4 approach. This method exhibits exceptional performance in classification accuracy, achieving an impressive score of 98.00%, with a low SD value of ±0.0156, and selecting a mere 18 genes. To account for the imbalanced class distributions and the trade-off between false positives and false negatives in the Ovarian cancer dataset, the F1 score is employed. Although the Ensemble_4 method does not secure the highest F1-score, it attains a highly competitive result of 93.70% compared to ensemble_2, which achieves an F1-score of 96.00%.

Overall, the Ensemble_4 approach once again demonstrates its remarkable efficacy in the gene selection process, yielding accurate outcomes with minimal gene selection. It consistently outperforms other methods across various evaluation measures, affirming its effectiveness in addressing the challenges posed by the Ovarian cancer dataset.

Table 6 displays the p-values derived from Ensemble_4 and Ensemble_2 for classification accuracy and the F1-score, respectively. The inclusion of ensemble_2 in the comparison is as well driven by its competitive performance in specific evaluation measures. The obtained p-values indicate a statistically significant difference between the mean results of the two methods for both measures. In both cases, the calculated p-values are less than 0.05, indicating statistically significant differences.

However, despite the outstanding performance exhibited by the Ensemble_4 method in the microarray cancer datasets, it is important to acknowledge the significance of selecting the appropriate optimization algorithm for a specific problem. The No Free Lunch (NFL) theorem, proposed by Wolpert & Macready, [59], emphasizes this aspect and has profound implications in the fields of optimization and machine learning. The NFL theorem posits that no single optimization algorithm can universally outperform all others across all possible optimization problems. This theorem underscores the understanding that there is no "one-size-fits-all" solution for optimization problems. The effectiveness of different algorithms can vary depending on various characteristics of the problem, such as the dimensionality of the search space, smoothness of the objective function, presence of constraints, and other factors.

Consequently, the NFL theorem serves as a motivating factor for researchers to continually develop new algorithms or enhance existing ones to address a diverse range of optimization problems, including the challenge of cancer gene selection. The theorem emphasizes the importance of exploring and advancing optimization algorithms to cater to the specific requirements and characteristics of different problem domains.

TABLE V. RESULTS OBTAINED FROM THE OVARIAN CANCER DATASET

Dataset	Algorithms	No. Features	Accuracy & SD	Precision	Recall	F1-Score
Ovarian Cancer	Ensemble_1 ChiSquare×SSA	21	94.20±0.0339	95.90	92.20	93.70±0.0400
	Ensemble_2 MutualInfo×SSA	19	96.30±0.0125	96.80	94.90	**96.00±0.0163**
	Ensemble_3 ((ChiSquare - MutualInfo) +SSA)	21	91.80±0.0155	94.60	89.20	90.10±0.0173
	Ensemble_4 ((ChiSquare + MutualInfo) +SSA)	**18**	**98.00±0.0156**	**98.50**	**97.00**	93.70±0.0400

TABLE VI. T-TEST RESULTS ON THE ACCURACY AND F1 SCORE USING THE TOP TWO METHODS FOR THE OVARIAN CANCER DATASET

Measures	Competing algorithms	P_values
Accuracy	Ensemble_2 Vs. Ensemble_4	0.000000051
F1-Score	Ensemble_2 Vs. Ensemble_4	0.0000000034

D. Ensemble_4 Versus Other Optimization Algorithms

To further strengthen the case of the performance of our proposed approach for cancer gene selection, we conducted a comparative analysis with four optimization algorithms previously reported in the literature. Our results, documented in Table 7, were compared with the following algorithms: IG-PIRCACO [48], and Ensemble approach [36].

TABLE VII. COMPARING ENSEMBLE_4 WITH OTHER OPTIMIZATION ALGORITHMS IN THE LITERATURE

Datasets	Citations/Methods	Accuracy Score (%)	F1 Score (%)	Number of Features
Ovarian	Proposed Approach	98.00	97.00	18
	(Ke et al., 2022)	98.42	-	-
	(Alrefai & Ibrahim, 2022)	**100.00**	**100.00**	**13**
Colon	Proposed Approach	**94.20**	**93.60**	**16**
	(Ke et al., 2022)	91.90	-	-
	(Alrefai & Ibrahim, 2022)	92.86	93.00	41

Note that the basis of comparison with the mentioned optimization algorithms is the use of NIAs in combination with one or more filter methods while using the Colon and or Ovarian cancer datasets for evaluation. Throughout the comparison, we ensured a fair and consistent evaluation by maintaining the parameter settings of all benchmark optimization algorithms as reported by their respective authors. From Table 7, the Ensemble_4 from previous experiments having the best performance in both datasets is compared against the mentioned optimization algorithms and is denoted as the 'Proposed Approach'. The classification accuracy, the number of selected genes, and the F1 score are the measures used for performance evaluation. The best results from each of the two categories are bolded and underlined. Furthermore, empty spaces within the table show the absence of the particular result in the corresponding paper.

In the case of the Ovarian cancer dataset, the study conducted by Alrefai & Ibrahim, [36] stands out by combining Particle Swarm Optimization with three filter methods: chi-square, information gain, and Relief. This approach yielded exceptional results across all performance measures, including a classification accuracy of 100%, an F1 score of 100%, and the identification of 13 relevant genes. However, our proposed approach achieves competitive performance with a classification accuracy of 98.00%, an F1 score of 100%, and the selection of 18 informative genes. Shifting focus to the Colon cancer dataset, our proposed approach surpasses all others by attaining the highest performance metrics. It achieves an impressive accuracy score of 94.20%, and an F1 score of 93.60%, and identifies 16 relevant genes. The superior performance of our proposed approach underscores its effectiveness in selecting an optimal subset of genes specifically tailored to the Colon cancer dataset.

The superior performance observed in the proposed methods can be attributed to the unique capabilities of the filter methods employed in our approach. Specifically, the chi-square method effectively captures dependencies on the target variable, while mutual information measures associations between features. By merging the selected features from these methods, we can leverage their complementary information and obtain a more comprehensive set of relevant features. This merging process also contributes to the reduction of redundancy by eliminating duplicates and highly correlated features, thereby enhancing the efficiency of subsequent analysis. Additionally, the SSA plays a crucial role in optimizing the gene selection process. Through iterative refinement of the feature subset, the SSA increases the likelihood of identifying an optimal subset that exhibits improved performance. The synergistic integration of these techniques maximizes the chances of selecting features that

capture essential information while minimizing the impact of noise and irrelevant genes.

To conclude this section, our proposed methods effectively utilize both filtering and wrapping techniques to determine their varying levels of performance. The combination of filter and wrapper methods can successfully identify relevant genes while eliminating redundant or irrelevant ones. Moreover, integrating mechanisms to address overfitting, a common challenge in machine learning and gene selection, can further improve performance. It's important to note that the superiority of a specific filter-wrapper combination can be influenced by the dataset's unique characteristics, such as dimensionality, noise levels, class imbalance, and feature interactions. Considering these factors, our methods aim to optimize gene selection performance by adapting to each dataset's specific characteristics and challenges. Custom-designed or adjusted combinations tailored to address the dataset's uniqueness have the potential to outperform more generic approaches. Recognizing the "no one-size-fits-all" principle in optimization problems, the selection of appropriate algorithms should be guided by the specific nature of the problem. By taking into account the dataset's requirements and tailoring the approach accordingly, we can maximize the effectiveness of the combination and optimize performance for the given problem.

V. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this research paper, we proposed four approaches for microarray gene selection that combines the filtering techniques of chi-square and mutual information with the optimization power of the SSA. The Colon and Ovarian cancer datasets were used as benchmark datasets for evaluating the performance of these methods. Our results demonstrated the effectiveness of these approaches in identifying a subset of relevant genes while minimizing redundancy and noise. By leveraging the complementary information captured by chi-square and mutual information, we were able to create a more comprehensive set of relevant genes. Furthermore, the iterative refinement process facilitated by the SSA enhanced the selection of an optimal subset with improved performance. The ensemble methods showcased promising results, with the Ensemble_4 which combines Chi-Square and Mutual Information methods for simultaneous gene selection before passing to SSA, emerging as the best performer across multiple evaluation measures in the Colon cancer dataset. This selected ensemble method i.e., Ensemble_4 was then compared with other optimization algorithms in the literature, further validating its superiority in gene selection. The success of our approach highlights the importance of combining multiple techniques in microarray gene selection. By integrating filtering methods and NIAs, we can leverage their respective strengths and mitigate their limitations, ultimately leading to more accurate and reliable gene selection results.

Future research directions may involve exploring different combinations of filter methods and optimization algorithms, as well as investigating the application of our approach to other datasets and diseases. Additionally, efforts can be directed toward enhancing the interpretability of the selected gene subsets, furthering our understanding of the underlying biological mechanisms.

REFERENCES

[1] S. S. Shekhawat, H. Sharma, S. Kumar, A. Nayyar, S. Member, and B. Qureshi, "bSSA : Binary Salp Swarm Algorithm With Hybrid Data Transformation for Feature Selection," vol. 9, pp. 14867–14882, 2021, doi: 10.1109/ACCESS.2021.3049547.

[2] O. A. Sarumi and C. K. Leung, "Adaptive Machine Learning Algorithm and Analytics of Big Genomic Data for Gene Prediction," Intell. Syst. Ref. Libr., vol. 206, pp. 103–123, 2022, doi: 10.1007/978-3-030-76732-7_5/COVER.

[3] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," Bioeng. 2023, Vol. 10, Page 173, vol. 10, no. 2, p. 173, Jan. 2023, doi: 10.3390/BIOENGINEERING10020173.

[4] K. Founta et al., "Gene targeting in amyotrophic lateral sclerosis using causality-based feature selection and machine learning," Mol. Med., vol. 29, no. 1, pp. 1–13, Dec. 2023, doi: 10.1186/S10020-023-00603-Y/FIGURES/3.

[5] R. Kundu, S. Chattopadhyay, E. Cuevas, and R. Sarkar, "AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets," Comput. Biol. Med., vol. 144, no. February, p. 105349, 2022, doi: 10.1016/j.compbiomed.2022.105349.

[6] R. Bandyopadhyay, A. Das Sharma, B. Dasgupta, A. Ghosh, C. Das, and S. Bose, "A New hybrid Feature selection-Classification model to Improve Cancer Sample Classification Accuracy in Microarray Gene Expression Data," 2023 Int. Conf. Comput. Electr. Commun. Eng., pp. 1–7, Jan. 2023, doi: 10.1109/ICCECE51049.2023.10085390.

[7] R. Layer, "Genetic Mutations Can Be Benign or Cancerous—a New Method to Differentiate Between Them Could Lead to Better Treatments," TheScientist. TheScientist, pp. 1–4, 2022. [Online]. Available: https://www.the-scientist.com/news-opinion/genetic-mutations-can-be-benign-or-cancerous-a-new-method-to-differentiate-between-them-could-lead-to-better-treatments-70077?utm_content=210388761&utm_medium=social&utm_source=twitter&hss_channel=tw-18198832

[8] M. Alzaqebah et al., "Memory based cuckoo search algorithm for feature selection of gene expression dataset," Informatics Med. Unlocked, vol. 24, p. 100572, 2021, doi: 10.1016/j.imu.2021.100572.

[9] E. H. Houssein, D. S. Abdelminaam, H. N. Hassan, M. M. Al-Sayed, and E. Nabil, "A Hybrid Barnacles Mating Optimizer Algorithm with Support Vector Machines for Gene Selection of Microarray Cancer Classification," IEEE Access, vol. 9, pp. 64895–64905, 2021, doi: 10.1109/ACCESS.2021.3075942.

[10] O. A. Alomari et al., "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators," Knowledge-Based Syst., vol. 223, p. 107034, 2021, doi: 10.1016/j.knosys.2021.107034.

[11] G. A. Sulley, J. Hamm, and M. M. Montemore, "Machine learning approach for screening alloy surfaces for stability in catalytic reaction conditions," J. Phys. Energy, vol. 5, no. 1, p. 015002, Nov. 2022, doi: 10.1088/2515-7655/ACA122.

[12] M. H. Nadimi-Shahraki, M. Banaie-Dezfouli, H. Zamani, S. Taghian, and S. Mirjalili, "B-MFO: A binary moth-flame optimization for feature selection from medical datasets," Computers, vol. 10, no. 11, pp. 1–18, 2021, doi: 10.3390/computers10110136.

[13] L. Meenachi and S. Ramakrishnan, "Metaheuristic Search Based Feature Selection Methods for Classification of Cancer," Pattern Recognit., vol. 119, p. 108079, 2021, doi: 10.1016/j.patcog.2021.108079.

[14] T. Thaher, H. Chantar, J. Too, M. Mafarja, H. Turabieh, and E. H. Houssein, "Boolean Particle Swarm Optimization with various Evolutionary Population Dynamics approaches for feature selection problems," Expert Syst. Appl., vol. 195, p. 116550, Jun. 2022, doi: 10.1016/J.ESWA.2022.116550.

[15] J. Rajeshwari and M. Sughasiny, "Modified Filter Based Feature Selection Technique for Dermatology Dataset Using Beetle Swarm Optimization," EAI Endorsed Trans. Scalable Inf. Syst., vol. 10, no. 2, pp. e1–e1, Jul. 2023, doi: 10.4108/EETSIS.VI.1998.

[16] M. Tubishat et al., "Dynamic Salp swarm algorithm for feature selection," Expert Syst. Appl., vol. 164, no. August 2020, p. 113873, 2021, doi: 10.1016/j.eswa.2020.113873.

[17] 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," vol. 19, no. July, pp. 13747–13781, 2022, doi: 10.3934/mbe.2022641.

[18] I. Jeremiah, M. Abdullahi, S. A. Yusuf, and I. H. Hassan, "Towards an Improved Particle Swarm Optimization for Feature Selection : A Survey," Sule Lamido Univ. J. Sci. Technol., vol. 6, no. 1, pp. 59–73, 2023, doi: oi.org/10.56471/slujst.v6i.354.

[19] K. A. Uthman, B.-A. Fadl Mutaher, and M. O. Suad, "A Survey on Feature Selection in Microarray Data: Methods Algorithms and Challenges," Int. J. Comput. Sci. Eng., no. November, pp. 106–116, 2020, doi: 10.26438/ijcse/v8i10.106116.

[20] N. Mohd Ali, R. Besar, and N. A. Nor, "Hybrid Feature Selection of Breast Cancer Gene Expression Microarray Data Based on Metaheuristic Methods: A Comprehensive Review," Symmetry (Basel)., vol. 14, no. 10, 2022, doi: 10.3390/sym14101955.

[21] E. H. Houssein, A. G. Gad, K. Hussain, and P. Nagaratnam, "Major Advances in Particle Swarm Optimization : Theory , Analysis , and Application," Swarm Evol. Comput., vol. 63, no. February, p. 100868, 2021, doi: 10.1016/j.swevo.2021.100868.

[22] M. Tubishat et al., "Dynamic Salp swarm algorithm for feature selection," Expert Syst. Appl., vol. 164, no. June 2020, p. 113873, 2021, doi: 10.1016/j.eswa.2020.113873.

[23] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," Appl. Intell. 2021 524, vol. 52, no. 4, pp. 4543–4581, Jul. 2021, doi: 10.1007/S10489-021-02550-9.

[24] Y. K. Saheed, "Effective dimensionality reduction model with machine learning classification for microarray gene expression data," Data Sci. Genomics, pp. 153–164, Jan. 2023, doi: 10.1016/B978-0-323-98352-5.00006-9.

[25] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," Brief. Bioinform., vol. 23, no. 2, Mar. 2022, doi: 10.1093/BIB/BBAB569.

[26] S. K. Prabhakar and S. W. Lee, "Transformation Based Tri-Level Feature Selection Approach Using Wavelets and Swarm Computing for Prostate Cancer Classification," IEEE Access, vol. 8, pp. 127462–127476, 2020, doi: 10.1109/ACCESS.2020.3006197.

[27] W. Ali and F. Saeed, "Hybrid Filter and Genetic Algorithm-Based Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data," Processes, vol. 11, no. 2, 2023, doi: 10.3390/pr11020562.

[28] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," 2021.

[29] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," Int. J. Cogn. Comput. Eng., vol. 1, pp. 78–97, Jun. 2020, doi: 10.1016/J.IJCCE.2020.11.001.

[30] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Improved intelligent water drop-based hybrid feature selection method for microarray data processing," Comput. Biol. Chem., vol. 103, p. 107809, Apr. 2023, doi: 10.1016/J.COMPBIOLCHEM.2022.107809.

[31] I. H. Hassan, A. Mohammed, Y. S. Ali, I. Jeremiah, and S. A. Abdulraheem, "Metaheuristic algorithms in text clustering," Compr. Metaheuristics, pp. 131–152, Jan. 2023, doi: 10.1016/B978-0-323-91781-0.00007-7.

[32] U. Ravindran and C. Gunavathi, "A survey on gene expression data analysis using deep learning methods for cancer diagnosis," Prog. Biophys. Mol. Biol., vol. 177, pp. 1–13, Jan. 2023, doi: 10.1016/J.PBIOMOLBIO.2022.08.004.

[33] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data: A survey," J. Biomed. Inform., vol. 117, no. March, p. 103764, 2021, doi: 10.1016/j.jbi.2021.103764.

[34] X. fang Song, Y. Zhang, D. wei Gong, and X. yan Sun, "Feature selection using bare-bones particle swarm optimization with mutual information," Pattern Recognit., vol. 112, p. 107804, 2021, doi: 10.1016/j.patcog.2020.107804.

[35] X. Deng, M. Li, L. Wang, and Q. Wan, "RFCBF: Enhance the Performance and Stability of Fast Correlation-Based Filter," Int. J. Comput. Intell. Appl., vol. 21, no. 2, Jun. 2022, doi: 10.1142/S1469026822500092/ASSET/IMAGES/LARGE/S1469026822500092FIGF6.JPEG.

[36] N. Alrefai and O. Ibrahim, "Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets," Neural Comput. Appl. 2022, pp. 1–16, Mar. 2022, doi: 10.1007/S00521-022-07147-Y.

[37] K. K. Ghosh et al., "Theoretical and Empirical Analysis of Filter Ranking Methods : Experimental Study on Benchmark DNA Microarray Data," Expert Syst. Appl., p. 114485, 2020, doi: 10.1016/j.eswa.2020.114485.

[38] M. Ayham, A. Alhafedh, and O. S. Qasim, "Two-Stage Gene Selection in Microarray Dataset Using Fuzzy Mutual Information and Binary Particle Swarm Optimization," no. January, 2019, doi: 10.5958/0973-9130.2019.00458.4.

[39] X. Li, J. Zhang, and F. Safara, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," Neural Process. Lett., no. 0123456789, 2021, doi: 10.1007/s11063-021-10491-0.

[40] Y. Zhang, Y. hu Wang, D. wei Gong, and X. yan Sun, "Clustering-guided particle swarm feature selection algorithm for high-dimensional imbalanced data with missing values," IEEE Trans. Evol. Comput., 2021, doi: 10.1109/TEVC.2021.3106975.

[41] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," Appl. Sci., vol. 8, no. 9, 2018, doi: 10.3390/app8091521.

[42] A. Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, and A. Makrariya, "A Review on Nature-Inspired Algorithms for Cancer Disease Prediction and Classification," 2023.

[43] 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, doi: 10.1016/j.advengsoft.2017.07.002.

[44] A. Sharma and R. Rani, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," Comput. Methods Programs Biomed., vol. 178, pp. 219–235, 2019, doi: 10.1016/j.cmpb.2019.06.029.

[45] S. Azadifar and A. Ahmadi, "A graph-based gene selection method for medical diagnosis problems using a many-objective PSO algorithm," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–16, 2021, doi: 10.1186/s12911-021-01696-3.

[46] S. K. Baliarsingh, S. Vipsita, and B. Dash, "A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm," Neural Comput. Appl., vol. 32, no. 12, pp. 8599–8616, 2020, doi: 10.1007/s00521-019-04355-x.

[47] A. Chaudhuri and T. P. Sahu, "A hybrid feature selection method based on Binary Jaya algorithm for micro-array data classification," Comput. Electr. Eng., vol. 90, no. November 2020, p. 106963, 2021, doi: 10.1016/j.compeleceng.2020.106963.

[48] L. Ke, M. Li, L. Wang, S. Deng, J. Ye, and X. Yu, "Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression tumor classification," Pattern Anal. Appl., 2022, doi: 10.1007/s10044-022-01117-9.

[49] W. Ma, X. Zhou, H. Zhu, L. Li, and L. Jiao, "A two-stage hybrid ant colony optimization for high-dimensional feature selection," Pattern Recognit., vol. 116, p. 107933, 2021, doi: 10.1016/j.patcog.2021.107933.

[50] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm.," Expert Syst. Appl., vol. 166, p. 114012, 2021, doi: 10.1016/j.eswa.2020.114012.

[51] A. Dabba, A. Tari, and S. Meftali, "Hybridization of Moth flame optimization algorithm and quantum computing for gene selection in

microarray data," J. Ambient Intell. Humaniz. Comput., vol. 12, no. 2, pp. 2731–2750, 2021, doi: 10.1007/s12652-020-02434-9.

[52] N. Almugren and H. M. Alshamlan, "New bio-marker gene discovery algorithms for cancer gene expression profile," IEEE Access, vol. 7, pp. 136907–136913, 2019, doi: 10.1109/ACCESS.2019.2942413.

[53] N. Almugren and H. Alshamlan, "FF-SVM: New FireFly-based Gene Selection Algorithm for Microarray Cancer Classification," 2019 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2019, 2019, doi: 10.1109/CIBCB.2019.8791236.

[54] I. Jeremiah, M. Abdullahi, S. A. Yusuf, M. Nuruddeen Idris, B. S. Garko, and M. Yusuf Haruna, "Integrating Local Search Methods in Metaheuristic Algorithms for Combinatorial Optimization: The Traveling Salesman Problem and its Variants," 2022 IEEE Niger. 4th Int. Conf. Disruptive Technol. Sustain. Dev., pp. 1–5, Apr. 2022, doi: 10.1109/NIGERCON54645.2022.9803003.

[55] J. O. Agushaka, A. E. Ezugwu, L. Abualigah, S. K. Alharbi, and H. A. E. W. Khalifa, "Efficient Initialization Methods for Population-Based Metaheuristic Algorithms: A Comparative Study," Arch. Comput. Methods Eng., vol. 30, no. 3, pp. 1727–1787, Apr. 2022, doi: 10.1007/S11831-022-09850-4/METRICS.

[56] M. Kalra, V. Kumar, M. Kaur, S. A. Idris, Ş. Öztürk, and H. Alshazly, "A novel binary emperor penguin optimizer for feature selection tasks," Comput. Mater. Contin., vol. 70, no. 3, pp. 6239–6255, 2022, doi: 10.32604/cmc.2022.020682.

[57] M. Y. Norfadzlia, A. K. Muda, S. F. Pratama, R. Carbo-Dorca, and A. Abraham, "Improved swarm intelligence algorithms with time-varying modified Sigmoid transfer function for Amphetamine-type stimulants drug classification," Chemom. Intell. Lab. Syst., vol. 226, 2022, doi: https://doi.org/10.1016/j.chemolab.2022.104574.

[58] S. G. William, "On student's 1908 article 'the probable error of a mean,'" J. Am. Stat. Assoc., vol. 103, no. 481, pp. 1–7, Mar. 1908, doi: 10.1198/016214508000000030.

[59] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Trans. Evol. Comput., vol. 1, no. 1, pp. 67–82, 1997, doi: 10.1109/4235.585893.